Essays in Econometrics: Nonparametrics and Robustness

by
Ben Deaner

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Economics and Statistics
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
May 2021

Author ....
Department of Economics
May 5th 2021

Certified by ....
Whitney Newey
Ford Professor of Economics, MIT
Thesis Supervisor

Certified by ....
Anna Mikusheva
Castle-Krob Career Development Associate Professor of Economics, MIT
Thesis Supervisor

Certified by ....
Jerry Hausman
John and Jennie S. MacDonald Professor of Economics, MIT
Thesis Supervisor

Accepted by ....
Amy Finkelstein
Chairman, Department Committee on Graduate Theses

Essays in Econometrics: Nonparametrics and Robustness

by

Ben Deaner

Submitted to the Department of Economics
on May 5th 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics and Statistics

Abstract

This thesis consists of three chapters. In each chapter I consider a particular problem in econometrics with implications for applied research, and in each case I attempt to solve that problem.

In Chapter 1 I consider the task of inferring causal effects when only 'proxy controls' are available. Proxy controls are informative proxies for unobserved confounding factors. For example, suppose we wish to estimate the causal impact of holding students back a grade on their future test scores. Academic ability is likely a confounding factor. While ability is not observed, early test scores may be used to proxy for ability. Under suitable conditions, nonparametric identification and estimation of treatment effects is possible in this setting. I present novel nonparametric identification results that motivate simple and 'well-posed' nonparametric estimation and inference methods for use with proxy controls.

My analysis applies to cross-sectional settings but is particularly well-suited to panel models. In panel settings, proxy control methods provide a novel approach to the difficult problem of identification with non-separable, general heterogeneity and fixed T. In panels, observations from different periods serve as proxies for unobserved heterogeneity and my key identifying assumptions follow from restrictions on the serial dependence structure.

I derive convergence rates for my estimator and construct uniform confidence bands with asymptotically correct size. I apply my methodology to two empirical settings. I estimate causal effects of grade retention on cognitive performance and I estimate consumer demand counterfactuals using panel data.

In Chapter 2 I show that nonparametric instrumental variables (NPIV) estimators are highly sensitive to misspecification: an arbitrarily small deviation from instrumental validity can lead to large asymptotic bias for a broad class of estimators. One can mitigate the problem by placing strong restrictions on the structural function in estimation. If the true function does not obey the restrictions then imposing them imparts bias. Therefore, there is a trade-off between the sensitivity to invalid instruments and bias from imposing excessive restrictions. In response, I present a method that allows researchers to empirically assess the sensitivity of their findings to misspecification. I apply my procedure to the empirical demand setting of Blundell(2007) and Horowitz (2011).

In Chapter 3 I consider methods for inference in dynamic discrete choice models that are robust to approximation error in the solution to the dynamic decision problem. Estimation and inference in dynamic discrete choice models often relies on approximation to lower the computational burden of dynamic programming. If it is not accounted for, the use of approximation can impart substantial bias in estimation and results in invalid confidence sets. I present a method for set estimation and inference that explicitly accounts for the use of approximation and is thus valid regardless of the approximation error. I show how one can account for the error from approximation at low computational cost. My methodology allows researchers to assess the estimation error due to approximation and thus more effectively manage the trade-off between bias and computational expedience. I provide simulation evidence to demonstrate the practicality of my approach.

JEL Codes C01, C10, C14

Thesis Supervisor: Whitney Newey
Title: Ford Professor of Economics, MIT

Thesis Supervisor: Anna Mikusheva
Title: Castle-Krob Career Development Associate Professor of Economics, MIT

Thesis Supervisor: Jerry Hausman
Title: John and Jennie S. MacDonald Professor of Economics, MIT

Acknowledgments

I owe a great debt to my advisors, Whitney Newey, Anna Mikusheva, and Jerry Hausman.

Whitney, my primary advisor, has provided invaluable support and advice over these six years. His suggestions, about the framing of ideas, about which features I should add to my papers and which I should omit, are always correct. Moreover, his contributions to econometrics made my work possible. Indeed, my second chapter considers a technique, nonparametric instrumental variables estimation, that Whitney invented, and my first chapter makes great use of both the theory and methodology he developed.

Anna is always a sparkling fount of encouragement and enthusiasm. Along with her detailed feedback on drafts, her advice for tackling the onerous economic publication process has been particularly helpful. Our Zoom calls during the pandemic are a great morale booster.

Jerry is always insightful and fun to talk with. I am ever astounded by the superhuman speed with which he is able to read and provide detailed written comments on every draft that I send him. Without his foundational contributions to panel data econometrics I do not think I could have written my first chapter.

Thanks to a number of friends for reading parts of this thesis at various stages. Thanks as well for countless discussions of this work, particularly those that consisted of my incomprehensible ramblings about whichever arcane mathematical problem I was working through on that particular week. In no particular order, Claire Lazar, Katharine Lauderdale, Liyang Sun, and Anat Kleiman, all took the time to read parts of this work and listened stoically while I ranted about it.

Thanks to the MIT economics department as a whole, particularly the econometricians. I calculate that I have presented around twelve times in departmental lunches and in each one received comments that led me to substantially improve my papers.

I would also like to thank my fellow students in the Economics PhD. It is a shame to have seen less of you this past year given the viral calamity scourging the globe, but I hope I will see more of you all in future.

Thanks also to my parents, who are ultimately responsible for all my successes. After all, without them I would literally not exist.

Contents

List of Figures

List of Tables

Chapter 1

Proxy Controls and Panel Data

A sizable portion of the empirical economist's working life is dedicated to diagnosing and accounting for confounding. A researcher engaged in this task often has in mind specific factors that plausibly explain the confounding. Ideally, a researcher would control for these factors, but they are often unavailable. We refer to factors that fully account for the confounding as 'perfect controls'. Academic ability, human capital, and preferences for consumables plausibly account for all the confounding in certain settings, but they are inherently unmeasurable. When perfect controls are unobserved, the researcher may have access to a number variables that are informative about these latent factors. Test scores are informative about academic ability and years of experience about human capital. With panel data, past observations can provide a wealth of information about the latent characteristics of that individual. For example, past consumption habits are likely informative about the individual's consumption preferences.

If a set of covariates is informative about an ideal unobserved set of perfect controls then we say those covariates are 'proxy controls'. A naive approach would treat the proxies as if they were perfect controls. For example, one could treat test scores as if they did in fact perfectly measure ability. However, if the proxies mis-measure the perfect controls, then controlling for the proxies in the conventional manner need not account for all the confounding and the resulting estimates would likely be asymptotically biased.

The problem of proxy controls has long been acknowledged in the Labor economics literature, particularly in the context of returns to schooling (for a survey see Section 2 of [9] and see [43] for a key early empirical example). Classical analyses assume additive linear specifications for the potential outcomes and the measurement error. Linearity may be implausible in some settings and precludes the study of nonlinearities and heterogeneity in treatment effects. An emerging literature considers proxy controls in non-linear and nonparametric settings.

We develop new nonparametric identification results in the context of proxy controls. We identify the conditional (on observed treatments) average structural function (CASF) and the conditional distribution of potential outcomes. Identification of the CASF is necessary if we wish to identify say, the effect of treatment on the treated and implies identification and implies identification of the average structural function and thus average treatment effects. We show that our characterization of the CASF is 'well-posed' under our identifying assumptions. Well-posedness is crucial for deriving simple and transparent convergence rates for estimation methods based on our identification results. We show that the problem of proxy controls is tied to causal analysis of panel models and use our general identification results with proxy controls to develop new nonparametric identification results for panel models with a fixed number of time periods.

Intuitively, our analysis treats identification and estimation with proxy controls as a measurement error problem. Proxy controls mis-measure a set of latent perfect controls. To account for the measurement error, the researcher divides the available proxy controls into two groups and, in effect, uses one group of proxy controls to instrument for the other. This approach resembles the standard strategy for dealing with classical measurement error in linear models when multiple measurements are available.

While our analysis applies in cross-sectional settings, our results are particularly well-suited to the context of panel data with fixed-T. In panels, observations from other time periods can be informative proxies for factors that explain the confounding (i.e., perfect controls). By definition, confounding factors are associated with treatments and potential outcomes, and so, if the confounding factors are persistent, past treatments and past outcomes must be informative about the confounding factors. We provide conditions on the serial dependence structure of the data and latent variables so that one can form proxy controls from past observations that satisfy the identifying assumptions of the general cross-sectional case.

We suggest non-parametric estimation and inference procedures based on our identification results. The procedures can be applied in both cross-sectional and panel settings. The estimation method is based on series regression, and therefore it also suggests a flexible parametric method if the number of series terms is simply held fixed rather than allowed to grow with the sample size. We establish consistency and a convergence rate for our estimator under our identifying assumptions and primitive conditions of the kind employed in the literature on standard non-parametric regression. We give conditions under which our estimator can be asymptotically approximated by a Gaussian process. We develop a method for constructing uniform confidence bands that is based on the multiplier bootstrap and show that the bands have asymptotically correct size.

The nonparametric proxy controls estimation and inference problem is highly nonstandard and thus requires novel asymptotic analysis. Estimation implicitly involves solving a conditional moment equality and then estimating a set of unknown linear functionals of the solution. Unlike in the Nonparametric Instrumental Variables (NPIV) literature ([4], [64]), the solution to our moment equality may not be identified nor even unique, and it generally does not have a structural interpretation. Therefore, we avoid placing any smoothness conditions directly upon the solution. In sum, our problem is distinct from other problems involving conditional moment equalities in that the problem is well-posed, involves unknown linear functionals of the solution, and the solution may not be unique nor smooth.

To demonstrate the usefulness of our methodology we apply it to two very different real-world data problems. We use data from the Panel Survey of Income Dynamics (PSID) to estimate a structural Engel curve for food. In this case our analysis is premised upon pre-determination and a Markov-type serial dependence restriction.

We revisit the empirical setting of [42] who use data from the Early Childhood Longitudinal Study of Kindergartners (ECLS-K) to estimate the causal impact of grade retention on the performance of US students in cognitive tests.

Related Literature

This paper contributes to an expanding body of recent research on the use of proxy controls in the biometrics literature. [62] consider the problem of nonparametric identification of the Average Structural Function (ASF) and the marginal distribution of potential outcomes in cross-sectional settings when controls are mis-measured. Our paper adds to this work in that we develop nonparametric estimation and inference methods under similar assumptions, we derive convergence rates for our estimator, prove asymptotic normality, and show our uniform confidence bands have asymptotically correct coverage. Key to the asymptotic results is our proof that the estimation problem is well-posed under the identifying assumptions. We also adapt the proxy controls approach to nonparametric causal inference in panel models. In addition, our cross-sectional identification results extend [62] in several ways. Firstly, we identify the conditional average structural function (CASF) and conditional distribution of potential outcomes, where conditioning is on the observed treatments. Identification of the CASF (conditional distribution of potential outcomes) implies identification of the ASF (marginal distribution of potential outcomes) but the converse does not hold. Secondly, we offer an alternative path to identification based on our Assumption 4.i. We also employ completeness conditions Assumption 3, which are, for practical purposes, significantly weaker than the corresponding assumptions in [62]. A key result in our paper is that under our identifying assumptions, the characterization of the CASF is well-posed. Well-posedness plays a crucial role in our estimation and inference results.[1]

A working paper, [63] considers parametric models with proxy controls under related assumptions to those in [62]. They consider models that involve a 'confounding

[1] Initial development of this work was independent of [62]. We became aware of their paper in the time between posting the first and second versions of this project on Arxiv.

bridge' with a known functional form, and they impose assumptions under which the confounding bridge is identified. They discuss parametric estimation and inference using standard Generalized Method of Moments. The confounding bridge roughly corresponds to the function γ in our Lemma 1.1.b, however this function plays a different role in our work. In our work the function is only defined indirectly, and need not be identified nor even unique, and our Theorem 1.1.b identifies the CASF when there is no confounding bridge. Moreover, our Lemma C.9 allows us to avoid assuming this function can be well approximated by some sieve basis in our asymptotic results.[2] [74] consider multiply robust estimation of the ATE with proxy controls in the special case in which confounders (perfect controls) and proxies are categorical (i.e., discretely distributed with finite support), whereas we allow for continuous, vector-valued confounders and proxies.

In recent subsequent work, [77] extends [62] to allow for time-varying treatment effects. [29] considers estimation with proxy controls and derives semiparametric efficiency bounds in the context of the proxy controls estimation problem under similar assumptions to our paper.[3]

[49] provide identification results for nonparametric and non-separable models with measurement error and present a related estimator. [4] Unlike [49], we do not require a normalization like mean- or median-unbiasedness of the mis-measured variables and we provide a simple, well-posed and constructive identification of causal objects and uncomplicated estimation and inference methods. We are able to achieve this because in our problem is more tractable than that of [49]. In our case the measurement error is only in control variables and not in treatment variables. We are uninterested in the causal effect of the latent perfect controls themselves.

[2]The first version of [63] appeared on Arxiv on the 15th of August 2018, roughly six weeks before the first version of our work was posted on the 30th of September 2018.

[3]Our treatment of the nonparametric estimation problem significantly predates [29]. The first version of our work ([32]) suggested a nonparametric estimator for the panel case and derived asymptotic results for this estimator. The second version, posted in December 2019 considered nonparametric estimation in the cross-sectional case, which amounts to a change in notation for the estimator.

[4]A working paper [69] employs results from [49] to achieve identification using proxy controls in a regression discontinuity setting.

The results of [49] have been used for estimation in panel models using a factor analytic approach. Notably in [51] and [12]. [79] applies their results to a panel model with noisy measurements of the covariates of interest. Our work differs from these in that we are not interested in recovering the distributions of latent unobservables. We assume our treatment variables are measured correctly, but we treat the problem somewhat artificially as a measurement error problem because lagged treatments or outcomes are noisy measurements of unobserved confounders.

More generally, our panel analysis follows a long line of work in which observations from other periods are used to account for unobserved heterogeneity. This approach is the basis of classic methods like those of [45], [46], and [11] and some more recent work that allows for nonlinearity like [40] and [37].

An extensive literature examines the effects of grade retention on cognitive and social success. For a meta-analysis see [55]. We build on the work of [42]. We use the cleaned data available with their paper and we estimate some of the same causal effects. Recent work to estimate consumer demand counterfactuals (in particular, structural Engel curves) in nonparametric/semi-parametric models includes the instrumental variables approach of [18] and the panel approach of [27]. For a short survey see [59].

1.1 General Model and Identification

Consider the following structural model:

$$Y = y_0(X, U) \tag{1.1.1}$$

Y is an observed dependent variable, X is a column vector of observables that represents the levels of assigned treatments, and U is a (potentially infinite-dimensional) vector that represents unobserved heterogeneity. The 'structural function' y_0 is not assumed to be of any particular parametric form.

The model above incorporates both cross-sectional and panel settings. In the panel case the model applies for a particular period t, that is, for a particular cross-sectional slice of the panel data. We could make the time-dependence explicit and rewrite the model above as $Y_t = y_{0,t}(X_t, U_t)$.

The structural function y_0 in (1.1.1) captures the causal effect of X on Y. For clarity, we situate our analysis in the potential outcomes framework. If for some unit $U = u$, then $y_0(x, u)$ is the unit's 'potential outcome' from treatment level x. That is, the outcome that would have been observed had the treatment of that unit been set to level x. Thus U captures all heterogeneity in the potential outcomes.

The focus of this paper is on the identification and estimation of conditional average potential outcomes, where we condition on the assigned treatments X. We refer to the function that returns the conditional average potential outcomes as the 'conditional average structural function' (CASF).

The CASF $\bar{y}$ is defined formally as follows:

$$\bar{y}(x_1|x_2) = E\big[y_0(x_1, U)|X = x_2\big]$$

In words, suppose we draw a unit at random from the sub-population who were assigned treatment $X = x_2$. Then the expected counterfactual outcome had the unit instead received treatment level x_1 is $\bar{y}(x_1|x_2)$.[5]

One may also be interested in identifying average potential outcomes conditional on the treatments as well as some additional variables S which could represent say, membership of a demographic sub-group. That is, one may wish to identify $E\big[y_0(x_1, U)|X = x_2, S = s\big]$. Our results extend straight-forwardly to this case. For instance, if S is discrete one can simply apply our analysis to the sub-population with S at some fixed s.[6]

By transforming the model, one can define an even richer set of counterfactual objects in terms of the CASF of the transformed model. For example, let y be some

[5] Note that if X is continuously distributed then $\bar{y}(x_1|x_2)$ is only uniquely defined for x_2 up to a set of F_X-measure zero, where F_X is the law of X.

[6] A previous working version of this paper explicitly incorporated additional conditioning variables S which could be continuous, discrete or a mixture of both. For ease of exposition we have dropped this feature from the current draft.

fixed scalar, let $\tilde{Y} = 1\{Y \leq y\}$, and let $\tilde{y}_0(x,u) = 1\{y_0(x,u) \leq y\}$. The transformed model is $\tilde{Y} = \tilde{y}_0(X,U)$. The conditional cumulative distribution function of the potential outcomes in the original model is the CASF of the transformed model:

$$P\big(y_0(x_1,U) \leq y | X = x_2\big) = E[\tilde{y}_0(x_1,U)|X = x_2]$$

A common approach to identification in the presence of confounding relies on the presence of what we term 'perfect controls'. A vector of perfect controls is an observable random vector W^*, so that conditioning on W^*, the treatments X and the heterogeneity in potential outcomes are independent. Formally, we write $U \perp\!\!\!\perp X|W^*$. We use this notation to denote conditional independence throughout the paper.

We give sufficient conditions for identification with observed perfect controls in Assumption 1 below.

Assumption 1 (Perfect Controls). i. $U \perp\!\!\!\perp X|W^*$ ii. The joint distribution $F_{(W^*,X)}$ is absolutely continuous with the product of the distributions F_{W^*} and F_X. iii. $E[|Y|] < \infty$ and for F_X-almost all x $E[|y_0(x,U)|] < \infty$.

Assumption 1.ii is a common support assumption. Loosely speaking it requires that when we condition on W^* the support of X does not change. We state the assumption in this way so that it can apply to both discretely and continuously distributed variables and to variables with distributions that are a mix of discrete and continuous parts. Assumption 1.iii is a weak regularity condition that implies the CASF is well-defined. Under Assumption 1 the CASF satisfies:

$$\bar{y}(x_1|x_2) = E\big[E[Y|W^*, X = x_1]\big|X = x_2\big]$$

Where the equality holds for F_X-almost all x_1 and x_2. If W^* is observed, then the RHS of the final equality depends only on the distribution of observables, thus the equation above identifies $\bar{y}(x_1|x_2)$. The characterization above is well-known but we formally state and prove it in Proposition C.1 in the supplementary materials.

When perfect controls W^* are unavailable the researcher may have access to proxy controls. By proxy controls, we mean variables that are informative about the perfect controls but which are not perfect controls themselves. Unlike the perfect controls, the proxies do not satisfy the unconfoundedness condition in Assumption 1.i., i.e., conditional on the proxies, potential outcomes and treatments are dependent. This failure of unconfoundedness typically arises because the proxies mis-measure the underlying perfect controls.

Application 1: The Causal Impact of Grade Retention. Suppose we are interested in the effect of being held back a grade on students' reading and math scores in tests taken at the end of elementary school. We have access to observational (that is, non-experimental) data. Estimation of the causal effect of grade retention is challenging because students that repeat a grade typically do so due to poor academic performance. Poor academic performance may reflect low academic ability which in turn may cause low achievement in future, regardless of whether the student is retained. Thus academic ability causes confounding.

It may be plausible that academic ability explains all the confounding: that a set of factors that capture academic ability are perfect controls. If these factors were observed, then assuming Assumptions 1.ii and 1.iii are satisfied, we could control for them in the usual way and achieve identification of causal objects. While academic ability is not observed by the researcher, suppose the data contains scores in some tests taken early in kindergarten and elementary, the scores on these tests may be informative about academic ability. Because these test scores mis-measure academic ability they are unlikely to satisfy unconfoundedness. These early test scores are thus 'proxy controls'.

Note that 'academic ability' could be quite broadly defined. One can think of academic ability as a set of factors that explain the propensity to perform well academically. These factors could include not just inherent features of the child, but also the extent to which the child's home environment is conducive to learning.

This setting is captured in the causal diagram in Sub-Figure 1.a below.

Application 2: The Causal Impact of the Household Budget on Food De-

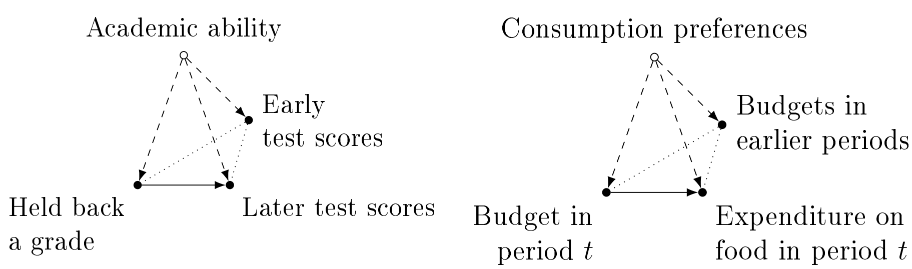
mand. Suppose we wish to estimate the causal effect of a household's total budget for non-durables on the household's demand for food. In particular, we are interested in the expected food expenditure in a period t if a household had their budget for nondurables in that period set exogenously to some quantity. We have access to observational panel data on household spending.

In this case, heterogeneity in household preferences is a key potential source of confounding. This is because total expenditure on non-durables is chosen by the household, and thus depends upon the household's underlying consumption preferences. These same preferences partially determine the household's expenditure on food. We may think that household preferences, along with the household's budget for nondurables, are the only source of confounding (at the individual level we may take prices to be exogenous), so factors that capture the household's preferences are a set of perfect controls.

While we cannot observe household preferences directly, the same argument that suggests preferences determine period t total nondurables expenditure also applies to periods prior to t. Thus the total budget the household allocated to nondurables in periods prior to t should be informative proxies for the household's preferences. This setting is captured in Sub-Figure 1.b.

Figure 1-1: Causal Diagrams for Applications 1 and 2

(a) Impact of Grade Retention (b) Impact of the Household Budget



Hollow nodes represent unobserved variables, and dashed directed lines indicate the existence of a causal effect of an unobserved variable on another variable. Filled nodes represent observed variables and directed lines from these variables represent their causal effects. The dotted lines represent possible causal relationships of an unspecified direction (including simultaneous causation or omitted variables that cause both). A causal diagram is associated with a nonparametric structural equations model and one can write potential outcomes (and thus the heterogeneity U) in terms of this model. See chapter 3 of [67].

In order to achieve identification, we suppose that the researcher has access to two vectors of proxies V and Z. In Application 1, our first empirical application, V and Z are scores in tests taken by each student early in their education. In particular, V is a set of scores on cognitive and behavioral tests taken in kindergarten and Z are scores on cognitive and behavioral tests taken early in elementary school. In Application 2, our second empirical application, V and Z are composed of lagged household expenditures from different periods.

Below we state additional assumptions under which we can achieve identification using proxy controls. The assumptions refer to the vector of perfect controls W^* for which V and Z act as proxies. Since W^* is unobserved, the assumptions can be understood to state that a vector of latent variables W^* exists that simultaneously satisfies all the conditions in our assumptions. Thus one can be agnostic as to the exact set of latent confounders W^*. For instance, one need not specify exactly what a vector of factors that capture academic ability should look like, we need only know that the test scores in V and Z are good proxies for these factors in the sense that V, Z, and W^* simultaneously satisfy the identifying assumptions.

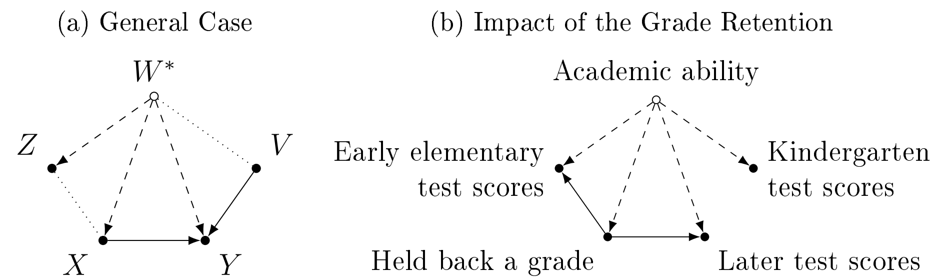
The proxy controls V and Z can be understood as measurements of W^* that are subject to non-classical (i.e., non-zero mean and non-additive) noise. To account for the measurement error, the researcher in effect uses the proxy controls in Z as instruments for the proxy controls in V.

As we discuss in Section 2, it may be useful to allow for the possibility that the vectors V and Z have components in common. In that case V and Z will be composed of overlapping sets of past observables. Our assumptions do not preclude this so long as the shared components are non-random conditional on W^* (e.g., if the shared components are also components of W^*). We denote the vector of shared components by $\bar{W}$ and refer to them in Assumption 3 below. If Z and V do not share components then one can ignore $\bar{W}$ in that assumption.

Assumption 2 (Conditional Independence). i. $U \perp\!\!\!\perp Z|(X, W^*)$ ii. $V \perp\!\!\!\perp (X, Z)|W^*$

The conditional independence restrictions in Assumption 2 and Assumption 1.i are implied by the causal diagram in Sub-Figure 2.a. Note this figure is similar to the causal structure (f) in [62]. Recall that the undirected dotted lines denote unrestricted causal relationships (causality could run in either direction or both, and additional confounding factors could affect both variables). There are other causal structures that imply the conditional independence restrictions, but we argue that both of our empirical applications are consistent with the structure in Sub-Figure 2.a.

Figure 1-2: Causal Diagrams for Assumptions 2 and 1.i



Recall that the dotted lines denote unrestricted causal relationships (i.e., causation could run in either or both directions and there can be additional unmeasured confounding between the variables). Assumption 1.i follows from the structure in Sub-Figure 2.a because we see W^* is the only source of confounding between treatments X and outcomes Y. Assumption 2.i states that after conditioning on treatments X and perfect controls W^*, any remaining variation in the proxies Z is independent of the heterogeneity in potential outcomes U. This holds in Sub-Figure 2.a because all confounding between Z and Y is due to W^*, and any causal connection between Z and Y is mediated by X. Assumption 2.ii states that any association between proxies V on the one hand, and Z and X on the other, is explained by the mutual dependence on W^*. We see this is Sub-Figure 2.a in that any causal path from V to X or to Z must run through W^*, and W^* is the only variable that can simultaneously cause both V on the one hand, and Z and X on the other.

Note that the proxies in V and Z are treated asymmetrically. The dotted line between X and Z indicates that the relationship between X and Z is unrestricted. Conversely, the proxies in V can directly affect outcomes, whereas those in Z cannot.

Sub-Figure 2.b contains a special case of the more general structure in Sub-Figure 2.a, where the variables are those in Application 1. Note that we need a student's kindergarten and early elementary test scores to be related only because they both reflect underlying academic ability. Once ability has been controlled for, remaining variation in the two sets of test scores should reflect factors like mood upon taking the test, and these factors should be unrelated between the two sets of tests.

In Application 1, the proxies in Z contain scores for tests taken early in elementary school. Some children are held back a grade in kindergarten and therefore treatment (being held back a grade) may have a causal effect on the proxies in Z. This is captured by the arrow from 'Held back a grade' to 'Early elementary test scores', i.e., from X to Z. This is compatible with Sub-Figure 2.a which leaves the relationship between X and Z unrestricted. Note that we could allow the kindergarten test scores to directly impact the outcome of interest (later test scores), but these cannot directly affect nor be affected by, the decision to hold a child back a grade. In the ECLS-K dataset that we use for our empirical analysis the cognitive and behavioral scores are not shared with the students, parents nor teachers, and thus they should not determine the decision to retain a child.[7], hence there is no causal effect of V on treatments X nor any causal effect of V on Y.

Assumption 3 (Informativeness). i. For $F_{(X,\bar{W})}$-almost all $(x, \bar{w})$, for any function $\delta \in L_2(F_{W^*|\bar{W}=\bar{w},X=x})$:

$$E[\delta(W^*)|X = x, Z = z] = 0 \iff \delta(w^*) = 0$$

Where the first equality above is $F_{Z|X=x,\bar{W}=\bar{w}}$-almost sure and the second is $F_{W^*|X=x,\bar{W}=\bar{w}}$-almost sure. ii. For $F_{(X,\bar{W})}$-almost all $(x, \bar{w})$, for any function $\delta \in L_2(F_{W^*|X=x,\bar{W}=\bar{w}})$:

$$E[\delta(W^*)|X = x, V = v] = 0 \iff \delta(w^*) = 0$$

Where the first equality above is $F_{V|X=x,\bar{W}=\bar{w}}$-almost sure and the second is $F_{W^*|X=x,\bar{W}=\bar{w}}$-

[7] This was confirmed by email with the ECLS study director.

almost sure.

Assumptions 3.i and 3.ii state, loosely speaking, that both V and Z are sufficiently informative about the unobserved perfect controls W^*. The informativeness conditions are in terms of 'completeness', or more precisely, L_2-completeness ([7]). Completeness is used to achieve identification in the non-parametric instrumental variables (NPIV) models of [64] and [4].

In the NPIV context, completeness is an instrumental relevance condition analogous to the rank condition in linear IV (see [64]). With this interpretation, 3.i states that conditional on any given value of assigned treatments, Z is a relevant instrument for W^*, and 3.ii states that conditioning on X, V is a relevant instrument for W^*. To provide further intuition, we show how these assumptions map into simple parametric cases in Appendix A.2.

In the linear IV case, the rank condition can only hold if the number of instruments exceeds the number of endogenous regressors, this is known as the 'order condition'. Analogously, statistical completeness cannot hold for certain classes of distributions unless this same order condition holds. [8] In our setting the relevant order condition is that each of the vectors V and Z be of a weakly larger dimension than W^*.

In the context of Application 1, the informativeness condition requires that the kindergarten test scores are relevant instruments for ability, and likewise early elementary test scores are relevant instruments for ability. The order condition then suggests that academic ability is captured by no more factors than there are test scores in each of the two sets.

Assumption 3.ii differs from the corresponding assumption in [62] (Condition 3 in their paper). Their Condition 3 requires that (conditional on a fixed X) Z is complete for V. Our Assumption 3.ii requires W^* is complete for V. From a practical standpoint this is an important distinction. The condition in [62] essentially requires that V be a relevant instrument for Z, which generally means V must have greater dimension than Z. W^* may be much of much lower dimension than Z in which case the requirement on the dimension of V is weaker. Moreover, if Z contains some

[8] One such class of distributions is the conditional Gaussian class discussed in [64].

component that is simply orthogonal noise, unrelated to anything else, this would violate their assumption but would generally not violate ours. [9]

In addition to Assumptions 1, 2 and 3 we require that either of two regularity conditions hold. These conditions are somewhat technical and so we resign them to Appendix A.1 along with further discussion. The regularity conditions, Assumptions 4.i and 4.ii, are used to establish the existence of functions φ and γ that satisfy conditional moment restrictions. These objects need not have a structural interpretation, and need not be unique. Instead, they should be understood as 'representers' in that their existence allows us to rewrite two other functions in a more convenient form. The existence results are stated in Lemma 1.1 below.

Lemma 1.1 refers to $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}$, the Radon-Nikodym derivative of $F_{V|X=x_2}$ with respect to $F_{V|X=x_1}$, which must exists under Assumptions 1.ii and 2.ii.[10] If V admits a probability density conditional on $X = x$ of $f(\cdot|x)$ then this equals $f(v|x_1)/f(v|x_2)$. We use Radon-Nikodym derivatives to allow for mixtures of finite and discrete random variables.

Lemma 1.1. *Suppose Assumptions 1, 2 and 3 hold. Then: a., Assumption 4.i implies that there exists a function φ so that for F_X-almost all x_1 and x_2, $E\big[\varphi(x_1, x_2, Z)^2|X = x_1\big] \leq C(x_1, x_2)$ with $C(x_1, x_2) < \infty$ and:*

$$E[\varphi(x_1, x_2, Z)|X = x_1, V] = \frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(V)$$

b., Assumption 4.ii implies that there exists a function γ that satisfies the inequality $E\big[\gamma(X, V)^2|X = x\big] \leq D(x)$ with $D(x) < \infty$ for F_X-almost all x and:

$$E[\gamma(X, V)|X, Z] = E[Y|X, Z]$$

[9] Assumption 3.ii is not weaker than their Condition 3 in a formal mathematical sense. However, our results would go through if Assumption 3.ii were weakened to hold only for $\delta \in L_2(F_{W^*|X=x,\bar{W}=\bar{w}})$ that can be written in the form $\delta(w^*) = E[\tilde{\delta}(Z)|X = x, W^* = w^*, \bar{W} = \bar{w}]$ for some $\tilde{\delta}$. When Assumption 3.i (or Condition 2 in [62]) holds, this weakened version of Assumption 3.ii is weaker than their condition 3 in a formal mathematical sense. We employ the slightly stronger assumption for ease of exposition, some earlier online versions of this work employ the weaker assumption.

[10] see Proposition C.2 in the supplementary materials for a formal proof.

We now state our first result, which presents two alternative characterizations of the CASF.

Theorem 1.1 (Identification). *Suppose Assumptions 1, 2, and 3 hold. Then:*

a. If 4.ii (and not necessarily 4.i) holds, there exists a function γ *with* $\gamma(x_1, \cdot) \in L_2(F_{V|X=x_1})$ *for* F_X*-almost all* x_1*, so that:*

$$E[Y - \gamma(X, V)|X, Z] = 0 \tag{1.1.2}$$

And for any such γ*, for* F_X*-almost all* x_1 *and* x_2*:*

$$\bar{y}(x_1|x_2) = E\big[\gamma(x_1, V)|X = x_2\big]$$

b. If 4.i (and not necessarily 4.ii) holds, there exists a function φ *so that* $\varphi(x_1, x_2, \cdot) \in L_2(F_{Z|X=x_1})$ *for* F_X*-almost all* x_1*, and:*

$$E[\varphi(x_1, x_2, Z)|X = x_1, V] = \frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(V) \tag{1.1.3}$$

And for any such φ*, for* F_X*-almost all* x_1 *and* x_2*:*

$$\bar{y}(x_1|x_2) = E\big[Y\varphi(x_1, x_2, Z)|X = x_1\big]$$

Theorem 1.1 characterizes the CASF in terms of observables and thus establishes identification. The characterization in Theorem 1.1.a closely resembles that of [62] but applies for the CASF rather than just the average structural function. The characterization in Theorem 1.1.b is, to the best of our knowledge, entirely new. In order to build intuition for these results it may be helpful to consider simple parametric examples, we provide some worked examples in Appendix A.2.

Theorem 1.2 below shows that the equations that characterize the CASF in 1.1.a and 1.1.b are well-posed.

Theorem 1.2 (Well-Posedness). *Suppose Assumptions 1, 2, and 3 hold, then: a. If 4.i (and not necessarily 4.ii) holds, for any* $\tilde{\gamma}(x_1, \cdot) \in L_2(F_{V|X=x_1})$*, and for*

F_X-almost all x_1 and x_2:

$$\begin{aligned}&\big(\bar{y}(x_1|x_2) - E[\tilde{\gamma}(x_1, V)|X = x_2]\big)^2 \\ \leq &C(x_1, x_2)E\bigg[\big(E[Y - \tilde{\gamma}(X, V)|X, Z]\big)^2\bigg|X = x_1\bigg]\end{aligned} \tag{1.1.4}$$

b. If 4.ii (and not necessarily 4.i) holds, for any $\tilde{\varphi}(x_1, x_2, \cdot) \in L_2(F_{Z|X=x_1})$, and for F_X-almost all x_1 and x_2:

$$\begin{aligned}&\big(\bar{y}(x_1|x_2) - E[Y\tilde{\varphi}(x_1, x_2, Z)|X = x_1]\big)^2 \\ \leq &D(x_1)E\bigg[\bigg(E[\tilde{\varphi}(X, x_2, Z)|X, V] - \frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(V)\bigg)^2\bigg|X = x_1\bigg]\end{aligned} \tag{1.1.5}$$

Theorems 1.1 and 1.2 suggest a two-step approach to estimation. Recall the moment condition in part a. of Theorem 1:

$$E[Y - \gamma(X, V)|X, Z] = 0 \tag{1.1.6}$$

The equation above is equivalent to an NPIV moment condition with endogenous regressors V, exogenous regressors X , and instruments Z. Suppose $\hat{\gamma}$ solves an empirical analogue of the moment condition (1.1.6). In a second step Theorems 1.1 and 1.2 suggests we estimate the CASF by $\hat{E}_V\big[\hat{\gamma}(x_1, V)|X = x_2\big]$, where '$\hat{E}_V$' denotes some empirical analogue of the conditional expectation and $\hat{\gamma}$ is treated as non-random in the expectation. The inequality in Theorem 1.2.a implies that if $\hat{\gamma}$ satisfies the population moment condition (1.1.6) with small error (in a mean-squared sense), then $E_V\big[\hat{\gamma}(x_1, V)|X = x_2\big]$ is close to the CASF $\bar{y}(x_1|x_2)$. If, in addition, $E_V\big[\hat{\gamma}(x_1, V)|X = x_2\big]$ is close to the sample analogue $\hat{E}_V\big[\hat{\gamma}(x_1, V)|X = x_2\big]$, then the latter provides a good estimate of the CASF. This motivates our estimator in Section 3.

Theorem 1.2.a suggests that the estimation problem based on Theorem 1.1.a is well-posed. Well-posedness plays a crucial role in our analysis. It allows us to derive simple convergence rates for our estimation method that are comparable to those in standard non-parametric regression and do not depend on any 'sieve-measure of ill-posedness' (see [18], [24]).

The well-posedness of our problem may be surprising because it is well-known that estimation of an NPIV regression function is generally ill-posed. However, we characterize the CASF as a linear functional (specifically a conditional mean) of an NPIV regression function. Estimation of a sufficiently smooth linear functional of an NPIV regression function is well-posed. In particular, the existence of a solution to (1.1.3) guarantees sufficient smoothness of the relevant linear functional. Because Assumption 4 ensures this existence, this assumption is crucial to the well-posedness. Existence conditions of a similar kind are shown to be closely related to root-n estimability ([4], [73], [53]) and to robust estimation ([33]) in NPIV.

One could also motivate an estimator based on part b. of Theorem 1.1. In particular one would first estimate $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(\cdot)$ and then solve for φ in an empirical analogue of equation (1.1.3). One would then plug the empirical solution $\hat{\varphi}$ into an empirical analogue of the conditional expectation $E_Z\big[Y\hat{\varphi}(x_1, x_2, Z)|X = x_1\big]$, where $\hat{\varphi}$ is treated as non-random in the expectation. One could also construct a doubly robust estimator based on a combination of both characterizations in Theorem 1.1. We intend to explore this avenue in future work.

1.2 Panel Models

The analysis in the previous section considers the model (1.1.1) which may apply in both cross-sectional and panel settings. In panel settings, observations from previous periods are a natural source of proxy controls. Recall that perfect controls explain confounding between treatments and potential outcomes. Thus (if there is confounding) perfect controls are associated with both treatments and outcomes. If the perfect controls are persistent, and explain confounding in each period, treatments and outcomes in other periods are associated with (i.e., informative about) the perfect controls. If the proxies V and Z contain treatments (and/or outcomes) from different periods, then the conditional independence restrictions in Assumptions 1.i and 2 can be understood in terms of the serial dependence structure.

In the panel setting, the data have a 'time' dimension and a 'unit' dimension. To apply our analysis in the panel case we rewrite the model (1.1.1) with time subscripts as $Y_t = y_{0,t}(X_t, U_t)$. Then for each group there is an associated draw of the random variables $(X_1, ..., X_T)$, $(U_1, ..., U_T)$ and a resulting sequence of outcomes $(Y_1, ..., Y_T)$. [11]

In the panel setting our goal is to identify and estimate for a particular value of t, $E[y_{0,t}(x_1, U_t)|X_t = x_2]$. This is the conditional average potential outcome at period t from treatment x_1 conditional on assignment of treatment x_2 at t. In this context Assumptions 1.i, 2.i and 2.ii state that $U_t \perp\!\!\!\perp (X_t, Z)|W^*$ and $V \perp\!\!\!\perp (X_t, Z)|W^*$ [12]

Note that W^* does not have a time-subscript. In Application 2 we consider W^* to be composed of two sets of components. These are $\tilde{W}^*$ which are time-invariant factors that represent aspects of household consumption preferences, and the treatments in a previous period $\bar{t}$ which are used to account for serial dependence in the treatments. Thus $\tilde{W}^*$ are analogous to the fixed effects in linear panel models but can be multi-dimensional and need not be additively separable in the structural function. More generally we could allow W^* to contain time-varying confounding factors.

Below we present two cases in which the conditional independence restrictions above follow from primitive conditions on the panel structure for appropriate choices of V and Z. In both cases V and Z are composed of lagged observables. These cases are not exhaustive, instead they should be understood as leading examples. Under different assumptions on the serial dependence one could justify say, proxy controls that include leads of observables as well as lags.

1.2.1 Markov Treatment Assignments and Predetermination

Suppose we are interested in the CASF at some fixed period t. Suppose that conditional on some (possibly period t-specific) latent variables $\tilde{W}^*$, the following condi-

[11] Note that this specification above allows for dynamic models with feedback, for example if X_t includes lags of Y_t.

[12] We have used that, $U_t \perp\!\!\!\perp (X_t, Z)|W^*$ is equivalent to the combination of $U_t \perp\!\!\!\perp X_t|W^*$ and $U_t \perp\!\!\!\perp Z|(X_t, W^*)$.

tional independence restriction holds:

$$U_t \perp\!\!\!\perp (X_1, ..., X_t) | \tilde{W}^* \tag{1.2.1}$$

In words, the condition above states that the history of treatments up to and including period t is only related to potential outcomes through some factors $\tilde{W}^*$. If $\tilde{W}^*$ is taken to represent some persistent latent factors, then the restriction is a non-parametric analogue of the 'predetermination' condition often employed in linear panel models. One justification for the assumption is as follows. Suppose we interpret $\tilde{W}^*$ to contain all persistent factors in the potential outcomes. Then any remaining variation in U_t represents shocks to potential outcomes. In this case the assumption states that the history of treatments up to and including time t is uninformative about these shocks. However, the assumption allows for the possibility that shocks to potential outcomes impact (or are otherwise associated with) future treatment assignments.

In the context of Application 2 (the causal effect of the household nondurables budget on food expenditure), U_t represents variation in the household's spending on food given their total budget for nondurables. Suppose we interpret $\tilde{W}^*$ as some fixed factors that explain household preferences (including say, permanent income), then remaining variation in U_t represents the time-varying component of preferences. If preferences vary over time due to events like the discovery of new preferred food products or the need to provide food for a new family member, then the assumption requires that these events were not anticipated by the household in earlier periods (or if they were, that it did not effect the total amount spent on nondurables). If the length of each time period is long, as it is in our dataset, then the ability to anticipate these preference shocks is likely reduced and the predetermination assumption more plausible. Note that predetermination allows for the possibility that these shocks (due to say, welcoming a new family member) affect future budgets for nondurables.

Suppose that conditional on the latent variables $\tilde{W}^*$, the regressors satisfy a first-order Markov dependence structure at some period $\bar{t}$. Formally:

$$(X_1, ..., X_{\bar{t}}) \perp\!\!\!\perp (X_{\bar{t}+1}, ..., X_T)|(\tilde{W}^*, X_{\bar{t}}) \tag{1.2.2}$$

That is, conditional on the latent variables $\tilde{W}^*$, the treatment assignments for periods strictly prior to the given period t are only only related to treatments after t through the treatment at t.

The length of time between periods may play a key role in justifying first-order Markov dependence. If periods represent long stretches of time then the assumption is easier to justify, because in that case direct causal effects of periods before the previous are less likely. For example, in the context of Application 2, we may worry that the household budgets in periods prior to $t-1$ might directly affect the household's budget in period t because they affect the credit scores of the householders. However, in the data we use to for this application the time between subsequent periods is two years, and so this is less likely than if say, periods lasted a month. Similarly, we may be concerned about trends in household income that span multiple periods. But on longer time-scales it is perhaps plausible that household income follows a random walk around a deterministic trend.

Proposition 2.1 below shows that in this setting Assumptions 1.i and 2 hold for V and Z composed of particular lagged treatments. Let $\bar{t}$ be some period between 1 and t.

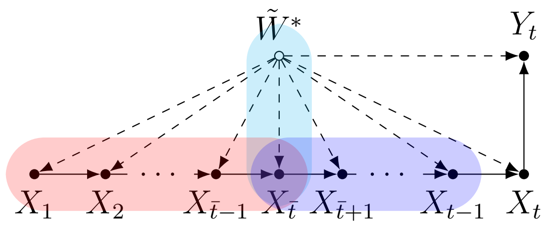
Proposition 2.1. *Suppose that (1.2.1) and (1.2.2) hold. Set* $V = (X_1, ..., X_{\bar{t}})$, $Z = (X_{\bar{t}}, ..., X_{t-1})$, *and* $W^* = (\tilde{W}^*, X_{\bar{t}})$. *Then Assumptions 1.i and 2 hold:* $U_t \perp\!\!\!\perp (X_t, Z)|W^*$ *and* $V \perp\!\!\!\perp (X_t, Z)|W^*$.

Note that we treat $X_{\bar{t}}$ as an observable perfect control, we therefore include it in both V and Z. Given the Markov structure, conditioning on the treatment at period $\bar{t}$ removes the dependence between V and Z.

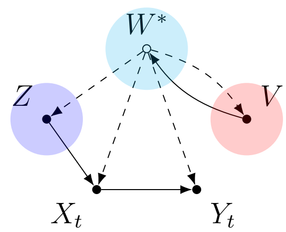
The causal diagrams in Figure 3 may clarify how the serial dependence restrictions imply the general proxy control conditional independence assumptions. In Sub-Figure 3.a is a causal diagram that implies (1.2.1) holds and (1.2.2) holds at period $\bar{t}$. The observations are grouped by color into corresponding variables V (red), Z (purple), and W^* (sky blue). With the observations grouped in this way, the causal structure in Sub-Figure 3.a collapses to the structure in Sub-Figure 3.b. The causal structure in Sub-Figure 3.b is a special case of Sub-Figure 2.a, where there is no causal link between V and Y, causation runs from Z to X, and W^* and V simultaneously cause each other (due to the shared presence of $X_{\bar{t}}$).

Figure 1-3: Conditional Independence in the Panel Case

(a) Pre-Determination and First-Order Markov Treatments



(b) Group Structure



One can extend our results straight-forwardly to the case of second- or third-order Markov dependence and higher orders. In this case, instead of Z and V overlapping by one period $\bar{t}$, they must overlap by two periods in the second-order case, three periods in the third-order case, and so on. Again, the treatments from periods in both V and Z must be included in W^*. If treatments are mutually independent given latent characteristics $\tilde{W}^*$ (this must be true for some latent characteristics is treatments are exchangeable over time), then Z and V need not overlap.

1.2.2 Markov Treatment Assignments and Heterogeneity

We now give conditions under which Z and V may be composed not only of treatment assignments from periods other than t, but also the outcomes from other periods. We strengthen the conditional independence restriction from the previous subsection:

$$U_t \perp\!\!\!\perp (X_1, ..., X_t, U_1, ..., U_{t-1}) | \tilde{W}^* \tag{1.2.3}$$

Loosely speaking, the above strengthens the pre-determination condition by imposing (conditional) serial independence of the shocks U_t.

We suppose that conditional on the latent variables $\tilde{W}^*$, both the treatment assignments and heterogeneity follow a joint first-order Markov dependence structure, at least at some period $\bar{t}$. Formally, conditional on $(\tilde{W}^*, X_{\bar{t}}, U_{\bar{t}})$:

$$(X_1, ..., X_{\bar{t}-1}, U_1, ..., U_{\bar{t}-1}) \perp\!\!\!\perp (X_{\bar{t}+1}, ..., X_t, U_{\bar{t}+1}, ..., U_t) \tag{1.2.4}$$

Finally, we assume (without much loss of generality) that $y_{0,t}(x, \cdot)$ is injective for all x. The following proposition shows that if we set V, Z and W^* much as in the previous subsection but now with lagged outcomes, then Assumptions 1.i and 2 hold.

Proposition 2.2. *Suppose (1.2.3), (1.2.4) and for all x, $y_{0,t}(x, \cdot)$ is injective. Set $W^* = (\tilde{W}^*, X_{\bar{t}}, Y_{\bar{t}})$ and set V and Z as follows:*

$$V = (X_1, ..., X_{\bar{t}}, Y_1, ..., Y_{\bar{t}})$$
$$Z = (X_{\bar{t}}, ..., X_{t-1}, Y_{\bar{t}}, ..., Y_{t-1})$$

Then Assumptions 1.i and 2 hold: $U_t \perp\!\!\!\perp (X_t, Z)|W^*$ and $V \perp\!\!\!\perp (X_t, Z)|W^*$.

1.2.3 Assumption 3 and the Order Condition in Panels

As we discuss in Section 1, Assumption 3 requires both V and Z be relevant instruments for the perfect controls W^*. In the Markov treatment assignment case of Subsection 2.2, both V and Z are likely to be strongly associated with the perfect controls $W^* = (\tilde{W}^*, X_{\bar{t}})$. Both Z and V contain $X_{\bar{t}}$ and treatment assignments for periods other than t. By predetermination, $\tilde{W}^*$ explains the confounding between $(X_1, ..., X_t)$ and U_t. If there is confounding in each period that is explained by the presence of those same variables $\tilde{W}^*$ then each component of V and Z ought to be informative about $\tilde{W}^*$. In Subsection 2.2, V and Z also contain outcomes, and so the case is even stronger because outcomes are associated shocks U_t (and hence $\tilde{W}^*$) by construction.

In the case of Markov treatment assignments and heterogeneity, outcomes can also be used as proxies. With more proxies the informativeness assumptions are weaker (if a set of proxies satisfy the informativeness assumptions then so does any set of proxies that contains that set). Outcome variables are generally correlated with confounding factors by construction, and so the additional proxies are likely to be informative about $\tilde{W}^*$. Thus we see there is a trade-off between the conditional independence assumptions and informativeness assumptions: the conditional independence restrictions are stronger when outcomes are used as proxies but the informativeness conditions are weaker.

Note that if T is large then there are more observations from different periods from which to form V and Z. V and Z are then more likely to satisfy Assumption 3. Recall the order condition discussed in Section 1: V and Z each be of a weakly larger dimension than W^*. Suppose we are interested in the CASF at the final period T. Consider the first-order Markov treatment assignment example and let $\bar{t}$ be $\lfloor T/2 \rfloor$ (the greatest integer less than $T/2$). If treatments are scalar, Z is of length $T - \lfloor t/2 \rfloor$ and V is of length $\lfloor T/2 \rfloor$. Therefore, the order condition requires that $\tilde{W}^*$ be of length at most $\lfloor T/2 \rfloor - 1$. If we are interested in the CASF at the final period T then $\tilde{W}^*$ must be of length no greater than $\lfloor T/2 \rfloor - 1$.

1.3 Estimation and Inference

In this section we describe our estimation and inference procedures and analyze their asymptotic properties. The key step in estimation corresponds to penalized sieve minimum distance (PSMD) estimation (see [24] and [25]). Inference is based on the multiplier bootstrap (see for example [17]). Our methods can be applied in panel settings or to cross-sectional data. To emphasize this generality we return to the notation in Section 1 in which we suppress time subscripts.

Because our method is sieve-based it has a natural parametric analogue in which the number of basis functions is kept fixed rather than allowed to grow with the sample size. We discuss some parametric analogues of the method in Appendix A.2.

Let $\{(Y_i, X_i, Z_i, V_i)\}_{i=1}^n$ be a sample of n observations of the variables Y, X, Z, and V. In the panel case, Y_i and X_i should be understood to come from one fixed period t. For each n let ϕ_n be a column vector of basis functions defined on the support of (X, V). The first stage of the procedure consists of non-parametric regression. The practitioner estimates regression functions g, π_n and α_n which are defined by:

$$\begin{aligned} g(x,z) &= E[Y|X=x, Z=z] \\ \pi_n(x,z) &= E[\phi_n(X,V)|X=x, Z=z] \\ \alpha_n(x_1,x_2) &= E[\phi_n(x_1,V)|X=x_2] \end{aligned}$$

The estimation of each function above can be carried out using a standard non-parametric regression method like local-linear regression or series least-squares. Denote estimates of the fitted values $g(X_i, Z_i)$, $\pi_n(X_i, Z_i)$, and $\alpha_n(x_1, x_2)$ by $\hat{g}_i$, $\hat{\pi}_{n,i}$, and $\hat{\alpha}_n(x_1, x_2)$ respectively.

Let $Pen(\cdot)$ be some penalty function. Let $\lambda_{0,n}$ be a positive scalar penalty parameter. In the second stage, the researcher evaluates a vector of coefficients $\hat{\theta}$ that minimize the penalized least-squares objective:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}_i - \hat{\pi}_{n,i}'\theta\right)^2 + \lambda_{0,n}Pen(\theta) \tag{1.3.1}$$

The estimate of the CASF is then given by:

$$\bar{y}(x_1|x_2) \approx \hat{\alpha}_n(x_1,x_2)'\hat{\theta} \tag{1.3.2}$$

Some of our asymptotic results pertain to a particular version of the method described above in which the first-stage regressions are carried out using series ridge, a ridge penalty is used in the second stage and there is sample-splitting between some of the first-stage regressions.

The series ridge version of our estimator is carried out as follows. Let $\mathcal{I}_g$, $\mathcal{I}_\pi$, and $\mathcal{I}_\alpha$ represent subsets of $\{1, 2, ..., n\}$ of size n_g, n_π, and n_α. Let $Pen(\cdot)$ be a ridge penalty so $Pen(\theta) = ||\theta||^2$. Let I denote the identity matrix and define $\hat{\Sigma}_{\lambda_{0,n}} = \frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}\hat{\pi}'_{n,i} + \lambda_{0,n}I$. Then the objective (1.3.1) is minimized by setting $\hat{\theta} = \hat{\Sigma}^{-1}_{\lambda_{0,n}}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}\hat{g}_i$.

We assume the basis functions $\phi_n(x,v)$ are multiplicatively separable in x and v. In particular there are length $k(n)$ and $l(n)$ vectors of functions, ρ_n and χ_n, so that $\phi_n(x,v) = \rho_n(v)\otimes\chi_n(x)$ where '$\otimes$' is the Kronecker product.

The first-stage regressions are then carried out using series ridge regression with penalty parameters $\lambda_{1,n}$, $\lambda_{2,n}$, and $\lambda_{3,n}$. Let ψ_n be a length-$m(n)$ vector of basis functions defined on the support of (X,Z) and let $\psi_{n,i} = \psi_n(X_i,Z_i)$. Similarly, let ζ_n be a length-$p(n)$ on (X,Z) and $\zeta_{n,i} = \zeta_n(X_i,Z_i)$. Similarly, let $\rho_{n,i} = \rho_n(V_i)$ and $\chi_{n,i} = \chi_n(X_i)$. For any λ define $\hat{\Omega}_\lambda = \frac{1}{n_\pi}\sum_{i\in\mathcal{I}_\pi}\psi_{n,i}\psi'_{n,i}+\lambda I$, $\hat{\Xi}_\lambda = \frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}\zeta'_{n,i}+\lambda I$, and $\hat{G}_\lambda = \frac{1}{n_\alpha}\sum_{i\in\mathcal{I}_\alpha}\chi_{n,i}\chi'_{n,i}+\lambda I$. We obtain estimates $\hat{g}_i$, $\hat{\pi}_{n,i}$, and $\hat{\alpha}_n(x_1,x_2)$ as follows:

$$\begin{aligned}
\hat{g}_i =& \zeta'_{n,i}\hat{\Xi}^{-1}_{\lambda_{1,n}}\frac{1}{n_g}\sum_{j\in\mathcal{I}_g}\zeta_{n,j}Y_j \qquad (1.3.3)\\
\hat{\pi}_{n,i} =& \big(\psi'_{n,i}\hat{\Omega}^{-1}_{\lambda_{2,n}}\frac{1}{n_\pi}\sum_{j\in\mathcal{I}_\pi}\psi_{n,j}\rho_{n,j}\big)\otimes\chi_{n,i}\\
\hat{\alpha}_n(x_1,x_2) =& \big(\chi_n(x_2)'\hat{G}^{-1}_{\lambda_{3,n}}\frac{1}{n_\alpha}\sum_{j\in\mathcal{I}_\alpha}\chi_{n,j}\rho_{n,j}\big)\otimes\chi_n(x_1)
\end{aligned}$$

1.3.1 Consistency and Convergence Rate

We prove the consistency of our estimator and derive convergence rates under primitive conditions similar to those common in the literature on standard non-parametric regression.

Let us introduce some additional notation. $||a||$ is the Euclidean norm of a vector a and $||A||$ is the spectral norm of a matrix A. For sequences of scalars a_n and b_n, $a_n \precsim b_n$ means $a_n/b_n = O(1)$ and $a_n \prec b_n$ means that $a_n/b_n = o(1)$. For sequences of random scalars $a_n \precsim_p b_n$ means $a_n/b_n = O_p(1)$ and $a_n \prec_p b_n$ means $a_n/b_n = o_p(1)$. We say $a_n(x) \precsim_p b_n(x)$ uniformly (over x) if $\sup_x a_n(x)/b_n(x) \precsim_p 1$.

For any $s, c > 0$ let $\Lambda^d_s(c)$ be the space of smooth functions defined as follows. For any vector $q = (q_1, q_2, ..., q_d) \in \mathbb{N}^d_0$, let D_q be the partial derivative operator.

That is, for any scalar function δ on $\mathbb{R}^d$, $D_q[\delta](r) = \frac{\partial^{q_1+q_2+\ldots+q_d}}{\partial^{q_1} r_1 \partial^{q_2} r_2 \ldots \partial^{q_{\dim(R)}} r_{\dim(R)}} \delta(r)$. Then $\delta \in \Lambda_s^d(c)$ if any only if, for any $q \in \mathbb{N}_0^d$ with $\sum_{k=1}^d q_k \leq \lfloor s \rfloor$, $D_q[\delta]$ exists and has magnitude bounded uniformly by c, and for all $\sum_{k=1}^d q_k = \lfloor s \rfloor$ and $r_1, r_2 \in \mathbb{R}^d$:

$$|D_q[\delta](r_1) - D_q[\delta](r_2)| \leq c||r_1 - r_2||^{s-\lfloor s \rfloor}$$

Let $\tilde{\psi}_{n,i} = \psi_{n,i} \otimes \chi_{n,i}$, and then let $\tilde{\Omega}_n = E\big[(\tilde{\psi}_{n,i}\tilde{\psi}'_{n,i})\big]$. Let us define matrices $Q_n = E\big[\rho_{n,i}\rho'_{n,i}\big]$, $G_n = E\big[\chi_{n,i}\chi'_{n,i}\big]$, $\Omega_n = E\big[\psi_{n,i}\psi'_{n,i}\big]$, and $\Xi_n = E\big[\zeta_{n,i}\zeta'_{n,i}\big]$.

Let us define the scalars $\xi_{\rho,n} = \operatorname{ess\,sup} ||Q_n^{-1/2}\rho_{n,i}||$, $\xi_{\chi,n} = \operatorname{ess\,sup} ||G_n^{-1/2}\chi_{n,i}||$, $\xi_{\psi,n} = \operatorname{ess\,sup} ||\Omega_n^{-1/2}\psi_{n,i}||$, $\xi_{\zeta,n} = \operatorname{ess\,sup} ||\Xi_n^{-1/2}\zeta_{n,i}||$, and $\xi_{\tilde{\psi},n} = \operatorname{ess\,sup} ||\tilde{\Omega}_n^{-1/2}\tilde{\psi}_{n,i}||$.

Assumption 5.1 (Bases). i. Q_n is non-singular. For any $s > 0$, $\ell_{\rho,n}(s) \prec 0$ and uniformly over $c > 0$ and $\delta \in \Lambda_s^{\dim(V)}(c)$:

$$\inf_{\beta \in \mathbb{R}^{k(n)}} E\big[\big(\delta(V) - \rho_n(V)'\beta\big)^2\big]^{1/2} \precsim c\ell_{\rho,n}(s)$$

ii. **Either** X is has finite discrete support and for sufficiently large n any scalar function on $\mathcal{X}$ is a linear transformation of χ_n, **or** G_n is non-singular, and for any $s > 0$, $\ell_{\chi,n}(s) \prec 0$ and uniformly over $c > 0$ and $\delta \in \Lambda_s^{\dim(X)}(c)$, if $\beta^*[\delta]$ minimizes $E\big[\big(\delta(X) - \chi_n(X)'\beta\big)^2\big]$ then:

$$\operatorname{ess\,sup} \big|\delta(X) - \chi_n(X)'\beta^*[\delta]\big| \precsim c\ell_{\chi,n}(s)$$

iii. Ω_n is non-singular, and for any $s > 0$, $\ell_{\psi,n}(s) \prec 0$ and uniformly over $c > 0$ and $\delta \in \Lambda_s^{\dim(X,Z)}(c)$, if $\beta^*[\delta]$ minimizes $E\big[\big(\delta(X,Z) - \psi_n(X,Z)'\beta\big)^2\big]$:

$$\operatorname{ess\,sup} \big|\delta(X,Z) - \psi_n(X,Z)'\beta^*[\delta]\big| \precsim c\ell_{\psi,n}(s)$$

iv. Ξ_n is non-singular, and for any $s > 0$, $\ell_{\zeta,n}(s) \prec 0$ and uniformly over $c > 0$ and $\delta \in \Lambda_s^{\dim(X,Z)}(c)$, if $\beta^*[\delta]$ minimizes $E\big[\big(\delta(X,Z) - \zeta_n(X,Z)'\beta\big)^2\big]$:

$$\operatorname{ess\,sup} \big|\delta(X,Z) - \zeta_n(X,Z)'\beta^*[\delta]\big| \precsim c\ell_{\zeta,n}(s)$$

Assumption 5.2 (Densities, Conditional Variance). i. The joint distribution $F_{(X,Z,V)}$ is absolutely continuous with the product of the marginals $F_{(X,Z)} \otimes F_V$. ii. The Radon-Nikodym derivative $\frac{dF_{(X,V)}}{dF_X \otimes F_V}$ is bounded above and away from zero. iii. X has finite discrete support **or** X is continuously distributed on support $\mathcal{X} \in \mathbb{R}^{\dim(X)}$, X admits a probability density f_X that is bounded above and away from zero, and there exist $b > 0$ and $\underline{r} > 0$, so that for any $x \in \mathcal{X}$ and $0 < b' \leq b$, $vol(B_{x,b'} \cap \mathcal{X}) \geq \underline{r} vol(B_{0,b'})$, where $B_{x,b'}$ is the Euclidean ball of radius b' centered at x and $vol(\cdot)$ returns the volume. iv. there exists $\bar{\sigma}_Y < \infty$ so that with probability 1, $E[Y^2|X,Z] \leq \bar{\sigma}_Y^2$.

Assumption 5.3 (Smoothness). In each case, for $F_{(X,Z,V)}$-almost all (x,z,v), i. $\frac{dF_{(X,Z,V)}}{F_{(X,Z)} \otimes dF_V}(x,z,\cdot) \in \Lambda_{s_1}^{\dim(V)}(c_1)$ and $\frac{dF_{(X,Z,V)}}{F_{(X,Z)} \otimes dF_V}(\cdot,z,v) \in \Lambda_{s_2}^{\dim(X)}(c_2)$ ii. $g(\cdot,\cdot) \in \Lambda_{s_3}^{\dim(X)}(c_3)$, iii. $\frac{dF_{(X,Z,V)}}{F_{(X,Z)} \otimes dF_V}(\cdot,\cdot,v) \in \Lambda_{s_4}^{\dim(X,Z)}(c_4)$, iv. Either X has finite discrete support or $\frac{dF_{(X,V)}}{dF_X \otimes F_V}(\cdot,v) \in \Lambda_{s_5}^{\dim(X,Z)}(c_5)$.

Assumption 5.4 (Sieve Growth). i. $\frac{\xi_{\chi,n}^2 \log(l(n))}{n_\alpha} \prec 1$, ii. $\frac{\xi_{\psi,n}^2 \log(m(n))}{n_\pi} \prec 1$, iii. $\frac{\xi_{\zeta,n}^2 \log(p(n))}{n_g} \prec 1$, iv. $\frac{\xi_{\psi,n}^2 \log(m(n)l(n))}{n_g} \prec 1$

Assumption 5.1 specifies the rate at which the basis functions can approximate smooth functions. Precise bounds for particular basis functions can be found in the approximation literature (see [34]). [13]

Assumption 5.2.i is a weak condition on the joint distribution of V and (X,Z). It is satisfied if, for example, V, X, and Z have a non-zero joint probability density on a rectangular support. Assumption 5.2.ii holds if, for example, X and V have joint probability density that is bounded above and away from zero on a rectangular support. The condition on the support of X in the continuous case in 5.2.iii is very weak condition on the boundary. It holds, for example, if $\mathcal{X}$ is rectangular. 5.2.iv is standard in the non-parametric regression literature.

Assumption 5.3 imposes that some reduced-form objects be smooth. Note the assumption does not directly impose smoothness on any structural objects (such as

[13]For example, if ρ_n, χ_n, ψ_n, and ζ_n are spline series, local polynomial partition series, or Cohen-Daubechies-Vial wavelets of order $s_0 \geq s$, then $\xi_{\psi,n}^2 \precsim m(n)$, $\xi_{\rho,n}^2 \precsim k(n)$, $\xi_{\chi,n}^2 \precsim l(n)$, $\xi_{\zeta,n}^2 \precsim p(n)$, $\ell_{\psi,n}(s) \precsim m(n)^{-s/\dim(X,Z)}$, $\ell_{\zeta,n}(s) \precsim p(n)^{-s/\dim(X,Z)}$, $\ell_{\rho,n}(s) \precsim k(n)^{-s/\dim(V)}$, and $\ell_{\chi,n}(s) \precsim l(n)^{-s/\dim(X)}$ (see for example [26]).

γ or φ from Theorem 1.1). Assumption 5.4 restricts the rate at which the numbers of basis functions can grow. This assumption allows us to apply Rudelson's matrix law of large numbers ([70]).

Theorem 3.1 establishes a rate of convergence for the estimator in terms of the first stage convergence rates. Theorem 3.1 refers to rates $R_{g,n}$, $R_{\pi,n}$, $R_{\pi,n}(x_1)$, and $R_{\alpha,n}(x_1,x_2)$ which are rates of convergence for $\hat{g}_i$, $\hat{\pi}_{n,i}$, and $\hat{\alpha}_n$ respectively. In particular we let $\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\left(\hat{g}_i-g_i\right)^2 \precsim_p R_{g,n}^2$ and $\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}|(\pi_{n,i}-\hat{\pi}_{n,i})'\theta|^2 \precsim_p R_{\pi,n}^2$.

Further, uniformly over F_X-almost all x_1 and x_2 and over all $\theta\in\mathbb{R}^{k(n)l(n)}$ with $E\left[|\phi_n(X,V)'\theta|^2\right]=1$:

$$\begin{aligned} E_Z\left[|(\pi_{n,i}-\hat{\pi}_{n,i})'\theta|^2\middle|X=x_1\right] &\precsim_p R_{\pi,n}(x_1)^2 \\ \left(\hat{\alpha}_n(x_1,x_2)-\alpha_n(x_1,x_2)\right)'\theta &\precsim_p R_{\alpha,n}(x_1,x_2) \end{aligned}$$

Finally, let $\underline{\mu}_n^2=\mu_{min}(Q_n)\mu_{min}(G_n)$ and $\bar{\mu}_n^2=\mu_{max}(Q_n)\mu_{max}(G_n)$ for $\mu_{min}(Q_n)$ and $\mu_{max}(Q_n)$ the smallest and largest eigenvalues of Q_n and likewise for G_n. Define $\xi_{\tilde{\Omega},n}(x_1)=||E[\tilde{\psi}_{n,i}\tilde{\psi}_{n,i}'|X=x_1]^{1/2}\tilde{\Omega}_n^{-1/2}||$ and define $\xi_{\Omega,n}(x_1)$ by $\xi_{\Omega,n}(x_1)=||E[\psi_{n,i}\psi_{n,i}'|X=x_1]^{1/2}\Omega_n^{-1/2}||$.

Theorem 3.1 (Convergence). *Suppose Assumptions 1-4 hold with $D(\cdot)$ and $C(\cdot,\cdot)$ in Assumption 4 uniformly bounded and that Assumptions 5.1.i-ii, 5.2.i-iii, 5.3.i-ii, 5.3.iv, and 5.4.iv hold. Let $\hat{\theta}$ be defined as in (1.3) and let $\phi_n(x,v)=\rho_n(v)\otimes\chi_n(x)$. For any first-stage estimators $\hat{g}_i$, and $\hat{\alpha}_{n,i}$ and any series estimator $\hat{\pi}_{n,i}$ of the form $\hat{\pi}_{n,i}=(\hat{\omega}_n'\psi_{n,i})\otimes\chi_{n,i}$ with $\hat{\omega}_n$ a matrix of estimated coefficients:*

$$\begin{aligned} &\bar{y}(x_1|x_2)-\hat{\alpha}_n(x_1,x_2)'\hat{\theta} \\ \precsim_p &\left(\frac{\bar{\mu}_n}{\underline{\mu}_n}+\frac{\bar{\mu}_n}{\lambda_{0,n}^{1/2}}(R_{\theta,n}+R_{\pi,n}+R_{g,n})\right)\left(R_{\pi,n}(x_1)+R_{\alpha,n}(x_1,x_2)\right) \\ &+\xi_{\tilde{\Omega},n}(x_1)(R_{\theta,n}+R_{\pi,n}+R_{g,n}+\frac{\lambda_{0,n}^{1/2}}{\underline{\mu}_n})+R_{\theta,n} \end{aligned}$$

Where $R_{\theta,n}=\ell_{\rho,n}(s_1)$ if X has finite discrete support and otherwise we have $R_{\theta,n}=\ell_{\rho,n}(s_1)+\left(\xi_{\rho,n}\ell_{\chi,n}(1)\right)^{\tilde{s}}$, for $\tilde{s}=\frac{\min\{s_2,s_3,1\}}{\min\{s_2,s_3,1\}+1}$.

Theorem 3.2 (First-Stage Rates). *Suppose Assumptions 5.1-5.4 all hold and $\hat{g}$, $\hat{\pi}_n$, and $\hat{\alpha}_n$ are the series ridge estimates given by (1.3.3), then:*

$$\begin{aligned}
R_{g,n} =& \sqrt{p(n)/n_g} + \ell_{\zeta,n}(s_3) + \lambda_{1,n}||\Xi_n^{-1}|| \\
R_{\pi,n} =& \xi_{\chi,n}\big(\sqrt{\xi_{\psi,n}^2 k(n) \wedge \xi_{\rho,n}^2 m(n)/n_\pi} + \ell_{\psi,n}(s_4)\big) + \lambda_{2,n}||\Omega_n^{-1}|| \\
R_{\pi,n}(x_1) =& \xi_{\Omega,n}(x_1)R_{\pi,n} + \big(1 - \xi_{\Omega,n}(x_1)\big)\xi_{\chi,n}\ell_{\psi,n}(s_4) \\
R_{\alpha,n}(x_1,x_2) =& \xi_{\chi,n}\big(1 + \sqrt{\xi_{\chi,n}^2 l(n)/n_\alpha}\big)\ell_{\chi,n}(s_5) + \lambda_{3,n}\xi_{\chi,n}^2||G_n^{-1}|| \\
&+ \xi_{\chi,n}^2\sqrt{\xi_{\chi,n}^2 k(n) \wedge \xi_{\rho,n}^2 l(n)/n_\alpha}
\end{aligned}$$

While our estimator is of the PSMD type, our convergence rate results differ markedly from those of PSMD estimators of the structural function in NPIV models. In particular, our results do not depend on a sieve-measure of ill-posedness. In fact, Theorem 3.1 suggests rates of convergence comparable to standard nonparametric regression are attainable. Suppose $\lambda_{0,n}$ is chosen to optimize the rate in Theorem 3.1, then rate simplifies to:[14]

$$\begin{aligned}
&\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta} \\
\precsim_p& \frac{\bar{\mu}_n}{\underline{\mu}_n}\big(R_{\pi,n}(x_1) + R_{\alpha,n}(x_1,x_2)\big) + \xi_{\tilde{\Omega},n}(x_1)(R_{\theta,n} + R_{\pi,n} + R_g) + R_{\theta,n}
\end{aligned}$$

For many commonly used bases $\frac{\bar{\mu}_n}{\underline{\mu}_n}$ is bounded above and away from zero under very weak conditions ([17]). This must be true in the case of discrete and finitely supported treatments. Furthermore, with discrete finitely supported treatments $\xi_{\tilde{\Omega},n}(x_1)$ is bounded above, and so with the penalties chosen optimally, the estimator converges as quickly as either the slowest first-stage nonparametric regression or $R_{\theta,n}$. Moreover, in the finite discrete treatment case $l(n)$, $\xi_{\Omega,n}(x_1)$, and $\xi_{\chi,n}$ are also bounded above uniformly over n, and $\ell_{\chi,n} = 0$ for large enough n. Consider the ridge series estimator using as basis functions all the interactions between indicators for the possible treat-

[14] The $\lambda_{0,n}$ that optimizes the rate in Theorem 3.1 is given below:

$$\lambda_{0,n} = \xi_{\tilde{\Omega},n}(x_1)^{-1}\underline{\mu}_n\bar{\mu}_n(R_{\theta,n} + R_{g,n} + R_{\pi,n})\big(R_{\pi,n}(x_1) + R_{\alpha,n}(x_1,x_2)\big)$$

ments and either spline series, local polynomial partition series, or Cohen-Daubechies-Vial wavelets of order $s_0 \geq s = \min\{s_1, s_3, s_4\}$. Then $\xi^2_{\psi,n} \precsim m(n)$, $\xi^2_{\rho,n} \precsim k(n)$, $\xi^2_{\zeta,n} \precsim p(n)$, $\ell_{\psi,n}(s) \precsim m(n)^{-s/\dim(Z)}$, $\ell_{\zeta,n}(s) \precsim p(n)^{-s/\dim(Z)}$, $\ell_{\rho,n}(s) \precsim k(n)^{-s/\dim(V)}$. If penalties are chosen optimally and each subsample grows at rate n, then the estimates converge at rate:

$$\sqrt{k(n)m(n)/n} + \sqrt{p(n)/n} + m(n)^{-s/\dim(V)} + p(n)^{-s/\dim(Z)} + k(n)^{-s/\dim(V)}$$

We can choose $k(n)$, $m(n)$, and $p(n)$ to optimize the rate above. The optimal rate is $n^{\frac{-s}{\dim(V)+\dim(Z)+2s}}$. Note that as the smoothness coefficient s goes to infinity this approaches the root-n parametric rate.

Theorems 3.1 and 3.2 together imply that if $\ell_{\rho,n}(s)$, $\ell_{\chi,n}(s)$, $\ell_{\psi,n}(s)$, and $\ell_{\zeta,n}(s)$ each go to zero exponentially quickly in s, and if the penalty parameters, number of basis functions, and sub-sample sizes are set optimally, then as the smoothness parameters approach infinity the convergence rate approaches $n^{-1/2}$.

The term $R_{\theta,n}$ in Theorem 3.1 accounts for the approximation error due to the use of a finite vector of basis functions ϕ_n. Remarkably, this term is guaranteed to converge quickly to zero under smoothness conditions on only reduced-form objects.[15] This allows us to avoid directly placing any smoothness restrictions on objects like γ and φ in Theorem 1.1 which do not generally have a simple structural interpretation.

Theorem 1.2 plays a key role in our analysis. Recall that Theorem 1.2 establishes that under our identifying assumptions, our estimation problem is well-posed. This is why a sieve measure of ill-posedness does not show up in the convergence rate and we are able to achieve rates comparable to standard nonparametric regression.

Let us briefly discuss the role of penalization. The first-stage rates in Theorem 3.2 are optimized by setting the first-stage penalties equal to zero or letting the penalties go to zero sufficiently quickly. Nonetheless, it may be useful to penalize the first-stage regressions in finite samples. By contrast, the second stage penalty parameter $\lambda_{0,n}$ that optimizes the rate in Theorem 3.1 goes to zero at a rate comparable to the

[15] see Lemma C.9 in the supporting materials, which may be of independent interest

convergence rates of the first stage regressions. The second-stage penalization in our problem is not used to regularize an ill-posed inverse problem (as it may be in NPIV). Instead, penalization in the second stage prevents a norm the mean-square norm of the function $\phi_n(x_1, \cdot)'\hat{\theta}$, given by $E_V[|\phi_n(x_1, V)'\hat{\theta}|^2]^{1/2}$, from blowing up too quickly. This norm matters because the final estimator $\hat{\alpha}_n(x_1, x_2)'\hat{\theta}$, is a fitted-value from regression of $\phi_n(x_1, V_i)'\hat{\theta}$ on X_i. The variance (conditional on $\hat{\theta}$) of the dependent variable is bounded by $E_V[|\phi_n(x_1, V)'\hat{\theta}|^2]^{1/2}$.

1.3.2 Asymptotic Normality

We now provide conditions under which the estimated CASF can be asymptotically approximated by a Gaussian process. The Gaussian approximation motivates a multiplier bootstrap method for constructing uniform confidence bands.

The results in this subsection apply specifically to the series ridge version of our estimator with sample splitting. Crucially, we require that π_n be estimated using an entirely separate sub-sample from that used to estimate g, formally $\mathcal{I}_g \cap \mathcal{I}_\pi = \emptyset$. We also require additional conditions stated below.

Assumption 5.5 (Inference). i. There is a sequence $R_{\mathcal{N},n} \prec 1$ that satisfies $p(n)^2 \xi_{\zeta,n} n_g^{-1/2} R_{\mathcal{N},n}^{-3} log(n_g) \to 0$. ii. $E[(Y_i - g_i)^2 | X_i, Z_i] \geq \underline{\sigma}_Y^2 > 0$ with probability 1. iii. $E[(Y_i - g_i)^3 | X_i, Z_i] \leq \bar{\kappa}_3 < \infty$ with probability 1. iv. $\frac{\sqrt{ln(n_g)} \xi_{\zeta,n} log(p(n))}{n_g} \prec 1$. v. $E[(Y_i - g_i)^4 | X_i, Z_i] \leq \bar{\kappa}_4 < \infty$ with probability 1.

Assumption 5.5 i. restricts the growth of $p(n)$ relative to n_g. This condition is used in our application of Yurinskii's coupling, which plays the role of a central limit theorem in our analysis. 5.5.ii helps us lower bound the variance of the Gaussian approximation. 5.5.iii bounds the conditional third moment of Y_i for use in Yurinskii's coupling. 5.5.iv slightly strengthens Assumption 5.4.iii, and 5.5.v bounds the fourth moment of Y_i.

To state Theorem 3.3 we must introduce some additional notation. Let $\tilde{\alpha}_n(x_1, x_2) = E\left[\varphi(x_1, x_2, Z_i)\hat{\pi}_{n,i} \middle| \mathcal{I}_\pi, X_i = x_1\right]$ where φ is the function in Theorem 1.1.ii. Note that the expectation in the definition of $\tilde{\alpha}_n$ is conditional on $\mathcal{I}_\pi$, by which we mean we condition on all the observations in the sub-sample $\mathcal{I}_\pi$. Let $\Gamma_n = E\big[\zeta_{n,i}\zeta_{n,i}'(Y_i - \zeta_{n,i}'\Xi_n^{-1}E[\zeta_{n,i}g_i])^2\big]$, $\bar{\Sigma}_n = E\big[\hat{\pi}_{n,i}\hat{\pi}_{n,i}'|\mathcal{I}_\pi\big]$, and $\bar{\Sigma}_{\lambda_{0,n}} = \bar{\Sigma}_n + \lambda_{0,n}I$. Then define the vector-valued function s_n by:

$$s_n(x_1,x_2)' = \tilde{\alpha}_n(x_1,x_2)'\bar{\Sigma}_{\lambda_{0,n}}^{-1}E\big[\hat{\pi}_{n,i}\zeta_{n,i}'\big|\mathcal{I}_\pi\big]\Xi_n^{-1}\Gamma_n^{1/2}$$

Define $b_n = ||E\big[\Xi_n^{-1/2}\zeta_{n,i}\tilde{\psi}_{n,i}'\tilde{\Omega}_n^{-1/2}\big]^{-1}||$, and let $\underline{c}_n$ be a deterministic function that satisfies:

$$\underline{c}_n(x_1,x_2) \precsim_p ||E\big[\Xi_n^{-1/2}\zeta_{n,i}\big(\hat{\pi}_{n,i}'\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}(x_1,x_2)\big)\big|\mathcal{I}_\pi\big]||$$

Theorem 3.3 (Asymptotic Normality). *Suppose Assumptions 1-5.5.iii all hold with $D(\cdot)$ and $C(\cdot,\cdot)$ in Assumption 4 uniformly bounded. Let $\hat{\theta}$ be defined as in (1.3) and $\hat{g}$, $\hat{\pi}_n$, and $\hat{\alpha}_n$ as in (1.3.3) with $\mathcal{I}_g \cap \mathcal{I}_\pi = \emptyset$. Then for each n there is a length-$k(n)l(n)$ random vector $\mathcal{N}_n \sim N(0,I)$ that is independent of the observations in sub-sample $\mathcal{I}_\pi$ so that uniformly:*

$$\begin{aligned}&\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{||s_n(x_1,x_2)||} - \frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||}\mathcal{N}_n \\ &\precsim_p R_{\mathcal{N},n} + \xi_{\zeta,n}\frac{p(n)}{n_g} + b_n(R_{\theta,n} + R_{\pi,n} + \ell_{\zeta,n}(s_3) + \lambda_{1,n}||\Xi_n^{-1}||) \\ &+\frac{\sqrt{n_g}}{\underline{c}_n(x_1,x_2)}R_n(x_1) \end{aligned} \tag{1.3.4}$$

The remainder term $R_n(x_1)$ is given by:

$$\begin{aligned}&R_{\theta,n} + \big(\frac{\bar{\mu}_n}{\underline{\mu}_n} + \frac{\bar{\mu}_n}{\lambda_{0,n}^{1/2}}(R_{\theta,n} + R_{\pi,n} + R_{g,n})\big)\big(R_{\pi,n}(x_1) + R_{\alpha,n}(x_1,x_2)\big) \\ &+\xi_{\tilde{\Omega},n}(x_1)\sqrt{\frac{p(n)}{n_g}}\big(\sqrt{\xi_{\zeta,n}^2 k(n)l(n)/n_g} + \big(\xi_{\tilde{\psi},n}^2 log\big(l(n)m(n)\big)/n_g\big)^{1/4}\big) \\ &+\xi_{\tilde{\Omega},n}(x_1)\lambda_{0,n}^{1/2}\underline{\mu}_n^{-1}\end{aligned}$$

Theorem 3.3 gives a rate at which the normalized estimation error can be approximated by a zero mean Gaussian process. The term b_n is often simple to bound, for example $b_n = 1$ whenever $\tilde{\Omega}_n^{-1/2}\tilde{\psi}_n(x,z)$ is a subvector of $\Xi_n^{-1/2}\zeta_n(x,z)$.

The sequence $\underline{c}_n(x_1,x_2)^2$ is (under our assumptions) a lower bound on the rate of the conditional asymptotic variance $||s_n(x_1,x_2)||^2$. Thus $\frac{\sqrt{n_g}}{\underline{c}_n(x_1,x_2)}R_n(x_1)$ is the ratio of an asymptotic bias term $\sqrt{n_g}R_n(x_1)$ and the rate of the asymptotic variance.[16] Thus, for $\frac{\sqrt{n_g}}{\underline{c}_n(x_1,x_2)}R_n(x_1)$ to converge to zero one must 'under-smooth'. Under-smoothing is a common strategy for inference in non-parametric models, to under-smooth one lets the bias decrease more quickly than the variance. $\underline{c}_n(x_1,x_2)$ is the norm of a length $p(n)$ vector and so, loosely speaking, letting $p(n)$ grow sufficiently quickly helps to ensure $\underline{c}_n(x_1,x_2)$ shrinks to zero slowly enough that $\frac{\sqrt{n_g}}{\underline{c}_n(x_1,x_2)}R_n(x_1) \to 0$.

1.3.3 Uniform Confidence Bands

If the RHS of (1.3.4) goes to zero and a consistent estimate of the asymptotic variance is available. Then Theorem 3.3 immediately implies that asymptotically valid pointwise inference on $\bar{y}(x_1|x_2)$ can be achieved in the usual way using a zero-mean Gaussian approximation. For uniform inference we describe a multiplier bootstrap procedure for constructing confidence bands that are asymptotically valid uniformly over all (x_1,x_2) in a set $\mathcal{X}$.

Let $\hat{\theta}$ be defined as in (1.3) and let $\hat{g}$, $\hat{\pi}_n$, and $\hat{\alpha}_n$ be given by the formulas in (1.3.3) with $\mathcal{I}_g \cap \mathcal{I}_\pi = \emptyset$. Let $Q_{b,i}$ for each $i \in \mathcal{I}_g$ and $b \in \{1,...,B\}$ be iid standard exponential random variables that are independent of the data. [17] The b^{th} multiplier bootstrap estimate is $\hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b$ where $\hat{\theta}_b$ is given by $\hat{\Sigma}^{-1}_{\lambda_{0,n}}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}\hat{g}_{i,b}$ where $\hat{g}_{i,b} = \zeta'_{n,i}\hat{\beta}_b$ for $\hat{\beta}_b$ is defined by:

$$\hat{\beta}_b = \big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g} Q_{b,i}\zeta_{n,i}\zeta'_{n,i} + \lambda_{1,n}I\big)^{-1}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g} Q_{b,i}\zeta_{n,i}Y_i$$

The goal of the bootstrap procedure is to approximate the distribution of $\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}$ using the distribution of $\hat{\alpha}_n(x_1,x_2)'\hat{\theta} - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b$. A size-$a$ bootstrap

[16]Strictly speaking $R_n(x_1)$ captures more than just bias. It captures some of the error of the Gaussian approximation which can also be due to noise in the first-stage estimation. We refer to it as bias because in the context of inference in our setting it plays a role analogous to bias in nonparametric regression.

[17]We follow ([17]) and use the standard exponential, but $Q_{b,i}$ may have any distribution with mean and variance both equal to 1 and $\max_{i\in\mathcal{I}_g}|Q_{b,i}| \precsim_p ln(n_g)$.

confidence band is an interval-valued random function $\hat{\Theta}_{1-a}$ of the form:

$$\hat{\Theta}_{1-a}(x_1,x_2) = \Big[\hat{\alpha}_n(x_1,x_2)'\hat{\theta} - \frac{\hat{\sigma}(x_1,x_2)}{\sqrt{n_g}}\hat{c}_{1-a}, \hat{\alpha}_n(x_1,x_2)'\hat{\theta} + \frac{\hat{\sigma}(x_1,x_2)}{\sqrt{n_g}}\hat{c}_{1-a}\Big]$$

Where $\hat{\sigma}(x_1,x_2) > 0$ is an estimate of $||s_n(x_1,x_2)||$ and $\hat{c}_{1-a}$ is a critical value. The uniform (over all pairs (x_1,x_2) in a set $\mathcal{X}$) critical value is the smallest scalar $c > 0$ that satisfies the inequality below:

$$\frac{1}{B}\sum_{b=1}^{B} 1\big\{\sup_{x_1,x_2\in\mathcal{X}} |\frac{\hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}}{\hat{\sigma}(x_1,x_2)/\sqrt{n_g}}| \leq c\big\} \geq 1-a$$

Theorem 3.4 (Uniform Inference). *Suppose Assumptions 1-5 all hold with $D(\cdot)$ and $C(\cdot,\cdot)$ in Assumption 4 uniformly bounded. Suppose that for some $\bar{r}_n \to 0$:*

$$\sup_{(x_1,x_2)\in\mathcal{X}} |R_{\mathcal{N},n} + \frac{\xi_{\zeta,n}p(n)}{\sqrt{n_g}} + b_n\big(R_{\theta,n} + R_{\pi,n}\big) + \frac{\sqrt{n_g}}{\underline{c}_n(x_1,x_2)}R_n(x_1) + b_n(\ell_{\zeta,n}(s_3) + \lambda_{1,n}||\Xi_n^{-1}||)| \prec_p \bar{r}_n$$

Where $R_{\mathcal{N},n}$, $R_{\theta,n}$, $R_{\pi,n}$ and $R_n(x_1)$ are as defined in Theorem 3.3. Further, suppose $\sup_{(x_1,x_2)\in\mathcal{X}} |\frac{s_n(x_1,x_2)}{||s_n(x_1,x_2)||}'\mathcal{N}_n| < \infty$ almost surely, $\sup_{(x_1,x_2)\in\mathcal{X}} |\frac{\hat{\sigma}(x_1,x_2)}{||s_n(x_1,x_2)||} - 1| \prec_p \bar{r}_n^2$, and:

$$E\big[\sup_{(x_1,x_2)\in\mathcal{X}} |\frac{s_n(x_1,x_2)}{||s_n(x_1,x_2)||}'\mathcal{N}_n|\big] \lesssim \bar{r}_n^{-1}$$

then if $B \to \infty$ sufficiently quickly with the sample size:

$$P\big(\bar{y}(x_1|x_2) \in \hat{\Theta}_{1-a}(x_1,x_2),\, \forall (x_1,x_2) \in \mathcal{X}\big) = 1 - a + o(1)$$

The conditions in Theorem 3.4 on the Gaussian process $\frac{s_n(x_1,x_2)}{||s_n(x_1,x_2)||}'\mathcal{N}_n$ are needed in order to apply results from [26]. In the case in which $\mathcal{X}$ is a finite set they hold trivially.

1.4 Empirical Applications

We apply our methodology to real data. In order to emphasize the applicability of our approach to both cross-sectional and panel models we present two separate empirical settings. In our first application we use cross-sectional variation to estimate causal effects, and in the second application we exploit the panel structure of the data. Estimation was carried out using the series ridge version of our estimator with sample-splitting. In all cases we use as our basis functions the first, second, and third powers of each variable and up to three-way interactions (this includes say, the interaction between the square of one variable and one other variable). In each case the penalty parameter is set to the square root of the size of the subsamples.

1.4.1 Causal Impact of Grade Retention

[42] examine the causal effect of being made to repeat a particular grade level on the cognitive development of US students. They use data from the ECLS-K panel study which contains panel data on the early cognitive development of US children. We use our methods to examine the effect of grade retention on the cognitive outcomes of children in the 1998-1999 kindergarten school year using cleaned data available with their paper. Following [42], we take our outcome variables to be the tests scores in reading and math when aged approximately eleven. Also in line with [42] our treatments are indicators for retention in kindergarten, 'early' (in first or second grade) and 'late' (in third or fourth grade). The cleaned data from [42] contains only students who are retained at most once in the sample period and no students who skip a grade.

The ECLS-K dataset contains scores that measure a student's behavioral and social skills and their scores on a range of cognitive tests at different ages. To account for the confounding effect of unmeasured ability (see the arguments in Section 1), [42] estimate a latent factor model with a particular structure. They assume that all confounding between grade retention and potential future cognitive test scores is due entirely to the presence of three latent factors representing different dimensions of ability. [42] then use test scores to recover the distribution of the latent factors and their loadings. They assume a particular multiplicative structure between the factors (which are time-invariant) and time-specific factor loadings in both their outcome and selection equations.

In our approach, W^* represents underlying propensity to perform well academically, and plays a role analogous to the latent factors in their analysis. Our methods allow us to avoid any strong assumptions on the factor structure. As we discuss in Section 1, we let the set of proxies V contain the student's scores on the cognitive and behavioral tests in kindergarten and Z contain the scores from early in elementary school (first or second grade). See Section 1 for discussion of our identifying assumptions in this context.

We apply the estimation method set out in Section 3 to estimate average effects of retention at different grades. Table 1 below presents our results, and corresponds roughly to Table 4 in [42].

Consider the average effects of treatment on the treated for each of the three treatments (retention in kindergarten, early in elementary, and late in elementary). These are highlighted in sky blue in Table 1. We estimate that retention in kindergarten and retention in first or second grade lowers the average scores for reading and math for those students who were in fact retained at these ages, we estimate that retention late in elementary raises scores in both reading and math. The figures for retention early in elementary are statistically significant at the 99% level, but the estimated effects of treatment on the treated for both retenition in kindergarten and late in elementary are statistically insignificant. These findings [42] who estimate mostly negative ATTs.

The first column in each table gives the counterfactual effects of retention at different ages for those students who were not retained at any of the ages covered in our data. We find a mix of positive and negative effects. Only two of these counterfactual estimates are statistically significant at the 95% level: the negative estimated effect of early elementary retention on the reading scores of those not retained, and the positive effect of retention in kindergarten on the math scores of

Table 1.1: Effects of Grade Retention on Cognitive Performance

(a) Reading

$n = 1948$ **Difference from non-retention:**	Observed retention status: Not retained	Retained kindergarten	Retained early	Retained late
Retained kindergarten	0.01	-0.06	-0.03	-0.01
	(0.04)	(0.04)	(0.05)	(0.04)
Retained early	-0.11	-0.13	-0.16	-0.03
	(0.04)	(0.02)	(0.03)	(0.02)
Retained late	-0.01	0.04	0.02	0.00
	(0.03)	(0.02)	(0.02)	(0.01)

(b) Math

$n = 1951$ **Difference from non-retention:**	Observed retention status: Not retained	Retained kindergarten	Retained early	Retained late
Retained kindergarten	0.08	-0.01	0.02	-0.02
	(0.04)	(0.03)	(0.03)	(0.03)
Retained early	-0.01	-0.09	-0.09	-0.07
	(0.04)	(0.03)	(0.03)	(0.03)
Retained late	-0.02	-0.03	0.01	0.01
	(0.04)	(0.04)	(0.04)	(0.03)

Estimates of the treatment effects for groups with different treatment statuses. Numbers in parentheses are standard errors calculated as the standard deviation of the estimates over 10,000 replications of the multiplier bootstrap method detailed in Section 3. The sample size n differs because for some individuals not all three outcomes are available.

those not retained. [42] estimate positive effects of all treatments on both the reading and math scores of those students who were not retained.

Table 2 compares estimates of the average effect of treatment on the treated under different approaches to estimation. In the first column are linear least-squares estimates of the average treatment effects when no additional covariates are included in the regression. This amounts to a simple difference in means. The estimated effects are all strongly negative and very statistically significant. In the second column, the kindergarten cognitive and behavioral scores are included as regressors in a linear specification, note that in every case the estimated negative effects are at least halved in magnitude compared to the case in which the scores are not included.

In the third column we apply our method with linear specifications (i.e., ρ_n, ψ_n, ζ_n, and χ_n return their arguments) and without any penalization, that is, $\lambda_{0,n} = \lambda_{1,n} = \lambda_{2,n} = \lambda_{3,n} = 0$. In every case the the estimated ATTs are more positive

Table 1.2: Average Effects of Treatment on the Treated

(a) Reading

Treatments:	Method: OLS, no proxy scores	OLS, kindergarten scores	Proxy controls, linear	Proxy controls, PSMD
Retained kindergarten	-0.22	-0.08	-0.04	-0.06
	(0.03)	(0.02)	(0.03)	(0.04)
Retained early	-0.31	-0.15	-0.11	-0.16
	(0.02)	(0.02)	(0.03)	(0.03)
Retained late	-0.19	-0.04	0.06	0.00
	(0.04)	(0.03)	(0.05)	(0.01)

(b) Math

Treatments:	Method: OLS, no proxy scores	OLS, kindergarten scores	Proxy controls, linear	Proxy controls, PSMD
Retained kindergarten	-0.22	-0.05	-0.01	-0.01
	(0.03)	(0.02)	(0.03)	(0.03)
Retained early	-0.33	-0.16	-0.06	-0.09
	(0.03)	(0.02)	(0.03)	(0.03)
Retained late	-0.24	-0.06	0.06	0.01
	(0.04)	(0.03)	(0.05)	(0.03)

Estimates of the average effect of treatment on the treated using various estimation methods. Numbers in parentheses are standard errors, calculated for the OLS estimates using the standard heteroskedasticity robust formula, and for the latter two columns using $10,000$ replications of the multiplier bootstrap method detailed in Section 3.

than when the kindergarten scores are treated as perfect controls. Finally, the last column contains the ATTs from the ridge series estimates in Table 1. Compared to the linear case the ATTs are more negative, although the estimated effect of retention in kindergarten on math outcomes is unchanged up to two decimal places. Note that the estimated effect of early retention on reading outcomes is slightly more negative using our method than it is in the linear specification with kindergarten scores treated as perfect controls.

The results in Table 2 are consistent with the notion that unmeasured ability biases the estimated ATTs downwards. Including kindergarten test scores as controls mitigates some of this bias, and the proxy controls method mitigates the bias further still, resulting in mostly less negative estimated ATTs.

The ECLS-K dataset includes additional covariates that measure features of the child, the child's school, and the child's household. Table 3 contain estimated ATTs when these additional covariates are included in each specification. Due to the large number of these additional covariates it is infeasible to include powers or interations of these terms, and so in our PSMD estimation we only allow these to enter linearly into each regression. [42] includes these covariates in their specification and so we include their results in the final column for comparison.

Again, moving from left to right across the first three columns of Table 3 we see increasingly positive estimates of all ATTs. The estimates for the non-linear proxy controls specifications are in some cases more positive and in some cases more negative than in the linear proxy controls specification. In this case the non-linear proxy controls estimates are always more positive than both corresponding OLS estimates.

Table 1.3: Average Effects of Treatment on the Treated (Additional Controls)

(a) Reading

Treatments:	**Method:** OLS, no proxy scores	OLS, kindergarten scores	Proxy controls, linear	Proxy controls, PSMD	Freuhwirth, Navarro & Takahashi
Retained kindergarten	-0.17	-0.09	-0.06	-0.05	-0.06
	(0.02)	(0.02)	(0.02)	(0.03)	(0.01)
Retained early	-0.23	-0.14	-0.12	-0.12	-0.11
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Retained late	-0.13	-0.04	0.00	-0.01	0.02
	(0.03)	(0.03)	(0.04)	(0.02)	(0.08)

(b) Math

Treatments:	**Method:** OLS, no proxy scores	OLS, kindergarten scores	Proxy controls, linear	Proxy controls, PSMD	Freuhwirth, Navarro & Takahashi
Retained kindergarten	-0.18	-0.07	-0.04	0.00	-0.06
	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)
Retained early	-0.24	-0.14	-0.07	-0.06	-0.10
	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)
Retained late	-0.17	-0.08	0.03	-0.02	-0.05
	(0.04)	(0.03)	(0.04)	(0.04)	(0.26)

Estimates of the average effect of treatment on the treated using various estimation methods. In each specification 36 additional covariates were allowed to enter linearly into each regression. These covariates are those other than the cognitive and behavioral scores listed in Table 1 of [42] and (following [42]) the square of age. Numbers in parentheses are standard errors, calculated for the OLS estimates using the standard heteroskedasticity robust formula, and for the third and fourth columns using $10,000$ replications of the multiplier bootstrap method detailed in Section 3.The figures in the final column are taken directly from [42].

1.4.2 Structural Engel Curve for Food

A household's Engel curve for a particular class of good captures the relationship between the share of the household's budget spent on that class and the total expenditure of the household. An Engel curve is 'structural' if it captures the effect of an exogenous change in total expenditure. Imagine an ideal experiment in which the household's total expenditure is chosen by a researcher using a random number generator and the household then chooses how to allocate that total expenditure between different classes of goods. Then the resulting relationship between the total expenditure and budget share is a structural Engel curve.

Nonparametric regression of the budget share spent on food and the total expenditure on certain classes of goods is unlikely to represent the average structural Engel curve. This is because, as we argue in Section 1, total expenditure is chosen by the household and thus depends upon the household's underlying consumption preferences. These same preferences partially determine the household's expenditure on food.

We estimate average structural Engel curves for food eaten at home using data from the Panel Study of Income Dynamics (PSID). The PSID study follows US households over a number years and record expenditure on various classes of goods. We use ten periods of data from the surveys carried out every two years between 1999 and 2017. We drop all households whose household heads are not married or cohabiting and drop all households for which we lack the full ten periods of data, leaving us with 840 households. We take as the total expenditure the sums of expenditures on food (both at home and away from home), housing, utilities, transportation, education, childcare and health-care.

We apply the approach to identification with fixed-T panels described in Section 2, in particular the Markov treatment assignment and predetermination case. Let X_t denote the total expenditure in period t, which is the treatment in this setting. We aim to estimate the average and conditional average structural Engel curve for period T. Let $\tilde{W}^*$ consist of factors that capture heterogeneity in household preferences. In

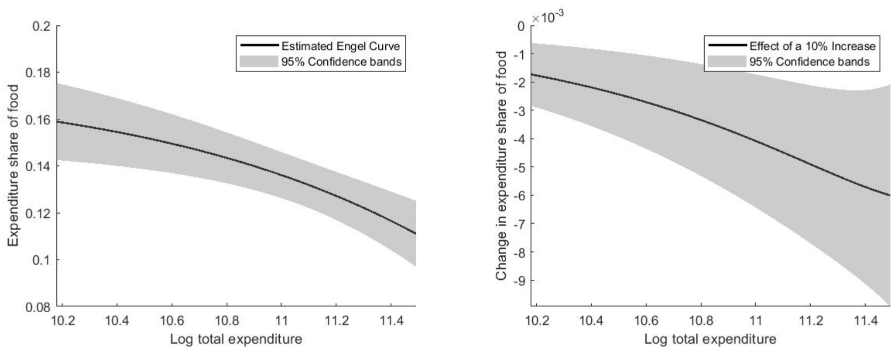
Section 2 we argue these assumptions are plauble in this setting, in particular because of the length of time between periods is long.

We set V and Z in line with the suggestions in Section 2. That is, $V = (X_1, ..., X_5)$ and $Z = (X_5, ..., X_9)$. The order condition here requires that the dimension of preferences $\tilde{W}^*$ be no greater than four.

Figure 1-4: Demand for Food

(a) Average Engel Curve for Food

(b) Change in Food Demand from a 10% Increase in Total Expenditure



Estimates are plotted at 100 points evenly spaced (in levels not logs) between the 10% and 90% quantiles of total expenditure. For Sub-Figure 4.a we estimate $\frac{1}{n}\sum_{i=1}^{n}\bar{y}(x, X_i)$ at each point x on the grid. For Sub-Figure 4.b we estimate $\bar{y}(1.1x, x) - \bar{y}(x, x)$ for each x. The uniform confidence bands are evaluated using $10,000$ replications of the multiplier bootstrap as detailed in Section 3. The pointwise standard errors were set equal to pointwise standard deviations over the bootstrap replications.

Sub-Figure 4.a plots our nonparametric estimate of the average over our sample of the structural Engel curve for food. The sub-figure shows a downward-sloping Engel curve that (with a log scale for total expenditure) is subtly concave. The downward slope of the curve suggests that food is a normal good, at least in aggregate.

Sub-Figure 4.b presents estimates of the average change in the budget share of food from an exogenous 10% increase in total expenditure broken down by the observed total expenditure. This is the difference of two conditional average structural Engel curves. In all cases the estimated change in expenditure share is negative, which again would be true of a normal good.

1.5 Conclusion

We present new results on identification, estimation, and inference with proxy controls in cross-sectional and panel settings. The present work raises a number of questions for future research. Firstly we have yet to explore a doubly robust approach to estimation based on both of the characterizations in Theorem 1.1. It is also unclear whether one can achieve valid inference without the need for the regularity conditions in Assumption 4. We conjecture that it is possible and we intend to address this in future work.

1.6 Appendix A.1: Regularity Conditions

First, let us introduce some additional notation. The vectors Z and V may share some common components. Denote the shared components by $\bar{W}$ and let $\tilde{V}$ and $\tilde{Z}$ contain the remaining components of V and Z respectively. Thus we can decompose $V = (\tilde{V}, \bar{W})$, $Z = (\tilde{Z}, \bar{W})$.

For each $(x, \bar{w})$ in the support of $(X, \bar{W})$ define a linear operator $A_{x,\bar{w}}$ that maps from $L_2(F_{\tilde{V}|X=x,\bar{W}=\bar{w}})$ to $L_2(F_{\tilde{Z}|X=x,\bar{W}=\bar{w}})$, by:

$$A_{x,\bar{w}}[\delta](\tilde{z}) = E[\delta(\tilde{V})|\tilde{Z} = \tilde{z}, X = x, \bar{W} = \bar{w}]$$

The adjoint of this linear operator $A^*_{x,\bar{w}} : L_2(F_{\tilde{Z}|X=x,\bar{W}=\bar{w}}) \to L_2(F_{\tilde{V}|X=x,\bar{W}=\bar{w}})$ is given by:

$$A^*_{x,\bar{w}}[\delta](\tilde{v}) = E[\delta(\tilde{Z})|\tilde{V} = \tilde{v}, X = x, \bar{W} = w]$$

Assumption A.1 (Compact Operator). The following holds for $F_{(X,\bar{W})}$-almost all $(x, \bar{w})$. Let 'F_{prod}' denote the product measure of $\tilde{V}$ and $\tilde{Z}$ conditional on $(X, \bar{W}) = (x, \bar{w})$.[18] The conditional joint measure $F_{(\tilde{V},\tilde{Z})|X=x,\bar{W}=\bar{w}}$ is absolutely continuous with respect to F_{prod} and the corresponding Radon-Nikodym derivative is square integrable

[18] In more conventional notation F_{prod} is equal to $F_{\tilde{V}|X=x,\bar{W}=\bar{w}} \otimes F_{\tilde{Z}|X=x,\bar{W}=\bar{w}}$.

with respect to F_{prod}:

$$\int \Big(\frac{dF_{(\tilde{V},\tilde{Z})|X=x,\bar{W}=\bar{w}}}{dF_{prod}}(\tilde{v},\tilde{z})\Big)^2 F_{prod}(d\tilde{v},d\tilde{z}) < \infty$$

Under Assumption A.1 for $F_{(X,\bar{W})}$-almost all $(x,\bar{w})$ there exists a unique singular system (indexed by $(x,\bar{w})$) $\{(u_k^{(x,\bar{w})}, v_k^{(x,\bar{w})}, \mu_k^{(x,\bar{w})})\}_{k=1}^{\infty}$ for $A_{x,\bar{w}}$. $\mu_k^{(x,\bar{w})}$ is the k^{th} singular value of $A_{x,\bar{w}}$. $u_k^{(x,\bar{w})}$ is a real-valued function that maps from the support of $\tilde{V}$ conditional on $(X,\bar{W})=(x,\bar{w})$. $v_k^{(x,\bar{w})}$ is a scalar valued function that maps from the support of $\tilde{Z}$ conditional on $(X,\bar{W})=(x,\bar{w})$. $u_k^{(x,\bar{w})}$ and $v_k^{(x,\bar{w})}$ are the k^{th} singular functions of the operator $A_{x,\bar{w}}$.[19]

Define functions $\{(u_k,v_k,\mu_k)\}_{k=1}^{\infty}$ as follows. For each k, and each pair $(x,\bar{w})$, $u_k(x,v) = u_k^{(x,\bar{w})}(\tilde{v})$ where $v=(\tilde{v},\bar{w})$ (recall that we can decompose $V=(\tilde{V},\bar{W})$). Similarly we let $v_k(x,z) = v_k^{(x,\bar{w})}(\tilde{z})$ where $z=(\tilde{z},\bar{w})$ and $\mu_k(x,\bar{w}) = \mu_k^{(x,\bar{w})}$.

Assumption 4 (Regularity). For both 4.i and 4.ii below assume the following: Assumptions A.1, 1.ii, and 2.ii hold, and for $F_{\bar{W}}$-almost all $\bar{w}$ and F_X-almost all x_1 and x_2, $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(V)$ and $E[Y|X,Z]$ have finite mean squares conditional on $(X,\bar{W}) = (x_1,\bar{w})$.

i. For some function $\tilde{C}$ with $E[\tilde{C}(x_1,x_2,\bar{W})|X=x_1] = C(x_1,x_2) < \infty$:

$$\sum_{k=1}^{\infty} \frac{1}{\mu_k(x_1,\bar{w})^2} E[\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(V) u_k(X,V)|X=x_1,\bar{W}=\bar{w}]^2 \le \tilde{C}(x_1,x_2,\bar{w})$$

ii. For some function $\tilde{D}$ with $E[\tilde{D}(X,\bar{w})|X=x] = D(x) < \infty$:

$$\sum_{k=1}^{\infty} \frac{1}{\mu_k(x,\bar{w})^2} E\big[E[Y|X,Z] v_k(X,Z)\big|X=x,\bar{W}=\bar{w}\big]^2 \le \tilde{D}(x,\bar{w})$$

Assumption 4 states that $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(V)$ and $E[Y|X,Z]$ have finite conditional mean squares so that the terms of the sums in 4.i and 4.ii are well-defined. 4.i and 4.ii each require that generalized Fourier coefficients go to zero sufficiently quickly. These restrictions can be understood as smoothness conditions (see [44]). Conditions of the

[19] See, e.g., [58] Theorem 15.16 and associated discussion.

same form are used elsewhere in the literature, for example in [30] and [62].

1.7 Appendix A.2: Parametric Examples

To build intuition it may be helpful to show how our analysis applies in some simple parametric settings. This also serves to demonstrate the connection of our methods to [43].

1.7.1 The Classical Additive Linear Case

Recall $Y = y_0(X, U)$, suppose y_0 is linear so that for a vector β:

$$y_0(x, u) = x'\beta + u$$

For simplicity we assume variables are zero mean so we can ignore intercepts. We assume without loss of generality that $E[W^* W^{*\prime}]$ is positive definite. We assume that the following conditional expectations are linear in the conditioning variables: $E[U|W^*, V]$, $E[V|W^*]$, $E[W^*|V, X]$, and $EW^*|Z, X]$. In particular, we assume: $U = \omega' W^* + \xi' V + \varepsilon$, $V = BW^* + \upsilon$, and $W^* = CZ + DX + \eta$. Where A, B, C, and D are matrices of parameters and $E[\varepsilon|W^*, V] = 0$, $E[\upsilon|W^*] = 0$, and $E[\eta|X, Z] = 0$. Finally, we assume that $E[ZZ'|X = x]$ is positive definite. Assumption 1.ii, the linearity assumptions, and Assumption 3.ii together imply that $B'B$ is non-singular, in which case we get the following linear IV model:

$$Y = X'\beta + V'\delta + \epsilon$$

$$V = FZ + GX + e$$

In the first equation $\delta = B(B'B)^{-1}\omega + \xi$ and $\epsilon = \varepsilon - \omega'(B'B)^{-1}B'\upsilon$. In the second equation $F = BC$, $G = BD$, and $e = B\eta + \upsilon$.

Assumption 2.i states that $U \perp\!\!\!\perp (X, Z)|W^*$ which implies $E[\varepsilon|X, Z] = 0$. Assumption 2.ii states $V \perp\!\!\!\perp (X, Z)|W^*$ and so $E[\upsilon|X, Z] = 0$. Therefore, $E[\epsilon|X, Z] = 0$

and $E[e|X,Z]=0$. One can show that for β to be identified in the linear IV model above, it suffices that there exists a matrix M so that $G = FM$. In our case this holds with $M = C'(CC')^{-1}D$ where invertibility of (CC') follows by Assumpton 3.i. Note that we do not need to identify δ to recover the CASF.

Thus β is identified and can be recovered by standard IV methods. This is the approach of [43]. From β we immediately get treatment effects (in the linear additive model treatments effects are the same for all individuals), and the CASF is given by $\bar{y}(x_1|x_2) = (x_1 - x_2)'\beta + E[Y|X = x_2]$. If the basis functions in our method consist only of linear terms and there is no penalization or sample-splitting, then our method reduces to that of [43]. To be precise, let $\phi_n(x,v) = (x',v')'$, $\psi_n(x,z) = \zeta_n(x,z) = (x',z')'$. Then our estimator $\hat{\theta}$ is equal to $(\hat{\beta}',\hat{\delta}')'$ where $\hat{\beta}$ and $\hat{\delta}$ are 2SLS estimates of β and δ in the linear model above. In this setting the conclusion of Lemma 1.1.b holds with $\gamma(x,v) = x'\beta + v'\delta$.

1.7.2 The Linear Multiplicative Case

Now let us consider the linear multiplicative case, with $y_0(x,u) = x'u_1 + u_2$, so that $u = (u_1,u_2)'$. Again let us assume linearity of the conditional means of U_1, U_2, and V in W^* and of W^* in Z and X. We have $Y = X'U_1 + U_2$, $U_1 = \omega_1 W^* + \varepsilon_1$, $U_2 = \omega_2 W^* + \varepsilon_2$, $V = BW^* + v$, and $W^* = CZ + DX + \eta$. Where $E[\varepsilon_1|W^*] = E[\varepsilon_2|W^*] = 0$, $E[v|W^*] = 0$, and $E[\eta|W^*,X] = 0$. In this case one can show that under our assumptions we get the following IV model:

$$\begin{aligned}
Y =& (X \otimes V)'\rho + V'\delta + \epsilon \\
V =& FZ + GX + e \\
X \otimes V =& H(X \otimes Z) + J(X \otimes X) + \tilde{e}
\end{aligned}$$

'$\otimes$' denotes the Kronecker product and $\rho = vec\big(\omega_1(B'B)^{-1}B'\big)$ where $vec(\cdot)$ returns the vectorization of its matrix argument. $\delta = B(B'B)^{-1}\omega_2$, F, G and e are defined as in the linear additive case. $H = I \otimes F$ where I is the identity matrix of appropriate size, and $J = I \otimes G$. Further:

$$\epsilon = -X'(B'B)^{-1}B'v + X'\varepsilon_1 - \omega_2(B'B)^{-1}B'v + \varepsilon_2$$

And $\tilde{e} = X \otimes e$. Under our Assumptions $E[e|X,Z] = E[\tilde{e}|X,Z] = 0$ and $E[\epsilon|X,Z] = 0$ and so the conclusion of Lemma 1.1.b holds with the function γ given by $\gamma(x,v) = (x \otimes v)'\rho + v'\delta$.

Consider the 2SLS estimates of ρ and δ in the model above where we treat Z, X, $X \otimes Z$, and $X \otimes X$ as instruments, and $(X \otimes V)$ and V as endogenous regressors. With no penalization nor sample-splitting then our PSMD estimate $\hat{\theta}$ is identical to these 2SLS estimates if we choose $\phi_n(x,v) = \big((x \otimes v)', v'\big)'$ and $\psi_n(x,z) = \zeta_n(x,z) = \big(z', x', (x \otimes z)', (x \otimes x)'\big)'$. In this case an appropriate choice for $\chi_n(x)$ would be $\big(x', (x \otimes x)'\big)$.

1.8 Appendix B: Proofs

Proof Theorem 1.1. Throughout the proof, statements involving x_1 and x_2 should be understood to hold for F_X-almost all x_1 and x_2.

Theorem 1.1.a: By Lemma 1.1, under Assumption 4.ii there exists a function γ with $E\big[\gamma(X,V)^2\big|X\big] < \infty$ so that $E\big[Y - \gamma(X,V)\big|X,Z\big] = 0$. Fix such a γ, by iterated expectations and Assumption 2.ii:

$$\begin{aligned} E[\gamma(X,V)|X,Z] &= E\big[E[\gamma(X,V)|X,W^*,Z]\big|X,Z\big] \\ &= E\big[E[\gamma(X,V)|X,W^*]\big|X,Z\big] \end{aligned}$$

And by iterated expectations and Assumption 2.i:

$$\begin{aligned} E[Y|X,Z] &= E[y_0(X,U)|X,Z] \\ &= E\big[E[y_0(X,U)|X,W^*,Z]\big|X,Z\big] \\ &= E\big[E[y_0(X,U)|X,W^*]\big|X,Z\big] \end{aligned}$$

And so:

$$E\big[E\big[y_0(X,U)\big|X,W^*\big] - E\big[\gamma(X,V)\big|X,W^*\big]\big|X,Z\big] = 0$$

But then by Assumption 3.i:

$$E[y_0(X,U)|X,W^*] = E[\gamma(X,V)|X,W^*]$$

By Assumption 1.i and 1.ii:

$$E[y_0(X,U)|X = x_1, W^*] = E[y_0(x_1,U)|X = x_2, W^*]$$

And by Assumption 2.ii:

$$E[\gamma(X,V)|X = x_1, W^*] = E[\gamma(x_1,V)|X = x_2, W^*]$$

And so:

$$E[y_0(x_1,U)|X = x_2, W^*] = E[\gamma(x_1,V)|X = x_2, W^*]$$

The LHS has finite expectation conditional on $X = x_2$ by Assumption 1.iii, and thus so does the RHS above. By iterated expectations:

$$E[y_0(x_1,U)|X = x_2] = E[\gamma(x_1,V)|X = x_2]$$

And by definition the LHS equals $\bar{y}(x_1|x_2)$.

Theorem 1.1.b: By Lemma 1.1, under Assumption 4.i there exists a function φ with $E[\varphi(x_1,x_2,Z)^2|X = x_1] < \infty$ so that:

$$E[\varphi(x_1,x_2,Z)|X = x_1, V] = \frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(V)$$

Fix such a φ. The next step refers to $\frac{dF_{W^*|X=x_2}}{dF_{W^*|X=x_1}}$, which is the Radon-Nikodym derivative of $F_{W^*|X=x_2}$ with respect to $F_{W^*|X=x_1}$. By the Radon-Nikodym theorem this exists if the two distributions are absolutely continuous which is in turn is implied by

Assumption 1.ii.

Note that under Assumption 2.ii:

$$E\Big[\frac{dF_{W^*|X=x_2}}{dF_{W^*|X=x_1}}(W^*)\Big|X=x_1, V=v\Big] = \frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(v)$$

To see this note that by properties of Radon-Nikodym derivatives:

$$E\Big[\frac{dF_{W^*|X=x_2}}{dF_{W^*|X=x_1}}(W^*)\Big|X=x_1, V=v\Big] = E\Big[\frac{dF_{V|W^*,X=x_1}}{dF_{V|X=x_1}}(v, W^*)\Big|X=x_2\Big] \tag{1.8.1}$$

By Assumption 2.ii $\frac{dF_{V|W^*,X=x_1}}{dF_{V|X=x_1}} = \frac{dF_{V|W^*,X=x_2}}{dF_{V|X=x_1}}$, and so:

$$\begin{aligned} E\Big[\frac{dF_{V|W^*,X=x_1}}{dF_{V|X=x_1}}(v, W^*)\Big|X=x_2\Big] =& E\Big[\frac{dF_{V|W^*,X=x_2}}{dF_{V|X=x_1}}(v, W^*)\Big|X=x_2\Big] \\ =& \frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(v) \end{aligned}$$

Where the final equality follows by the properties of the Radon-Nikodym derivative. Further, by iterated expectations and Assumption 2.ii:

$$\begin{aligned} E[\varphi(x_1, x_2, Z)|X=x_1, V] =& E\Big[E[\varphi(x_1, x_2, Z)|W^*, X, V]\Big|X=x_1, V\Big] \\ =& E\Big[E[\varphi(x_1, x_2, Z)|W^*, X]\Big|X=x_1, V\Big] \end{aligned}$$

So we have:

$$E\Big[\frac{dF_{W^*|X=x_2}}{dF_{W^*|X=x_1}}(W^*) - E[\varphi(x_1, x_2, Z)|W^*, X]\Big|X=x_1, V\Big] = 0$$

By Assumption 3.ii this implies:

$$E[\varphi(x_1, x_2, Z)|W^*, X=x_1] = \frac{dF_{W^*|X=x_2}}{dF_{W^*|X=x_1}}(W^*) \tag{1.8.2}$$

By iterated expectations and Assumptions 1.i and 1.ii:

$$
\begin{aligned}
E[y_0(x_1,U)|X=x_2] &= E\big[E[y_0(x_1,U)|X=x_2,W^*]\big|X=x_2\big] \\
&= E\big[E[y_0(X,U)|X=x_1,W^*]\big|X=x_2\big] \qquad (1.8.3)
\end{aligned}
$$

And by the Radon-Nikodym theorem:

$$
\begin{aligned}
&E\big[E[Y|X=x_1,W^*]\big|X=x_2\big] \\
=&E\big[\frac{dF_{W^*|X=x_2}}{dF_{W^*|X=x_1}}(W^*)E[y_0(X,U)|X,W^*]\big|X=x_1\big]
\end{aligned}
$$

Substituting for the LHS by (1.8.3) and for the Radon-Nikodym derivative on the RHS by (1.8.2) we get that $E[y_0(x_1,U)|X=x_2]$ is equal to:

$$
E\big[E[\varphi(x_1,x_2,Z)|W^*,X]E[y_0(X,U)|X,W^*]\big|X=x_1\big]
$$

Next note that:

$$
\begin{aligned}
&E\big[E[\varphi(x_1,x_2,Z)|W^*,X]E[y_0(X,U)|X,W^*]\big|X=x_1\big] \\
=&E\big[\varphi(x_1,x_2,Z)E[y_0(X,U)|X,W^*]\big|X=x_1\big] \\
=&E\big[\varphi(x_1,x_2,Z)E[y_0(X,U)|X,W^*,Z]\big|X=x_1\big] \\
=&E\big[\varphi(x_1,x_2,Z)y_0(X,U)\big|X=x_1\big] \\
=&E\big[\varphi(x_1,x_2,Z)Y\big|X=x_1\big]
\end{aligned}
$$

Where the first equality follows by iterated expectations, the second by Assumption 2.i, the third by iterated expectations and the final by the definition of Y. Combining we get:

$$
E[y_0(x_1,U)|X=x_2] = E\big[\varphi(x_1,x_2,Z)Y\big|X=x_1\big]
$$

$\square$

Proof Theorem 1.2. **Theorem 1.2.a:** By The Radon-Nikodym theorem, for any $\tilde{\gamma}$

with $E[\tilde{\gamma}(X,V)^2|X=x]$ finite for F_X-almost all x, we have:

$$E\big[\tilde{\gamma}(x_1,V)\big|X=x_2\big] = E\Big[\tilde{\gamma}(X,V)\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(V)\Big|X=x_1\Big]$$
$$= E\big[\tilde{\gamma}(X,V)E[\varphi(x_1,x_2,Z)|X,V]\big|X=x_1\big]$$

Where φ satisfies conclusion a. of Lemma 1.1. By iterated expectations:

$$E\big[\tilde{\gamma}(X,V)E[\varphi(x_1,x_2,Z)|X,V]\big|X=x_1\big]$$
$$= E\big[E[\tilde{\gamma}(X,V)|X,Z]\varphi(x_1,x_2,Z)\big|X=x_1\big]$$

Under the conditions of 1.2.a, Theorem 1.1.b holds, and so combining the characterization of $\bar{y}(x_1|x_2)$ in Theorem 1.1.b and the equation above we get:

$$\bar{y}(x_1|x_2) - E[\tilde{\gamma}(x_1,V)|X=x_2]$$
$$= E\big[\big(Y - E[\tilde{\gamma}(X,V)|X,Z]\big)\varphi(x_1,x_2,Z)\big|X=x_1\big]$$

Applying Cauchy-Schwartz:

$$(\bar{y}(x_1|x_2) - E[\tilde{\gamma}(x_1,V)|X=x_2])^2$$
$$\leq E[\varphi(x_1,x_2,Z)^2|X=x_1]E\big[\big(Y - E[\tilde{\gamma}(X,V)|X,Z]\big)^2\big|X=x_1\big]$$

By Lemma 1.1 $E[\varphi(x_1,x_2,Z)^2|X=x_1] \leq C(x_1,x_2)$.

Theorem 1.2.b: By iterated expectations for any function $\tilde{\varphi}$ with finite conditional mean square $E[\tilde{\varphi}(x_1,x_2,Z)^2|X=x_1] \leq \infty$, for F_X-almost all x_1 and x_2, we have:

$$E[\tilde{\varphi}(x_1,x_2,Z)Y|X=x_1] = E\big[\tilde{\varphi}(x_1,x_2,Z)E[Y|X,Z]\big|X=x_1\big]$$

Letting γ satisfy the equations in conclusion b. of Lemma 1.1 we can substitute out

$E[Y|X,Z]$ in the above to get:

$$
\begin{aligned}
&E[\tilde{\varphi}(x_1,x_2,Z)Y|X=x_1]\\
=&E\big[\tilde{\varphi}(x_1,x_2,Z)E[\gamma(X,V)|X,Z]\big|X=x_1\big]\\
=&E\big[E[\tilde{\varphi}(x_1,x_2,Z)|X,V]\gamma(X,V)\big|X=x_1\big]
\end{aligned}
\tag{1.8.4}
$$

Where the second equality follows by iterated expectations. Recall that under the conditions of part a. of the Theorem 1.1:

$$
\bar{y}(x_1|x_2)=E[\gamma(x_1,V)|X=x_2] \tag{1.8.5}
$$

By the Radon-Nikodym theorem:

$$
E[\gamma(x_1,V)|X=x_2]=E\big[\gamma(X,V)\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(V)\big|X=x_1\big]
$$

Combining with (1.8.4) and (1.8.5) we get:

$$
\begin{aligned}
&\bar{y}(x_1|x_2)-E[\tilde{\varphi}(x_1,x_2,Z)Y|X=x_1]\\
=&E\big[\gamma(X,V)\big(\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(V)-E[\tilde{\varphi}(x_1,x_2,Z)|X,V]\big)\big|X=x_1\big]
\end{aligned}
$$

By Cauchy-Schwartz:

$$
\begin{aligned}
&(\bar{y}(x_1|x_2)-E[\tilde{\varphi}(x_1,x_2,Z)Y|X=x_1])^2\\
=&E[\gamma(X,V)^2|X=x_1]\\
\times&E\big[\big(\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(v)-E[\tilde{\varphi}(x_1,x_2,Z)|X,V]\big)^2\big|X=x_1\big]
\end{aligned}
$$

By Lemma 1.1 with $E\big[\gamma(x_1,V)^2|X=x_1\big]\leq D(x_1)$. $\square$

The proofs of Propositions 2.1 and 2.2 below use the following three facts about conditional independence. Let W_1, W_2, W_3, and W_4 be random variables. We have:

$$W_1 \perp\!\!\!\perp W_2|(W_3,W_4) \implies W_1 \perp\!\!\!\perp (W_2,W_3)|(W_3,W_4) \tag{1.8.6}$$

$$W_1 \perp\!\!\!\perp (W_2,W_3)|W_4 \implies W_1 \perp\!\!\!\perp (W_2,W_3)|(W_3,W_4) \tag{1.8.7}$$

$$W_1 \perp\!\!\!\perp (W_2,W_3)|W_4 \implies W_1 \perp\!\!\!\perp W_2|W_4 \tag{1.8.8}$$

Proof of Proposition 2.1. By supposition:

$$(X_1,...,X_{\lfloor t/2\rfloor-1}) \perp\!\!\!\perp (X_{\lfloor t/2\rfloor+1},...,X_T)|(\tilde{W}^*,X_{\lfloor t/2\rfloor}) \tag{1.8.9}$$

$$U_t \perp\!\!\!\perp (X_1,...,X_t)|\tilde{W}^* \tag{1.8.10}$$

Using (1.8.7), (1.8.10) implies $U_t \perp\!\!\!\perp (X_1,...,X_t)|(\tilde{W}^*,X_{\lfloor t/2\rfloor})$ Applying (1.8.8) we get $U_t \perp\!\!\!\perp (X_{\lfloor t/2\rfloor},...,X_t)|(\tilde{W}^*,X_{\lfloor t/2\rfloor})$. Substituting the definitions of Z and W^* gives $U_t \perp\!\!\!\perp (X_t,Z)|W^*$.

Twice applying (1.8.6) to (1.8.9) implies that:

$$(X_1,...,X_{\lfloor t/2\rfloor}) \perp\!\!\!\perp (X_{\lfloor t/2\rfloor},...,X_t)|(\tilde{W}^*,X_{\lfloor t/2\rfloor})$$

Substituting the definitions of V, Z and W^* we get $V \perp\!\!\!\perp (X_t,Z)|W^*$. □

Proof of Proposition 2.2. By Supposition:

$$U_t \perp\!\!\!\perp (X_1,...,X_t,U_1,...,U_{t-1})|\tilde{W}^* \tag{1.8.11}$$

and conditional on $(\tilde{W}^*,X_{\lfloor t/2\rfloor},U_{\lfloor t/2\rfloor})$:

$$(X_1,...,X_{\lfloor t/2\rfloor-1},U_1,...,U_{\lfloor t/2\rfloor-1}) \perp\!\!\!\perp (X_{\lfloor t/2\rfloor+1},...,X_t,U_{\lfloor t/2\rfloor+1},...,U_t) \tag{1.8.12}$$

For all x and $u_1 \neq u_2$, $y_{0,t}(x,u_1) \neq y_{0,t}(x,u_2)$. So from (1.8.11):

$$U_t \perp\!\!\!\perp (X_1,...,X_t,Y_1,...,Y_{t-1})|\tilde{W}^*$$

Applying (1.8.7) we get:

$$U_t \perp\!\!\!\perp (X_1, ..., X_t, Y_1, ..., Y_{t-1}) | (\tilde{W}^*, X_{\lfloor t/2 \rfloor}, Y_{\lfloor t/2 \rfloor})$$

Applying (1.8.8):

$$U_t \perp\!\!\!\perp (X_{\lfloor t/2 \rfloor}, ..., X_t, Y_{\lfloor t/2 \rfloor}, ..., Y_{t-1}) | (\tilde{W}^*, X_{\lfloor t/2 \rfloor}, Y_{\lfloor t/2 \rfloor})$$

Substituting the definitions of Z and W^* gives $U_t \perp\!\!\!\perp (X_t, Z)|W^*$.

For all x and $u_1 \neq u_2$, $y_{0,t}(x, u_1) \neq y_{0,t}(x, u_2)$, so (1.8.12) implies that conditional on $(\tilde{W}^*, X_{\lfloor t/2 \rfloor}, Y_{\lfloor t/2 \rfloor})$:

$$(X_1, ..., X_{\lfloor t/2 \rfloor - 1}, Y_1, ..., Y_{\lfloor t/2 \rfloor - 1}) \perp\!\!\!\perp (X_{\lfloor t/2 \rfloor + 1}, ..., X_t, Y_{\lfloor t/2 \rfloor + 1}, ..., Y_{t-1})$$

Applying (1.8.6) we get that conditional on $(\tilde{W}^*, X_{\lfloor t/2 \rfloor}, Y_{\lfloor t/2 \rfloor})$:

$$(X_1, ..., X_{\lfloor t/2 \rfloor}, Y_1, ..., Y_{\lfloor t/2 \rfloor}) \perp\!\!\!\perp (X_{\lfloor t/2 \rfloor}, ..., X_t, Y_{\lfloor t/2 \rfloor}, ..., Y_{t-1})$$

Substituting the definitions of V, Z and W^* we see the above is equivalent to $V \perp\!\!\!\perp (X_t, Z)|W^*$. □

Proof of Theorem 3.1. From Theorem 1.2, for any function $\tilde{\gamma}$ that satisfies the condition $E[\tilde{\gamma}(X, V)^2 | X = x] < \infty$, for F_X-almost all x, we have :

$$E[\tilde{\gamma}(x_1, V)|X = x_2] = E\big[\varphi(x_1, x_2, Z) E\big[\tilde{\gamma}(X, V)\big| X = x_1, Z\big]\big| X = x_1\big]$$

Where φ satisfies the conclusion of Lemma 1.1.a. From the above, and the definitions of α_n and π_n, it follows that:

$$\alpha_n(x_1, x_2) = E\big[\varphi(x_1, x_2, Z)\pi_n(x_1, Z)\big| X = x_1\big]$$

By Theorem 1.1.b:

$$\bar{y}(x_1|x_2) = E\big[\varphi(x_1,x_2,Z)g(x_1,Z)\big|X = x_1\big]$$

Let $\tilde{\alpha}_n(x_1,x_2) = E_Z\big[\varphi(x_1,x_2,Z)\hat{\pi}_n(x_1,Z)\big|X = x_1\big]$. By Lemma C.9 there is a sequence $\{\theta_n\}_{n=1}^{\infty}$ with $E\big[|\phi_n(X,V)'\theta_n|^2\big]$ bounded above over n so that, uniformly over $F_{(X,Z)}$-almost all (x,z), $|g(x,z) - \pi_n(x,z)'\theta_n| \precsim R_\theta$. Adding and subtracting terms we get:

$$\begin{aligned}
&\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta} \\
=&\tilde{\alpha}_n(x_1,x_2)'(\theta_n - \hat{\theta}) \\
+&E\big[\varphi(x_1,x_2,Z)\big(g(x_1,Z) - \pi_n(x_1,Z)'\theta_n\big)\big|X = x_1\big] \\
+&E_Z\big[\varphi(x_1,x_2,Z)\big(\pi_n(x_1,Z) - \hat{\pi}_n(x_1,Z)\big)\big|X = x_1\big]'(\theta_n - \hat{\theta}) \\
+&\big(\alpha_n(x_1,x_2) - \hat{\alpha}_n(x_1,x_2)\big)'\hat{\theta}
\end{aligned}$$

$E\big[\varphi(x_1,x_2,Z)^2|X = x_1\big] \leq C(x_1,x_2)$, so applying Cauchy-Schwartz and the triangle inequality:

$$\begin{aligned}
&\big|\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big) - \tilde{\alpha}_n(x_1,x_2)'(\theta_n - \hat{\theta})\big| \\
\leq&C(x_1,x_2)^{1/2}E\big[|g_n(x_1,Z) - \pi_n(x_1,Z)'\theta_n|^2\big|X = x_1\big]^{1/2} \\
+&C(x_1,x_2)^{1/2}E_Z\big[\big|\big(\pi_n(x_1,Z) - \hat{\pi}_n(x_1,Z)\big)'(\theta_n - \hat{\theta})\big|^2\big|X = x_1\big]^{1/2} \\
+&||\big(\hat{\alpha}_n(x_1,x_2) - \alpha_n(x_1,x_2)\big)'\hat{\theta}||
\end{aligned}$$

By the definition of the rate $R_{\pi,n}(x_1)$:

$$\begin{aligned}
&E_Z\big[\big|\big(\pi_n(x_1,Z) - \hat{\pi}_n(x_1,Z)\big)'(\theta_n - \hat{\theta})\big|^2\big|X = x_1\big]^{1/2} \\
\precsim_p&R_{\pi,n}(x_1)E\big[|\phi_n(V,X)'(\theta_n - \hat{\theta})|^2\big]^{1/2}
\end{aligned}$$

And:

$$\big(\hat{\alpha}_n(x_1,x_2) - \alpha_n(x_1,x_2)\big)'\hat{\theta} \precsim_p R_{\alpha,n}(x_1,x_2)E\big[|\phi_n(V,X)'\theta_n|^2\big]^{1/2}$$

Lemma C.3 implies that $E\big[|\phi_n(V,X)'\hat{\theta}|^2\big]^{1/2} \precsim \frac{\bar{\mu}_n}{\underline{\mu}_n} + \frac{\bar{\mu}_n R_n}{\lambda_{0,n}^{1/2}}$ and implies the same bound on the growth of $E\big[|\phi_n(V,X)'(\theta_n - \hat{\theta})|^2\big]^{1/2}$, where R_n satisfies $\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}(\hat{g}_i - \hat{\pi}'_{n,i}\theta_n)^2 \precsim_p R_n^2$. By the triangle inequality and since $E\big[|\phi_n(V,X)'\theta_n|^2\big]^{1/2} \precsim 1$ the condition on R_n is satisfied by $R_n = R_{\theta,n} + R_{g,n} + R_{\pi,n}$, so we get:

$$\begin{aligned}
&\big|\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big) - \tilde{\alpha}_n(x_1,x_2)'(\theta_n - \hat{\theta})\big| \\
\precsim_p & R_{\theta,n} + \big(\frac{\bar{\mu}_n}{\underline{\mu}_n} + \frac{\bar{\mu}_n(R_{\theta,n} + R_{g,n} + R_{\pi,n})}{\lambda_{0,n}^{1/2}}\big)\big(R_{\pi,n}(x_1) + R_{\alpha,n}(x_1,x_2)\big)
\end{aligned} \tag{1.8.13}$$

Uniformly when $C(x_1,x_2)$ is uniformly bounded. We now derive a rate for the term $\tilde{\alpha}_n(x_1,x_2)'(\theta_n - \hat{\theta})$. Recall that $\hat{\theta} = \hat{\Sigma}^{-1}_{\lambda_{0,n}}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}\hat{g}_i$. Adding and subtracting terms:

$$\begin{aligned}
\tilde{\alpha}_n(x_1,x_2)'(\hat{\theta} - \theta_n) =& \tilde{\alpha}_n(x_1,x_2)'\hat{\Sigma}^{-1}_{\lambda_{0,n}}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}(\hat{g}_i - g_i) \\
&+ \tilde{\alpha}_n(x_1,x_2)'\hat{\Sigma}^{-1}_{\lambda_{0,n}}\hat{\Sigma}_n^{1/2}\tilde{r}_n - \tilde{\alpha}_n(x_1,x_2)'\lambda_{0,n}\hat{\Sigma}^{-1}_{\lambda_{0,n}}\theta_n
\end{aligned} \tag{1.8.14}$$

Where $\tilde{r}_n$ is defined by:

$$\tilde{r}_n = \hat{\Sigma}_n^{-1/2}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}(g_i - \pi'_{n,i}\theta_n) + \hat{\Sigma}_n^{-1/2}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}(\pi_{n,i} - \hat{\pi}_{n,i})'\theta_n$$

By the triangle inequality and definition of the operator norm:

$$\begin{aligned}
||\tilde{r}_n|| \leq & ||\hat{\Sigma}_n^{-1/2}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}(g_i - \pi'_{n,i}\theta_n)|| \\
&+ ||\hat{\Sigma}_n^{-1/2}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}(\pi_{n,i} - \hat{\pi}_{n,i})'\theta_n||
\end{aligned}$$

By the properties of least squares projection and Markov's inequality:

$$\begin{aligned}
||\hat{\Sigma}_n^{-1/2}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}(g_i - \pi'_{n,i}\theta_n)|| \leq & \big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}(g_i - \pi'_{n,i}\theta_n)^2\big)^{1/2} \\
& \precsim_p E\big[(g_i - \pi'_{n,i}\theta_n)^2\big]^{1/2} \precsim_p R_{\theta,n}
\end{aligned}$$

Similarly, by properties of least squares projection we get:

$$||\hat{\Sigma}_n^{-1/2}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}(\pi_{n,i}-\hat{\pi}_{n,i})'\theta_n|| \leq \big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\big((\pi_{n,i}-\hat{\pi}_{n,i})'\theta_n\big)^2\big)^{1/2}$$
$$\lesssim_p R_{\pi,n}$$

And so $||\tilde{r}_n|| \lesssim_p R_{\theta,n} + R_{\pi,n}$. Next, note that $||\hat{\Sigma}_{\lambda_{0,n}}^{-1/2}|| \leq \lambda_{0,n}^{-1/2}$ and so, using $E\big[|\phi_n(X,V)'\theta_n|^2\big] \lesssim 1$:

$$\lambda_{0,n}||\hat{\Sigma}_{\lambda_{0,n}}^{-1/2}\theta_n|| \leq \lambda_{0,n}||\hat{\Sigma}_{\lambda_{0,n}}^{-1}||^{1/2}||\theta_n|| \leq \lambda_{0,n}^{1/2}||\theta_n||$$
$$\leq \lambda_{0,n}^{1/2}\underline{\mu}_n^{-1}E\big[|\phi_n(X,V)'\theta_n|^2\big]^{1/2} \lesssim \lambda_{0,n}^{1/2}\underline{\mu}_n^{-1}$$

Substituting the rates derived above for $||\tilde{r}_n||$ and $\lambda_{0,n}||\hat{\Sigma}_{\lambda_{0,n}}^{-1/2}\theta_n||$ into (1.8.14) and applying Cauchy-Schwartz:

$$\tilde{\alpha}_n(x_1,x_2)'(\hat{\theta}-\theta_n) - \tilde{\alpha}_n(x_1,x_2)'\hat{\Sigma}_{\lambda_{0,n}}^{-1}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}(\hat{g}_i-g_i)$$
$$\lesssim_p ||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||(R_{\theta,n}+R_{\pi,n})$$
$$+||\hat{\Sigma}_{\lambda_{0,n}}^{-1/2}\tilde{\alpha}_n(x_1,x_2)||\lambda_{0,n}^{1/2}\underline{\mu}_n^{-1} \tag{1.8.15}$$

Again, by the properties of least squares:

$$||\hat{\Sigma}_n^{-1/2}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}(\hat{g}_i-g_i)|| \leq \big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}(\hat{g}_i-g_i)^2\big)^{1/2} \lesssim_p R_g$$

Combining by the triangle inequality we get:

$$\tilde{\alpha}_n(x_1,x_2)'(\hat{\theta}-\theta_n) \lesssim_p ||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||(R_{\theta,n}+R_{\pi,n}+R_g)$$
$$+||\hat{\Sigma}_{\lambda_{0,n}}^{-1/2}\tilde{\alpha}_n(x_1,x_2)||\lambda_{0,n}^{1/2}\underline{\mu}_n^{-1} \tag{1.8.16}$$

From Lemma C.8 we have $||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||^2 \lesssim_p C(x_1,x_2)\xi_{\tilde{\Omega},n}^2(x_1)$ and also that $||\hat{\Sigma}_{\lambda_{0,n}}^{-1/2}\tilde{\alpha}_n(x_1,x_2)||^2 \lesssim_p C(x_1,x_2)\xi_{\tilde{\Omega},n}^2(x_1)$. Combining with (1.8.16) and (1.8.13) gives the conclusion. □

Proof of Theorem 3.2. Follows immediately from Lemmas C.5, C.6, and C.7. $\square$

Proof of Theorem 3.3. Recall from the proof of Theorem 3.1 that $||\hat{\Sigma}_{\lambda_{0,n}}^{-1/2}\tilde{\alpha}(x_1,x_2)|| \precsim_p \xi_{\tilde{\Omega},n}(x_1)$ and the rate above is uniform if $C(x_1,x_2)$ from 4.i is uniformly bounded. Combining this with (1.8.13) and (1.8.15):

$$\begin{aligned}
&\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big) - \tilde{\alpha}(x_1,x_2)'\hat{\Sigma}_{\lambda_{0,n}}^{-1}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}(\hat{g}_i - g_i) \\
&\precsim_p R_{\theta,n} + \big(\frac{\bar{\mu}_n}{\underline{\mu}_n} + \frac{\bar{\mu}_n(R_{\theta,n}+R_{g,n}+R_{\pi,n})}{\lambda_{0,n}^{1/2}}\big)\big(R_{\pi,n}(x_1) + R_{\alpha,n}(x_1,x_2)\big) \\
&+||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||(R_{\theta,n}+R_{\pi,n}) + \xi_{\tilde{\Omega},n}(x)\lambda_{0,n}^{1/2}\underline{\mu}_n^{-1}
\end{aligned} \tag{1.8.17}$$

Our estimator $\hat{g}_i$ is given by $\hat{g}_i = \zeta_{n,i}'\hat{\beta}$ with $\hat{\beta} = \hat{\Xi}_{\lambda_{1,n}}^{-1}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}Y_i$. Let $\tilde{\beta} = \hat{\Xi}_n^{-1}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}Y_i$, let $\bar{\Sigma}_{\lambda_{0,n}} = E[\hat{\pi}_{n,i}\hat{\pi}_{n,i}'|\mathcal{I}_\pi] + \lambda_{0,n}I$ and lastly define β_n by $\beta_n = \Xi_n^{-1}E[\zeta_{n,i}g_i]$. We can decompose:

$$\begin{aligned}
&\tilde{\alpha}(x_1,x_2)'\hat{\Sigma}_{\lambda_{0,n}}^{-1}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}(\hat{g}_i - g_i) \\
=&\tilde{\alpha}(x_1,x_2)'\bar{\Sigma}_{\lambda_{0,n}}^{-1}E\big[\hat{\pi}_{n,i}\zeta_{n,i}'\big|\mathcal{I}_\pi\big]\Xi_n^{-1}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}(r_{n,i}+\epsilon_i) \\
&+\tilde{\alpha}(x_1,x_2)'(\hat{\Sigma}_{\lambda_{0,n}}^{-1}\hat{\Sigma}_n^{1/2}\hat{r}_{1,n} + \hat{\Sigma}_{\lambda_{0,n}}^{-1}\hat{\Sigma}_n^{1/2}\hat{r}_{2,n}) \\
&+\tilde{\alpha}(x_1,x_2)'\big(\bar{\Sigma}_n^{-1/2}\hat{r}_{3,n}\Xi_n^{1/2}(\tilde{\beta}-\beta_n) + \bar{\Sigma}_{\lambda_{0,n}}^{-1}E[\hat{\pi}_{n,i}\zeta_{n,i}'|\mathcal{I}_\pi]\Xi_n^{-1/2}\hat{r}_{4,n}\big)
\end{aligned}$$

Where $\hat{r}_{1,n}$, $\hat{r}_{2,n}$, $\hat{r}_{3,n}$, and $\hat{r}_{4,n}$, are given by the formulas below:

$$\begin{aligned}
\hat{r}_{1,n} &= \hat{\Sigma}_n^{-1/2}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}(\zeta_{n,i}'\beta_n - g_i) \\
\hat{r}_{2,n} &= \hat{\Sigma}_n^{-1/2}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}\zeta_{n,i}'(\hat{\beta}_n - \tilde{\beta}_n) \\
\hat{r}_{3,n} &= \bar{\Sigma}_n^{1/2}\big(\hat{\Sigma}_{\lambda_{0,n}}^{-1}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}\zeta_{n,i}' - \bar{\Sigma}_{\lambda_{0,n}}^{-1}E\big[\hat{\pi}_{n,i}\zeta_{n,i}'\big|\mathcal{I}_\pi\big]\big)\Xi_n^{-1/2} \\
\hat{r}_{4,n} &= \Xi_n^{1/2}(\tilde{\beta} - \beta_n) - \Xi_n^{-1/2}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}(r_{n,i}+\epsilon_i)
\end{aligned}$$

And so, by the triangle inequality and Cauchy-Schwartz:

$$
\begin{aligned}
&\tilde{\alpha}(x_1,x_2)'\hat{\Sigma}^{-1}_{\lambda_{0,n}}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}(\hat{g}_i-g_i)\\
&-\tilde{\alpha}(x_1,x_2)'\bar{\Sigma}^{-1}_{\lambda_{0,n}}E\big[\hat{\pi}_{n,i}\zeta'_{n,i}\big|\mathcal{I}_\pi\big]\Xi_n^{-1}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}(r_{n,i}+\epsilon_i)\\
\lesssim_p&||\hat{\Sigma}_n^{1/2}\hat{\Sigma}^{-1}_{\lambda_{0,n}}\tilde{\alpha}(x_1,x_2)||\big(||\hat{r}_{1,n}||+||\hat{r}_{2,n}||\big)\\
&+||\bar{\Sigma}_n^{-1/2}\tilde{\alpha}(x_1,x_2)||\cdot||\hat{r}_{3,n}||\cdot||\Xi_n^{1/2}(\tilde{\beta}-\beta_n)||\\
&+||\Xi_n^{-1/2}E[\zeta_{n,i}\hat{\pi}'_{n,i}|\mathcal{I}_\pi]\bar{\Sigma}^{-1}_{\lambda_{0,n}}\tilde{\alpha}(x_1,x_2)||\cdot||\hat{r}_{4,n}|| \qquad (1.8.18)
\end{aligned}
$$

From Lemma C.8 we get the following two rates:

$$||\bar{\Sigma}_n^{-1/2}\tilde{\alpha}_n(x_1,x_2)||^2\lesssim C(x_1,x_2)\xi^2_{\bar{\Omega},n}(x_1)$$

$$||\hat{\Sigma}_n^{1/2}\hat{\Sigma}^{-1}_{\lambda_{0,n}}\tilde{\alpha}_n(x_1,x_2)||\lesssim_p||\bar{\Sigma}_n^{1/2}\bar{\Sigma}^{-1}_{\lambda_{0,n}}\tilde{\alpha}_n(x_1,x_2)||$$

Now let $r_{n,i}=g_i-\zeta'_{n,i}\beta_n$ and $\epsilon_i=Y_i-g_i$. Then adding and subtracting terms and using the definition of the operator norm:

$$
\begin{aligned}
&||\Xi_n^{1/2}(\tilde{\beta}-\beta_n)||\\
\leq&||\Xi_n^{1/2}\hat{\Xi}_n^{-1}\Xi_n^{1/2}||\cdot||\Xi_n^{-1/2}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}(r_{n,i}+\epsilon_i)||\lesssim_p\sqrt{\frac{p(n)}{n_g}} \qquad (1.8.19)
\end{aligned}
$$

Where the final line above follows because $||\Xi_n^{1/2}\hat{\Xi}_n^{-1}\Xi_n^{1/2}||\lesssim_p 1$ from Assumption 5.4.iii and Lemma C.1 and $||\Xi_n^{-1/2}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}(r_{n,i}+\epsilon_i)||\lesssim_p\sqrt{\frac{p(n)}{n_g}}$ which is shown in Lemma C.7. Now we derive rates for $||\hat{r}_{1,n}||$, $||\hat{r}_{2,n}||$, $||\hat{r}_{3,n}||$, and $||\hat{r}_{4,n}||$.

Rate for $||\hat{r}_{1,n}||$: By Assumptions 5.1.iii and 5.3.ii, $|\zeta'_{n,i}\beta_n-g_i|\lesssim_p\ell_{\zeta,n}(s_3)$ almost surely. And so:

$$||\hat{r}_{1,n}||\leq\Big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}(\zeta'_{n,i}\beta_n-g_i)^2\Big)^{1/2}\lesssim_p\ell_{\zeta,n}(s_3) \qquad (1.8.20)$$

Rate for $||\hat{r}_{2,n}||$: By the definition of the operator norm $||\hat{r}_{2,n}||$ is less than:

$$||\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\Sigma}_n^{-1/2}\hat{\pi}_{n,i}\zeta_{n,i}\hat{\Xi}_n^{-1/2}|| \cdot ||\hat{\Xi}_n^{-1/2}(\hat{\Xi}_{\lambda_{2,n}}^{-1} - \hat{\Xi}_n^{-1})\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}Y_i||$$

By properties of least-squares projection, for any θ:

$$||\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\Sigma}_n^{-1/2}\hat{\pi}_{n,i}\zeta_{n,i}'\hat{\Xi}_n^{-1/2}\theta|| \leq \big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}(\zeta_{n,i}'\hat{\Xi}_n^{-1/2}\theta)^2\big)^{1/2} = ||\theta||$$

And so $||\hat{\Sigma}_n^{-1/2}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}\zeta_{n,i}'\Xi_n^{-1/2}|| \leq 1$. Further, note that:

$$\begin{aligned}
&||\hat{\Xi}_n^{1/2}(\hat{\Xi}_{\lambda_{1,n}}^{-1} - \hat{\Xi}_n^{-1})\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}Y_i|| \\
\leq &||\hat{\Xi}_n^{1/2}\hat{\Xi}_{\lambda_{1,n}}^{-1}\hat{\Xi}_n^{1/2} - I|| \cdot ||\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\Xi}_n^{-1/2}\zeta_{n,i}Y_i|| \\
\leq &||\hat{\Xi}_n^{1/2}\hat{\Xi}_{\lambda_{1,n}}^{-1}\hat{\Xi}_n^{1/2} - I||\big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}Y_i^2\big)^{1/2}
\end{aligned}$$

By Assumption 5.2.iv $\big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}Y_i^2\big)^{1/2} \precsim_p E[Y_i^2] \leq \bar{\sigma}_Y \precsim 1$, and note that:

$$\begin{aligned}
&||\hat{\Xi}_n^{1/2}\hat{\Xi}_{\lambda_{1,n}}^{-1}\hat{\Xi}_n^{1/2} - I|| \\
\leq &\lambda_{1,n}||\hat{\Xi}_n^{1/2}\hat{\Xi}_{\lambda_{1,n}}^{-1}\hat{\Xi}_n^{1/2}|| \cdot ||\Xi_n^{-1}|| \cdot ||\Xi_n^{1/2}\hat{\Xi}_n^{-1}\Xi_n^{1/2}||
\end{aligned}$$

Note that $||\hat{\Xi}_n^{1/2}\hat{\Xi}_{\lambda_{1,n}}^{-1}\hat{\Xi}_n^{1/2}|| \leq 1$, and by Assumption 5.4.iii and Lemma C.1 we have $||\Xi_n^{-1/2}\hat{\Xi}_n\Xi_n^{-1/2}|| \precsim_p 1$, and so $||\hat{\Xi}_n^{1/2}\hat{\Xi}_{\lambda_{1,n}}^{-1}\hat{\Xi}_n^{1/2} - I|| \precsim_p \lambda_{1,n}||\Xi_n^{-1}||$. Combining we get:

$$||\hat{r}_{2,n}|| \precsim_p \lambda_{1,n}||\Xi_n^{-1}|| \tag{1.8.21}$$

Rate for $||\hat{r}_{3,n}||$**:** By the triangle inequality and definition of the operator norm:

$$\begin{aligned}||\hat{r}_{3,n}|| \leq &\big(||\bar{\Sigma}_n^{1/2}\bar{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2} - \bar{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2}|| \\ &\times ||E\big[\bar{\Sigma}_n^{-1/2}\hat{\pi}_{n,i}\zeta_{n,i}'\Xi_n^{-1/2}\big|\mathcal{I}_\pi\big]||\big) \\ &+\big(||\bar{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2}|| \cdot ||E\big[\bar{\Sigma}_n^{-1/2}\hat{\pi}_{n,i}\zeta_{n,i}'\Xi_n^{-1/2}\big|\mathcal{I}_\pi\big] \\ &- \frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\bar{\Sigma}_n^{-1/2}\hat{\pi}_{n,i}\zeta_{n,i}'\Xi_n^{-1/2}||\big)\end{aligned}$$

From Lemma C.8:

$$||\bar{\Sigma}_n^{1/2}\bar{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2} - \bar{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2}|| \precsim_p (\xi_{\tilde{\psi},n}^2 log\big(l(n)m(n)\big)/n_g)^{1/4}$$

Because the RHS above converges to zero by Assumption 5.4.iv, this also implies $||\bar{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2}|| \precsim_p 1$. Applying Lemma C.4 conditional on the sample $\mathcal{I}_\pi$:

$$\begin{aligned}&||E\big[\bar{\Sigma}_n^{-1/2}\hat{\pi}_{n,i}\zeta_{n,i}'\Xi_n^{-1/2}\big|\mathcal{I}_\pi\big] - \frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\bar{\Sigma}_n^{-1/2}\hat{\pi}_{n,i}\zeta_{n,i}'\Xi_n^{-1/2}||^2 \\ \precsim_p &\frac{1}{n_g}E\big[||\bar{\Sigma}_n^{-1/2}\hat{\pi}_{n,i}||^2||\Xi_n^{-1/2}\zeta_{n,i}||^2\big|\mathcal{I}_\pi\big] \precsim_p \frac{1}{n_g}\xi_{\zeta,n}^2 k(n)l(n)\end{aligned}$$

In all:

$$||\hat{r}_{3,n}|| \precsim_p \Big(\sqrt{\xi_{\zeta,n}^2 k(n)l(n)/n_g} + \big(\xi_{\tilde{\psi},n}^2 log\big(l(n)m(n)\big)/n_g\big)^{1/4}\Big) \tag{1.8.22}$$

Rate for $||\hat{r}_{4,n}||$**:** Adding and subtracting terms and applying the definition of the operator norm we get:

$$\begin{aligned}&|\Xi_n^{1/2}(\tilde{\beta} - \beta_n) - \Xi_n^{-1/2}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}(r_{n,i}+\epsilon_i)| \\ \leq &||\Xi_n^{1/2}(\hat{\Xi}_n^{-1} - \Xi_n^{-1})\Xi_n^{1/2}|| \cdot ||\Xi_n^{-1/2}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}(r_{n,i}+\epsilon_i)|| \precsim_p \xi_{\zeta,n}\frac{p(n)}{n_g}\end{aligned}$$

Where the final line above follows because by Assumption 5.4.iii and Lemma C.1 $||\Xi_n^{1/2}(\hat{\Xi}_n^{-1} - \Xi_n^{-1})\Xi_n^{1/2}|| \precsim_p \sqrt{\xi_{\zeta,n}^2 p(n)/n_g}$ and we have already shown above that

$||\Xi_n^{-1/2}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}(r_{n,i}+\epsilon_i)|| \precsim_p \sqrt{\frac{p(n)}{n_g}}$. And so:

$$||\hat{r}_{4,n}|| \precsim_p \xi_{\zeta,n}\frac{p(n)}{n_g} \tag{1.8.23}$$

Combine previous steps: Together, (1.8), (1.8.17),(1.8.18), (1.8), (1.8.20), (1.8.21), (1.8.22), and (1.8.23) give:

$$\begin{aligned}
&\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)\\
&-\tilde{\alpha}(x_1,x_2)'\bar{\Sigma}_{\lambda_{0,n}}^{-1}E\big[\hat{\pi}_{n,i}\zeta_{n,i}'\big|\mathcal{I}_\pi\big]\Xi_n^{-1}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}(r_{n,i}+\epsilon_i)\\
\precsim_p&||\bar{\Sigma}_n^{1/2}\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||\\
&\times(R_{\theta,n}+R_{\pi,n}+\ell_{\zeta,n}(s_3)+\lambda_{1,n}||\Xi_n^{-1}||)\\
&+\xi_{\zeta,n}\frac{p(n)}{n_g}||\Xi_n^{-1/2}E\big[\zeta_{n,i}\hat{\pi}_{n,i}'\big|\mathcal{I}_\pi\big]\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}(x_1,x_2)||\\
&+R_n(x_1)
\end{aligned} \tag{1.8.24}$$

For $R_n(x_1)$ defined as in the statement of the theorem.

Apply Yurinskii's coupling for Gaussian approximation Yurinskii's coupling (see e.g., Theorem 10 in [68] or [17]), states that for any $\delta > 0$ and sequence of independent, zero-mean length-K_n vectors $a_{n,i}$ with finite third moments, there a length-K_n multivariate Guassian $\mathcal{N}_n$ with mean zero and same covariance matrix as $\frac{1}{\sqrt{n}}\sum_{i=1}^n a_{n,i}$ so that:

$$P\big[||\frac{1}{\sqrt{n}}\sum_{i=1}^n a_{n,i} - \mathcal{N}_n|| > 3\delta\big] \leq C_0K_nq_nn^{-1/2}\delta^{-3}\big(1+\frac{|-log(K_nq_n\delta^{-3}n^{-1/2})|}{K_n}\big)$$

C_0 is a finite constant and $q_n = E[\frac{1}{\sqrt{n}}\sum_{i=1}^n||a_{n,i}||^3]$. To apply this in our case we take the probability space conditional on the sample $\mathcal{I}_\pi$ so that $\mathcal{N}_n$ is independent of $\mathcal{I}_\pi$ and so we take the index i over $\mathcal{I}_g$ so that n is replaced by n_g in the above. For each $i\in\mathcal{I}_g$, let $a_{n,i} = \Gamma^{-1/2}\zeta_{n,i}(r_{n,i}+\epsilon_i)$ and note that $E[a_{n,i}] = 0$ and $E[a_{n,i}a_{n,i}'] = I$.

Furthermore K_n is replaced by $p(n)$ and the average third moment q_n satisfies:

$$\begin{aligned}
q_n &= E\big[||\Gamma_n^{-1/2}\zeta_{n,i}(r_{n,i}+\epsilon_i)||^3\big|\mathcal{I}_\pi\big] \\
&\leq ||\Xi_n^{1/2}\Gamma_n^{-1}\Xi_n^{1/2}||^{1/2}E\big[||\Xi_n^{-1/2}\zeta_{n,i}||^2\big]\operatorname{ess\,sup}||\Xi_n^{-1/2}\zeta_{n,i}|| \\
&\times \big(\operatorname{ess\,sup}|r_{n,i}| + \operatorname{ess\,sup} E[\epsilon_i^3|X_i,Z_i]^{1/3}\big)^3 \\
&\precsim p(n)\xi_{\psi,n}
\end{aligned}$$

Where we have used that by Assumption 5.5.iii, $\operatorname{ess\,sup} E[\epsilon_i^3|X_i,Z_i]^{1/3} \precsim 1$. Assumption 5.5.i states that $p(n)^2\xi_{\zeta,n}n_g^{-1/2}R_{\mathcal{N},n}^{-3}log(n_g) \to 0$ and so for sufficiently large n:

$$\begin{aligned}
&p(n)q_n n_g^{-1/2}\delta^{-3}R_{\mathcal{N},n}^{-3}\big(1+\frac{|-log(p(n)q_n n_g^{-1/2}\delta^{-3}R_{\mathcal{N},n}^{-3})|}{p(n)}\big) \\
\leq &p(n)q_n n_g^{-1/2}\delta^{-3}R_{\mathcal{N},n}^{-3}\big(1+\frac{log(n_g)}{2p(n)}\big) \prec 1
\end{aligned}$$

So using the Yurinskii coupling gives us that for all δ:

$$P\big[||\frac{1}{\sqrt{n_g}}\sum_{i\in\mathcal{I}_g}\Gamma_n^{-1/2}\zeta_{n,i}(r_{n,i}+\epsilon_i) - \mathcal{N}_n|| > 3R_{\mathcal{N},n}\delta\big|\mathcal{I}_\pi\big] \prec_p 1$$

Where the multivariate normal $\mathcal{N}_n$ has the same variance covariance matrix as the random vector $\frac{1}{\sqrt{n_g}}\sum_{i\in\mathcal{I}_g}\Gamma_n^{-1/2}\zeta_{n,i}(r_{n,i}+\epsilon_i)$ which is simply the identity matrix. Because all of the expressions inside the probability are independent of the observations in sub-sample $\mathcal{I}_\pi$ the above implies:

$$P\big[||\frac{1}{\sqrt{n_g}}\sum_{i\in\mathcal{I}_g}\Gamma_n^{-1/2}\zeta_{n,i}(r_{n,i}+\epsilon_i) - \mathcal{N}_n|| > 3R_{\mathcal{N},n}\delta\big] \prec_p 1$$

Then using the definition of s_n and (1.8.24) :

$$\begin{aligned}
&\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2)-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{||s_n(x_1,x_2)||}-\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||}\mathcal{N}_n\\
&\precsim_p R_{\mathcal{N},n}+\frac{R_n(x_1)}{||s_n(x_1,x_2)||}\\
&\quad+\xi_{\psi,n}\frac{m(n)}{n_g}\frac{||\tilde{\alpha}(x_1,x_2)'\bar{\Sigma}_{\lambda_{0,n}}^{-1/2}E\big[\hat{\pi}_{n,i}\zeta_{n,i}'\big|\mathcal{I}_\pi\big]\Xi_n^{-1/2}||}{||s(x_1,x_2)||}\\
&\quad+\frac{||\bar{\Sigma}_n^{1/2}\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||}{||s(x_1,x_2)||}(R_{\theta,n}+R_{\pi,n}+\ell_{\zeta,n}(s_3)+\lambda_{1,n}||\Xi_n^{-1}||)
\end{aligned}$$

Now note that:

$$||s(x_1,x_2)||\geq||\Xi_n^{1/2}\Gamma_n^{-1/2}||^{-1}||E\big[\Xi_n^{-1/2}\zeta_{n,i}\hat{\pi}_{n,i}'\big|\mathcal{I}_\pi\big]\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}(x_1,x_2)||$$

Recall $\Gamma_n=E\big[\zeta_{n,i}\zeta_{n,i}'(r_{n,i}+\epsilon_i)^2\big]$, then:

$$||\Xi_n^{1/2}\Gamma_n^{-1}\Xi_n^{1/2}||^2\geq\operatorname{ess\,inf}E[\epsilon_i^2|X_i,Z_i]$$

By Assumption 5.5.ii, $\operatorname{ess\,inf}E[\epsilon_i^2|X_i,Z_i]\succsim_p 1$ so $||\Xi_n^{1/2}\Gamma_n^{-1}\Xi_n^{1/2}||\succsim 1$. The matrix $\Xi_n^{1/2}\Gamma_n^{-1/2}$ is square and so $||\Xi_n^{1/2}\Gamma_n^{-1/2}||=||\Xi_n^{1/2}\Gamma_n^{-1}\Xi_n^{1/2}||^{1/2}$, hence:

$$||s(x_1,x_2)||\succsim_p||E\big[\Xi_n^{-1/2}\zeta_{n,i}\hat{\pi}_{n,i}'\big|\mathcal{I}_\pi\big]\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}(x_1,x_2)||$$

Furthermore, using that $\hat{\pi}_n(x,z)=\big(\hat{\omega}_n'\psi_n(x,z)\big)\otimes\chi_n(x)$:

$$\begin{aligned}
&||E\big[\Xi_n^{-1/2}\zeta_{n,i}\hat{\pi}_{n,i}'\big|\mathcal{I}_\pi\big]\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}(x_1,x_2)||\\
=&||E\big[\Xi_n^{-1/2}\zeta_{n,i}\tilde{\psi}_{n,i}'\tilde{\Omega}_n^{-1/2}\big]\tilde{\Omega}_n^{1/2}vec\big(\hat{\omega}_n\iota(\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}(x_1,x_2)\big)||\\
\geq&||E\big[\Xi_n^{-1/2}\zeta_{n,i}\tilde{\psi}_{n,i}'\tilde{\Omega}_n^{-1/2}\big]^{-1}||^{-1}||\tilde{\Omega}_n^{1/2}vec\big(\hat{\omega}_n\iota(\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}(x_1,x_2)\big)||\\
\geq&||E\big[\Xi_n^{-1/2}\zeta_{n,i}\tilde{\psi}_{n,i}'\tilde{\Omega}_n^{-1/2}\big]^{-1}||^{-1}||\bar{\Sigma}_n^{1/2}\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}(x_1,x_2)||
\end{aligned}$$

By definition $||E\big[\Xi_n^{-1/2}\zeta_{n,i}\tilde{\psi}_{n,i}'\tilde{\Omega}_n^{-1/2}\big]^{-1}|| \precsim b_n$, and so we have:

$$\begin{aligned}&\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{||s_n(x_1,x_2)||} - \frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||}\mathcal{N}_n \\ &\precsim_p R_{\mathcal{N},n} + \xi_{\zeta,n}\frac{p(n)}{n_g} + b_n(R_{\theta,n} + R_{\pi,n} + \ell_{\zeta,n}(s_3) + \lambda_{1,n}||\Xi_n^{-1}||) \\ &\quad + \frac{\sqrt{n_g}R_n(x_1)}{||E\big[\Xi_n^{-1/2}\zeta_{n,i}\big(\hat{\pi}_{n,i}'\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}(x_1,x_2)\big)\big|\mathcal{I}_\pi\big]||}\end{aligned}$$

And by definition $||E\big[\Xi_n^{-1/2}\zeta_{n,i}\big(\hat{\pi}_{n,i}'\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}(x_1,x_2)\big)\big|\mathcal{I}_\pi\big]|| \succsim_p \underline{c}_n(x_1,x_2)$, which gives the result. $\square$

Proof of Theorem 3.4. Define $Y_{i,b} = \sqrt{Q_{b,i}}Y_i$, define $\zeta_{n,i,b} = \sqrt{Q_{b,i}}\zeta_{n,i}$, and let $\hat{\Xi}_{\lambda_{1,n},b} = \hat{\Xi}_{n,b} + \lambda_{1,n}I$ for $\hat{\Xi}_{n,b} = \frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\zeta_{n,i,b}\zeta_{n,i,b}'$. The bootstrap estimator $\hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b$ differs from the estimator $\hat{\alpha}_n(x_1,x_2)'\hat{\theta}$ only in that $\hat{g}_i$ in the formula for $\hat{\theta}$ is replaced by $\hat{g}_{i,b} = \zeta_{n,i}'\hat{\beta}_b$ where $\hat{\beta}_b$ is defined by $\hat{\beta}_b = \hat{\Xi}_{\lambda_{1,n},b}^{-1}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\zeta_{n,i,b}Y_{i,b}$.

Let $r_{n,i} = g_i - \zeta_{n,i}'\beta_n$ where $\beta_n = \Xi_n^{-1}E[\zeta_{n,i}Y_i]$ and $\epsilon_i = Y_i - g_i$. Note then that $Y_{i,b} = \zeta_{n,i,b}'\beta_n + \sqrt{Q_{b,i}}r_{n,i} + \sqrt{Q_{b,i}}\epsilon_i$.

This decomposition can take the place of the analogous decomposition of Y_i in the proofs of Lemmas 3.7 and Theorem 3.3. $\{Q_{b.i}\}_{i=1}^{n_g}$ are independent of the data and $E[Q_{b.i}] = 1$, so $E[\zeta_{n,i,b}\zeta_{n,i,b}'] = \Xi_n$ and using Assumption 5.2.iv, $E[(\sqrt{Q_{b,i}}\epsilon_i)^2|X_i,Z_i] \leq \bar{\sigma}_Y^2$. Further, $E\big[\zeta_{n,i,b}(\sqrt{Q_{b,i}}r_{n,i} + \sqrt{Q_{b,i}}\epsilon_i)\big] = 0$. $\max_{i\in\mathcal{I}_g}|Q_{b,i}| \precsim_p ln(n_g)$, and so:

$$\begin{aligned}\max_{i\in\mathcal{I}_g}||E[\zeta_{n,i,b}\zeta_{n,i,b}']^{-1/2}\zeta_{n,i,b}|| &= \max_{i\in\mathcal{I}_g}||\sqrt{\frac{Q_{b,i}}{E[Q_{b,i}]}}E[\zeta_{n,i}\zeta_{n,i}']^{-1/2}\zeta_{n,i}|| \\ &\leq \xi_{\zeta,n}\sqrt{\max_{i\in\mathcal{I}_g}|Q_{b,i}|} \precsim_p \xi_{\zeta,n}\sqrt{ln(n_g)}\end{aligned}$$

Given the above it is clear that the analysis of the non-bootstrap estimator applies unchanged for the bootstrap estimator with the exception that the rate $\xi_{\zeta,n}$ is replaced by $\xi_{\zeta,n}\sqrt{ln(n_g)}$. Applying Lemma C.1 and Assumption 5.5.iv this gives $||\Xi_n^{1/2}\hat{\Xi}_{n,b}^{-1}\Xi_n^{1/2} - I|| \precsim_p \sqrt{ln(n_g)\xi_{\zeta,n}^2 p(n)/n_g} \prec 1$, which is slower than the non-bootstrap equivalent by a factor $\sqrt{ln(n_g)}$.

Following the steps in Lemma C.7 for the bootstrap estimator and we get:

$$(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\big(\hat{g}_{i,b}-g_i\big)^2\big)^{1/2}\precsim_p\sqrt{\frac{p(n)}{n_g}}+\ell_{\zeta,n}(s_3)+\lambda_{1,n}||\Xi_n^{-1}||=R_g$$

Where R_g is given in Theorem 3.2. And following steps in Theorem 3.3 for the bootstrap estimator we get:

$$\begin{aligned}&\big(\bar{y}(x_1|x_2)-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)\\&-\tilde{\alpha}(x_1,x_2)'\bar{\Sigma}_{\lambda_{0,n}}^{-1}E\big[\hat{\pi}_{n,i}\zeta_{n,i}'\big|\mathcal{I}_\pi\big]\Xi_n^{-1}\sum_{i\in\mathcal{I}_g}Q_{b,i}\zeta_{n,i}(r_{n,i}+\epsilon_i)\\&\precsim_p||\bar{\Sigma}_n^{1/2}\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||(R_{\theta,n}+R_{\pi,n}+\ell_{\zeta,n}(s_3)+\xi_{\tilde{\Omega},n}(x_1)\lambda_{2,n}||\Xi_n^{-1}||)\\&+ln(n_g)\xi_{\zeta,n}\frac{p(n)}{n_g}||\tilde{\alpha}(x_1,x_2)'\bar{\Sigma}_{\lambda_{0,n}}^{-1/2}E\big[\hat{\pi}_{n,i}\psi_{n,i}'\big|\mathcal{I}_\pi\big]\Xi_n^{-1/2}||+R_n(x_1)\end{aligned}$$

Note that this differs from the rate (1.8.24) derived in the proof of Theorem 3.3 only in the multiplication by $ln(n_g)$ in the second line due to the slower convergence of $||\Xi_n^{1/2}\hat{\Xi}_{n,b}^{-1}\Xi_n^{1/2}-I||$. Subtracting (1.8.24) from the above:

$$\begin{aligned}&\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta}-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)\\&-\tilde{\alpha}(x_1,x_2)'\bar{\Sigma}_{\lambda_{0,n}}^{-1}E\big[\hat{\pi}_{n,i}\zeta_{n,i}'\big|\mathcal{I}_\pi\big]\Xi_n^{-1}\sum_{i\in\mathcal{I}_g}(Q_{b,i}-1)\zeta_{n,i}(r_{n,i}+\epsilon_i)\\&\precsim_p||\bar{\Sigma}_n^{1/2}\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||(R_{\theta,n}+R_{\pi,n}+\ell_{\zeta,n}(s_3)+\xi_{\tilde{\Omega},n}(x_1)\lambda_{2,n}||\Xi_n^{-1}||)\\&+ln(n_g)\xi_{\zeta,n}\frac{p(n)}{n_g}||\tilde{\alpha}(x_1,x_2)'\bar{\Sigma}_{\lambda_{0,n}}^{-1/2}E\big[\hat{\pi}_{n,i}\psi_{n,i}'\big|\mathcal{I}_\pi\big]\Xi_n^{-1/2}||+R_n(x_1)\end{aligned}\tag{1.8.25}$$

The Yurinskii's coupling argument in Theorem 3.3 can be applied with

$$\frac{1}{\sqrt{n_g}}\sum_{i\in\mathcal{I}_g}\Gamma_n^{-1/2}(Q_{b,i}-1)\zeta_{n,i}(r_{n,i}+\epsilon_i)$$

replacing $\frac{1}{\sqrt{n_g}}\sum_{i\in\mathcal{I}_g}\Gamma_n^{-1/2}\zeta_{n,i}(r_{n,i}+\epsilon_i)$ and conditioning on the whole data $\mathcal{I}$ rather than just the sub-sample $\mathcal{I}_\pi$. The average third moment conditional on the data is stochastically bounded by the unconditional average third moment in Theorem 3.3, that is:

$$E\big[||\Gamma_n^{-1/2}(Q_{b,i}-1)\zeta_{n,i}(r_{n,i}+\epsilon_i)||^3\big|\mathcal{I}\big] \precsim_p E\big[||\Gamma_n^{-1/2}\zeta_{n,i}(r_{n,i}+\epsilon_i)||^3\big]$$

Then applying the steps in Theorem 3.3 we get:

$$P\big[||\frac{1}{\sqrt{n_g}}\sum_{i\in\mathcal{I}_g}\Gamma_n^{-1/2}(Q_{b,i}-1)\zeta_{n,i}(r_{n,i}+\epsilon_i)-\hat{\mathcal{N}}_n||>3R_{\mathcal{N},n}\delta\big|\mathcal{I}\big] \prec_p 0$$

Where $\hat{\mathcal{N}}_n$ is a multivariate Guassian independent of the data with covariance matrix $\Gamma_n^{-1/2}\tilde{\Gamma}_n\Gamma_n^{-1/2}$ where:

$$\tilde{\Gamma}_n = E\big[\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}(Q_{b,i}-1)^2\zeta_{n,i}\zeta_{n,i}'(r_{n,i}+\epsilon_i)^2|\mathcal{I}\big]$$

which equals $\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}\zeta_{n,i}'(r_{n,i}+\epsilon_i)^2$. We can define a multivariate Gaussian random vector $\tilde{\mathcal{N}}_n$ that is independent of the data and has identity covariance matrix so that $\hat{\mathcal{N}}_n = (\Gamma_n^{-1/2}\tilde{\Gamma}_n\Gamma_n^{-1/2})^{1/2}\tilde{\mathcal{N}}_n$. The remaining arguments in Theorem 3.3 go through unchanged and so uniformly over F_X-almost all x_1 and x_2:

$$\begin{aligned}
&\frac{\sqrt{n_g}\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta}-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)-s_n(x_1,x_2)'(\Gamma_n^{-1/2}\tilde{\Gamma}_n\Gamma_n^{-1/2})^{1/2}\tilde{\mathcal{N}}_n}{||s_n(x_1,x_2)||}\\
\precsim_p & R_{\mathcal{N},n}+ln(n_g)\xi_{\zeta,n}\frac{p(n)}{n_g}+b_n(R_{\theta,n}+R_{\pi,n}+\ell_{\zeta,n}(s_3)+\lambda_{1,n}||\Xi_n^{-1}||)\\
&+\frac{b_n\sqrt{n_g}R_n(x_1)}{\underline{c}_n(x_1,x_2)}
\end{aligned}$$

By Lemma C.4:

$$\begin{aligned}
&||\Gamma_n^{-1/2}\tilde{\Gamma}_n\Gamma_n^{-1/2}-I||^2\\
=&||\frac{1}{n_g}\big(\sum_{i\in\mathcal{I}_g}\Gamma_n^{-1/2}\zeta_{n,i}\zeta_{n,i}'(r_{n,i}+\epsilon_i)^2\Gamma_n^{-1/2}-I\big)||^2\\
\precsim_p&\frac{1}{n_g}E\big[||\Gamma_n^{-1/2}\zeta_{n,i}(r_{n,i}+\epsilon_i)||^4\big]\\
\precsim_p&\frac{1}{n_g}\xi_{\zeta,n}^2p(n)\big(\operatorname{ess\,sup} r_{n,i}^4+\operatorname{ess\,sup}[\epsilon_i^4|X_i,Z_i]\big)\precsim_p\frac{1}{n_g}\xi_{\zeta,n}^2p(n)
\end{aligned}$$

Where we have used that $||\Gamma_n^{-1/2}\Xi_n^{1/2}||^2 \leq ||\Xi_n^{1/2}\Gamma_n^{-1}\Xi_n^{1/2}|| \precsim_p 1$ as shown in the proof of Theorem 3.3, and $\operatorname{ess\,sup}|r_{n,i}| \prec 0$ from Assumptions 5.1.iv and 5.3.ii, and $\operatorname{ess\,sup} E[\epsilon_i^4|X_i, Z_i] < \infty$ by Assumption 5.5.v. For any symmetric positive definite matrix A, $||A^{1/2} - I|| \leq ||A - I||$, and so:

$$||(\Gamma_n^{-1/2}\tilde{\Gamma}_n\Gamma_n^{-1/2})^{1/2} - I|| \precsim_p \sqrt{\frac{\xi_{\zeta,n}^2 p(n)}{n_g}}$$

It follows then that:

$$\begin{aligned}||\tilde{\mathcal{N}}_n - (\Gamma_n^{-1/2}\tilde{\Gamma}_n\Gamma_n^{-1/2})^{1/2}\tilde{\mathcal{N}}_n|| &\leq ||(\Gamma_n^{-1/2}\tilde{\Gamma}_n\Gamma_n^{-1/2})^{1/2} - I|| \cdot ||\tilde{\mathcal{N}}_n|| \\ &\precsim_p \sqrt{\frac{\xi_{\zeta,n}^2 p(n)^2}{n_g}}\end{aligned}$$

Where we have used that $||\tilde{\mathcal{N}}_n||^2 \precsim_p p(n)$ by Markov's inequality. And so uniformly over F_X-almost all x_1 and x_2:

$$\begin{aligned}&\frac{\sqrt{n_g}\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta} - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)}{||s_n(x_1,x_2)||} - \frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||}\tilde{\mathcal{N}}_n \\ &\precsim_p R_{\mathcal{N},n} + \sqrt{\frac{\xi_{\zeta,n}^2 p(n)^2}{n_g}} + ln(n_g)\xi_{\zeta,n}\frac{p(n)}{n_g} \\ &+ b_n(R_{\theta,n} + R_{\pi,n} + \ell_{\zeta,n}(s_3) + \lambda_{1,n}||\Xi_n^{-1}||) + \frac{b_n\sqrt{n_g}R_n(x_1)}{\underline{c}_n(x_1,x_2)}\end{aligned} \tag{1.8.26}$$

Using the condition on $\bar{r}_n$ we thus have that for some multivariate Guassian random vector $\tilde{\mathcal{N}}_n$, independent of the whole data $\mathcal{I}$ and with identity variance-covariance matrix:

$$\sup_{(x_1,x_2)\in\mathcal{X}} |\frac{\sqrt{n_g}\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta} - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)}{||s_n(x_1,x_2)||_2} - \frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\tilde{\mathcal{N}}_n| \prec_p \bar{r}_n \tag{1.8.27}$$

By Theorem 3.3 and since $\xi_{\zeta,n}\frac{p(n)}{n_g} \precsim \sqrt{\frac{\xi_{\zeta,n}^2 p(n)^2}{n_g}}$ we also have for some other multivariate Guassian random vector $\mathcal{N}_n$, independent of the sub-sample $\mathcal{I}_\pi$ with identity variance-covariance matrix:

$$\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2)-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{||s_n(x_1,x_2)||_2}-\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n|\prec_p \bar{r}_n \tag{1.8.28}$$

Let $e\hat{r}r_n$ denote the quantity on the LHS of (1.8.28) above and define the scalar $R_{\sigma,n}$ by $R_{\sigma,n}=\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{||s_n(x_1,x_2)||_2}{\hat{\sigma}_n(x_1,x_2)}-1|$. Note that:

$$\begin{aligned}&\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2)-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{\hat{\sigma}_n(x_1,x_2)}-\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n|\\ \leq& R_{\sigma,n}\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n|+(1+R_{\sigma,n})e\hat{r}r_{1,n}\prec_p \bar{r}_n\end{aligned} \tag{1.8.29}$$

Where we have used that $R_{\sigma,n}\prec_p \bar{r}_n^2\prec 1$, $e\hat{r}r_{1,n}\prec_p \bar{r}_n$, and by Markov's inequality:

$$\sup_{(x_1,x_2)\in\mathcal{X}}\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n\precsim_p E\big[\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n|\big]$$

which is $O(\bar{r}_n^{-1})$ by supposition. Call the LHS of the first inequality in (1.8.29) $e\tilde{r}r_{1,n}$. Now, for any random scalars a, b, and c:

$$|P(a\leq c)-P(b\leq c)|\leq P(|b-c|\leq|a-b|)$$

It follows that for any deterministic sequence c_n:

$$\begin{aligned}&\big|P\big(\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2)-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{\hat{\sigma}_n(x_1,x_2)}|\leq c_n|\mathcal{I}_\pi\big)\\ -&P\big(\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||}\mathcal{N}_n|\leq c_n|\mathcal{I}_\pi\big)\big|\\ \leq& P\big(\big|\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||}\mathcal{N}_n|-c_n\big|\leq e\tilde{r}r_{1,n}|\mathcal{I}_\pi\big)\\ \precsim_p& P\big(\big|\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||}\mathcal{N}_n|-c_n\big|\leq e\tilde{r}r_{1,n}\big)\end{aligned}$$

Where the final step follows by Markov's inequality. We have that $e\tilde{r}r_{1,n}\prec_p \bar{r}_n$ and so:

$$
\begin{aligned}
&P\big(\big|\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{s_n(x_1,x_2)'}{\hat{\sigma}_n(x_1,x_2)}\mathcal{N}_n| - c_n\big| \leq e\tilde{r}r_{1,n}\big)\\
\precsim &P\big(\big|\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{s_n(x_1,x_2)'}{\hat{\sigma}_n(x_1,x_2)}\mathcal{N}_n| - c\big| \leq \bar{r}_n\big)
\end{aligned}
$$

So in all:

$$
\begin{aligned}
&\big|P\big(\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{\hat{\sigma}_n(x_1,x_2)}| \leq c_n\big|\mathcal{I}_\pi\big)\\
-&P\big(\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||}\mathcal{N}_n| \leq c_n\big|\mathcal{I}_\pi\big)\big|\\
\precsim_p &P\big(\big|\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||}\mathcal{N}_n| - c_n\big| \leq \bar{r}_n\big)
\end{aligned}
\tag{1.8.30}
$$

By similar reasoning, from 1.8.27 similar steps give:

$$
\begin{aligned}
&\big|P\big(\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{\sqrt{n_g}\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta} - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)}{\hat{\sigma}_n(x_1,x_2)}| \leq c_n\big|\mathcal{I}\big)\\
-&P\big(\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n| \leq c_n\big|\mathcal{I}_\pi\big)\big|\\
\precsim_p &P\big(\big|\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n| - c_n\big| \leq \bar{r}_n\big)
\end{aligned}
\tag{1.8.31}
$$

In the above we have used that:

$$
\begin{aligned}
&P\big(\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\tilde{\mathcal{N}}_n| \leq c_n\big|\mathcal{I}\big)\\
=&P\big(\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n| \leq c_n\big|\mathcal{I}_\pi\big)
\end{aligned}
$$

By supposition, with probability 1 we have $|\frac{s_n(x_1,x_2)}{||s_n(x_1,x_2)||_2}'\mathcal{N}_n| < \infty$ for every pair $(x_1,x_2) \in \mathcal{X}$, and $E\big[\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{s_n(x_1,x_2)}{||s_n(x_1,x_2)||_2}'\mathcal{N}_n|\big] \precsim_p \bar{r}_n^{-1}$. By Corollary 2.1 in [26] this implies:

$$
\sup_{c\in\mathbb{R}} P\big(\big|\sup_{(x_1,x_2)\in\mathcal{X}}|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n| - c\big| \leq \bar{r}_n\big) \prec 1
$$

So from (1.8.30):

$$\Big|P\Big(\sup_{(x_1,x_2)\in\mathcal{X}}\Big|\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2)-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{\hat{\sigma}_n(x_1,x_2)}\Big|\leq c_n\Big|\mathcal{I}_\pi\Big)$$
$$-P\Big(\sup_{(x_1,x_2)\in\mathcal{X}}\Big|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n\Big|\leq c_n\Big|\mathcal{I}_\pi\Big)\Big|\prec_p 1 \tag{1.8.32}$$

And also using (1.8.31):

$$\Big|P\Big(\sup_{(x_1,x_2)\in\mathcal{X}}\Big|\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2)-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{\hat{\sigma}_n(x_1,x_2)}\Big|\leq c_n\Big|\mathcal{I}_\pi\Big)$$
$$-P\Big(\sup_{(x_1,x_2)\in\mathcal{X}}\Big|\frac{\sqrt{n_g}\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta}-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)}{\hat{\sigma}_n(x_1,x_2)}\Big|\leq c_n\Big|\mathcal{I}\Big)\Big|$$
$$\prec_p 1$$

By Glivenko-Cantelli, if $B\to\infty$ sufficiently quickly then the above implies:

$$\frac{1}{B}\sum_{b=1}^{B}1\Big\{\sup_{(x_1,x_2)\in\mathcal{X}}\Big|\frac{\sqrt{n_g}\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta}-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)}{\hat{\sigma}(x_1,x_2)}\Big|\leq c_n\Big\}$$
$$-P\Big(\sup_{(x_1,x_2)\in\mathcal{X}}\Big|\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2)-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{\hat{\sigma}_n(x_1,x_2)}\Big|\leq c_n\Big|\mathcal{I}_\pi\Big)\prec_p 1 \tag{1.8.33}$$

(1.8.33) and (1.8.32) together imply:

$$\frac{1}{B}\sum_{b=1}^{B}1\Big\{\sup_{(x_1,x_2)\in\mathcal{X}}\Big|\frac{\sqrt{n_g}\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta}-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)}{\hat{\sigma}(x_1,x_2)}\Big|\leq c_n\Big\}$$
$$-P\Big(\sup_{(x_1,x_2)\in\mathcal{X}}\Big|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n\Big|\leq c_n\Big|\mathcal{I}_\pi\Big)\prec_p 1 \tag{1.8.34}$$

$\sup_{(x_1,x_2)\in\mathcal{X}}\Big|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n\Big|$ is continuously distributed conditional on the sub-sample $\mathcal{I}_\pi$, and so for any $\tilde{a}\in(0,1)$ and any n there is some $\tilde{c}$ so that:

$$P\Big(\sup_{(x_1,x_2)\in\mathcal{X}}\Big|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n\Big|\leq\tilde{c}\Big|\mathcal{I}_\pi\Big)=\tilde{a}$$

It then follows from (1.8.34) that for any scalar $\eta\in(0,a\vee|a-1|)$ there must exist a sequence of critical values $\bar{c}_n$ so that:

$$\frac{1}{B}\sum_{b=1}^{B} 1\Big\{ \sup_{(x_1,x_2)\in\mathcal{X}} \Big|\frac{\sqrt{n_g}\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta} - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)}{\hat{\sigma}(x_1,x_2)}\Big| \leq \bar{c}_n \Big\} = 1 - a + \frac{1}{2}\eta + o_p(1)$$

And a sequence of critical values $\underline{c}_n$ so that:

$$\frac{1}{B}\sum_{b=1}^{B} 1\Big\{ \sup_{(x_1,x_2)\in\mathcal{X}} \Big|\frac{\sqrt{n_g}\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta} - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)}{\hat{\sigma}(x_1,x_2)}\Big| \leq \underline{c}_n \Big\} = 1 - a - \frac{1}{2}\eta + o_p(1)$$

Recall that $\hat{c}_{1-a}$ is the smallest c that satisfies:

$$\frac{1}{B}\sum_{b=1}^{B} 1\Big\{ \sup_{x_1,x_2\in\mathcal{X}} \Big|\frac{\hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}}{\hat{\sigma}(x_1,x_2)/\sqrt{n_g}}\Big| \leq \hat{c}_{1-a} \Big\} \leq 1 - a$$

So with probability approaching 1, $\hat{c}_{1-a} \in [\underline{c}_n, \bar{c}_n]$. Using (1.8.33) this then implies that with probability approaching 1:

$$P\Big(\sup_{(x_1,x_2)\in\mathcal{X}} \Big|\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{\hat{\sigma}_n(x_1,x_2)}\Big| \leq \hat{c}_{1-a} \Big| \mathcal{I}_\pi \Big) \in [1 - a - \eta, 1 - a + \eta]$$

Since η can be set to be arbitrarily small, it follows that:

$$P\Big(\sup_{(x_1,x_2)\in\mathcal{X}} \Big|\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{\hat{\sigma}_n(x_1,x_2)}\Big| \leq \hat{c}_{1-a} \Big| \mathcal{I}_\pi \Big) = 1 - a + o_p(1)$$

Using the definition of $\hat{\Theta}_{1-a}(x_1,x_2)$ the above implies:

$$P\big(\bar{y}(x_1|x_2) \in \hat{\Theta}_{1-a}(x_1,x_2), \forall (x_1,x_2) \in \mathcal{X} \big| \mathcal{I}_\pi\big) = 1 - a + o_p(1)$$

Convergence in probability of a bounded (in magnitude) random variable implies convergence of the mean, so:

$$P\big(\bar{y}(x_1|x_2) \in \hat{\Theta}_{1-a}(x_1,x_2), \forall (x_1,x_2) \in \mathcal{X}\big) = 1 - a + o(1)$$

$\square$

1.9 Appendix C: Additional Results and Supporting Lemmas

Proposition C.1. *Under Assumption 1:*

$$\bar{y}(x_1|x_2) = E\big[E[Y|W^*, X = x_1]\big|X = x_2\big]$$

Proof. By iterated expectations:

$$\bar{y}(x_1|x_2) = E\big[E[y_0(x_2, U)|W^*, X = x_2]\big|X = x_2\big]$$

By Assumption 1.i and the definition of y_0:

$$E[y_0(x_1, U)|W^* = w^*, X = x_2] = E[Y|W^* = w^*, X = x_1]$$

Under Assumption 1.ii $E[Y|W^* = w^*, X = x_1]$ is well-defined for $F_{W^*|X=x_2}$-almost all w^* (rather than just $F_{W^*|X=x_1}$-almost all w^*). So we can substitute to get:

$$\bar{y}(x_1|x_2) = E\big[E[Y|W^*, X = x_1]\big|X = x_2\big]$$

□

Proposition C.2. *Under Assumption 1.ii and 2.ii the Radon-Nikodym derivative* $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}$ *exists.*

Proof. Let $F_{(X,V)}[\mathcal{A}]$ be the probability that (X, V) is in the set $\mathcal{A}$. Note that:

$$\begin{aligned} F_{(X,V)}[\mathcal{A}] &= E\big[P\big[(X, V) \in \mathcal{A}\big|X, W^*\big]\big] \\ &= \int P[(x, V) \in \mathcal{A}\big|W^* = w^*\big] F_{(X,W^*)}\big(d(x, w^*)\big) \end{aligned}$$

Where the first equality follows by iterated expectations and the second by Assumption 2.ii. Let $F_X \otimes F_V[\mathcal{A}]$ be the product measure of $\mathcal{A}$. We have:

$$\begin{aligned} F_X \otimes F_V[\mathcal{A}] &= \int P[(x,V) \in \mathcal{A}] F_X(dx) \\ &= \int P[(x,V) \in \mathcal{A} | W^* = w^*] F_X \otimes F_{W^*}\big(d(x,w^*)\big) \end{aligned}$$

Where $F_X \otimes F_{W^*}$ is the product measure of F_X and F_{W^*}. By Assumption 1.ii, the measure $F_{(X,W^*)}$ is non-zero on precisely the sets for which $F_X \otimes F_{W^*}$ is non-zero. Since $P[(x,V) \in \mathcal{A}|W^* = w^*]$ is weakly positive it follows that $F_{(X,V)}[\mathcal{A}]$ is strictly positive if and only if $F_X \otimes F_V[\mathcal{A}]$ is strictly positive. Since this holds for any $\mathcal{A}$, $F_X \otimes F_V$ and $F_{(X,V)}$ are absolutely continuous. Existence of $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}$ then follows by the Radon-Nikodym theorem. $\square$

Lemma 1.1 proved below is an application of Theorems 15.16 and 15.18 (Picard) in [58].

Proof of Lemma 1.1. Let $A_{x_1,\bar{w}}$, $A^*_{x_1,\bar{w}}$, $\bar{W}$, $\tilde{W}^*$, $\tilde{V}$ and $\tilde{Z}$ be defined as in Appendix A.1. Suppose that for $F_{\bar{W}}$-almost all $\bar{w}$ and F_X-almost all x_1 and x_2 the function $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}\big((\cdot,\bar{w})\big)$ is in the range of $A^*_{x_1,\bar{w}}$. That is, there exists a function $\tilde{\varphi}_{x_1,x_2,\bar{w}} \in L_2(F_{\tilde{Z}|X=x_1,\bar{W}=\bar{w}})$ so that:

$$E[\tilde{\varphi}_{x_1,x_2,\bar{w}}(\tilde{Z})|X = x_1, \bar{W} = \bar{w}, \tilde{V} = \tilde{v}] = \frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}\big((\tilde{v},\bar{w})\big)$$

For $F_{\tilde{V}|X=x_1,\bar{W}=\bar{w}}$-almost all $\tilde{v}$. Further suppose that for $F_{\bar{W}|X=x_1}$-almost all $\bar{w}$, has $\tilde{\varphi}_{x_1,x_2,\bar{w}}$ has $L_2(F_{\tilde{Z}|X=x_1,\bar{W}=\bar{w}})$-norm bounded by $\sqrt{\tilde{C}(x_1,x_2,\bar{w})}$, that is:

$$E\big[\tilde{\varphi}_{x_1,x_2,\bar{w}}(\tilde{Z})^2|X = x_1, \bar{W} = \bar{w}\big] \leq \tilde{C}(x_1,x_2,\bar{w})$$

and $E\big[\tilde{C}(x_1,x_2,\bar{W})\big|X = x_1\big] \leq C(x_1,x_2)$. Letting $\varphi(x_1,x_2,(\tilde{z},\bar{w})) = \tilde{\varphi}_{x_1,x_2,\bar{w}}(\tilde{z})$, the conclusion a. of the lemma holds.

Similarly, the following implies conclusion b. For $F_{\bar{W}}$-almost all $\bar{w}$ and F_X-almost all x the following conditions hold. The function $E[Y|Z = (\cdot,\bar{w}), X = x]$ is in the range of $A_{x,\bar{w}}$ and that the solution $\tilde{\gamma}_{x,\bar{w}}$ has $L_2(F_{\tilde{V}|X=x,\bar{W}=\bar{w}})$-norm bounded by

$\sqrt{\tilde{D}(x,\bar{w})}$ and $E\big[\tilde{D}(x,\bar{w})\big|X=x\big] \leq D(x)$.

To establish the relevant functions are in the ranges of the desired operators we apply Theorem 15.18 (Picard) in [58]. This states the following. Let $T: H_1 \to H_2$ be a compact linear operator from a Hilbert space H_1 to a Hilbert space H_2 with singular system $\{(u_k^{(T)}, v_k^{(T)}, \mu_k^{(T)})\}_{k=1}^{\infty}$. Then $\delta \in H_2$ is in the range of T if and only if δ is in the orthogonal complement of the null space of the adjoint T^* and for some $c < \infty$:

$$\sum_{k=1}^{\infty} \frac{1}{(\mu_k^{(T)})^2} |\langle \delta, v_k^{(T)} \rangle|^2 \leq c \tag{1.9.1}$$

In which case the solution f with smallest norm has norm $\sqrt{c}$. $\langle \cdot, \cdot \rangle$ is the inner product of H_2.

To apply Picard's theorem and show $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}\big((\cdot,\bar{w})\big)$ is in the range of $A^*_{x_1,\bar{w}}$ with norm of a solution weakly less than $\sqrt{\tilde{C}(x_1,x_2,\bar{w})}$ we need to show: a. $A^*_{x_1,\bar{w}}$ is compact, b. that $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}\big((\cdot,\bar{w})\big) \in L_2(F_{\tilde{V}|X=x_1,\bar{W}=\bar{w}})$, and c. that $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}\big((\cdot,\bar{w})\big)$ is in the orthogonal complement of $A_{x_1,\bar{w}}$, and d. (1.9.1) holds for $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}\big((\cdot,\bar{w})\big)$ in place of δ and for the relevant Hilbert space and operator $A^*_{x_1,\bar{w}}$ and with c bounded by $\tilde{C}(x_1,x_2,\bar{w})$.

To show $E[Y|Z=(\cdot,\bar{w}),X=x]$ is in the range of $A_{x,\bar{w}}$ so that there is a solution with norm bounded by $\sqrt{\tilde{D}(x,\bar{w})}$, we need to show: that e. $A_{x,\bar{w}}$ is compact, f. $E[Y|Z=(\cdot,\bar{w}),X=x] \in L_2(F_{\tilde{Z}|X=x,\bar{W}=\bar{w}})$, g. that the regression function $E[Y|Z=(\cdot,\bar{w}),X=x]$ is in the orthogonal complement of $A^*_{x,\bar{w}}$, h. (1.9.1) holds for $E[Y|Z=(\cdot,\bar{w}),X=x]$ in place of δ and for the relevant Hilbert space and operator $A_{x,\bar{w}}$ and with c bounded by $\tilde{D}(x,\bar{w})$

First we show that Assumptions A.1 implies that the operator $A_{x,\bar{w}}$ and its adjoint $A^*_{x,\bar{w}}$ are compact and therefore have unique singular systems (points a. and e. above). To see this first note that:

$$\begin{aligned} A_{x,\bar{w}}[\delta](\tilde{z}) &= E[\delta(\tilde{V})|\tilde{Z}=\tilde{z},X=x,\bar{W}=\bar{w}] \\ &= \int \frac{dF_{(\tilde{V},\tilde{Z})|X=x,\bar{W}=\bar{w}}}{dF_{prod}}(\tilde{v},\tilde{z})\delta(\tilde{w}^*)F_{\tilde{V}|X=x,\bar{W}=\bar{w}}(d\tilde{v}) \end{aligned}$$

Thus $A_{x,\bar{w}} : L_2(F_{\tilde{V}|X=x,\bar{W}=\bar{w}}) \to L_2(F_{\tilde{Z}|X=x,\bar{W}=\bar{w}})$ is an integral operator with kernel $\frac{dF_{(\tilde{V},\tilde{Z})|X=x,\bar{W}=\bar{w}}}{dF_{prod}}$ and Assumption A.1 states that the kernel is square integral with respect to the product measure of $F_{\tilde{V}|X=x,\bar{W}=\bar{w}}$ and $F_{\tilde{Z}|X=x,\bar{W}=\bar{w}}$. This implies that the operator $A_{x,\bar{w}}$ is Hilbert-Schmidt and therefore compact (see for example Section 3.3.1 of [76]). Compactness of an operator implies compactness of its adjoint (alternatively we could simply repeat the steps above for $A^*_{x,\bar{w}}$). If $A_{x,\bar{w}}$ is compact then by Theorem 15.16 of [58] it admits a singular system. Note that the singular system of the adjoint $A^*_{x,\bar{w}}$ is the same as for $A_{x,\bar{w}}$ but with the roles of the singular functions $u_k^{(x,\bar{w})}$ and $v_k^{(x,\bar{w})}$ switched for each k. Thus we have shown a. and e. hold.

Next note that the first part of Assumptions 4.i states that

$$E\big[\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}\big((\tilde{V},\bar{W})\big)^2\big|X=x_1\big] < \infty$$

or in other words:

$$\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}\big((\cdot,\bar{w})\big) \in L_2(F_{\tilde{V}|X=x_1,\bar{W}=\bar{w}})$$

Similarly the first part of 4.ii is equivalent to:

$$E[Y|Z=(\cdot,\bar{w}),X=x] \in L_2(F_{\tilde{Z}|X=x_1,\bar{W}=\bar{w}})$$

Thus 4.i and 4.ii imply point b. and f. respectively.

Now let us show that under Assumptions 1 and 2, c. and g. hold, that is the functions $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}\big((\cdot,\bar{w})\big)$ and $E[Y|Z=(\cdot,\bar{w}),X=x]$ are in the orthogonal complements of the null spaces of operators $A_{x_1,\bar{w}}$ and $A^*_{x,\bar{w}}$ respectively.

Under Assumption 2.ii, for F_X-almost all x_1 and x_2, and F_V-almost all v:

$$E\big[\frac{dF_{W^*|X=x_2}}{dF_{W^*|X=x_1}}(W^*)\big|X=x_1,V=v\big] = \frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(v)$$

For intermediate steps that show the above see (1.8.1) in the proof of Theorem 1.1 in Appendix B. Now, let a function δ be in the null space of $A_{x_1,\bar{w}}$, that is $F_{\tilde{Z}|X=x_1,\bar{W}=\bar{w}}$-

almost surely:

$$E[\delta(\tilde{V})|X = x_1, \bar{W} = \bar{w}, \tilde{Z} = z] = 0$$

Then by iterated expectations and Assumption 2.ii:

$$\begin{aligned}
&E[\delta(\tilde{V})|X = x_1, \bar{W} = \bar{w}, \tilde{Z} = z] \\
=&E\big[E[\delta(\tilde{V})|X, W^*]\big|X = x_1, \bar{W} = \bar{w}, \tilde{Z} = z\big] \\
=&0
\end{aligned}$$

And so by Assumption 3.i $F_{W^*|X=x_1,\bar{W}=\bar{w}}$-almost surely:

$$E[\delta(\tilde{V})|X = x_1, W^* = w^*] = 0$$

But then we see that the $L_2(F_{\tilde{V}|X=x_1,\bar{W}=\bar{w}})$-inner product of δ and $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}\big((\cdot, \bar{w})\big)$ is zero:

$$\begin{aligned}
&E\big[\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(V)\delta(\tilde{V})\big|X = x_1, \bar{W} = \bar{w}\big] \\
=&E\big[E\big[\frac{dF_{W^*|X=x_2}}{dF_{W^*|X=x_1}}(W^*)\big|X, V\big]\delta(\tilde{V})\big|X = x_1, \bar{W} = \bar{w}\big] \\
=&E\big[\frac{dF_{W^*|X=x_2}}{dF_{W^*|X=x_1}}(W^*)E[\delta(\tilde{V})|X = x_1, W^*]\big|X = x_1, \bar{W} = \bar{w}\big] \\
=&0
\end{aligned}$$

Where the first equality follows by substituting for $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}$, the second by iterated expectations and the third because $E[\delta(\tilde{V})|X = x_1, W^*]$ is $F_{W^*|X=x_1,\bar{W}=\bar{w}}$-almost surely zero. Since the inner product is zero for any δ in the null-space of $A_{x_1,\bar{w}}$, by definition $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}\big((\cdot, \bar{w})\big)$ is in the orthogonal complement of the null-space. Thus c. holds.

Next note that Assumption 2.i and iterated expectations implies:

$$E\big[E[Y|W^*, X]\big|X = x, Z = z\big] = E[Y|X = x, Z = z]$$

Let a function δ be in the null space of $A^*_{x,\bar{w}}$, that is, for $F_{\tilde{V}|X=x,\bar{W}=\bar{w}}$-almost all $\tilde{v}$:

$$E[\delta(\tilde{Z})|X = x, \bar{W} = \bar{w}, \tilde{V} = \tilde{v}] = 0$$

Then by iterated expectations and Assumption 2.ii:

$$\begin{aligned}
&E[\delta(\tilde{Z})|X = x, \bar{W} = \bar{w}, \tilde{V} = \tilde{v}] \\
=&E\big[E[\delta(\tilde{Z})|X, W^*]\big|X = x, \bar{W} = \bar{w}, \tilde{V} = \tilde{v}\big] \\
=&0
\end{aligned}$$

And so by Assumption 3.ii $F_{W^*|X=x,\bar{W}=\bar{w}}$-almost surely:

$$E[\delta(\tilde{V})|X = x, W^* = w^*] = 0$$

But then we see that the $L_2(F_{\tilde{Z}|X=x,\bar{W}=\bar{w}})$-inner product of δ and the regression function $E[Y|X = x, Z = (\cdot, \bar{w})]$ is zero:

$$\begin{aligned}
&E\big[E[Y|X, Z]\delta(\tilde{Z})\big|X = x, \bar{W} = \bar{w}\big] \\
=&E\big[E\big[E[Y|X, W^*]\big|X, Z\big]\delta(\tilde{Z})\big|X = x, \bar{W} = \bar{w}\big] \\
=&E\big[E[Y|X, W^*]E[\delta(\tilde{Z})|X = x, W^*]\big|X = x, \bar{W} = \bar{w}\big] \\
=&0
\end{aligned}$$

And so $E[Y|X = x, Z = (\cdot, \bar{w})]$ is in the null space of $A^*_{x,\bar{w}}$. Thus g. holds.

Finally, points d. and h. One can then see that for each given $(x_1, \bar{w})$ Assumption 4.i. is precisely the condition (1.9.1) where $\frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}\big((\cdot, \bar{w})\big)$ is the function δ, the inner-product is that of the space $L_2(F_{\tilde{V}|X=x,\bar{W}=\bar{w}})$, and the singular values , $\mu_k^{(T)}$ and functions $v_k^{(T)}$ are those of $A^*_{x,\bar{w}}$. In particular, $\mu_k^{(T)}$ and $v_k^{(T)}$ are given by $\mu_k^{(x,\bar{w})} = \mu_k(x, \bar{w})$ and $u_k^{(x,\bar{w})} = u_k(x, (\cdot, \bar{w}))$. Moreover, c is replaced by $\tilde{C}(x_1, x_2, \bar{w})$. For each given $(x, \bar{w})$ Assumption 4.ii is precisely the condition (1.9.1) with δ given by $E[Y|Z = (\cdot, \bar{w}), X = x]$, the inner-product that of $L_2(F_{\tilde{Z}|X=x,\bar{W}=\bar{w}})$, and the singular system that of $A_{x,\bar{w}}$ and c equal to $\tilde{D}(x, \bar{w})$.

$\square$

In this appendix we provide proofs for the results in Section 3. Throughout, expectations with a subscript only integrate over the random variables in that subscript, the expectation treats all other random variables as fixed. For example, $E_Z[ZX] = \int zXF_Z(dz) = E[Z]X$ and $E_Z[(\pi(x,Z)'\hat{\theta})^2|X=x] = \int(\pi(x,z)'\hat{\theta})^2F_{Z|X=x}(dz)$.

Lemma C.1 below simply lists some consequences of Rudelson's matrix law of large numbers ([70]) proxy which are used throughout subsequent proofs.

Lemma C.1. *Suppose that for each n, $\{a_{n,i}\}_{i=1}^n$ is a sequence of independent length-$q(n)$ random vectors so that $A_n = \frac{1}{n}\sum_{i=1}^n E[a_{n,i}a_{n,i}']$ is nonsingular. If, for each i and n, $ess\sup ||A_n^{-1/2}a_{n,i}|| \leq \xi_n$ almost surely $\frac{\xi_n^2 log(q(n))}{n} \prec 1$ then letting $\hat{A}_n = \frac{1}{n}\sum_{i=1}^n a_{n,i}a_{n,i}'$:*

$$||A_n^{-1/2}\hat{A}_nA_n^{-1/2} - I|| \precsim_p \sqrt{\frac{\xi_n^2 log(q(n))}{n}} \prec_p 1$$

Where I is the identity matrix of dimension $q(n)$. Moreover, under the same conditions:

$$||A_n^{1/2}\hat{A}_n^{-1}A_n^{1/2} - I|| \precsim_p ||A_n^{-1/2}\hat{A}_nA_n^{-1/2} - I||$$

Furthermore, we have $||A_n^{-1/2}\hat{A}_nA_n^{-1/2}|| \precsim_p 1$, $||A_n^{1/2}\hat{A}_n^{-1}A_n^{1/2}|| \precsim_p 1$, $||A_n^{-1/2}\hat{A}_n^{1/2}|| \precsim_p 1$, $||A_n^{1/2}\hat{A}_n^{-1/2}|| \precsim_p 1$, $||\hat{A}_n^{-1/2}A_n^{1/2}|| \precsim_p 1$, $||\hat{A}_n^{1/2}A_n^{-1/2}|| \precsim_p 1$, and uniformly over all $\lambda \geq 0$, $||A_n^{1/2}(\hat{A}_n + \lambda I)^{-1}A_n^{1/2}|| \precsim_p 1$.

Proof. The first result follows immediately from Rudelson's matrix LLN ([70]). By the triangle inequality:

$$||A_n^{-1/2}\hat{A}_nA_n^{-1/2}|| \leq ||A_n^{-1/2}\hat{A}_nA_n^{-1/2} - I|| + 1 \precsim_p 1$$

Next note that for any nonsingular matrix A with $||A-I|| < 1$ we have $||A^{-1} - I|| \leq \frac{||A-I||}{1-||A-I||}$. It follows that if $||A_n^{-1/2}\hat{A}_nA_n^{-1/2} - I|| \prec_p 1$, then $||A_n^{1/2}\hat{A}_n^{-1}A_n^{1/2} - I|| \precsim_p ||A_n^{-1/2}\hat{A}_nA_n^{-1/2} - I||$. Again, applying the triangle inequality $||A_n^{1/2}\hat{A}_n^{-1}A_n^{1/2}|| \precsim_p 1$. Next note that for any matrix A we must have $||A|| \leq ||A'A||^{1/2}$ and if A is square then

$||A|| = ||A'||$, and so, under the same conditions $||A_n^{-1/2}\hat{A}_n^{1/2}|| \precsim_p 1$, $||A_n^{1/2}\hat{A}_n^{-1/2}|| \precsim_p 1$, $||\hat{A}_n^{-1/2}A_n^{1/2}|| \precsim_p 1$, and finally $||\hat{A}_n^{1/2}A_n^{-1/2}|| \precsim_p 1$. Now note that:

$$\begin{aligned}
||A_n^{1/2}(\hat{A}_n + \lambda I)^{-1}A_n^{1/2}|| &= ||(A_n^{-1/2}\hat{A}_n A_n^{-1/2} + \lambda A_n^{-1})^{-1}|| \\
&\leq ||(A_n^{-1/2}\hat{A}_n A_n^{-1/2})^{-1}|| \\
&= ||A_n^{1/2}\hat{A}_n^{-1}A_n^{1/2}|| \precsim_p 1
\end{aligned}$$

Where the inequality in the second line holds because λA_n^{-1} and $A_n^{-1/2}\hat{A}_n A_n^{-1/2}$ are positive definite with probability approaching 1 and for any positive definite matrices A and B we have $||(A+B)^{-1}|| \leq ||A^{-1}||$. $\square$

Lemma C.2. *Let Assumptions 5.1.i, 5.1.ii, and 5.2.ii hold and that ϕ_n is of the form $\phi_n(x, v) = \rho_n(v) \otimes \chi_n(x)$. Then there exist constants $\underline{c} > 0$ and $\bar{c} < \infty$ so that for all n and $\theta \in \mathbb{R}^{k(n)l(n)}$:*

$$\underline{\mu}_n^2 \underline{c}||\theta||^2 \leq E\big[|\phi_n(X, V)'\theta|^2\big] \leq \bar{\mu}_n \bar{c}||\theta||^2$$

where $\underline{\mu}_n^2 = \mu_{min}(Q_n)\mu_{min}(Q_n)$ and $\bar{\mu}_n^2 = \mu_{max}(Q_n)\mu_{max}(Q_n)$ for $\mu_{min}(Q_n)$ and $\mu_{max}(Q_n)$ respectively the smallest and largest eigenvalues of Q_n and likewise for G_n. Furthermore, for all n, all $\theta \in \mathbb{R}^{k(n)l(n)}$ and F_X-almost all x:

$$E\big[|\phi_n(x, V)'\theta|^2\big] \leq \frac{\xi_{\chi,n}^2}{\underline{c}} E\big[|\phi_n(X, V)'\theta|^2\big]$$

Proof. Applying the separability of ϕ_n and properties of the Kronecker product we have that for any $\theta \in \mathbb{R}^{k(n)l(n)}$:

$$\begin{aligned}
E\big[|\phi_n(X, V)'\theta|^2\big] &= E\big[\theta'\phi_n(X, V)\phi_n(X, V)'\theta\big] \\
&= E\big[\theta'\big[\big(\rho_n(V)\rho_n(V)'\big) \otimes \big(\chi_n(X)\chi_n(X)'\big)\big]\theta\big] \qquad (1.9.2)
\end{aligned}$$

By Assumption 5.2.ii there exist constants $\underline{c} > 0$ and $\bar{c} < \infty$ so that so that $\underline{c} \leq$

$\frac{dF_{(X,V)}}{dF_X \otimes F_V}(x,v) \leq \bar{c}$ for $F_{(X,V)}$-almost all (x,v). Therefore, for any $\theta \in \mathbb{R}^{k(n)l(n)}$:

$$\begin{aligned}
&\underline{c}\theta'\big(E\big[\rho_n(V)\rho_n(V)'\big] \otimes E\big[\chi_n(X)\chi_n(X)'\big]\big)\theta \\
\leq &E\big[\theta'\big[\big(\rho_n(V)\rho_n(V)'\big) \otimes \big(\chi_n(X)\chi_n(X)'\big)\big]\theta\big] \\
\leq &\bar{c}\theta'\big(E\big[\rho_n(V)\rho_n(V)'\big] \otimes E\big[\chi_n(X)\chi_n(X)'\big]\big)\theta
\end{aligned}$$

By elementary properties of the Kronecker product, $A \otimes B$ is non-singular if and only if A and B are each non-singular. By Assumptions 5.1.i and 5.1.ii, $E\big[\rho_n(V)\rho_n(V)'\big]$ and $E\big[\chi_n(X)\chi_n(X)'\big]$ are non-singular and thus so is their Kronecker product. Substituting (1.9.2) the above implies:

$$\begin{aligned}
&\underline{c}||\big(E\big[\rho_n(V)\rho_n(V)'\big] \otimes E\big[\chi_n(X)\chi_n(X)'\big]\big)^{-1}||^{-1}||\theta||^2 \\
\leq &E\big[|\phi_n(X,V)'\theta|^2\big] \\
\leq &\bar{c}||E\big[\rho_n(V)\rho_n(V)'\big] \otimes E\big[\chi_n(X)\chi_n(X)'\big]|| \cdot ||\theta||^2
\end{aligned}$$

By elementary properties of the Kronecker product, the smallest eigenvalue of $A \otimes B$ is the product of the smallest eigenvalues of A and B and the largest eigenvalue of $A \otimes B$ is the product of the largest eigenvalues of the two matrices. Therefore:

$$\underline{\mu}_n^2\underline{c}||\theta||^2 \leq E\big[|\phi_n(X,V)'\theta|^2\big] \leq \bar{\mu}_n^2\bar{c}||\theta||^2$$

Now for the second statement of the lemma. Let ι be the function that maps a length-$k(n)l(n)$ column vector θ to a $k(n)$-by-$l(n)$ matrix $\tilde{\theta}$ so that the (j,k) entry of $\tilde{\theta}$ is the $(j-1)l(n)+k$-th entry of θ. Then for any length-$k(n)l(n)$ column vector θ, $\big(\rho_n(v) \otimes \chi_n(x)\big)'\theta = \rho_n(v)'\iota(\theta)\chi_n(x)$. Note that for any $\theta \in \mathbb{R}^{k(n)l(n)}$:

$$\begin{aligned}
E\big[\big(\rho_n(V)'\iota(\theta)\chi_n(x)\big)^2\big] &= E\big[||\big(\chi_n(x)\chi_n(x)'\big)^{1/2}\iota(\theta)'\rho_n(V)||^2\big] \\
&\leq ||\big(\chi_n(x)\chi_n(x)'\big)^{1/2}G_n^{-1/2}||^2 \\
&\times E\big[||G_n^{1/2}\iota(\theta)'\rho_n(V)||^2\big]
\end{aligned}$$

Again, by Assumption 5.2.ii there exists $\bar{c} < \infty$ and $\underline{c} > 0$ so that $\underline{c} \leq \frac{dF_{(X,V)}}{dF_X \otimes F_V}(x, v) \leq \bar{c}$ for $F_{(X,V)}$-almost all (x, v), and so:

$$\begin{aligned} E\big[||G_n^{1/2}\iota(\theta)'\rho_n(V)||^2\big] &= \int |\chi_n(x)'\iota(\theta)'\rho_n(v)|^2 F_X \otimes F_V\big(d(x,v)\big) \\ &\leq \frac{1}{\underline{c}} E\big[|\chi_n(X)'\iota(\theta)'\rho_n(V)|^2\big] \end{aligned}$$

And further, note that:

$$||\big(\chi_n(x)\chi_n(x)'\big)^{1/2} G_n^{-1/2}||^2 = ||G_n^{-1/2}\chi_n(x)||^2 \leq \xi_{\chi,n}^2$$

Combining we get:

$$E\big[\big(\rho_n(V)'\iota(\theta)\chi_n(x)\big)^2\big] \leq \frac{1}{\underline{c}}\xi_{\chi,n}^2 E\big[\big(\rho_n(V)'\iota(\theta)\chi_n(X)\big)^2\big]$$

Or equivalently:

$$E\big[|\phi_n(x, V)'\theta|^2\big] \leq \frac{1}{\underline{c}}\xi_{\chi,n}^2 E\big[|\phi_n(X, V)'\theta|^2\big]$$

□

Lemma C.3. *Let Assumptions 5.1.i-iii, 5.2.i-ii, 5.3.i, 5.3.iii, and 5.4.i-iii hold. Suppose* $E\big[|\phi_n(X, V)'\theta_n|^2\big]$ *is bounded above uniformly over* n *and suppose that* $\frac{1}{n_g}\sum_{i \in \mathcal{I}_g}(\hat{g}_i - \hat{\pi}_{n,i}'\theta_n)^2 \precsim_p R_n^2$*, then:*

$$E_{(X,Z)}\big[|\phi_n(X, V)'\hat{\theta}|^2\big] \precsim_p \frac{\bar{\mu}_n}{\underline{\mu}_n} + \frac{\bar{\mu}_n R_n}{\lambda_{0,n}^{1/2}}$$

Where $\bar{\mu}_n$ *and* $\underline{\mu}_n$ *are defined as in Lemma C.2. Similarly:*

$$E_{(X,Z)}\big[|\phi_n(X, V)'(\hat{\theta} - \theta_n)|^2\big] \precsim_p \frac{\bar{\mu}_n}{\underline{\mu}_n} + \frac{\bar{\mu}_n R_n}{\lambda_{0,n}^{1/2}}$$

Proof. Recall that $\hat{\theta}$ is defined by:

$$\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^{K(n)}} \frac{1}{n_g} \sum_{i \in \mathcal{I}_g} (\hat{g}_i - \hat{\pi}'_{n,i}\theta)^2 + \lambda_{0,n}||\theta||^2$$

and so we have:

$$\begin{aligned}
\lambda_{0,n}||\hat{\theta}||^2 &\leq \frac{1}{n_g} \sum_{i \in \mathcal{I}_g} (\hat{g}_i - \hat{\pi}'_{n,i}\hat{\theta})^2 + \lambda_{0,n}||\hat{\theta}||^2 \\
&\leq \frac{1}{n_g} \sum_{i \in \mathcal{I}_g} (\hat{g}_i - \hat{\pi}'_{n,i}\theta_n)^2 + \lambda_{0,n}||\theta_n||^2
\end{aligned}$$

And so:

$$||\hat{\theta}||^2 \leq ||\theta_n||^2 + \lambda_{0,n}^{-1} \frac{1}{n_g} \sum_{i \in \mathcal{I}_g} (\hat{g}_i - \hat{\pi}'_{n,i}\theta_n)^2$$

Using Lemma C.2 the above implies:

$$\begin{aligned}
E_{(X,Z)}\big[|\phi_n(X,V)'\hat{\theta}|^2\big] \leq& \frac{\bar{\mu}_n^2}{\underline{\mu}_n^2} \frac{\bar{c}}{\underline{c}} E\big[|\phi_n(X,V)'\theta_n|^2\big] \\
&+ \bar{\mu}_n^2 \lambda_{0,n}^{-1} \frac{1}{n_g} \sum_{i \in \mathcal{I}_g} (\hat{g}_i - \hat{\pi}'_{n,i}\theta_n)^2
\end{aligned}$$

Where we have used that $E\big[|\phi_n(X,V)'\theta_n|^2\big] \precsim 1$. Using the triangle inequality, $\frac{1}{n_g}\sum_{i \in \mathcal{I}_g} (\hat{g}_i - \hat{\pi}'_{n,i}\theta_n)^2 \precsim_p R_n^2$ implies:

$$E_{(X,Z)}\big[|\phi_n(X,V)'\hat{\theta}|^2\big]^{1/2} \precsim_p \frac{\bar{\mu}_n}{\underline{\mu}_n} + \frac{\bar{\mu}_n R_n}{\lambda_{0,n}^{1/2}}$$

By the triangle inequality:

$$\begin{aligned}
&E_{(X,V)}\Big[\big(\phi_n(X,V)'(\hat{\theta} - \theta_n)\big)^2\Big]^{1/2} \\
\leq& E\Big[\big(\phi_n(X,V)'\theta_n\big)^2\Big]^{1/2} + E_{(X,V)}\Big[\big(\phi_n(X,V)'\hat{\theta}\big)^2\Big]^{1/2}
\end{aligned}$$

And so, since $\frac{\bar{\mu}_n}{\underline{\mu}_n} \geq 1$ and $E\big[|\phi_n(X,V)'\theta_n|^2\big] \precsim 1$ we also have:

$$E_{(X,V)}\big[\big(\phi_n(X,V)'(\hat{\theta}-\theta_n)\big)^2\big]^{1/2} \precsim_p \frac{\bar{\mu}_n}{\underline{\mu}_n} + \frac{\bar{\mu}_n R_n}{\lambda_{0,n}^{1/2}}$$

□

The following Lemma just restates a well-known law of large numbers for matrices of a particular form.

Lemma C.4. *Let $a_{n,i}$ and $b_{n,i}$ be jointly iid random column vectors and suppose $E\big[||a_{n,i}||^2||b_{n,i}||^2\big] < \infty$ for each n. Then:*

$$||\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(a_{n,i}b_{n,i}' - E[a_{n,i}b_{n,i}']\big)||^2 \precsim_p E\big[||a_{n,i}||^2||b_{n,i}||^2\big]$$

Proof. Let $[a_{n,i}]_k$ be the k^{th} component of the vector $a_{n,i}$. Note that:

$$\begin{aligned}
&||\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(a_{n,i}b_{n,i}' - E[a_{n,i}b_{n,i}']\big)||^2 \\
\leq &\sum_{k=1}^{K_n}||\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big([a_{n,i}]_k b_{n,i} - E\big[[a_{n,i}]_k b_{n,i}\big]\big)||^2
\end{aligned}$$

Then by Markov's inequality and then using that $(a_{n,i}, b_{n,i})$ are iid we get:

$$\begin{aligned}
&\sum_{k=1}^{K_n}||\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big([a_{n,i}]_k b_{n,i} - E\big[[a_{n,i}]_k b_{n,i}\big]\big)||^2 \\
\precsim_p &E\big[\sum_{k=1}^{K_n}||\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big([a_{n,i}]_k b_{n,i} - E\big[[a_{n,i}]_k b_{n,i}\big]\big)||^2\big] \\
= &\sum_{k=1}^{K_n}E\big[||\big([a_{n,i}]_k b_{n,i} - E\big[[a_{n,i}]_k b_{n,i}\big]\big)||^2\big] \\
\leq &E\big[||a_{n,i}||^2||b_{n,i}||^2\big]
\end{aligned}$$

□

Lemma C.5. *Suppose Assumptions 5.1.i-iii, 5.2.i-ii, 5.3.i, 5.3.iii, and 5.4.i-iii hold* *and let $\ell_{\psi,n}(s_4) \prec 1$. Let $\hat{\pi}_n(x,z)$ have the formula:*

$$\hat{\pi}_n(x,z) = \big(\psi_n(x,z)'\hat{\Omega}^{-1}_{\lambda_{2,n}} \frac{1}{n_\pi} \sum_{i \in \mathcal{I}_\pi} \psi_{n,i}\rho'_{n,i}\big)' \otimes \chi_n(x)$$

Then uniformly over all $\theta \in \mathbb{R}^{k(n)l(n)}$ with $E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2} = 1$ and F_X-almost all x:

$$E\big[||\big(\pi_n(x_1,Z) - \hat{\pi}_n(x_1,Z)\big)'\theta||^2\big|X = x_1\big]^{1/2} \precsim_p R_{\pi,n}(x_1)$$

and $\big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g} |(\pi_{n,i} - \hat{\pi}_{n,i})'\theta|^2\big)^{1/2} \precsim_p R_{\pi,n}$. $R_{\pi,n}(x_1)$ is given by:

$$\begin{aligned} R_{\pi,n}(x_1) =& \xi_{\Omega,n}(x)\xi_{\chi,n} \min\{\sqrt{\xi^2_{\psi,n}k(n)/n_\pi}, \sqrt{\xi^2_{\rho,n}m(n)/n_\pi}\} \\ &+ \xi_{\Omega,n}(x)\lambda_{2,n}||\Omega_n^{-1}|| + \xi_{\chi,n}\ell_{\psi,n}(s_4) \end{aligned}$$

and $R_{\pi,n}$ is given by:

$$\begin{aligned} R_{\pi,n} =& \xi_{\chi,n} \min\{\sqrt{\xi^2_{\psi,n}k(n)/n_\pi}, \sqrt{\xi^2_{\rho,n}m(n)/n_\pi}\} \\ &+ \lambda_{2,n}||\Omega_n^{-1}|| + \xi_{\chi,n}\ell_{\psi,n}(s_4) \end{aligned}$$

Proof. For each n and i define the length-$k(n)l(n)$ column vector $\epsilon_{n,i}$ by the formula $\epsilon_{n,i} = \rho_{n,i} - E\big[\rho_n(V)\big|X_i, Z_i\big]$. By construction, it must be the case that $E[Q_n^{-1/2}\epsilon_{n,i}|X_i,Z_i] = 0$. Further, note that:

$$\begin{aligned} E\big[||Q_n^{-1/2}\epsilon_{n,i}||^2\big] \leq& 2E\big[||Q_n^{-1/2}\rho_{n,i}||^2\big] + 2E\big[||E\big[Q_n^{-1/2}\rho_n(V)\big|X_i,Z_i\big]||^2\big] \\ \leq& 4E\big[||Q_n^{-1/2}\rho_{n,i}||^2\big] = 4k(n) \end{aligned} \tag{1.9.3}$$

The first inequality above follows by the definition of $\epsilon_{n,i}$ and Young's inequality and the second follows by positivity of the variance. The equality then follows by definition of Q_n. Next, note that with probability 1:

$$\begin{aligned}||Q_n^{-1/2}\epsilon_{n,i}|| &\leq ||Q_n^{-1/2}\rho_{n,i}|| + ||E\big[Q_n^{-1/2}\rho_n(V)\big|X_i, Z_i\big]|| \\ &\leq \operatorname{ess\,sup}||Q_n^{-1/2}\rho_n(V)|| + \operatorname{ess\,sup}||E\big[Q_n^{-1/2}\rho_n(V)\big|X, Z\big]|| \\ &\leq 2\operatorname{ess\,sup}||Q_n^{-1/2}\rho_n(V)|| \leq 2\xi_{\rho,n} \end{aligned} \tag{1.9.4}$$

Where the first inequality follows by the triangle inequality and definition of $\epsilon_{n,i}$, and the third inequality holds because Jensen's inequality gives:

$$||E\big[Q_n^{-1/2}\rho_n(V)\big|X, Z\big]||^2 \leq E\big[||Q_n^{-1/2}\rho_n(V)||^2\big|X, Z\big]$$

and the RHS above is clearly bounded by $\operatorname{ess\,sup}||Q_n^{-1/2}\rho_n(V)||$. For each n define the $m(n)$-by-$k(n)$ matrix β_n by:

$$\beta_n = \Omega_n^{-1}E\big[\psi_n(X, Z)E\big[\rho_n(V)\big|X, Z\big]'\big]$$

Then define r_n by $r_n(x, z) = E\big[\rho_n(V)\big|X = x, Z = z\big] - \beta_n'\psi_n(x, z)$ Note that $E[\psi_{n,i}r_n(X_i, Z_i)'] = 0$. By the Radon-Nikodym Theorem, Assumption 5.2.i implies that for any $\theta \in \mathbb{R}^{k(n)}$:

$$E\big[\rho_n(V)'\theta|X = x, Z = z\big] = E\big[\rho_n(V)'\theta\frac{dF_{(X,Z,V)}}{d(F_{(X,Z)} \otimes F_V)}(x, z, V)\big]$$

By Assumption 5.3.iii, for F_V-almost all v, $\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot, \cdot, v) \in \Lambda_{s_4}^{\dim(X,Z)}(c_4)$. So for any vector $q \in \mathbb{N}_0^{\dim(X,Z)}$ with $||q||_1 \leq \lfloor s_4 \rfloor$, the corresponding partial derivative $D_q\big[\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot, \cdot, v)\big](x, z)$ exist and has magnitude less than c_4 uniformly over $F_{(X,Z,V)}$-almost all (x, z, v). By the dominated convergence theorem we can differentiate under the integral to get:

$$
\begin{aligned}
&\big|D_q\big[E\big[\rho_n(V)'\theta|X=\cdot,Z=\cdot\big]\big](x,z)\big| \\
=&\big|E\big[\rho_n(V)'\theta D_q\big[\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot,\cdot,V)\big](x,z)\big]\big| \\
\leq&E\big[|\rho_n(V)'\theta|^2\big]^{1/2}\operatorname{ess\,sup}|D_q\big[\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot,\cdot,V)\big](x,z)| \\
\leq&c_4||Q_n^{1/2}\theta||
\end{aligned}
\tag{1.9.5}
$$

Moreover, for any $||q||_1=\lfloor s_4\rfloor$, we have:

$$
\begin{aligned}
&\big|D_q\big[\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot,\cdot,v)\big](x_1,z_1) \\
&-D_q\big[\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot,\cdot,v)\big](x_2,z_2)\big| \\
\leq&c_4\big(||x_1-x_2||^2+||z_1-z_2||^2\big)^{\frac{s_4-\lfloor s_4\rfloor}{2}}
\end{aligned}
$$

Again, differentiating under the integral:

$$
\begin{aligned}
&\big|D_q\big[E\big[\rho_n(V)'\theta|X=x_1,Z=z_1\big]\big] \\
&-D_q\big[E\big[\rho_n(V)'\theta|X=x_2,Z=z_2\big]\big]\big| \\
=&\big|E\big[\rho_n(V)'\theta\big(D_q\big[\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot,\cdot,V)\big](x_1,z_1) \\
&-D_q\big[\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot,\cdot,V)\big](x_2,z_2)\big)\big]\big| \\
\leq&E\big[|\rho_n(V)'\theta|^2\big]^{1/2}\operatorname{ess\,sup}\big|D_q\big[\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot,\cdot,V)\big](x_1,z_1) \\
&-D_q\big[\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot,\cdot,V)\big](x_2,z_2)\big| \\
\leq&c_4||Q_n^{1/2}\theta||\big(||x_1-x_2||^2+||z_1-z_2||^2\big)^{\frac{s_4-\lfloor s_4\rfloor}{2}}
\end{aligned}
\tag{1.9.6}
$$

(1.9.5) and (1.9.6) together imply:

$$
(x,z)\mapsto E\big[\rho_n(V)'\theta|X=x,Z=z\big]\in\Lambda_{s_4}^{\dim(X,Z)}(c_4||Q_n^{1/2}\theta||)
$$

Using Assumption 5.1.iii, the above implies that uniformly over all $\theta \in \mathbb{R}^{k(n)}$ and $F_{(X,Z)}$-almost all (x, z), $\frac{r_n(x,z)'\theta}{||Q_n^{1/2}\theta||} \precsim \ell_{\psi,n}(s_4)$, which in turn implies that (uniformly):

$$||Q_n^{-1/2} r_n(x, z)|| \precsim \ell_{\psi,n}(s_4) \tag{1.9.7}$$

Now decompose:

$$\pi_n(x, z)'\theta = \psi_n(x, z)'\beta_n \iota(\theta)\chi(x) + r_n(x, z)'\iota(\theta)\chi(x)$$

Recall that for $\theta \in \mathbb{R}^{k(n)l(n)}$, $\phi_n(x, v)'\theta = \rho_n(v)'\iota(\theta)\chi_n(x)$, where ι is defined as in Lemma C.3. Substituting the above and using the formulas for π_n and $\hat{\pi}_n$ we get:

$$\begin{aligned}
&\big(\pi_n(x, z) - \hat{\pi}_n(x, z)\big)'\theta \\
=&r_n(x, z)'\iota(\theta)\chi_n(x) \\
-&\psi_n(x, z)'\hat{\Omega}^{-1}_{\lambda_{2,n}} \frac{1}{n_\pi} \sum_{i \in \mathcal{I}_\pi} \psi_{n,i}(\epsilon_{n,i} + r_{n,i})'\iota(\theta)\chi_n(x) \\
-&\psi_n(x, z)'\hat{\Omega}^{-1}_{\lambda_{2,n}} \lambda_{2,n}\beta_n \iota(\theta)\chi_n(x)
\end{aligned}$$

Where $r_{n,i} = r_n(X_i, Z_i)$. By the triangle inequality:

$$\begin{aligned}
&E_Z\big[\big(\pi_n(x, Z)'\theta - \hat{\pi}_n(x, Z)'\theta\big)^2 \big| X = x\big]^{1/2} \\
\leq &E[|r_n(x, Z)'\iota(\theta)\chi_n(x)|^2 | X = x]^{1/2} \\
+&||\tilde{\Omega}_n(x)^{1/2}\hat{\Omega}^{-1}_{\lambda_{2,n}} \frac{1}{n_\pi} \sum_{i \in \mathcal{I}_\pi} \psi_{n,i}(\epsilon_{n,i} + r_{n,i})'\iota(\theta)\chi_n(x)|| \\
+&||\tilde{\Omega}_n(x)^{1/2}\hat{\Omega}^{-1}_{\lambda_{2,n}} \lambda_{2,n}\beta_n \iota(\theta)\chi_n(x)||
\end{aligned} \tag{1.9.8}$$

Taking (1.9.7) and applying Cauchy-Schwartz, uniformly over $\theta \in \mathbb{R}^{k(n),l(n)}$ and $F_{(X,Z)}$-almost all (x, z) and F_X-almost all x_1:

$$\begin{aligned}
|r_n(x, z)'\iota(\theta)\chi_n(x_1)| &\leq ||r_n(x, z)Q_n^{-1/2}|| \cdot ||Q_n^{1/2}\iota(\theta)\chi_n(x)|| \\
&\precsim E\big[|\phi_n(X, V)'\theta|^2\big]^{1/2} \xi_{\chi,n}\ell_{\psi,n}(s_4)
\end{aligned} \tag{1.9.9}$$

Where the final line above follows by Lemma C.3, which states that:

$$||Q_n^{-1/2}\iota(\theta)\chi_n(x)|| \precsim E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}$$

(1.9.9) implies (again uniformly):

$$E[|r_n(x,Z)'\iota(\theta)\chi_n(x)|^2|X=x]^{1/2} \precsim E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}\ell_{\psi,n}(s_4)$$

Using the definition of the operator norm and recalling that $\xi_{\Omega,n}(x) = ||\tilde{\Omega}_n(x)^{1/2}\Omega_n^{1/2}||$ we get:

$$\begin{aligned}
&||\tilde{\Omega}_n(x)^{1/2}\hat{\Omega}_{\lambda_{2,n}}^{-1}\frac{1}{n_\pi}\sum_{i\in\mathcal{I}_\pi}\psi_{n,i}(\epsilon_{n,i}+r_{n,i})'\iota(\theta)\chi_n(x)|| \\
\leq&\xi_{\Omega,n}(x)||\Omega_n^{1/2}\hat{\Omega}_{\lambda_{2,n}}^{-1}\Omega_n^{1/2}|| \\
\times&||\frac{1}{n_\pi}\sum_{i\in\mathcal{I}_\pi}(\Omega_n^{-1/2}\psi_{n,i})(\epsilon_{n,i}+r_{n,i})'Q_n^{-1/2}||\cdot||Q_n^{1/2}\iota(\theta)\chi_n(x))||
\end{aligned}$$

Recall that $E\big[\psi_{n,i}(\epsilon_{n,i}+r_{n,i})'\big]=0$ and the data are iid, so applying Lemma C.4 we get:

$$\begin{aligned}
&||\frac{1}{\sqrt{n_\pi}}\sum_{i\in\mathcal{I}_\pi}(\Omega_n^{-1/2}\psi_{n,i})(\epsilon_{n,i}+r_{n,i})'Q_n^{-1/2}|| \\
\precsim_p&E\big[||\Omega_n^{-1/2}\psi_{n,i}||^2||Q_n^{-1/2}(\epsilon_{n,i}+r_{n,i})||^2\big]^{1/2} \\
\precsim_p&E\big[||\Omega_n^{-1/2}\psi_{n,i}||^2||Q_n^{-1/2}\epsilon_{n,i}||^2\big]^{1/2} \\
&+2E\big[||\Omega_n^{-1/2}\psi_{n,i}||^2||Q_n^{-1/2}r_{n,i}||^2\big]^{1/2} \\
\precsim_p&\min\big\{\xi_{\psi,n}E\big[||Q_n^{-1/2}\epsilon_{n,i}||\big],\xi_{\rho,n}E\big[||\Omega_n^{-1/2}\psi_{n,i}||\big] \\
&+\operatorname{ess\,sup}||Q_n^{-1/2}r_n(X,Z)||E\big[||\Omega_n^{-1/2}\psi_{n,i}||\big] \\
\precsim_p&\min\{\sqrt{\xi_{\psi,n}^2k(n)},\sqrt{\xi_{\rho,n}^2m(n)}\}+\sqrt{m(n)}\ell_{\psi,n}(s_4) \\
\precsim_p&\min\{\sqrt{\xi_{\psi,n}^2k(n)},\sqrt{\xi_{\rho,n}^2m(n)}\}
\end{aligned} \tag{1.9.10}$$

The second inequality follows from Young's inequality, the third by the Holder inequality, the fourth by (1.9.3), (1.9.4), and (1.9.7). The final equality follows because

$\xi_{\rho,n} \gtrsim 1$ and $\ell_{\psi,n}(s_4) \prec 1$. In all:

$$\begin{aligned}
&E\big[||\tilde{\Omega}_n(x)^{1/2}\hat{\Omega}^{-1}_{\lambda_{2,n}}\frac{1}{n_\pi}\sum_{i\in\mathcal{I}_\pi}\psi_{n,i}(\epsilon_{n,i}+r_{n,i})'\iota(\theta)\chi_n(x)||\big] \\
\precsim_p&\frac{2\xi_{\Omega,n}||Q_n^{1/2}\iota(\theta)\chi_n(x)||\cdot||\Omega_n^{1/2}\hat{\Omega}^{-1}_{\lambda_{2,n}}\Omega_n^{1/2}||}{\sqrt{n_\pi}} \\
&\times\min\big\{\sqrt{\xi^2_{\psi,n}k(n)},\sqrt{\xi^2_{\rho,n}m(n)}\big\}
\end{aligned}$$

By Lemma C.2, we have $||\Omega_n^{1/2}\hat{\Omega}^{-1}_{\lambda_{2,n}}\Omega_n^{1/2}|| \precsim_p 1$, and again from Lemma C.3 we get, $||Q_n^{-1/2}\iota(\theta)\chi_n(x)|| \precsim E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}$ and so, uniformly over $\theta\in\mathbb{R}^{k(n)l(n)}$ and F_X-almost all x:

$$\begin{aligned}
&E\big[||\tilde{\Omega}_n(x)^{1/2}\hat{\Omega}^{-1}_{\lambda_{2,n}}\frac{1}{n_\pi}\sum_{i\in\mathcal{I}_\pi}\psi_{n,i}(\epsilon_{n,i}+r_{n,i})'\iota(\theta)\chi_n(x)||\big] \\
\precsim_p&E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}\xi_{\Omega,n}(x)\big(\sqrt{\xi^2_{\psi,n}k(n)/n_\pi}\vee\sqrt{\xi^2_{\rho,n}m(n)/n_\pi}\big)
\end{aligned}$$

Finally, consider the term $||\tilde{\Omega}_n(x)^{1/2}\hat{\Omega}^{-1}_{\lambda_{2,n}}\lambda_{2,n}\beta_n\iota(\theta)\chi_n(x)||$. Note that:

$$\begin{aligned}
&||\tilde{\Omega}_n(x)^{1/2}\hat{\Omega}^{-1}_{\lambda_{2,n}}\lambda_{2,n}\beta_n\iota(\theta)\chi_n(x)|| \\
\leq&\lambda_{2,n}\xi_{\Omega,n}||\Omega_n^{1/2}\hat{\Omega}^{-1}_{\lambda_{2,n}}\Omega_n^{1/2}||\cdot||\Omega_n^{-1}||\cdot||\Omega_n^{1/2}\beta_n\iota(\theta)\chi_n(x)||
\end{aligned}$$

By Lemma C.2 $||\Omega_n^{1/2}\hat{\Omega}^{-1}_{\lambda_{2,n}}\Omega_n^{1/2}|| \precsim_p 1$. Furthermore:

$$\begin{aligned}
||\Omega_n^{1/2}\beta_n\iota(\theta)\chi_n(x)||^2 &= E\big[|\psi_n(X,Z)\beta_n\iota(\theta)\chi_n(x)|^2\big] \\
&\leq E\big[|E[\rho_n(V)|X,Z]'\iota(\theta)\chi_n(x)|^2\big] \\
&\leq E\big[|\rho_n(V)'\iota(\theta)\chi_n(x)|^2\big] \\
&= ||Q^{1/2}\iota(\theta)\chi_n(x)||^2 \\
&\precsim \xi^2_{\chi,n}E\big[|\phi_n(X,V)'\theta|^2\big]
\end{aligned}$$

Where the first inequality holds by the properties of least-squares projection, the second inequality by positivity of the variance, and the final inequality by Lemma

C.3. Thus we have uniformly over $\theta \in \mathbb{R}^{k(n)l(n)}$ and F_X-almost all x:

$$||\tilde{\Omega}_n(x)^{1/2}\hat{\Omega}_{\lambda_{2,n}}^{-1}\lambda_{2,n}\beta_n\iota(\theta)\chi_n(x)|| \leq O(E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}||\Omega_n^{-1}||\lambda_{2,n}\xi_{\Omega,n}\xi_{\chi,n})$$

Substituting the results above into (1.9.8) we get that uniformly over $\theta \in \mathbb{R}^{k(n)l(n)}$ and F_X-almost all x:

$$\begin{aligned}
&E_Z\big[\big(\pi_n(x,z)'\theta - \hat{\pi}_n(x,z)'\theta\big)^2\big|X=x\big]^{1/2} \\
&\lesssim_p E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}\xi_{\Omega,n}\big(\sqrt{\xi_{\psi,n}^2 k(n)/n_\pi} \wedge \sqrt{\xi_{\rho,n}^2 m(n)/n_\pi}\big) \\
&+ E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}\big(\lambda_{2,n}\xi_{\Omega,n}||\Omega_n^{-1}|| + \ell_{\psi,n}(s_4)\big)
\end{aligned}$$

Next we consider $\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}|(\pi_{n,i} - \hat{\pi}_{n,i})'\theta|^2$. By the triangle inequality, Cauchy-Schwartz and the definition of the operator norm:

$$\begin{aligned}
&\big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}|(\pi_{n,i} - \hat{\pi}_{n,i})'\theta|^2\big)^{1/2} \\
\leq&\big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}|r_{n,i}'\iota(\theta)\chi_{n,i}|^2\big)^{1/2} \\
&+\big(||\hat{\Omega}_n^{1/2}\Omega_n^{-1/2}|| \cdot ||\Omega_n^{1/2}\hat{\Omega}_{\lambda_{2,n}}^{-1}\Omega_n^{1/2}|| \\
&\times||\Omega_n^{-1/2}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\psi_{n,j}(\epsilon_{n,i}+r_{n,i})'Q_n^{1/2}|| \cdot ||Q_n^{-1/2}\iota(\theta)\chi_{n,i}||\big) \\
&+\big(||\hat{\Omega}_n^{1/2}\Omega_n^{-1/2}|| \cdot ||\Omega_n^{1/2}\hat{\Omega}_{\lambda_{2,n}}^{-1}\Omega_n^{1/2}|| \cdot ||\Omega_n^{-1}|| \\
&\times||\lambda_{2,n}\Omega_n^{1/2}\beta_n Q_n^{1/2}|| \cdot ||Q_n^{-1/2}\iota(\theta)\chi_{n,i}||\big)
\end{aligned}$$

By (1.9.9):

$$\big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}|r_{n,i}'\iota(\theta)\chi_{n,i}|^2\big)^{1/2} \lesssim_p ||Q_n^{1/2}\iota(\theta)\chi_n(x)||\xi_{\chi,n}\ell_{\psi,n}(s_4)$$

By Lemma C.2 we have $||\hat{\Omega}_n^{1/2}\Omega_n^{-1/2}|| \leq ||\Omega_n^{-1/2}\hat{\Omega}_n\Omega_n^{-1/2}||^{1/2} \lesssim_p 1$ and also by Lemma

C.2 $||\Omega_n^{1/2}\hat{\Omega}_{\lambda_{2,n}}^{-1}\Omega_n^{1/2}|| \precsim_p 1$, combining this, the above, and (1.9.10) we get:

$$
\begin{aligned}
&\Big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}|(\pi_{n,i}-\hat{\pi}_{n,i})'\theta|^2\Big)^{1/2} \\
&\precsim_p E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}\min\{\sqrt{\xi_{\psi,n}^2 k(n)/n_\pi},\sqrt{\xi_{\rho,n}^2 m(n)/n_\pi}\} \\
&+E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}\big(\ell_{\psi,n}(s_4)+\lambda_{2,n}||\Omega_n^{-1}||\big)
\end{aligned}
$$

□

Lemma C.6. *Suppose Assumptions 5.1.iv, 5.2.i-ii, 5.3.iv, and 5.4.iv hold. Let* $\hat{\alpha}_n(x_1,x_2)=\big(\chi_n(x_2)'\hat{G}_{\lambda_{3,n}}^{-1}\frac{1}{n_\alpha}\sum_{i\in\mathcal{I}_\alpha}\chi_{n,i}\rho_{n,i}'\big)'\otimes\chi_n(x_1)$, *then uniformly over* $\theta\in\mathbb{R}^{k(n)l(n)}$ *and* F_X*-almost all* x_1 *and* x_2*:*

$$
\begin{aligned}
&\big(\alpha_n(x_1,x_2)-\hat{\alpha}_n(x_1,x_2)\big)'\theta \\
&\precsim_p E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}\big(1+\sqrt{\xi_{\chi,n}^2 l(n)/n_\alpha}\big)\ell_{\chi,n}(s_5) \\
&+E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}^2\min\{\sqrt{\xi_{\chi,n}^2 k(n)/n_\alpha},\sqrt{\xi_{\rho,n}^2 l(n)/n_\alpha}\} \\
&+\lambda_{3,n}\xi_{\chi,n}^2 E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}||G_n^{-1}||
\end{aligned}
$$

Proof. We follow similar steps to Lemma C.5. For each n and i define the length-$k(n)$ column vector $\epsilon_{n,i}=\rho_{n,i}-E\big[\rho_n(V)\big|X_i\big]$. By construction $E[\epsilon_{n,i}|X_i]=0$. By a similar argument to that in the first part of Lemma C.4:

$$
\begin{aligned}
E\big[||Q_n^{-1/2}\epsilon_{n,i}||^2\big] &\leq 2E\big[||Q_n^{-1/2}\rho_{n,i}||^2\big]+2E\big[||E\big[Q_n^{-1/2}\rho_n(V)\big|X_i\big]||^2\big] \\
&\leq 4E\big[||Q_n^{-1/2}\rho_{n,i}||^2\big] \\
&=4k(n) \qquad (1.9.11)
\end{aligned}
$$

By a similar argument to Lemma C.4, with probability 1:

$$
\begin{aligned}
||Q_n^{-1/2}\epsilon_{n,i}|| &\leq ||Q_n^{-1/2}\rho_{n,i}||+||E\big[Q_n^{-1/2}\rho_n(V)\big|X_i\big]|| \\
&\leq \operatorname{ess\,sup}||Q_n^{-1/2}\rho_n(V)||+\operatorname{ess\,sup}||E\big[Q_n^{-1/2}\rho_n(V)\big|X\big]|| \\
&\leq 2\xi_{\rho,n} \qquad (1.9.12)
\end{aligned}
$$

For each n define the $p(n)$-by-$k(n)$ matrix μ_n by:

$$\mu_n = G_n^{-1} E\big[\chi_n(X) E\big[\rho_n(V)\big|X\big]'\big]$$

Then define r_n by:

$$r_n(x) = E\big[\rho_n(V)\big|X = x\big] - \mu_n' \psi_n(x)$$

Note that $E[\psi_{n,i} r_n(X_i)'] = 0$. By the Radon-Nikodym Theorem, Assumption 5.2.i implies that for any $\theta \in \mathbb{R}^{k(n)}$:

$$E\big[\rho_n(V)'\theta|X = x\big] = E\big[\rho_n(V)'\theta \frac{dF_{(X,V)}}{dF_X \otimes F_V}(x, V)\big]$$

By Assumption 5.3.iii, $\frac{dF_{(X,V)}}{dF_X \otimes F_V}(x, V) \in \Lambda_{s_5}^{\dim(X)}(c_5)$, so following steps analogous to those in Lemma C.4 we get:

$$x \mapsto E\big[\rho_n(V)'\theta|X = x\big] \in \Lambda_{s_5}^{\dim(X)}(c_5 ||Q_n^{1/2}\theta||)$$

Using Assumption 5.1.iii, the above implies that uniformly over $\theta \in \mathbb{R}^{k(n)}$ and F_X-almost all x, $\frac{r_n(x)'\theta}{||Q_n^{1/2}\theta||} \precsim \ell_{\chi,n}(s_5)$, which in turn implies:

$$||r_n(x) Q_n^{-1/2}|| \precsim \ell_{\chi,n}(s_5) \tag{1.9.13}$$

Recall that for any $\theta \in \mathbb{R}^{k(n)l(n)}$, $\phi_n(x, v)'\theta = \rho_n(v)'\iota(\theta)\chi_n(x)$, where ι is defined as in Lemma C.3. Now decompose:

$$\alpha_n(x_1, x_2)'\theta = \chi_n(x_2)'\mu_n \iota(\theta)\chi(x_1) + r_n(x_2)'\iota(\theta)\chi(x_1)$$

Substituting the above and using the formulas for α_n and $\hat{\alpha}_n$ we get:

$$\begin{aligned}
&\big(\alpha_n(x_1,x_2)-\hat{\alpha}_n(x_1,x_2)\big)'\theta\\
=&r_n(x_2)'\iota(\theta)\chi_n(x_1)\\
-&\chi_n(x_2)'\hat{G}^{-1}_{\lambda_{3,n}}\frac{1}{n_\alpha}\sum_{i\in\mathcal{I}_\alpha}\chi_{n,i}(\epsilon_{n,i}+r_{n,i})'\iota(\theta)\chi_n(x_1)\\
-&\chi_n(x_2)'\hat{G}^{-1}_{\lambda_{3,n}}\lambda_{3,n}\mu_n\iota(\theta)\chi_n(x_1)
\end{aligned}$$

Where $r_{n,i}=r_n(X_i)$. By the triangle inequality, Cauchy-Schwartz, and the definition of the operator norm, the above gives:

$$\begin{aligned}
&\big|\big(\alpha_n(x_1,x_2)-\hat{\alpha}_n(x_1,x_2)\big)'\theta\big|\\
\leq&|r_n(x_2)'\iota(\theta)\chi_n(x_1)|\\
+&\big(||G_n^{-1/2}\chi_n(x_2)||\cdot||G_n^{1/2}\hat{G}^{-1}_{\lambda_{3,n}}G_n^{1/2}||\\
\times&||\frac{1}{n_\alpha}\sum_{i\in\mathcal{I}_\alpha}G_n^{-1/2}\chi_{n,i}(\epsilon_{n,i}+r_{n,i})'\iota(\theta)\chi_n(x_1)||\big)\\
+&|\chi_n(x_2)'\hat{G}^{-1}_{\lambda_{3,n}}\lambda_{3,n}\mu_n\iota(\theta)\chi_n(x_1)| \qquad (1.9.14)
\end{aligned}$$

From Lemma C.3 implies $||Q_n^{1/2}\iota(\theta)\chi_n(x)||^2\precsim\xi^2_{\chi,n}E\big[|\phi_n(X,V)'\theta|^2\big]$ uniformly. Combining with (1.9.13) and applying Cauchy-Schwartz:

$$\begin{aligned}
|r_n(x_2)'\iota(\theta)\chi_n(x_1)|&\leq||r_n(x_2)Q_n^{-1/2}||\cdot||Q_n^{1/2}\iota(\theta)\chi_n(x)||\\
&\precsim E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}\ell_{\chi,n}(s_5) \qquad (1.9.15)
\end{aligned}$$

Uniformly over $\theta\in\mathbb{R}^{k(n)l(n)}$ and F_X-almost all x. From (1.9.15), for the first term we immediately get:

$$E[|r_n(x,Z)'\iota(\theta)\chi_n(x)|^2|X=x]^{1/2}\precsim E\big[|\phi_n(X,V)'\theta|^2\big]\xi_{\chi,n}\ell_{\chi,n}(s_5)$$

Next note that:

$$
\begin{aligned}
&||\frac{1}{n}\sum_{i=1}^{n} G_n^{-1/2}\chi_{n,i}(\epsilon_{n,i}+r_{n,i})'\iota(\theta)\chi_n(x_1)|| \\
\leq &||Q_n^{1/2}\iota(\theta)\chi_n(x)|| \cdot ||\frac{1}{n}\sum_{i=1}^{n} G_n^{-1/2}\chi_{n,i}(\epsilon_{n,i}+r_{n,i})'Q_n^{-1/2}||
\end{aligned}
$$

Recall that $E\big[\chi_{n,i}(\epsilon_{n,i}+r_{n,i})'\big]=0$ and the data are iid, so applying Lemma C.4:

$$
\begin{aligned}
&||\frac{1}{\sqrt{n_\alpha}}\sum_{i\in\mathcal{I}_\alpha} G_n^{-1/2}\chi_{n,i}(\epsilon_{n,i}+r_{n,i})'Q_n^{-1/2}||^2 \\
\precsim_p &E\big[||G_n^{-1/2}\chi_{n,i}||^2||(\epsilon_{n,i}+r_{n,i})Q_n^{-1/2}||^2\big] \\
\leq &2E\big[||G_n^{-1/2}\chi_{n,i}||^2||\epsilon_{n,i}Q_n^{-1/2}||^2\big] \\
+&2E\big[||G_n^{-1/2}\chi_{n,i}||^2||Q_n^{-1/2}r_{n,i}||^2\big] \\
\leq &2\min\big\{\xi_{\chi,n}^2 E\big[||\epsilon_{n,i}Q_n^{-1/2}||^2\big], \xi_{\rho,n}^2 E\big[||G_n^{-1/2}\chi_{n,i}||^2\big]\big\} \\
+&2\operatorname{ess\,sup}||Q_n^{-1/2}r_n(X)||^2 E\big[||G_n^{-1/2}\chi_{n,i}||^2\big]\big\} \\
\precsim &\min\{\xi_{\chi,n}^2 k(n), \xi_{\rho,n}^2 l(n)\} + l(n)\ell_{\chi,n}(s_5)^2
\end{aligned}
$$

The first inequality follows by Young's inequality, the next by Holder inequality and then the rate by (1.9.11), (1.9.12), and (1.9.13). So uniformly over $\theta\in\mathbb{R}^{k(n)l(n)}$ and F_X-almost all x:

$$
\begin{aligned}
&||\frac{1}{n}\sum_{i=1}^{n} G_n^{-1/2}\chi_{n,i}(\epsilon_{n,i}+r_{n,i})'\iota(\theta)\chi_n(x_1)|| \\
\precsim_p &||Q_n^{1/2}\iota(\theta)\chi_n(x)||\min\{\sqrt{\xi_{\chi,n}^2 k(n)/n_\alpha}, \sqrt{\xi_{\rho,n}^2 l(n)/n_\alpha}\} \\
+&||Q_n^{1/2}\iota(\theta)\chi_n(x)||\ell_{\chi,n}(s_5)\sqrt{l(n)/n_\alpha}
\end{aligned} \tag{1.9.16}
$$

Next let us consider the term $|\chi_n(x_2)'\hat{G}_{\lambda_{3,n}}^{-1}\lambda_{3,n}\mu_n\iota(\theta)\chi_n(x_1)|$. Note that:

$$
\begin{aligned}
&|\chi_n(x_2)'\hat{G}_{\lambda_{3,n}}^{-1}\lambda_{3,n}\mu_n\iota(\theta)\chi_n(x_1)| \\
\leq &\lambda_{3,n}\xi_{\chi,n}||G_n^{1/2}\hat{G}_{\lambda_{3,n}}^{-1}G_n^{1/2}|| \cdot ||G_n^{-1}|| \cdot ||G_n^{1/2}\mu_n\iota(\theta)\chi_n(x)||
\end{aligned}
$$

By Lemma C.2 $||G_n^{1/2}\hat{G}_{\lambda_{3,n}}^{-1}G_n^{1/2}|| \precsim_p 1$. Furthermore:

$$\begin{aligned}
||G_n^{1/2}\mu_n\iota(\theta)\chi_n(x)||^2 &= E\big[|\chi_n(X)'\mu_n\iota(\theta)\chi_n(x)|^2\big] \\
&\leq E\big[|E[\rho_n(V)|X]'\iota(\theta)\chi_n(x)|^2\big] \\
&\leq E\big[|\rho_n(V)'\iota(\theta)\chi_n(x)|^2\big] \\
&= ||Q_n^{1/2}\iota(\theta)\chi_n(x)||^2 \qquad (1.9.17)
\end{aligned}$$

Where the first inequality holds by the properties of least-squares projection, the second inequality by positivity of the variance. Thus we have uniformly over $\theta \in \mathbb{R}^{k(n)}$ and F_X-almost all x:

$$|\chi_n(x_2)'\hat{G}_{\lambda_{3,n}}^{-1}\lambda_{3,n}\mu_n\iota(\theta)\chi_n(x_1)| \precsim_p \lambda_{3,n}||Q_n^{1/2}\iota(\theta)\chi_n(x)||^2||G_n^{-1}||\xi_{\chi,n}$$

By Lemma C.3, $||Q_n^{1/2}\iota(\theta)\chi_n(x)|| \precsim E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}$, and finally note that $||G_n^{-1/2}\chi_n(x_2)|| \leq \xi_{\chi,n}$. Substituting these and also (1.9.15), (1.9.16), and (1.9.17) into (1.9.14) we get that uniformly over $\theta \in \mathbb{R}^{k(n)l(n)}$ and F_X-almost all x:

$$\begin{aligned}
&\big(\alpha_n(x_1,x_2) - \hat{\alpha}_n(x_1,x_2)\big)'\theta \\
\precsim_p &E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}\big(1+\sqrt{\xi_{\chi,n}^2 l(n)/n_\alpha}\big)\ell_{\chi,n}(s_5) \\
&+E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}^2\min\{\sqrt{\xi_{\chi,n}^2 k(n)/n_\alpha},\sqrt{\xi_{\rho,n}^2 l(n)/n_\alpha}\} \\
&+\lambda_{3,n}\xi_{\chi,n}E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}||G_n^{-1}||\xi_{\chi,n}
\end{aligned}$$

□

Lemma C.7. *Suppose Assumptions 5.1.iv, 5.2.iv, 5.3.ii and 5.4.iii hold. Let* $\hat{g}_i = \zeta_{n,i}'\hat{\Xi}_{\lambda_{1,n}}^{-1}\sum_{j\in\mathcal{I}_g}\zeta_{n,j}Y_j$*, then:*

$$\Big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\big(\hat{g}_i - g_i\big)^2\Big)^{1/2} \precsim_p \sqrt{\frac{p(n)}{n_g}} + \ell_{\zeta,n}(s_3) + \lambda_{1,n}||\Xi_n^{-1}||$$

Proof. Let $r_{n,i} = g_i - \zeta'_{n,i}\beta_n$ where $\beta_n = \Xi_n^{-1}E[\zeta_{n,i}Y_i]$ and $\epsilon_i = Y_i - g_i$. Then:

$$\hat{g}_i - g_i = \zeta'_{n,i}\hat{\Xi}_{\lambda_{1,n}}^{-1}\sum_{j\in\mathcal{I}_g}\zeta_{n,j}(\epsilon_j + r_{n,j}) - r_{n,i} - \lambda_{1,n}\zeta'_{n,i}\hat{\Xi}_{\lambda_{1,n}}^{-1}\beta_n$$

And so by the triangle inequality and the definition of the operator norm:

$$\begin{aligned}
&\Big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\big(\hat{g}_i - g_i\big)^2\Big)^{1/2} \\
\leq&||\hat{\Xi}_n^{1/2}\Xi_n^{-1/2}|| \cdot ||\Xi_n^{1/2}\hat{\Xi}_{\lambda_{1,n}}^{-1}\Xi_n^{1/2}|| \cdot ||\Xi_n^{-1/2}\sum_{j\in\mathcal{I}_g}\zeta_{n,j}(\epsilon_j + r_{n,j})|| \\
+&\Big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}r_{n,i}^2\Big)^{1/2} \\
+&\lambda_{1,n}||\hat{\Xi}_n^{1/2}\Xi_n^{-1/2}|| \cdot ||\Xi_n^{1/2}\hat{\Xi}_{\lambda_{1,n}}^{-1}\Xi_n^{1/2}|| \cdot ||\Xi_n^{-1}|| \cdot ||\Xi_n^{1/2}\beta_n||
\end{aligned}$$

By Lemma C.1 and Assumption 5.4.iii $||\hat{\Xi}_n^{1/2}\Xi_n^{-1/2}|| \precsim_p 1$ and $||\Xi_n^{1/2}\hat{\Xi}_{\lambda_{1,n}}^{-1}\Xi_n^{1/2}|| \precsim_p 1$. By Assumption 5.1.iv and 5.3.ii, $\operatorname{ess\,sup}|r_{n,i}| \precsim \ell_{\zeta,n}(s_3)$. By Markov's inequality and because the data are iid and $E[\zeta_{n,i}(r_{n,i} + \epsilon_i)] = 0$ we have:

$$\begin{aligned}
&||\Xi_n^{-1/2}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}(r_{n,i} + \epsilon_i)||^2 \precsim_p \frac{1}{n_g}E\big[||\Xi_n^{-1/2}\zeta_{n,i}(r_{n,i} + \epsilon_i)||^2\big] \\
\leq&\frac{1}{n_g}E\big[||\Xi_n^{-1/2}\zeta_{n,i}||^2\big]\big(\operatorname{ess\,sup} r_{n,i} + \operatorname{ess\,sup} E[\epsilon_i^2|X_i, Z_i]\big) \\
\precsim&\frac{p(n)}{n_g}\big(\ell_{\zeta,n}(s_3) + \bar{\sigma}_Y^2\big) \precsim \frac{p(n)}{n_g}
\end{aligned}$$

Where we have used $\operatorname{ess\,sup} r_{n,i} \precsim \ell_{\zeta,n}(s_3)$ and $\operatorname{ess\,sup} E[\epsilon_i^2|X_i, Z_i] \leq \bar{\sigma}_Y^2$ which follows from Assumption 5.2.iv. Finally note that:

$$||\Xi_n^{1/2}\beta_n||^2 = E\big[(\zeta'_{n,i}\beta_n)^2\big] \leq E\big[g_{n,i}^2\big] \leq E\big[Y_i^2\big] \leq \bar{\sigma}_Y^2$$

Where the first inequality follows by properties of least squares projection. In all we get:

$$(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\big(\hat{g}_i-g_i\big)^2\big)^{1/2}\precsim_p\sqrt{\frac{p(n)}{n_g}}+\ell_{\zeta,n}(s_3)+\lambda_{1,n}||\Xi_n^{-1}||$$

$\square$

Lemma C.8. *Suppose Assumption 1-4 and 5.4.iv hold. Suppose that $\hat{\pi}_n(x,z)$ is given by $\hat{\pi}_n(x,z)=\big(\hat{\omega}_n'\psi_n(x,z)\big)\otimes\chi_n(x)$ for some vector of coefficients $\hat{\omega}_n$, then:*

$$||\bar{\Sigma}_n^{-1/2}\hat{\Sigma}_n\bar{\Sigma}_n^{-1/2}-I||^2\precsim_p\sqrt{\xi^2_{\tilde{\psi},n}log\big(l(n)m(n)\big)/n_g}\tag{1.9.18}$$

Where $\bar{\Sigma}_n=E_{(X,Z)}[\hat{\pi}_n(X,Z)\hat{\pi}_n(X,Z)']$, and further:

$$||\bar{\Sigma}_n^{1/2}\bar{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2}-\bar{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2}||^2\precsim_p\sqrt{\xi^2_{\tilde{\psi},n}log\big(l(n)m(n)\big)/n_g}\tag{1.9.19}$$

Furthermore:

$$||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||\precsim_p||\bar{\Sigma}_n^{1/2}\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||\tag{1.9.20}$$

In addition, we have that $||\bar{\Sigma}_n^{-1/2}\tilde{\alpha}_n(x_1,x_2)||^2\precsim C(x_1,x_2)\xi^2_{\tilde{\Omega},n}(x_1)$, and similarly that $||\hat{\Sigma}_{\lambda_{0,n}}^{-1/2}\tilde{\alpha}_n(x_1,x_2)||^2\precsim_p C(x_1,x_2)\xi^2_{\tilde{\Omega},n}(x_1)$, and:

$$||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||^2\precsim_p C(x_1,x_2)\xi^2_{\tilde{\Omega},n}(x_1)$$

Where $C(x_1,x_2)$ is as in the Assumption 4.i and we define $\bar{\Sigma}_{\lambda_{0,n}}=\bar{\Sigma}_n+\lambda_{0,n}I$.

Proof. **First we show (1.9.18).** Let the function ι be defined as in Lemma C.2, so that for any $\theta\in\mathbb{R}^{k(n)l(n)}$, $\hat{\pi}_{n,i}'\theta=\big(\hat{\omega}_n'\psi_n(x,z)\big)'\iota(\theta)\chi_n(x)$, then using the definitions of then for any $\theta \in \mathbb{R}^{k(n)l(n)}$ we see:

$$
\begin{aligned}
||\hat{\Sigma}_n^{1/2}\bar{\Sigma}_n^{-1/2}\theta|| &= \frac{1}{n_g}\sum_{i\in\mathcal{I}_g}|\hat{\pi}_{n,i}'\bar{\Sigma}_n^{-1/2}\theta|^2 \\
&= \frac{1}{n_g}\sum_{i\in\mathcal{I}_g}|\big(\hat{\omega}_n'\psi_n(x,z)\big)'\iota(\bar{\Sigma}_n^{-1/2}\theta)\chi_n(x)|^2 \\
&= \frac{1}{n_g}\sum_{i\in\mathcal{I}_g}|\big(\chi_n(x)\otimes\psi_n(x,z)\big)'vec\big(\hat{\omega}_n\iota(\bar{\Sigma}_n^{-1/2}\theta)\big)|^2 \\
&= \frac{1}{n_g}\sum_{i\in\mathcal{I}_g}|\tilde{\psi}_n(x,z)'vec\big(\hat{\omega}_n\iota(\bar{\Sigma}_n^{-1/2}\theta)\big)|^2 \\
&= ||\hat{\tilde{\Omega}}_n^{1/2}vec\big(\hat{\omega}_n\iota(\bar{\Sigma}_n^{-1/2}\theta)\big)|| \\
&\leq ||\hat{\tilde{\Omega}}_n^{1/2}\tilde{\Omega}_n^{-1/2}||\cdot||\tilde{\Omega}_n^{1/2}vec\big(\hat{\omega}_n\iota(\bar{\Sigma}_n^{-1/2}\theta)\big)|| \\
&= ||\hat{\tilde{\Omega}}_n^{1/2}\tilde{\Omega}_n^{-1/2}||\cdot||\theta|| \qquad (1.9.21)
\end{aligned}
$$

Where $vec(\cdot)$ returns the vectorization of its matrix argument (i.e., returns a single column vector formed by stacking the transposed rows of its argument). The third equality above follows because $vec(ABC) = (C'\otimes A)vec(B)$. The above implies $||\hat{\Sigma}_n^{1/2}\bar{\Sigma}_n^{-1/2}|| \leq ||\hat{\tilde{\Omega}}_n^{1/2}\tilde{\Omega}_n^{-1/2}||$. By similar reasoning:

$$
\begin{aligned}
||\bar{\Sigma}_n^{1/2}\hat{\Sigma}_n^{-1/2}\theta|| &= E\big[|\bar{\pi}_{n,i}'\hat{\Sigma}_n^{-1/2}\theta|^2\big|\mathcal{I}_\pi\big] \\
&= E\big[|\big(\hat{\omega}_n'\psi_n(x,z)\big)'\iota(\hat{\Sigma}_n^{-1/2}\theta)\chi_n(x)|^2\big|\mathcal{I}_\pi\big] \\
&= E\big[|\big(\chi_n(x)\otimes\psi_n(x,z)\big)'vec\big(\hat{\omega}_n\iota(\hat{\Sigma}_n^{-1/2}\theta)\big)|^2\big|\mathcal{I}_\pi\big] \\
&= E\big[|\tilde{\psi}_n(x,z)'vec\big(\hat{\omega}_n\iota(\hat{\Sigma}_n^{-1/2}\theta)\big)|^2\big|\mathcal{I}_\pi\big] \\
&= ||\tilde{\Omega}_n^{1/2}vec\big(\hat{\omega}_n\iota(\hat{\Sigma}_n^{-1/2}\theta)\big)|| \\
&\leq ||\tilde{\Omega}_n^{1/2}\hat{\tilde{\Omega}}_n^{-1/2}||\cdot||\hat{\tilde{\Omega}}_n^{1/2}vec\big(\hat{\omega}_n\iota(\hat{\Sigma}_n^{-1/2}\theta)\big)|| \\
&= ||\tilde{\Omega}_n^{1/2}\hat{\tilde{\Omega}}_n^{-1/2}||\cdot||\theta||
\end{aligned}
$$

So $||\bar{\Sigma}_n^{1/2}\hat{\Sigma}_n^{-1/2}|| \leq ||\tilde{\Omega}_n^{1/2}\hat{\tilde{\Omega}}_n^{-1/2}||$. One can show that for any non-singular square matrix A:

$$
||(A'A-I)||^2 \leq ||A||^2\max\{||A||^2-1,0\} + (1-||A^{-1}||^{-2})||\theta||^2
$$

And so:

$$\begin{aligned}
||\bar{\Sigma}_n^{-1/2}\hat{\Sigma}_n\bar{\Sigma}_n^{-1/2} - I||^2 \leq & ||\hat{\Sigma}_n^{1/2}\bar{\Sigma}_n^{-1/2}||^2 \max\{||\hat{\Sigma}_n^{1/2}\bar{\Sigma}_n^{-1/2}||^2 - 1, 0\} \\
& + ||\bar{\Sigma}_n^{1/2}\hat{\Sigma}_n^{-1/2}||^{-2}(||\bar{\Sigma}_n^{1/2}\hat{\Sigma}_n^{-1/2}||^2 - 1) \\
\leq & ||\hat{\tilde{\Omega}}_n^{1/2}\tilde{\Omega}_n^{-1/2}||^2 \big| ||\hat{\tilde{\Omega}}_n\tilde{\Omega}_n^{-1/2}||^2 - 1 \big| \\
& + \big| ||\tilde{\Omega}_n^{1/2}\hat{\tilde{\Omega}}_n^{-1/2}||^2 - 1 \big| \\
\leq & ||\hat{\tilde{\Omega}}_n^{1/2}\tilde{\Omega}_n^{-1/2}||^2 ||\tilde{\Omega}_n^{-1/2}\hat{\tilde{\Omega}}_n\tilde{\Omega}_n^{-1/2} - I|| \\
& + ||\tilde{\Omega}_n^{1/2}\hat{\tilde{\Omega}}_n^{-1/2}||^{-2} ||\tilde{\Omega}_n^{1/2}\hat{\tilde{\Omega}}_n^{-1}\tilde{\Omega}_n^{1/2} - I|| \\
\precsim_p & \sqrt{\xi^2_{\tilde{\psi},n} log\big(l(n)m(n)\big)/n_g}
\end{aligned}$$

The second last line follows by the reverse triangle inequality and because for a square matrix A, $||A'A|| = ||A||^2$. The final line follows from Assumption 5.4.iv and Lemma C.1.the triangle inequality. By Assumption 5.4.iv the final term on the last line above is $o(1)$ and so by the reverse triangle inequality $||\bar{\Sigma}_n^{1/2}\hat{\Sigma}_n^{-1}\bar{\Sigma}_n^{1/2}|| \precsim_p 1$ and hence $||\hat{\Sigma}_n^{1/2}\bar{\Sigma}_n^{-1/2}|| \precsim_p 1$.

Now we show (1.9.19). Note that:

$$||\bar{\Sigma}_n^{-1/2}\hat{\Sigma}_n\bar{\Sigma}_n^{-1/2} - I||^2 = ||\bar{\Sigma}_n^{-1/2}\hat{\Sigma}_{\lambda_{0,n}}\bar{\Sigma}_n^{-1/2} - \bar{\Sigma}_n^{-1/2}\bar{\Sigma}_{\lambda_0}\bar{\Sigma}_n^{-1/2}||^2$$

For matrices A and B with $||B^{-1}|| \cdot ||A - B|| < 1$ one can show that:

$$||A^{-1} - B^{-1}|| \leq \frac{||B^{-1}||^2||A - B||}{1 - ||B^{-1}|| \cdot ||A - B||}$$

Since $||\bar{\Sigma}_n^{1/2}\bar{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2}|| \leq 1$ we get from (1.9.18):

$$||\bar{\Sigma}_n^{1/2}\bar{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2} - \bar{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2}|| \precsim_p (\xi^2_{\tilde{\psi},n} log\big(l(n)m(n)\big)/n_g)^{1/4}$$

By Assumption 5.4.iv and the triangle inequality this implies $||\bar{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2}|| \precsim_p 1$.

Now let us show(1.9.20). By the definition of the operator norm:

$$||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)|| \leq ||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_{\lambda_{0,n}}\bar{\Sigma}_n^{-1/2}|| \cdot ||\bar{\Sigma}_n^{1/2}\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||$$

Using the definitions of the matrices involved:

$$\begin{aligned}||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_{\lambda_{0,n}}\bar{\Sigma}_n^{-1/2}|| &= ||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2} + \lambda_{0,n}(\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{-1/2})|| \\ &\leq ||\hat{\Sigma}_n^{1/2}\bar{\Sigma}_n^{-1/2}|| \cdot ||\bar{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2}|| \\ &+ \lambda_{0,n}||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{-1/2}||\end{aligned}$$

By (1.9.18) and (1.9.19), the terms on the second line above are $O_p(1)$. Again, using definitions of the matrices involved:

$$\begin{aligned}||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{-1/2}|| &= ||(\hat{\Sigma}_n + \lambda_{0,n}I)^{-1}\hat{\Sigma}_n^{1/2}\bar{\Sigma}_n^{-1/2}|| \\ &\leq \lambda_{0,n}^{-1}||\hat{\Sigma}_n^{1/2}\bar{\Sigma}_n^{-1/2}|| \precsim_p \lambda_{0,n}^{-1}\end{aligned}$$

And so $||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_{\lambda_{0,n}}\bar{\Sigma}_n^{-1/2}|| \precsim_p 1$ and thus:

$$||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)|| \precsim_p ||\bar{\Sigma}_n^{1/2}\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||$$

Now the final three statements in the lemma. Applying Cauchy-Schwartz and using $E[\varphi(x_1,x_2,Z)^2|X=x_1] \leq C(x_1,x_2)$ we get that for any $\theta \in \mathbb{R}^{k(n)l(n)}$:

$$|\tilde{\alpha}_n(x_1,x_2)'\theta|^2 \leq C(x_1,x_2)E_Z\big[\big(\hat{\pi}_n(x_1,Z)'\theta\big)^2\big|X=x_1\big]$$

Using similar steps to (1.9.21):

$$
\begin{aligned}
&E_Z\big[\big(\hat{\pi}_n(x_1, Z)'\theta\big)^2\big|X = x_1\big] \\
=&||\tilde{\Omega}_n(x_1)^{1/2} vec\big(\hat{\omega}_n \iota(\theta)\big)||^2 \\
\leq&||\tilde{\Omega}_n(x_1)^{1/2}\tilde{\Omega}_n^{-1/2}||E_{(X,Z)}\big[|\tilde{\psi}_n(X, Z)' vec\big(\hat{\omega}_n \iota(\theta)\big)|^2\big] \\
\leq&\xi^2_{\tilde{\Omega},n}(x_1)E_{(X,Z)}\big[\big(\hat{\pi}_n(X, Z)'\theta\big)^2\big] \\
=&\xi^2_{\tilde{\Omega},n}(x_1)||\bar{\Sigma}_n^{1/2}\theta||^2
\end{aligned}
$$

Thus we get:

$$
|\tilde{\alpha}_n(x_1, x_2)'\bar{\Sigma}_n^{-1/2}\theta|^2 = C(x_1, x_2)\xi^2_{\tilde{\Omega},n}(x_1)||\theta||^2
$$

Since the above holds for all θ $||\bar{\Sigma}_n^{-1/2}\tilde{\alpha}_n(x_1, x_2)||^2 \precsim C(x_1, x_2)\xi^2_{\tilde{\Omega},n}(x_1)$. This then implies:

$$
\begin{aligned}
||\hat{\Sigma}_n^{1/2}\hat{\Sigma}^{-1}_{\lambda_{0,n}}\tilde{\alpha}_n(x_1, x_2)||^2 \leq& C(x_1, x_2)\xi^2_{\tilde{\Omega},n}(x_1)||\hat{\Sigma}_n^{1/2}\hat{\Sigma}^{-1}_{\lambda_{0,n}}\bar{\Sigma}_n^{1/2}||^2 \\
\precsim_p& C(x_1, x_2)\xi^2_{\tilde{\Omega},n}(x_1)
\end{aligned}
$$

Where the final line follows by $||\hat{\Sigma}_n^{1/2}\hat{\Sigma}^{-1}_{\lambda_{0,n}}\bar{\Sigma}_n^{1/2}||^2 \precsim_p 1$ shown earlier in the proof. Finally, again using the second inequality above and $||\hat{\Sigma}_n^{-1/2}\bar{\Sigma}_n^{1/2}||^2 \precsim_p 1$ shown earlier:

$$
\begin{aligned}
||\hat{\Sigma}^{-1/2}_{\lambda_{0,n}}\tilde{\alpha}_n(x_1, x_2)||^2 \leq& C(x_1, x_2)\xi^2_{\tilde{\Omega},n}(x_1)||\hat{\Sigma}_n^{1/2}\hat{\Sigma}^{-1}_{\lambda_{0,n}}\hat{\Sigma}_n^{1/2}||||\hat{\Sigma}_n^{-1/2}\bar{\Sigma}_n^{1/2}||^2 \\
\precsim_p& C(x_1, x_2)\xi^2_{\tilde{\Omega},n}(x_1)
\end{aligned}
$$

□

Lemma C.9. *Suppose Assumptions 1, 2, and 3 hold, Assumption 4.ii holds with $D(X)$ bounded above by a constant with probability 1, and Assumptions 5.1.i, 5.1.ii, 5.2, 5.3.i, and 5.3.ii hold. Then there is a sequence $\{\theta_n\}_{n=1}^{\infty}$ with $E\big[|\phi_n(X, V)'\theta_n|^2\big]$*

bounded above uniformly over n so that, uniformly over $F_{(X,Z)}$-almost all (x,z):

$$|g(x,z)-\pi_n(x,z)'\theta_n| \precsim \ell_{\rho,n}(s_1)$$

if X has finite discrete support, and otherwise:

$$|g(x,z)-\pi_n(x,z)'\theta_n| \precsim \ell_{\rho,n}(s_1)+\big(\xi_{\rho,n}\ell_{\chi,n}(1)\big)^{\tilde{s}}$$

Where $\tilde{s}=\frac{\min\{s_2,s_3,1\}}{\min\{s_2,s_3,1\}+1}$ and the above is uniform over $F_{(X,Z)}$-almost all (x,z).

Proof. First note that for any length-$k(n)l(n)$ column vector θ_n, there is a $l(n)$-by-$k(n)$ matrix $\tilde{\theta}_n$ so that $\big(\rho_n(V)\otimes\chi_n(X)\big)'\theta_n=\chi_n(X)'\tilde{\theta}_n\rho_n(V)$ and vice-versa. In particular, we can let the entry of $\tilde{\theta}_n$ in the j^{th} row and k^{th} column be the $(j-1)l(n)+k$-th entry of θ_n. For convenience we will find a $l(n)$-by-$k(n)$ matrix rather than a length-$k(n)l(n)$ vector.

We will now show that for each n, there exists a vector-valued function β_n so that $E\big[\big(\beta_n(X)'\rho_n(V)\big)^2\big]^{1/2}\precsim 1$ and:

$$\operatorname{ess\,sup}\big|g(X,Z)-\beta_n(X)'E\big[\rho_n(V)\big|X,Z\big]\big| \precsim \ell_{\rho,n}(s_1)$$

Combining Assumptions 5.1.i and 5.3.i, uniformly over $F_{(X,Z)}$-almost all (x,z):

$$\inf_B E\big[\big(\frac{dF_{(X,Z,V)}}{d(F_{(X,Z)}\otimes F_V)}(x,z,V)-\rho_n(V)'B\big)^2\big]^{1/2}\precsim \ell_{\rho,n}(s_1)$$

This implies that there exist sequences of functions B_n and r_n so that for $F_{(X,Z,V)}$-almost all (x,z,v):

$$\frac{dF_{(X,Z,V)}}{d(F_{(X,Z)}\otimes F_V)}(x,z,v)=\rho_n(v)'B_n(x,z)+r_n(x,z,v) \tag{1.9.22}$$

Where $E\big[r_n(x,z,V)^2\big]^{1/2}\precsim \ell_{\rho,n}(s_1)$. By Assumptions 1, 2, 3 and 4.ii there is a γ with $E[\gamma(x,V)^2|X=x]^{1/2}\leq D(x)$ so that $g(x,z)=E[\gamma(x,V)|X=x,Z=z]$ (see Lemma 1.1). Thus, for any vector-valued function of appropriate dimension $\beta_n(x)$ we get for

$F_{(X,Z)}$-almost all (x, z):

$$\begin{aligned} & g(x,z) - \beta_n(x)'E\big[\rho_n(V)\big|X = x, Z = z\big] \\ =& E\big[\gamma(x,V) - \beta_n(x)'\rho_n(V)\big|X = x, Z = z\big] \\ =& E\Big[\big(\gamma(x,V) - \beta_n(x)'\rho_n(V)\big)\frac{dF_{(X,Z,V)}}{d(F_{(X,Z)} \otimes F_V)}(x,z,V)\Big] \end{aligned}$$

Where the second equality follows by the definition of the Radon-Nikodym derivative. Substituting (1.9.22), the RHS above becomes:

$$\begin{aligned} & \big(E\big[\gamma(x,V)\rho_n(V)'\big] - \beta_n(x)'E\big[\rho_n(V)\rho_n(V)'\big]\big)B_n(x,z) \\ +& E\big[\big(\gamma(x,V) - \beta_n(x)'\rho_n(V)\big)r_n(x,z,V)\big] \end{aligned} \tag{1.9.23}$$

Assuming $E\big[\rho_n(V)\rho_n(V)'\big]$ is non-singular, we can set $\beta_n(x)$ so that:

$$\beta_n(x) = E\big[\rho_n(V)\rho_n(V)'\big]^{-1}E\big[\rho_n(V)\gamma(x,V)\big]$$

Substituting into (1.9.23) the first term disappears and we get:

$$\begin{aligned} & g(x,z) - \beta_n(x)'E\big[\rho_n(V)\big|X = x, Z = z\big] \\ =& E\big[\big(\gamma(x,V) - \beta_n(x)'\rho_n(V)\big)r_n(x,z,V)\big] \end{aligned}$$

By Cauchy-Schwartz:

$$\begin{aligned} & \big|E\big[\big(\gamma(x,V) - \beta_n(x)'\rho_n(V)\big)r_n(x,z,V)\big]\big| \\ \leq& E\Big[\big(\gamma(x,V) - \beta_n(x)'\rho_n(V)\big)^2\Big]^{1/2}E\big[r_n(x,z,V)^2\big]^{1/2} \\ \precsim& \ell_{\rho,n}(s_1)E\Big[\big(\gamma(x,V) - \beta_n(x)'\rho_n(V)\big)^2\Big]^{1/2} \end{aligned}$$

Note that $\beta_n(x)'\rho_n(v)$ is a least $L_2(F_V)$-norm projection of $\gamma(x,V)$ onto $\rho_n(V)$ and so:

$$\begin{aligned}&E\big[\big(\gamma(x,V)-\beta_n(x)'\rho_n(V)\big)^2\big]\\ \leq& E\big[\gamma(x,V)^2\big]=E\big[\gamma(x,V)^2\frac{dF_X\otimes F_V}{dF_{(X,V)}}(x,V)\big|X=x\big]\leq\frac{\bar{D}^2}{\underline{c}}\end{aligned}$$

Where $\bar{D}<\infty$ is an almost-sure upper bound on $D(X)$ which exists by supposition, and $\underline{c}>0$ is a lower bound on $\frac{dF_{(X,V)}}{dF_X\otimes F_V}$ which exists by Assumption 5.2.ii. Note that by properties of the least-squares projection we also have:

$$E\big[\big(\beta_n(x)'\rho_n(V)\big)^2\big]^{1/2}\leq\bar{D} \tag{1.9.24}$$

This in turn implies $E\big[\big(\beta_n(x)'\rho_n(V)\big)^2|X=x\big]\leq\frac{\bar{D}^2}{\underline{c}}$, and so:

$$E\big[\big(\beta_n(X)'\rho_n(V)\big)^2\big]^{1/2}\leq\frac{\bar{D}}{\sqrt{\underline{c}}}\precsim 1 \tag{1.9.25}$$

In all, we get that uniformly over $F_{(X,Z)}$-almost all (x,z):

$$\big|g(x,z)-\beta_n(x)'E\big[\rho_n(V)\big|X=x,Z=z\big]\big|\precsim\ell_{\rho,n}(s_1) \tag{1.9.26}$$

Now consider the case of X with discrete finite support. Both $\beta_n(X)$ and $\chi_n(X)$ then have discrete finite support and by Assumption 5.1.ii for n sufficiently large any function defined on $\mathcal{X}$ is a linear transformation of χ_n. Therefore, there exists a matrix θ_n so that for F_X-almost all x, $\chi_n(x)'\theta_n=\beta_n(x)$. And so from (1.9.26) we immediately get:

$$E\big[\big(g(X,Z)-\chi_n(x)'\theta_nE\big[\rho_n(V)\big|X,Z\big]\big)^2\big|X=x\big]^{1/2}\precsim\ell_{\rho,n}(s_1)$$

Moreover, from (1.9.25) we get $E\big[\big(\chi_n(X)'\theta_n\rho_n(V)\big)^2\big]^{1/2}\precsim 1$, and we are done.

The case of continuously distributed X requires more work. The function β_n defined above may not be smooth, to address this we first show that we can smooth-out β_n without incurring too much additional approximation error and then show we can approximate the smoothed out function by a linear combination of the basis functions that compose χ_n. Let $\{b_n\}_{n=1}^{\infty}$ be a sequence of strictly positive scalars with $b_n \to 0$ and for each n, define a linear operator M_n by:

$$M_n[\delta](x) = \frac{\int_{\mathcal{X}\cap B_{x,b_n}} \delta(x')dx'}{\int_{\mathcal{X}\cap B_{x,b_n}} dx'}$$

Where B_{x,b_n} denotes the Euclidean ball in $\mathbb{R}^{\dim(X)}$ of radius b_n centered at x. Note that under the Assumption 5.2.ii, for sufficiently large n there exists $\underline{r} > 0$ so that for all $x \in \mathcal{X}$, $1 \geq \frac{\int_{\mathcal{X}\cap B_{x,b_n}} dx'}{\int_{B_{0,b_n}} dx'} \geq \underline{r}$. We will use M_n to smooth out β, in particular let $\tilde{\beta}_n$ be the smoothed analogue of β which is given by $\tilde{\beta}_n(x) = M_n[\beta(X)](x)$. It is not difficult to see that:

$$\sup_{x\in\mathcal{X}} ||E\big[\rho_n(V)\rho_n(V)'\big]^{1/2}\tilde{\beta}_n(x)|| \leq \sup_{x\in\mathcal{X}} ||E\big[\rho_n(V)\rho_n(V)'\big]^{1/2}\beta_n(x)||^2$$

We will show $E\big[|\tilde{\beta}_n(X)'\rho_n(V)|^2\big]$ is bounded uniformly over n. Note that:

$$\begin{aligned}
\int_{\mathcal{X}} |\tilde{\beta}_n(x)'\rho_n(v)|^2 dx &= \int_{\mathcal{X}} \big|M_n[\beta_n(X)'\rho_n(v)](x)\big|^2 dx \\
&\leq \int_{\mathcal{X}}\int_{\mathcal{X}} \frac{1\{||x'-x|| \leq b_n\}}{\int_{\mathcal{X}\cap B_{x,b_n}} dx'} |\beta_n(x')'\rho_n(v)|^2 dx' dx \\
&\leq \frac{1}{\underline{r}} \int_{\mathcal{X}}\int_{\mathcal{X}} \frac{1\{||x'-x|| \leq b_n\}}{\int_{B_{0,b_n}} dx'} dx |\beta_n(x')'\rho_n(v)|^2 dx' \\
&\leq \frac{1}{\underline{r}} \int_{\mathcal{X}} |\beta_n(x)'\rho_n(v)|^2 dx
\end{aligned}$$

Where the first equality uses the definition of $\tilde{\beta}_n$, the subsequent inequality follows by Jensen's inequality, and the next inequality by swapping the order of integration (valid by Tonelli's theorem as the integrand is positive) and using $\frac{\int_{\mathcal{X}\cap B_{x,b_n}} dx'}{\int_{B_{0,b_n}} dx'} \geq \underline{r}$. The final inequality follows from $\int_{\mathcal{X}} \frac{1\{||x'-x|| \leq b_n\}}{\int_{B_{0,b_n}} dx'} dx \leq 1$. Now applying the upper and lower bounds $\bar{f} < \infty$ and $\underline{f} > 0$ on the density of X, which exist by Assumption

5.2.iii, the above implies:

$$E\big[|\tilde{\beta}_n(X)'\rho_n(v)|^2\big] \leq \frac{\bar{f}}{\underline{r}\underline{f}} E\big[|\beta_n(X)'\rho_n(v)|^2\big]$$

Finally, integrating both sides above over v against the measure F_V and applying the upper and lower bounds $\bar{c} < \infty$ and $\underline{c} > 0$ on $\frac{dF_{(X,V)}}{dF_X \otimes F_V}$, which exist by Assumption 5.2.ii, we get:

$$E\big[|\tilde{\beta}_n(X)'\rho_n(V)|^2\big] \leq \frac{\bar{c}\bar{f}}{\underline{c}\underline{r}\underline{f}} E\big[|\beta_n(X)'\rho_n(V)|^2\big] \leq \frac{\bar{c}\bar{f}}{\underline{c}^2\underline{r}\underline{f}} \bar{D}^2 \tag{1.9.27}$$

Where the final inequality follows by (1.9.25). Now we will show that the function $\tilde{\beta}_n(\cdot)'E\big[\rho_n(V)\big|X = x, Z = z\big]$ is Lipschitz continuous (i.e., an element of $\Lambda_1^{\dim(X)}(c)$ for some $c > 0$). With some work one can show that for any function δ with $|\delta(x)| \leq c$ for all $x \in \mathcal{X}$:

$$|M_n[\delta](x_1) - M_n[\delta](x_2)| \leq \frac{2}{\underline{r}} c \frac{\dim(X)}{b_n} ||x_1 - x_2|| \tag{1.9.28}$$

Now, we will upper bound the function $\beta_n(\cdot)'E\big[\rho_n(V)\big|X = x, Z = z\big]$. Note that (1.9.24) is equivalent to $||E\big[\rho_n(V)\rho_n(V)\big]^{1/2}\beta_n(x)|| \leq \bar{D}$. Using the this and the definition of $\xi_{\rho,n}$ we get:

$$\begin{aligned}
&\big|\beta_n(x_1)'E\big[\rho_n(V)\big|X = x_2, Z = z\big]\big| \\
\leq &||E\big[\rho_n(V)\rho_n(V)\big]^{1/2}\beta_n(x_1)|| \\
\times &||E\big[\rho_n(V)\rho_n(V)\big]^{-1/2}E\big[\rho_n(V)\big|X = x_2, Z = z\big]|| \\
\leq &\bar{D} \operatorname{ess\,sup} ||E\big[\rho_n(V)\rho_n(V)\big]^{-1/2}\rho_n(V)|| \leq \bar{D}\xi_{\rho,n}
\end{aligned}$$

Where 'ess sup' is the essential supremum over the distribution of V. Using the upper bound above, and applying (1.9.28) we get:

$$
\begin{aligned}
&|\tilde{\beta}_n(x_1)'E[\rho_n(V)|X=x,Z=z]-\tilde{\beta}(x_2)'E[\rho_n(V)|X=x,Z=z]|\\
=&|M_n\big[\beta_n(\cdot)'E[\rho_n(V)|X=x,Z=z]\big](x_1)\\
-&M_n\big[\beta_n(\cdot)'E[\rho_n(V)|X=x,Z=z]\big](x_2)|\\
\leq&\tilde{c}\frac{\xi_{\rho,n}}{b_n}||x_1-x_2||
\end{aligned}
$$

And thus $\tilde{\beta}_n(\cdot)'E[\rho_n(V)|X=x,Z=z]\in\Lambda_1^{\dim(X)}(\tilde{c}\frac{\xi_{\rho,n}}{b_n})$, where $\tilde{c}$ is some constant that is independent of n.

Next we show that replacing β_n with $\tilde{\beta}_n$ does not lose us much in terms of approximation error. Adding and subtracting terms and applying the triangle inequality we get:

$$
\begin{aligned}
&|g(x,Z)-\tilde{\beta}_n(x)'E\big[\rho_n(V)\big|X=x,Z\big]|\\
\leq&|\frac{\int_{\mathcal{X}\cup B_{x,b_n}}\big(g(x',Z)-\beta_n(x')'E\big[\rho_n(V)\big|X=x',Z\big]\big)dx'}{\int_{\mathcal{X}\cup B_{x,b_n}}dx'}|\\
+&\big|g(x,Z)-M_n\big[g(\cdot,Z)\big]\big|\\
+&|\frac{\int_{\mathcal{X}\cup B_{x,b_n}}\beta_n(x')'\big(E\big[\rho_n(V)\big|X=x,Z\big]-E\big[\rho_n(V)\big|X=x',Z\big]dx'}{\int_{\mathcal{X}\cup B_{x,b_n}}dx'}|
\end{aligned}
$$

For the first term on the RHS of the inequality note that:

$$
\begin{aligned}
&|(\int_{\mathcal{X}\cup B_{x,b_n}}dx')^{-1}\int_{\mathcal{X}\cup B_{x,b_n}}g(x',Z)-\beta_n(x')'E\big[\rho_n(V)\big|X=x',Z\big]dx'|\\
&\leq\sup_{x'\in\mathcal{X}}\big|g(x',Z)-\beta_n(x')'E\big[\rho_n(V)\big|X=x',Z\big]\big|\\
&\precsim\ell_{\rho,n}(s_1)
\end{aligned}
$$

Where, for the last step we have used (1.9.26). Next, it is easy to see that for any $\delta\in\Lambda_s^{\dim(X)}(c)$ we have $|\delta(x)-M_n[\delta](x)|\leq cb_n^{\min\{s,1\}}$. And so, using Assumption 5.3.ii, the second term satisfies:

$$\left|g(x,z) - M_n\left[g(z,\cdot)\right]\right| \leq c_3 b_n^{\min\{s_3,1\}}$$

For notational convenience define:

$$q(v,x,x',z) = \frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(x,z,v) - \frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(x',z,v)$$

For the third term, note that:

$$\begin{aligned}
&|\int_{\mathcal{X}\cap B_{x,b_n}} \beta_n(x')'\big(E\big[\rho_n(V)\big|X=x, Z=z\big] \\
&- E\big[\rho_n(V)\big|X=x', Z=z\big]dx'| \\
=&|\int_{\mathcal{X}\cap B_{x,b_n}} E\big[\beta_n(x')'\rho_n(V)q(V,x,x',z)\big]dx'| \\
\leq&\int_{\mathcal{X}\cap B_{x,b_n}} \big|E\big[\beta_n(x')'\rho_n(V)q(V,x,x',z)\big]\big|dx' \\
\leq&\int_{\mathcal{X}\cap B_{x,b_n}} E\big[\big(\beta_n(x')'\rho_n(V)\big)^2\big]^{1/2}E\big[q(V,x,x',z)^2\big]^{1/2}dx' \\
\leq&\bar{D}\int_{\mathcal{X}\cap B_{x,b_n}} E\big[q(V,x,x',z)^2\big]^{1/2}dx'
\end{aligned}$$

Where the first equality follows by the Radon-Nikodym theorem, the subsequent inequality by Jensen's inequality, the second inequality by Cauchy-Schwartz, and the final inequality by (1.9.24). Note that by the reverse triangle inequality:

$$\begin{aligned}
&\big|E\big[q(V,x,x_1,z)^2\big]^{1/2} - E\big[q(V,x,x_2,z)^2\big]^{1/2}\big| \\
\leq&E\big[\big|q(V,x,x_1,z) - q(V,x,x_2,z)\big|^2\big]^{1/2} \\
=&E\Big[\Big(\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(x_1,z,V) - \frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(x_2,z,V)\Big)^2\Big]^{1/2}
\end{aligned}$$

And by Assumption 5.3.i the final term above is bounded by $c_2||x_1 - x_2||^{\min\{s_2,1\}}$. So we get:

$$
\begin{aligned}
&|\frac{\int_{\mathcal{X}\cap B_{x,b_n}} \beta_n(x')'\big(E\big[\rho_n(V)\big|X=x,Z\big]-E\big[\rho_n(V)\big|X=x',Z\big]dx'}{\int_{\mathcal{X}\cap B_{x,b_n}} dx'}| \\
\leq & c_2 b_n^{\min\{s_2,1\}}
\end{aligned}
$$

And so in all:

$$
\begin{aligned}
&|g(x,z)-\tilde{\beta}_n(x)'E\big[\rho_n(V)\big|X=x,Z=z\big]| \\
\leq & \frac{\tilde{c}_1}{\sqrt{\underline{c}}}\bar{D}\ell_{\rho,n}(s_1)+c_1 b_n^{\min\{s_2,1\}}+c_2 b_n^{\min\{s_3,1\}}
\end{aligned}
$$

Now, let $G_n = E\big[\chi_n(X)\chi_n(X)'\big]$ and define the matrix θ_n by:

$$
\theta_n = G_n^{-1}E\big[\chi_n(X)'\tilde{\beta}_n(X)'\big]
$$

Recall that $\tilde{\beta}_n(\cdot)'E[\rho_n(V)|X=x,Z=z]\in\Lambda_1^{\dim(X)}(\tilde{c}\frac{\xi_{\rho,n}}{b_n})$, by Assumption 5.1.ii we get:

$$
\begin{aligned}
&\big|\chi_n(x)'\theta_n E[\rho_n(V)|X=x,Z=z]-\tilde{\beta}_n(x)'E[\rho_n(V)|X=x,Z=z]\big| \\
=&\big|\chi_n(x)'G_n^{-1}E\big[\chi_n(X)'\tilde{\beta}_n(X)'E[\rho_n(V)|X=x,Z=z]\big] \\
&-\tilde{\beta}_n(x)'E[\rho_n(V)|X=x,Z=z]\big| \precsim \frac{\xi_{\rho,n}}{b_n}\ell_{\chi,n}(1)
\end{aligned}
$$

And so, by the triangle inequality:

$$
\begin{aligned}
&|g(x,z)-\chi_n(x)'\theta_n E\big[\rho_n(V)\big|X=x,Z=z\big]| \\
\leq&|g(x,z)-\tilde{\beta}_n(x)'E\big[\rho_n(V)\big|X=x,Z=z\big]| \\
+&\big|\chi_n(x)'\theta_n E[\rho_n(V)|X=x,Z=z]-\tilde{\beta}_n(x)'E[\rho_n(V)|X=x,Z=z]\big| \\
\precsim&\ell_{\rho,n}(s_1)+b_n^{\min\{s_2,1\}}+b_n^{\min\{s_3,1\}}+\frac{\xi_{\rho,n}}{b_n}\ell_{\chi,n}(1)
\end{aligned}
$$

Choosing b_n rate-optimally the above gives:

$$\begin{aligned}&|g(x,z) - \chi_n(x)'\theta_n E\big[\rho_n(V)\big|X=x, Z=z\big]| \\ \leq &O\big(\ell_{\rho,n}(s_1) + \big(\xi_{\rho,n}\ell_{\chi,n}(1)\big)^{\tilde{s}}\big)\end{aligned}$$

Where $\tilde{s} = \frac{\min\{s_2,s_3,1\}}{\min\{s_2,s_3,1\}+1}$. Finally, By properties of least squares projection:

$$\begin{aligned}E\big[|\chi_n(X)'\theta_n\rho_n(v)|^2\big] &= E\big[\big|\chi_n(X)'R_n^{-1}E\big[\chi_n(X)'\tilde{\beta}_n(X)'\rho_n(v)\big]\big|^2\big] \\ &\leq E\big[|\tilde{\beta}_n(X)'\rho_n(v)|^2\big]\end{aligned}$$

And so:

$$\int |\chi_n(x)'\theta_n\rho_n(v)|^2 F_X \otimes F_V\big(d(x,v)\big) \leq \int |\tilde{\beta}_n(x)'\rho_n(v)|^2 F_X \otimes F_V\big(d(x,v)\big)$$

Using the upper and lower bounds $\bar{c} < \infty$ and $\underline{c} > 0$ on $\frac{dF_{(X,V)}}{dF_X \otimes F_V}$, which exist by Assumption 5.2.ii, this implies:

$$E\big[|\chi_n(X)'\theta_n\rho_n(V)|^2\big] \leq \frac{\bar{c}}{\underline{c}} E\big[|\tilde{\beta}_n(X)'\rho_n(V)|^2\big] \leq \frac{\bar{c}^2\bar{f}}{\underline{c}^3\underline{r}\underline{f}}\bar{D}^2 \lesssim 1$$

Where the final inequality follows by (1.9.27). So $E\big[|\chi_n(X)'\theta_n\rho_n(V)|^2\big]$ is bounded above uniformly over n, and we are done. $\square$

Chapter 2

Nonparametric Instrumental Variables Estimation Under Misspecification

Instrumental validity can be difficult to defend. The assumption rules out not only confounding between outcomes and instruments, but also any direct causal effect of the instruments on outcomes. Thus instruments may not be valid even if they are assigned by an ideal randomized experiment. To justify the use of instrumental variables (IV) methods, applied researchers usually argue not that instrumental validity holds exactly, but that any deviation from validity is small.[1] 'Small' does not imply 'inconsequential', and so it is important to assess the sensitivity of IV methods to a modest deviation from full instrumental validity.

In this work we consider the sensitivity of Nonparametric Instrumental Variables (NPIV) estimators ([64], [4]) to a failure of instrumental validity. We show that NPIV estimators of the structural function are highly sensitive to invalid instruments. For a broad class of NPIV estimators, an arbitrarily small deviation from instrumental validity can impart a large asymptotic bias. In some cases arbitrarily large. This sensitivity is a consequence of the 'ill-posedness' of the conditional moment restriction that identifies the structural function.

[1] See [28] for further discussion.

A researcher may be interested in estimating a continuous functional of the structural function rather than the function itself. Then robust estimation may be possible. We provide necessary and sufficient conditions under which a continuous linear functional can be estimated robustly.

One can achieve some degree of robustness by imposing strong smoothness conditions in estimation. For example, imposing a fixed bound on some derivatives of the structural function. This comes at a price: the estimates are inconsistent if the structural function is not sufficiently smooth. Many analyses of NPIV estimators assume the structural function is smooth in order to derive convergence rates, this is not enough for robustness, the estimators must be constructed so that they obey the given smoothness restrictions.[2] Even with these restrictions NPIV estimators are less robust than nonparametric regression and parametric IV estimators.

Our results suggest the need to supplement NPIV estimation with empirical sensitivity analysis. Methods for sensitivity analysis are increasingly popular in econometrics and include local methods (e.g., [8], [13], [19]), and global methods (e.g., [6], [61], [66])). We develop a method for global sensitivity analysis in which the researcher estimates the identified set under a relaxed instrumental validity assumption. Thus researchers can assess which of their findings are robust to some failure of instrumental validity. Estimation of the identified set under weakened assumptions is used elsewhere for sensitivity analysis, for example in [61]. The set estimation problem is well-posed, and we derive convergence rates under primitive conditions.

We apply our methods to the empirical setting shared by [18] and [47] who use NPIV methods to estimate shape-invariant Engel curves using data from The British Family Expenditure Survey. We use our methodology to assess which features of the structural Engel curve for food can be inferred robustly.

[2] Two NPIV methods which impose smoothness conditions are the procedures of [64] and [18]. A number of prominent NPIV procedures do not constrain the estimates to be smooth. For example, the methods described in [23], [30], [44], and [47]. Some analyses of NPIV estimation are general enough to incorporate both estimators that do and do not impose smoothness, for example [24].

2.1 The Sensitivity of NPIV Estimators

NPIV estimation is a flexible, nonparametric alternative to linear IV. It has been studied extensively, for example in [64], [4], [24], [30], [44], and [47]. NPIV models assume that outcomes Y can be written in the form $Y = h_0(X) + U$, where X is a vector of endogenous regressors and U is a latent variable. The deterministic function h_0 is known as the 'structural function' and usually has a causal interpretation.

NPIV models relax the linearity assumption of linear IV but retain additive separability. Suppose we interpret $h_0(x) + u$ as the potential outcome from treatment x for an individual with U equal to u. Then the effect of changing treatment from x_1 to x_2 is the same for all individuals.

NPIV is 'nonparametric' because the structural function h_0 is not assumed to take any particular parametric form. Instead, it is an unknown element of a function space. In particular, $h_0 \in \mathcal{B}_X$, where $\mathcal{B}_X$ is an infinite-dimensional Banach space with norm $||\cdot||_{\mathcal{B}_X}$.

In NPIV, h_0 is identified from a conditional moment restriction which corresponds to 'instrumental validity'. It is equivalent to $E[U|Z] = 0$.

Assumption 1.1 (Valid Instruments). $E[Y - h_0(X)|Z] = 0$

Implicit in Assumption 1.1 is that $E[Y|Z]$ exists and is finite. It is useful to think of $E[Y|Z]$ as a function g_0 of Z, so that $g_0(Z) \equiv E[Y|Z]$. We assume g_0 belongs to a Banach space $\mathcal{B}_Z$ with norm $||\cdot||_{\mathcal{B}_Z}$.

In addition, one may place a priori restrictions on the structural function. This can be captured in the assumption that h_0 belongs to a restricted parameter space $\mathcal{H} \subseteq \mathcal{B}_X$.

Assumption 1.2 (Parameter Space). $h_0 \in \mathcal{H} \subseteq \mathcal{B}_X$

Assumptions 1.1 and 1.2 are structural assumptions in the sense that one cannot cannot confirm that they hold by looking at the joint distribution of the observables. The following two non-structural conditions are also imposed throughout the NPIV literature.

Assumption 1.3 ($\mathcal{H}$-Completeness). For any $h \in \mathcal{H}$, $E[h(X)|Z] = 0 \iff h(X) = 0$

Assumption 1.4 (Compactness). Define $\mathbb{A} : \mathcal{B}_X \to \mathcal{B}_Z$ by the equation $\mathbb{A}[h](Z) = E[h(X)|Z]$. $\mathbb{A}$ is a compact linear operator from $\mathcal{B}_X$ into $\mathcal{B}_Z$.

Assumption 1.3 imposes a type of statistical completeness. Completeness plays a role analogous to the rank condition for identification in linear IV ([64]).[3] If Assumption 1.3 holds, then Assumptions 1.1 and 1.2 point identify h_0.

Assumption 1.4 does not aid identification, but it allows one to apply powerful tools from functional analysis to the NPIV estimation problem. The assumption holds under weak primitive conditions on the joint distribution of X and Z ([38], [47]).

Using $\mathbb{A}$ defined in Assumption 1.4, and g_0, we can rewrite the NPIV moment condition in Assumption 1.1 as $\mathbb{A}[h_0] = g_0$. Suppose Assumption 1.3 holds and $\mathcal{H}$ contains an open $||\cdot||_{\mathcal{B}_X}$-ball, then $\mathbb{A}$ is invertible on its range $R(\mathbb{A})$.[4] Denoting the inverse by $\mathbb{A}^{-1}$, we have $h_0 = \mathbb{A}^{-1}[g_0]$.

It is tempting to estimate h_0 by replacing $\mathbb{A}$ and g_0 with empirical analogues. This would be ill-advised. Suppose that $\mathbb{A}$ is known and we replace g_0 with a consistent estimate $\hat{g}_n \in R(\mathbb{A})$ (the subscript 'n' denotes the sample size). The resulting estimate, $\mathbb{A}^{-1}[\hat{g}_n]$, is generally inconsistent. This is because Assumption 1.4 implies $\mathbb{A}^{-1}$ is everywhere discontinuous. In fact, for any function g in the range of $\mathbb{A}$ and $\epsilon > 0$:

$$\sup_{g' \in R(\mathbb{A}) : ||g - g'||_{\mathcal{B}_Z} \leq \epsilon} ||\mathbb{A}^{-1}[g] - \mathbb{A}^{-1}[g']||_{\mathcal{B}_X} = \infty \tag{2.1.1}$$

So even if $\hat{g}_n$ is within a tiny ϵ distance of g_0, $\mathbb{A}^{-1}[\hat{g}_n]$ may be arbitrarily far from $\mathbb{A}^{-1}[g_0] = h_0$. This is the 'ill-posedness' of NPIV estimation.

To tackle ill-posedness, one may 'regularize' the problem. Consider a sequence of continuous operators $\{\mathbb{Q}_n\}_{n=1}^{\infty}$. Each operator $\mathbb{Q}_n$ is called a 'regularized inverse' of $\mathbb{A}$. Because $\mathbb{Q}_n$ is continuous, $\mathbb{Q}_n[\hat{g}_n]$ is not so sensitive to noise in $\hat{g}_n$, but replacing $\mathbb{A}^{-1}$ with $\mathbb{Q}_n$ generally imparts bias. For this bias to go to zero, $\mathbb{Q}_n$ must converge

[3] For some work on statistical completeness see [7], [20], [22], [35], [39], and [50].

[4] $R(\mathbb{A})$ is the set of elements $g \in \mathcal{B}_X$ such that there is an $h \in \mathcal{B}_X$ with $g = \mathbb{A}[h]$.

pointwise to $\mathbb{A}^{-1}$.[5] That is, the degree of regularization must be reduced. As a result, the sensitivity of $\mathbb{Q}_n[\hat{g}_n]$ to error in $\hat{g}_n$ increases. As the sample size grows, the estimation error in $\hat{g}_n$ diminishes and so it is worth raising the sensitivity in order to reduce the bias from regularization. If the degree of regularization is reduced slowly enough then $\mathbb{Q}_n[\hat{g}_n] \to^p h_0$ (see [30]).

In NPIV the operator $\mathbb{A}$ is not a priori known. As such, each regularized inverse must be estimated empirically. Let $\hat{\mathbb{Q}}_n$ estimate $\mathbb{Q}_n$. A typical NPIV estimator takes the form $\hat{h}_n = \hat{\mathbb{Q}}_n[\hat{g}_n]$.

2.1.1 Introducing Misspecification

We now consider the behavior of NPIV under misspecification. We replace Assumption 1.1 with a relaxed moment condition.

Assumption 1.1* **(Nearly Valid Instruments).** Define $u_0 \in \mathcal{B}_Z$ by $u_0(Z) \equiv [Y - h_0(X)|Z]$. We have $||u_0||_{\mathcal{B}_Z} \leq b$.

The scalar b in Assumption 1.1* captures the degree of misspecification. If $b = 0$ Assumption 1.1* reduces to Assumption 1.1 and instruments are valid. If b is small but non-zero, then the moment $E[Y - h_0(X)|Z]$ need not equal zero, but must be close to zero, with distance measured by $||\cdot||_{\mathcal{B}_Z}$.

To measure the sensitivity of an NPIV estimator to instrumental endogeneity, we consider the largest possible (un-scaled) asymptotic bias under Assumptions 1.1*, 1.2, 1.3, and 1.4. We fix parameters that are not directly related to misspecification. In particular, we fix the structural function h_0 and $\mu_{XZ\eta}$ the joint probability distribution of X, Z, and $\eta \equiv Y - E[Y|Z]$. The worst case asymptotic bias of an estimator $\hat{h}_n$ for a given b is then:

$$bias_{\hat{h}_n}(b) = \sup_{u_0 \in R(\mathbb{A}):\, ||u_0||_{\mathcal{B}_Z} \leq b} \inf \left[\epsilon : P(||\hat{h}_n - h_0||_{\mathcal{B}_X} \leq \epsilon) \to 1\right]$$

The expression above does not require that $\hat{h}_n$ has a probability limit. The existence

[5] That is, for any fixed $g \in R(\mathbb{A})$, $||\mathbb{Q}_n[g] - \mathbb{A}^{-1}[g]||_{\mathcal{B}_X} \to 0$.

of a probability limit for an NPIV estimator may be difficult to establish under misspecification. Sufficient conditions for the existence of a probability limit for misspecified sieve minimum distance estimators can be found in [5]. If $\hat{h}_n$ does have a probability limit (for any u_0 in the relevant set) then:

$$bias_{\hat{h}_n}(b) = \sup_{u_0 \in R(\mathbb{A}) :\, ||u_0||_{\mathcal{B}_Z} \leq b} \ \operatorname*{plim}_{n\to\infty} ||\hat{h}_n - h_0||_{\mathcal{B}_X}$$

The above is a case of the 'maximum bias' discussed in [52], albeit extended to an estimators whose value is a function.

The worst-case asymptotic bias captures the sensitivity of the estimator to a deviation from instrumental validity. We say an NPIV estimator is 'robust' if the worst-case asymptotic bias is small whenever the degree of misspecification b is small. Formally, $\hat{h}_n$ is robust if $\lim_{b\to 0} bias_{\hat{h}_n}(b) = 0$.

Theorem 2.1.1. *Fix $\mu_{XZ\eta}$ so that Assumptions 1.3 and 1.4 hold. Suppose that if Assumption 1.1 holds and $h_0 \in int(\mathcal{H})$ then $\hat{h}_n$ consistently estimates h_0, that is, $\underset{n\to\infty}{plim} ||\hat{h}_n - h_0||_{\mathcal{B}_X} = 0$. If $h_0 \in int(\mathcal{H})$, $\hat{h}_n$ is not robust. More precisely, let $r > 0$ be the radius of the largest open ball in $\mathcal{H}$ centered at h_0. For any $b > 0$, $bias_{\hat{h}_n}(b) \geq r$, and if $\mathcal{H} = \mathcal{B}_X$ then $bias_{\hat{h}_n}(b) = \infty$.*

Theorem 1.1 applies to any NPIV estimator that is consistent under under correct specification. The theorem states that if Assumptions 1.3 and 1.4 hold, and the true structural function lies in the interior of the parameter space, such an estimator is non-robust.

The theorem gives a lower bound for the worst-case asymptotic bias. The bound depends on the shape of the parameter space around the true structural function h_0. If we require consistency over the whole function space, i.e., $\mathcal{H} = \mathcal{B}_X$, then an arbitrarily small deviation from instrumental validity can lead to arbitrarily large asymptotic bias.

In finite samples, the sensitivity of an NPIV estimator to invalid instruments depends on the degree of regularization. Recall that regularization reduces the sensitivity to estimation error in $\hat{g}_n$ which estimates $g_0 = \mathbb{A}[h_0]$ under correct specification.

The presence of invalid instruments drives a wedge between g_0 and $\mathbb{A}[h_0]$, and thus between $\hat{g}_n$ and $\mathbb{A}[h_0]$. Thus invalid instruments act as an additional source of error in $\hat{g}_n$, and regularization controls the sensitivity to this error. To ensure an NPIV estimator is consistent, it is usually necessary to reduce the degree of regularization as the sample size grows. As the sample grows and the regularization is reduced, the estimator becomes more sensitive to invalid instruments.

A closely related result concerns the size of the identified set under Assumption 1.1*. Under Assumption 1.1*, h_0 is generally only set identified and the identified set takes the form $[h \in \mathcal{H} : ||\mathbb{A}[h] - g_0||_{\mathcal{B}_Z} \leq b]$.

Theorem 2.1.2. *Suppose Assumptions 1.3 and 1.4 hold and $\mathbb{A}^{-1}[g_0] \in int(\mathcal{H})$. Then for any $b > 0$, the diameter of the identified set under Assumptions 1.1* and 1.2 is at least twice the radius of the largest open ball in $\mathcal{H}$ centered at $\mathbb{A}^{-1}[g_0]$.*[6]

Theorem 1.2 suggests that even if $b > 0$ is very small, the identified set may be large. The diameter of the identified set depends crucially on the shape of $\mathcal{H}$ around $\mathbb{A}^{-1}[g_0]$. Thus, without strong restrictions on the parameter space $\mathcal{H}$, identification in NPIV models is very fragile: only an arbitrarily small relaxation of the NPIV moment condition is the difference between point identification and a large identified set. [7]

2.1.2 Strong Restrictions on the Parameter Space

The parameter space $\mathcal{H}$ plays a key role in Theorem 1.1. The theorem only applies when the structural function lies in the interior of $\mathcal{H}$. Some parameter spaces have an empty interior, of particular interest are compact parameter spaces which are employed extensively in the NPIV literature (e.g., in [64], [4], [18], [39], [72]). In NPIV settings compact parameter spaces are typically spaces of smooth functions, i.e., functions with some derivatives bounded by a known quantity.

If $\mathcal{H}$ is compact one can generally construct an estimator that is consistent under Assumptions 1.1 and 1.2 and also robust to the failure of instrumental validity.

[6]The diameter of a subset of a metric space is the smallest upper bound on the distance between any two points in the subset.

[7]We thank an anonymous referee for suggesting this interpretation of our results.

However, if $h_0 \notin \mathcal{H}$ (say, the structural function is not smooth enough), then such an estimator is inconsistent.

Recall the NPIV moment condition can be written as $\mathbb{A}[h_0] = g_0$. Given $h_0 \in \mathcal{H}$, this implies $\mathbb{A}[h_0] = P_Z[g_0]$ where P_Z is a projection from $\mathcal{B}_Z$ to $\mathbb{A}[\mathcal{H}]$ (the image of $\mathcal{H}$ under $\mathbb{A}$).[8] If Assumptions 1.1-1.3 hold there is a unique element of $\mathcal{H}$ that satisfies this equation. Let $\mathbb{A}_{\mathcal{H}}$ be the restriction of $\mathbb{A}$ to $\mathcal{H}$. Then the unique solution in $\mathcal{H}$ can be written as $\mathbb{A}_{\mathcal{H}}^{-1} P_Z[g_0]$.

Theorem 2.1.3. *Fix $\mu_{XZ\eta}$ so Assumptions 1.2, 1.3, and 1.4 hold. Let $\mathcal{H}$ be a compact subset of $\mathcal{B}_X$ and let P_Z be a uniformly continuous projection onto $\mathbb{A}[\mathcal{H}]$. Suppose $\hat{h}_n$ satisfies $||\hat{h}_n - \mathbb{A}_{\mathcal{H}}^{-1} P_Z[g_0]||_{\mathcal{B}_X} \to^p 0$. If $h_0 \in \mathcal{H}$ then $\hat{h}_n$ is robust. That is, $\lim_{b \to 0} bias_{\hat{h}_n}(b) = 0$.*

Theorem 1.3 requires that the estimator $\hat{h}_n$ has limit $\mathbb{A}_{\mathcal{H}}^{-1} P_Z[g_0] \in \mathcal{H}$. So if $\mathcal{H}$ contains functions with second derivatives bounded by a given constant, $\hat{h}_n$ must be constrained so that it too obeys this bound (at least in the limit), regardless of the data generating process. Thus the robustness in Theorem 1.3 comes as a price: if $h_0 \notin \mathcal{H}$, say the second derivatives of h_0 are too large, then the estimator is inconsistent.

$bias_{\hat{h}_n}(b)$ is identified, so the conditions of Theorem 1.3 imply the identified set under Assumptions 1.1* and 1.2 shrinks to a point as $b \to 0$.

Theorem 1.3 shows that the worst-case asymptotic bias goes to zero with b, but it does not give a rate. We show that even if nonparametric smoothness is imposed in estimation, NPIV estimators are generally less robust than linear IV estimators or nonparametric regression estimators.

Theorem 2.1.4. *Fix $\mu_{XZ\eta}$ so that Assumptions 1.3 and 1.4 hold. Suppose that $\mathcal{H} \subset \mathcal{B}_X$ is compact, convex, symmetric, and infinite dimensional and there exists $\alpha \in (0,1)$ so that $\frac{1}{\alpha} h_0 \in \mathcal{H}$.*[9] *Suppose $\hat{h}_n$ is consistent for h_0 whenever Assumptions 1.1 and 1.2 hold. Then $\lim_{b \to 0} bias_{\hat{h}_n}(b)/b = \infty$.*

[8] A projection P_Z from $\mathcal{B}_Z$ to $\mathbb{A}[\mathcal{H}]$ is a function with domain $\mathcal{B}_Z$ and range $\mathbb{A}[\mathcal{H}]$ so that for any $g \in \mathbb{A}[\mathcal{H}]$, $P_Z[g] = g$.

[9] A subset $\mathcal{H}$ of a vector space is symmetric if $h \in \mathcal{H}$ implies $-h \in \mathcal{H}$.

Theorem 1.4 states that $bias_{\hat{h}_n}(b)$ goes to zero at a strictly slower rate than b. This contrasts with standard non-parametric regression and linear IV in which the worst-case asymptotic bias is at most proportional to b. The conditions Theorem 1.4 places on $\mathcal{H}$ hold for the compact infinite-dimensional spaces typically used in the NPIV literature including all of those considered by [41].

2.1.3 Continuous Functionals

The sensitivity results in Subsections 1.1 and 1.2 apply to estimation of the structural function itself. If the object of interest is instead a smooth linear functional of the structural function then it may be possible to construct estimates that a) are consistent under instrumental validity without any a priori restrictions on the true structural function and also b) are robust to the failure of instrumental validity.

The estimation of functionals of the structural function in NPIV models is analyzed extensively in the literature ([4], [73], [53], and others). Following [73], we let $\mathcal{B}_X = L_2(\mu_X)$.

[73] consider linear functionals of the form $\mathbb{L}[h] = E[w(X)h(X)]$, where $w \in L_2(\mu_X)$. By the Reisz representation theorem, any continuous linear functional on $L_2(\mu_X)$ can be written in this form for some w.

Let $\hat{l}_n$ be an estimator of $\mathbb{L}[h_0]$. Fix h_0 and $\mu_{XZ\eta}$. The worst-case asymptotic bias of $\hat{l}_n$ given $u_0 \in R(\mathbb{A})$ and $||u_0||_{L_2(\mu_Z)} \leq b$ is:

$$bias_{\hat{l}_n}(b) = \sup_{u_0 \in R(\mathbb{A}):\, ||u_0||_{L_2(\mu_Z)} \leq b} \inf \left[\epsilon : P(|\hat{l}_n - \mathbb{L}[h_0]|) \leq \epsilon) \to 1\right]$$

[73] show that $E[w(X)h(X)]$ is estimable at rate $\sqrt{n}$ only if there exists a function $\alpha \in L_2(\mu_Z)$ (where μ_Z is the probability measure of Z) so that:

$$w(X) = E[\alpha(Z)|X] \tag{2.1.2}$$

Theorem 1.5 shows that this condition is necessary and sufficient for robust estimation of $\mathbb{L}[h_0]$ to be possible without restricting the parameter space.

Theorem 2.1.5. *Fix $\mu_{XZ\eta}$ so Assumptions 1.3 and 1.4 hold. Suppose that for any $h_0 \in L_2(\mu_X)$ if Assumption 1.1 holds then $|\hat{l}_n - \mathbb{L}[h_0]| \to^p 0$. a., If there exists $\alpha \in L_2(\mu_Z)$ with $w(X) = E[\alpha(Z)|X]$, then $bias_{\hat{l}_n}(b) \leq b||\alpha||_{L_2(\mu_Z)}$. b., If no such α exists, $bias_{\hat{l}_n}(b) = \infty$ for all $b > 0$.*

2.2 Empirical Sensitivity Analysis

The results in Section 1 show that estimation of the structural function entails a trade-off. Strong restrictions on the structural function can reduce the sensitivity to a failure of instrumental validity, but if the restrictions are too strong then imposing them imparts bias. Theorem 1.4 shows that even restricted estimators of the structural function are less robust than standard nonparametric regression or linear IV estimators. Robust estimation of a smooth linear functional is possible but only under a condition on the functional that could be difficult to verify in practice.[10]

These results may leave applied researchers concerned that their estimation results are not robust. To help researchers assess the robustness of their findings, we present a method for empirical sensitivity analysis in NPIV models. Our procedure is based on estimation of the identified set under the weakened instrumental validity condition in Assumption 1.1*. The approach is similar in spirit to [61] who analyze the sensitivity of treatment effect estimates to a small failure of unconfoundedness by weakening the assumptions used for point identification.

Let Θ be the identified set under Assumptions 1.1* and 1.2. Formally:

$$\Theta = [h \in \mathcal{H} : ||\mathbb{A}[h] - g_0||_{\mathcal{B}_X} \leq b]$$

Consider a linear functional $\mathbb{L} : \mathcal{B}_X \to \mathbb{R}$. The identified set for $\mathbb{L}[h_0]$ is $[\mathbb{L}[h] : h \in \Theta]$. For the choices of $\mathcal{B}_Z$ and $\mathcal{H}$ we consider, this set is an interval.[11] Of particular interest is the identified set for the value of h_0 at a point x, this is the set $[h(x) : h \in \Theta]$ and

[10] In fact, we conjecture that for a given w, the existence of $\alpha \in L_2(\mu_Z)$ with $w(X) = E[\alpha(Z)|X]$ is not testable.

[11] See Proposition 2.1 in the Appendix B for a proof.

corresponds to $\mathbb{L} : h \mapsto h(x)$. Denote the lower and upper endpoints for the interval by $\underline{\theta}_{\mathbb{L}}$ and $\bar{\theta}_{\mathbb{L}}$. We wish to estimate these end-points.

Consistent estimation of $\underline{\theta}_{\mathbb{L}}$ and $\bar{\theta}_{\mathbb{L}}$ allows us to construct an estimator of $\mathbb{L}[h_0]$ that achieves minimal worst-case asymptotic bias. An estimator achieves minimal worst-case asymptotic bias if and only if it converges in probability to $\frac{1}{2}(\underline{\theta}_{\mathbb{L}} + \bar{\theta}_{\mathbb{L}})$. So if $\hat{\underline{\theta}}_{\mathbb{L}}$ and $\hat{\bar{\theta}}_{\mathbb{L}}$ converge to $\underline{\theta}_{\mathbb{L}}$ and $\bar{\theta}_{\mathbb{L}}$, then $\frac{1}{2}(\hat{\underline{\theta}}_{\mathbb{L}} + \hat{\bar{\theta}}_{\mathbb{L}})$ achieves minimal worst-case asymptotic bias.

Assumption 1.1* depends on the norm of the space $\mathcal{B}_Z$ and Assumption 1.2 on the choice of $\mathcal{H}$. We assume that these spaces are chosen so that the two assumptions can be written in terms of inequality constraints.

Assumption 2.1. *i.* $||\cdot||_{\mathcal{B}_Z}$ *is the essential supremum norm. Thus Assumption* 1.1* *states that if* $h = h_0$*, then with probability* 1*:*

$$|E[Y - h(X)|Z]| \leq b \tag{2.2.1}$$

ii. $\mathcal{H}$ *is the set of uniformly bounded functions* h *so that:*

$$\mathbb{T}[h](x) \leq C(x), \forall x \in \mathcal{X} \tag{2.2.2}$$

Where $\mathcal{X}$ *is the support of* X*,* $\mathbb{T}$ *is a known linear functional from* $\mathcal{B}_X$ *to* $\mathcal{B}_X^d$*, and* C *is a known function in* $\mathcal{B}_X^d$*.*

Under Assumption 2.1 $\underline{\theta}_{\mathbb{L}}$ and $\bar{\theta}_{\mathbb{L}}$ are defined as follows:

$$\underline{\theta}_{\mathbb{L}} = \inf_{h \in \mathcal{B}_X} \mathbb{L}[h] \text{ subject to (2.2.1) and (2.2.2).}$$

$$\bar{\theta}_{\mathbb{L}} = \sup_{h \in \mathcal{B}_X} \mathbb{L}[h] \text{ subject to (2.2.1) and (2.2.2).}$$

2.2.1 Selecting b

The bound b in Assumption 1.1* determines the size of the identified set. We suggest the researcher estimate the identified set for a range of choices for b corresponding to mild, moderate, and severe misspecification. To determine whether b corresponds to say, 'mild' misspecification, it is helpful to compare b to a reference quantity. This approach is common in the empirical sensitivity literature. For example, in ([61]) misspecification arises due to selection on unobservables. ([61]) consider misspecification to be mild if the degree of selection on unobservables is small compared to the selection on observables, the latter of which is estimable. A similar approach is taken by [54], [6], and [66] among others.

We suggest one compare b with the variation in outcomes after accounting for the effect of the instruments. More precisely, we suggest one let b be the difference between the $0.5+\tau$ and $0.5-\tau$ quantiles of $Y - E[Y|Z]$ for various choices of $\tau > 0$. The quantiles can be estimated in a first stage. A heuristic argument in favor of this approach is as follows. By the definition of u_0 we can write:

$$Y = E[h_0(X)|Z] + u_0(Z) + \epsilon, \quad E[\epsilon|Z] = 0$$

$u_0(Z)$ represents any undesired channel by which the instruments are associated with the outcomes. For example, $u_0(Z)$ might capture the direct causal effect of the instruments on Y. Imagine we could remove this channel, say by eliminating the direct causal effect of Z on Y. Let $q_\epsilon(\cdot)$ return the quantiles of ϵ. Let $b = q_\epsilon(0.5+\tau)-q_\epsilon(0.5-\tau)$, then (2.2.1) states that the impact on Y of removing the channel $u_0(Z)$ is no greater than the impact of moving the residual variation in Y (after accounting for the effect of the instruments) from the $0.5-\tau$ to the $0.5+\tau$ quantile.

One must still select τ, but it may be easier to understand what constitutes a 'large' or 'small' shift in quantiles. In our empirical setting we take $b = q_\epsilon(0.6)-q_\epsilon(0.4)$ to correspond to severe misspecification: the undesired channel can impact Y by as much a shift in residual variation from the 40th to the 60th percentile.

2.2.2 Set estimation

To estimate $\underline{\theta}_{\mathbb{L}}$ and $\bar{\theta}_{\mathbb{L}}$ we must replace the optimization problems that define these quantities with feasible ones. Instead of optimizing over $\mathcal{B}_X$, we optimize over a finite-dimensional sieve space. We replace the conditional moment in (2.2.1) with a non-parametric estimate. Finally, we enforce the inequalities (2.2.1) and (2.2.2) only on finite grids.

Let Φ_n be a length-K_n column vector of basis functions on $\mathcal{X}$. In a first stage, one nonparametrically regresses Y and $\Phi_n(X)$ on Z to obtain estimates $\hat{g}_n$ and $\hat{\Pi}_n$ of $g_0(Z) = E[Y|Z]$ and $\Pi_n(Z) = E[\Phi_n(X)|Z]$. Let $\mathcal{X}_n$ and $\mathcal{Z}_n$ be finite grids in the supports of X and Z. We replace the constraints (2.2.1) and (2.2.2) with:

$$|\hat{g}_n(z) - \hat{\Pi}_n'(z)\beta| \leq b, \, \forall z \in \mathcal{Z}_n \tag{2.2.3}$$

$$\mathbb{T}[\Phi_n'](x)\beta \leq C(x), \, \forall x \in \mathcal{X}_n \tag{2.2.4}$$

Where $\mathbb{T}[\Phi_n'](x)$ be the d-by-K_n matrix whose k^{th} column is $\mathbb{T}[\Phi_{n,k}](x)$ where $\Phi_{n,k}$ is the k^{th} component of Φ_n. Let $\mathbb{L}[\Phi_n]$ be the column-vector whose k^{th} entry is $\mathbb{L}[\Phi_{n,k}]$. The estimates of $\bar{\theta}_{\mathbb{L}}$ and $\underline{\theta}_{\mathbb{L}}$ are respectively:

$$\hat{\bar{\theta}}_{\mathbb{L},n} = \max_{\beta \in \mathbb{R}^{K_n}} \mathbb{L}[\Phi_n]'\beta \;\text{ subject to (2.2.3) and (2.2.4).}$$

$$\hat{\underline{\theta}}_{\mathbb{L},n} = \min_{\beta \in \mathbb{R}^{K_n}} \mathbb{L}[\Phi_n]'\beta \;\text{ subject to (2.2.3) and (2.2.4).}$$

These are linear programming problems with K_n scalar parameters and $2|\mathcal{Z}_n| + d|\mathcal{X}_n|$ constraints.

2.2.3 Consistency and Convergence Rates

Let $||\cdot||_2$ be the Euclidean norm. For any $h \in \mathcal{B}_X$, $|h|_\infty = \sup_{x \in \mathcal{X}} |h(x)|$. For $g \in \mathcal{B}_Z$, $|g|_\infty$ is the essential supremum norm of g.[12]

A vector-valued function f on a domain $\mathcal{V}$ is of Hölder smoothness class $s \in (0, 1]$ with constant ξ if and only if:

$$\sup_{v_1, v_2 \in \mathcal{V}: v_1 \neq v_2} \frac{||f(v_1) - f(v_2)||_2}{||v_1 - v_2||_2^s} = \xi$$

[12] The essential supremum norm of g is the smallest number that exceeds $|g(Z)|$ with probability 1.

Let $\lfloor s \rfloor$ denote the largest integer less than s. f is of Hölder smoothness class $s > 1$ with constant ξ if and only if all the derivatives of each entry of f of order weakly less than $\lfloor s \rfloor$ are uniformly bounded by ξ and all the derivatives of order exactly $\lfloor s \rfloor$ are of Hölder smoothness class $s - \lfloor s \rfloor$ with constant ξ. A function is Lipschitz continuous with constant ξ if it is of Hölder smoothness class 1 with constant ξ.

Let $D_{1,n} = \sup_{x_1 \in \mathcal{X}} \min_{x_2 \in \mathcal{X}_n} ||x_1 - x_2||_2$ and $D_{2,n} = \sup_{z_1 \in \mathcal{Z}} \min_{z_2 \in \mathcal{Z}_n} ||z_1 - z_2||_2$. Let $C_n = 1 + \sup_{\beta \in \mathbb{R}^{K_n}} \frac{||\beta||_2}{|\Phi_n' \beta|_\infty}$.

Assumption 2.2. $\mathbb{T}[h](x) \leq C(x)$ implies $|h(x)| \leq \bar{c}$ for some $0 < \bar{c} < \infty$, and for some $\underline{c} > 0$, $C(x) \geq \underline{c}$ for all $x \in \mathcal{X}$.

Assumption 2.3. There is a sequence of positive scalars $a_n \to 0$ so that:

$$|\hat{g}_n - g_0|_\infty + \sup_{\beta \in \mathbb{R}^{K_n} : \Phi_n' \beta \in \mathcal{H}} |(\hat{\Pi}_n - \Pi_n)' \beta|_\infty = O_p(a_n)$$

Assumption 2.4. There is a sequence of positive scalars $\kappa_n \to 0$ so that for any $h \in \mathcal{H}$ there exists $\beta_n \in \mathbb{R}^{K_n}$ with $|\Phi_n' \beta_n - h|_\infty \leq \kappa_n$ and:

$$\mathbb{T}[\Phi_n'](x) \beta_n \leq \mathbb{T}[h](x), \, \forall x \in \mathcal{X}$$

Assumption 2.5. i. Φ_n, C, and each row of $\mathbb{T}[\Phi_n']$ are Lipschitz continuous with constant at most ξ_n. ii. With probability approaching 1 both $\hat{g}_n$ and $\hat{\Pi}_n$ are Lipschitz continuous with constant at most G_n. iii. $D_{1,n}, D_{2,n} \to 0$, $C_n \xi_n D_{1,n} \to 0$ and $C_n G_n D_{2,n} \to 0$.

Assumption 2.2 implies that $\mathcal{H}$ is bounded and convex. Assumption 2.3 quantifies the estimation error in $\hat{g}_n$ and $\hat{\Pi}_n$. 2.4 quantifies the error from the replacing $\mathcal{B}_X$ with a sieve space. Theorems 2.2 and 2.3 establish primitive conditions for Assumptions 2.3 and 2.4. Assumption 2.5 controls the error from the use of grids $\mathcal{X}_n$ and $\mathcal{Z}_n$.

Theorem 2.2.1. *Suppose Assumptions 2.1-2.5 hold and there exists $h \in \mathcal{B}_X$ so that (2.2.1) and (2.2.2) are slack, then uniformly over all linear functionals $\mathbb{L}$ with operator* *norms less than some fixed constant:*

$$|\hat{\underline{\theta}}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}}| = O_p(a_n + \kappa_n + C_n \xi_n D_{1,n} + C_n G_n D_{2,n}) = o_p(1)$$
$$|\hat{\bar{\theta}}_{\mathbb{L}} - \bar{\theta}_{\mathbb{L}}| = O_p(a_n + \kappa_n + C_n \xi_n D_{1,n} + C_n G_n D_{2,n}) = o_p(1)$$

In particular, the above holds uniformly over the functionals of the form $\mathbb{L}[h] = h(x)$ for each x in the support of X.

Theorem 2.1 demonstrates the well-posedness of the set estimation problem. The first-stage rate a_n is not premultiplied by some growing factor like a 'sieve measure of ill-posedness' ([18]).

$D_{1,n}$ and $D_{2,n}$ are small when the grids $\mathcal{X}_n$ and $\mathcal{Z}_n$ are dense. a_n and κ_n are independent of the grids, so if the grids grow dense quickly enough, the rates in the Theorem reduce to $a_n + \kappa_n$. This suggests the grids should be made as dense as computationally feasible.

If the dimension of the sieve space K_n grows quickly then κ_n converges rapidly to zero. Theorem 2.2 establishes a rate a_n that is independent of K_n. This suggests that under the conditions of Theorem 2.2, K_n should be made as large as is computational constraints allow.

We now provide primitive conditions for Assumption 2.3. Define series estimators $\hat{g}_n$ and $\hat{\Pi}_n$. Let Ψ_n be a length-L_n column vector of basis functions on $\mathcal{Z}$ and define $\hat{Q}_n = \frac{1}{n}\sum_{i=1}^{n} \Psi_n(Z_i)\Psi_n(Z_i)'$:

$$\hat{g}_n(z) = \Psi_n(z)'\hat{Q}_n^{-1}\frac{1}{n}\sum \Psi_n(Z_i)Y_i \tag{2.2.5}$$
$$\hat{\Pi}_n(z) = \Psi_n(z)'\hat{Q}_n^{-1}\frac{1}{n}\sum \Psi_n(Z_i)\Phi_n(X_i)' \tag{2.2.6}$$

In the Assumptions below $\mathcal{N}(\mathcal{H}, |\cdot|_\infty, \delta)$ is the smallest number of $|\cdot|_\infty$-balls of radius δ that can cover $\mathcal{H}$ and $\dim(Z)$ is the dimension of Z.

Assumption 2.6. i. The eigenvalues of $Q_n = E[\Psi_n(Z_i)\Psi_n(Z_i)']$ are bounded uniformly above and away from zero. ii. $\mathcal{Z}$ is bounded and the distribution of X given Z

admits a conditional density $f_{X|Z}$ so that $\forall x \in \mathcal{X}$, $f_{X|Z}(x,\cdot)$ is of Hölder smoothness class $s>0$ with constant at most $\bar{\ell}$.

Assumption 2.7. For any $s>0$ there is a sequence $R_n(s) \to 0$ so that for any $g \in \mathcal{B}_Z$ of Hölder smoothness class s with constant ξ:

$$\sup_{z\in\mathcal{Z}} |g(z) - \Psi_n(z)'Q_n^{-1}E[\Psi_n(Z)g(z)]| \leq \xi R_n(s)$$

ii. For all $z \in \mathcal{Z}$, $||\Psi_n(z)||_2 \leq \bar{\xi}_n$. $\alpha_n(z) = \frac{\Psi_n(z)}{||\Psi_n(z)||}$ is Lipschitz continuous with constant ℓ_n. iii. $\int_0^1 \sqrt{log\mathcal{N}(\mathcal{H}, |\cdot|_\infty, u)}du < \infty$. iv. $\frac{\bar{\xi}_n^2 log(L_n)}{n} \to 0$

Assumption 2.8. i. The function $g_0(Z) = E[Y|Z]$ is of Hölder smoothness class $s>0$. For $m>2$, $E\big[|Y - E[Y]|^m \big| Z\big] < \infty$, $\bar{\xi}_n^{2m/(m-2)} log(L_n)/n = O(1)$, $L_n log(L_n)/(n^{1-2/m}) = O(1)$ and $L_n^{2-2/\dim(Z)}/n = O(1)$. ii. $log(\ell_n) = O\big(log(L_n)\big)$, $\bar{\xi}_n = O(\sqrt{L_n})$ and $R_n(s) = O(L_n^{-s_0(s)/\dim(Z)})$ for some function $s_0 : \mathbb{R}_{++} \to \mathbb{R}_{++}$ and $R_n(s) = O(L_n^{-1/2})$.

Assumption 2.6.i is standard. Smoothness of the conditional density in 2.6.ii ensures that for any $h \in \mathcal{B}_X$, $\mathbb{A}[h]$ is smooth. Assumptions 2.7.i and 2.7.ii. can be verified for common basis functions. 2.7.iii is a condition on the metric entropy of $\mathcal{H}$, such conditions are commonplace in the sieve estimation literature (see [21]). Spaces of smooth functions like those used in our empirical application typically obey the condition (see [78] Chapter 5). Assumption 2.7.iv allows us to apply Rudelson's matrix law of large numbers ([70]).

Assumption 2.8.i allows us to apply results from [17] to derive a convergence rate for $\hat{g}_n$. 2.8.ii can be verified for a given choice of basis functions, for example if $s \geq 1/2$, then the Assumption holds for the one-dimensional B-spline case in Section 3.

Theorem 2.2.2. *Suppose Assumptions 2.1.ii, 2.2, and 2.6-2.8 hold. Define* $\hat{g}_n$ *and*

$\hat{\Pi}_n$ *as in (2.2.5) and (2.2.6). Then uniformly over sequences* $\{K_n\}_{n=1}^{\infty}$*:*

$$
\begin{aligned}
&|\hat{g}_n - g_0|_\infty + \sup_{\beta \in \mathbb{R}^{K_n}:\, \Phi_n' \beta \in \mathcal{H}} |(\hat{\Pi}_n - \Pi_n)'\beta|_\infty \\
=&O_p\Bigg(\sqrt{\frac{L_n log(L_n)}{n}} + L_n^{-s_0(s)/\dim(Z)}\Bigg) = o_p(1)
\end{aligned}
$$

If the conditions of the theorem hold, then setting L_n optimally Assumption 2.3 holds with $a_n = \left(\frac{log(n)}{n}\right)^{s_0(s)/(2s_0(s)+\dim(Z))}$.

Finally, we show Assumption 2.4 holds with $\kappa_n = O(K_n^{-\frac{1}{2}})$ for the setting in our empirical application.

Theorem 2.2.3. *Let $\mathcal{H}$ contain functions that map from a closed interval $[a,b]$ to $[0,1]$ so that any $h \in \mathcal{H}$ is twice-differentiable with $|\frac{\partial^2}{\partial x^2} h|_\infty \leq c$. Let Φ_n be a vector of s_0-order B-spline basis functions with K_n knot points evenly spaced between $[a,b]$. If $s_0 \geq 3$, Assumption 2.4 holds with $\kappa_n = O(K_n^{-\frac{1}{2}})$.*

2.2.4 Extensions

In some settings multiple vectors of instruments are available, each corresponding to a different moment condition. A researcher may wish to allow for different degrees of misspecification in each moment condition. Moreover, one may believe that $E[Y - h_0(X)|Z]$ is close to zero for some values of Z but but not others. One may also wish to allow $E[Y - h_0(X)|Z]$ to be say, more strongly positive than negative.

Our results extend to the more flexible set estimation problem that incorporates the relaxations above. For $k = 1, .., K$ let Z_k be a vector of instruments and let $b_{1,k}$ and $b_{2,k}$ be known functions. We replace the constraint (2.2.1) with a set of constraints:

$$
b_{1,k}(Z) \leq E[Y - h(X)|Z_k] \leq b_{2,k}(Z), k = 1, ..., K \tag{2.2.7}
$$

(2.2.1) is a special case of the above in which $K = 1$ and $b_{1,1}(Z) = -b_{2,1}(Z) = b$. To estimate of $\bar{\theta}_{\mathbb{L}}$ and $\underline{\theta}_{\mathbb{L}}$ with the constraints above, we replace the constraint (2.2.3) in the feasible problem with:

$$b_{1,k}(z) \leq \hat{g}_{n,k}(z) - \hat{\Pi}'_{n,k}(z)\beta \leq b_{2,k}(z), \forall z \in \mathcal{Z}_{n,k}, k = 1, ..., K$$

Where $\hat{g}_{n,k}$ and $\hat{\Pi}_{n,k}$ are first-stage nonparametric estimates of $g_{0,k}(Z_k) = E[Y|Z_k]$ and $\Pi_{n,k}(Z_k) = E[\Phi_n|Z_k]$, and $\mathcal{Z}_{n,k}$ is a finite subset of the support of Z_k. Theorem 2.0 in the appendix provides a rate of convergence for $\hat{\bar{\theta}}_{\mathbb{L}}$ and $\underline{\hat{\theta}}_{\mathbb{L}}$ in this more general set estimation problem. Theorem 2.1 is a special case of the more general theorem.

2.3 An Empirical Application

To demonstrate the usefulness of our methods we apply them to setting in Section 5.1 of [47]. [47] estimates a shape-invariant Engel curve for food using data from the British Family Expenditure Survey.[13] Horowitz's application is in turn based on [18] who also carry out NPIV estimation of shape-invariant Engel curves and use the same data.

From [47]: "The data are 1655 household-level observations from the British Family Expenditure Survey. The households consist of married couples with an employed head-of-household between the ages of 25 and 55 years."

[18] and [47] aim to estimate 'structural' Engel curves. Suppose a researcher were to exogenously allocate to a household a particular budget for non-durable goods. A structural Engel curve h_0 measures the share of that budget the household chooses to allocate to a class of goods as a function of the budget size.

In observational settings, the share of household wealth allocated to nondurables is decided by the household. Therefore, the budget for nondurables depends on the household's preferences. These preferences also determine the allocation of the budget to classes of non-durable goods. Thus expenditure on non-durable goods is endogenous.

To tackle this problem [18] and [47] use household income as an instrument for

[13] We made use of the data file that accompanies [47] and adapted the accompanying code in order to evaluate Horowitz's estimator and B-spline bases for our own methods.

nondurables expenditure. If variation in income is unrelated to preferences, then income is a valid instrument. However, household demographics, like the age and education of the householders, may have a causal effect on income and tastes. If these demographics are not controlled for then income is unlikely to be valid. [18] and [47] can only control for some coarse demographic variables.[14] Therefore, in this setting instrumental validity may not hold exactly and it is of interest to observe what empirical findings are robust to some failure of instrumental validity.

In this setting Y is the share of total expenditure on non-durables that a household spends on food. X is the logarithm of the household's total expenditure and Z is the logarithm of household income. We take $\mathcal{H}$ to be the set of functions that map from $\mathcal{X}$ to the unit interval and have second derivative bounded in magnitude by a constant c. We present results for $c = 2$ and $c = 5$, for comparison, the magnitude of the second derivatives of Horowitz's estimated structural function never exceed 0.5.

Following Subsection 2.1 we set $b = \hat{q}_\epsilon(0.5 + \tau) - \hat{q}_\epsilon(0.5 - \tau)$ for different values of τ. In particular we consider $\tau = 0.01, 0.05, 0.1$ to correspond to mild, moderate, and severe misspecification. $\hat{q}_\epsilon$ is an estimate of q_ϵ and is calculated by taking empirical quantiles of $Y_i - \hat{g}_n(Z_i)$. These values of τ correspond to $b = 0.0043, 0.024, 0.046$.

Following [47] we let Φ_n be fourth-order (cubic) B-spline basis functions with evenly-spaced knot points.[15] We carry out our first-stage estimates using series regression onto cubic B-splines. Motivated by the results in Theorem 2.2 we set K_n to be large, specifically we let $K_n = 10$. The grid $\mathcal{X}_n$ consists of 100 evenly spaced points between the smallest and largest observed values of X. The grid $\mathcal{Z}_n$ consists of 100 evenly spaced points between the 0.005 and 0.995 quantiles of Z.

Figure 4.1 contains the results of our procedure. The figure contains six subfigures each corresponding to a different set of values for τ and c. The lower and upper dotted lines represent $\hat{\underline{\theta}}_{h\mapsto h(x)}$ and $\hat{\bar{\theta}}_{h\mapsto h(x)}$, the upper and lower end points of the identified set for $h_0(x)$. The thick black line represents the half-way point between the end points, which (as we discuss in Section 2) is a point estimator with minimal

[14] Both papers control for the demographics by analyzing a homogeneous sub-sample. [18] incorporate additional dummy variable controls.

[15] See [31] for a practical introduction to B-splines.

worst-case asymptotic bias. The thin black line is the estimator from [47].

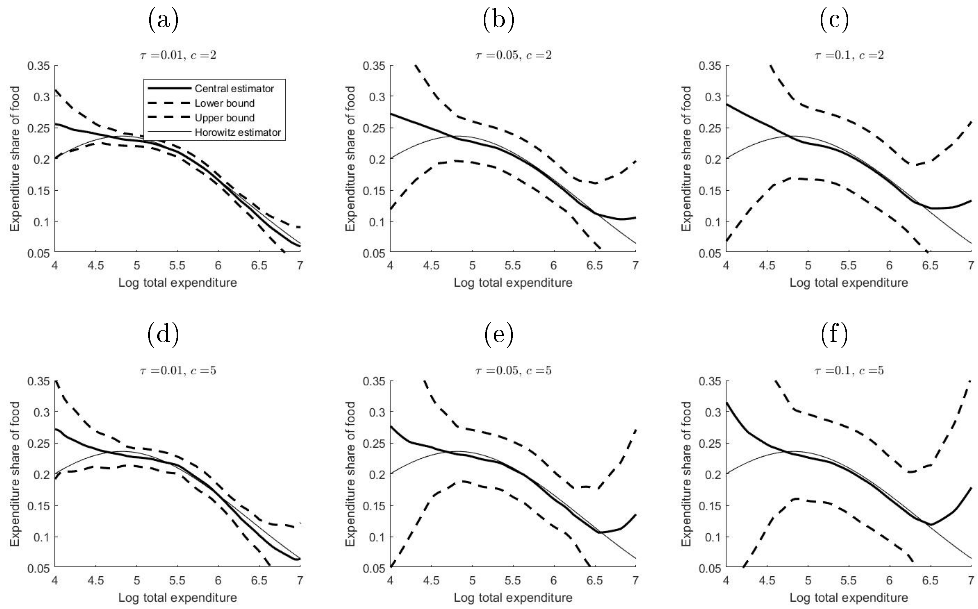


Figure 2-1: Set-Estimated Engel Curves

Results for various τ and c. Lower and upper dotted lines represent the end points of the identified set for $h_0(x)$ for each x. The thick black line is half-way between the dotted lines. The thin black line is the estimate in [47].

The results in Section 1 suggest that if the bound c on the second derivatives is loose then the identified set for $h_0(x)$ will be large, even if τ (and hence b) is small. This is clear in Figure 4.1 which shows that for a given τ, the intervals are wider when $c = 5$ than when $c = 2$.

Apart from in the severely misspecified case of $\tau = 0.1$ (Sub-figures (c) and (f)), the figures suggest a generally downward slope in the Engel curve for medium values of total expenditure. More precisely, the lower envelope at log-total expenditure of 5 exceeds the upper envelope at 6, which implies the Engel curve has decreased between these two values. None of the results in Figure 4.1 provide evidence in favor of an upward sloping Engel curve for low values of total expenditure as found by [47] and so we conclude then that this finding is not robust to even a mild failure of instrumental validity.

2.4 Conclusions

We demonstrate that NPIV estimates of the structural function can be highly sensitive to misspecification in the form of invalid instruments. If the interest is in smooth linear functionals, or strong smoothness restrictions are imposed in estimation, the situation may be more nuanced. This motivates a method for empirical sensitivity analysis based on estimation of the identified set under a relaxed instrumental validity assumption.

In future work we may generalize our sensitivity results to a broader class of conditional moment restriction models. The non-robustness of NPIV estimators is tied to the ill-posedness of the NPIV estimating equation and a range of other nonparametric conditional moment restriction models are ill-posed.

2.5 Appendix A: Proofs of Results in Section 1

Let $\mathcal{H} \subseteq \mathcal{B}_X$, $h_0 \in \mathcal{H}$ and $b \geq 0$. The b-modulus of continuity ([25]) of $\mathbb{A} : \mathcal{B}_X \to \mathcal{B}_Z$ in $\mathcal{H}$ at a point $h_0 \in \mathcal{H}$ is:

$$\omega(b, h_0, \mathcal{H}) = \sup_{h \in \mathcal{H} \colon ||\mathbb{A}[h] - \mathbb{A}[h_0]||_{\mathcal{B}_Z} \leq b} ||h - h_0||_{\mathcal{B}_X}$$

Lemma 1.1. *Fix $\mu_{XZ\eta}$ and suppose Assumptions 1.3 and 1.4 hold. Let $\hat{h}_n$ be an estimator so that if $h_0 \in int(\mathcal{H})$, and $u_0 = 0$ then $\underset{n\to\infty}{plim}||\hat{h}_n - h_0||_{\mathcal{B}_X} = 0$. Then $bias_{\hat{h}_n}(b) \geq \omega(b, h_0, \mathcal{H})$.*

Proof. Suppose $g_0 \in \mathbb{A}[\mathcal{H}]$. If $u_0 = 0$ then $h_0 = \mathbb{A}^{-1}[g_0]$ and so by consistency, for any $g_0 \in \mathbb{A}[\mathcal{H}]$, $\underset{n\to\infty}{\text{plim}}||\hat{h}_n - \mathbb{A}^{-1}[g_0]||_{\mathcal{B}_X} = 0$.

Now suppose $u_0 \in \mathbb{A}[\mathcal{H} - h_0]$ and $||u_0||_{\mathcal{B}_Z} \leq b$. Let $h_{u_0} = h_0 + \mathbb{A}^{-1}[u_0]$, then $h_{u_0} \in \mathcal{H}$, $||\mathbb{A}[h_0] - \mathbb{A}[h_{u_0}]||_{\mathcal{B}_Z} \leq b$. Further, $g_0 = \mathbb{A}[h_0] + u_0 = \mathbb{A}[h_{u_0}]$ so $g_0 \in \mathbb{A}[\mathcal{H}]$ and thus $\underset{n\to\infty}{\text{plim}}||\hat{h}_n - h_{u_0}||_{\mathcal{B}_X} = 0$, then by the triangle inequality:

$$\underset{n\to\infty}{\text{plim}}||\hat{h}_n - h_0||_{\mathcal{B}_X} \geq ||h_{u_0} - h_0||_{\mathcal{B}_X}$$

This holds for any $u_0 \in \mathbb{A}[\mathcal{H} - h_0]$ with $||u_0||_{\mathcal{B}_Z} \leq b$, or equivalently every $h_{u_0} \in \mathcal{H}$ with $||\mathbb{A}[h_0] - \mathbb{A}[h_{u_0}]||_{\mathcal{B}_Z} \leq b$. Therefore:

$$\sup_{u_0 \in \mathbb{A}[\mathcal{H}-h_0]:\, ||u_0||_{\mathcal{B}_Z} \leq b} \operatorname*{plim}_{n\to\infty} ||\hat{h}_n - h_0||_{\mathcal{B}_X} \geq \sup_{h \in \mathcal{H}:\, ||\mathbb{A}[h] - \mathbb{A}[h_0]||_{\mathcal{B}_Z} \leq b} ||h - h_0||_{\mathcal{B}_X}$$

Since $\mathbb{A}[\mathcal{H} - h_0] \subseteq R(A)$ the above implies:

$$\sup_{u_0 \in R(A):\, ||u_0||_{\mathcal{B}_Z} \leq b} \operatorname*{plim}_{n\to\infty} ||\hat{h}_n - h_0||_{\mathcal{B}_X} \geq \sup_{h \in \mathcal{H}:\, ||\mathbb{A}[h] - \mathbb{A}[h_0]||_{\mathcal{B}_Z} \leq b} ||h - h_0||_{\mathcal{B}_X}$$

Applying the definitions of the worst-case asymptotic bias and the modulus of continuity gives the result.

□

Lemma 1.2. *Let $\mathbb{A}$ be an infinite-dimensional linear operator from $\mathcal{B}_X$ to $\mathcal{B}_Z$ that satisfies Assumptions 1.3 and 1.4. Suppose $h \in int(\mathcal{H})$. Let $r > 0$ be the radius of an open ball in $int(\mathcal{H})$ centered at h. Then for any $b > 0$, $\omega(b, h, \mathcal{H}) \geq r$.*

Proof. Because $\mathcal{H}$ contains an open ball and $\mathbb{A}$ is linear, Assumption 1.3 implies that $\mathbb{A}$ is invertible. By Assumption 1.4, $\mathbb{A}$ is a compact infinite-dimensional operator, and so $\mathbb{A}^{-1}$ is unbounded (see Theorem 15.4 in [58]). Thus for any $b > 0$ and $0 < \epsilon < r$, there exists $h' \in \mathcal{B}_X$ with $||h'||_{\mathcal{B}_X} = 1$ and $||\mathbb{A}[h']||_{\mathcal{B}_Z} \leq \frac{b}{r-\epsilon}$. Define h'' by $h'' = h + (r-\epsilon)h'$. Note that $||h - h''||_{\mathcal{B}_X} = r - \epsilon < r$ and so (because there is an open ball in $\mathcal{H}$ centered at h of radius r), $h'' \in \mathcal{H}$. Using linearity of $\mathbb{A}$ and the properties of norms:

$$||\mathbb{A}[h] - \mathbb{A}[h'']||_{\mathcal{B}_Z} = (r - \epsilon)||\mathbb{A}[h']||_{\mathcal{B}_Z} \leq b$$

So $||\mathbb{A}[h] - \mathbb{A}[h'']||_{\mathcal{B}_Z} \leq b$, $||h - h''||_{\mathcal{B}_X} = r - \epsilon$, and $h'' \in \mathcal{H}$. So:

$$\omega(b, h, \mathcal{H}) = \sup_{h' \in \mathcal{H}:\, ||\mathbb{A}[h] - \mathbb{A}[h']||_{\mathcal{B}_Z} \leq b} ||h - h'||_{\mathcal{B}_X} \geq r - \epsilon$$

Since this holds for all $0 < \epsilon < r$ it also holds for $\epsilon = 0$. □

Proof of Theorem 1.1. The first two statements follow immediately from Lemmas 1.1

and 1.2. The final statement follows from the fact that the Banach space $\mathcal{B}_X$ must contain an open ball centered at h_0 of arbitrarily large radius. $\square$

Proof of Theorem 1.2. The identified set Θ, under Assumptions 1.1* and 1.2, is the set of functions $h \in \mathcal{B}_X$ that satisfy the assumptions. So:

$$\Theta = [h \in \mathcal{B}_X : \, h \in \mathcal{H}, ||g_0 - \mathbb{A}[h]|| \leq b]$$

Suppose $\mathbb{A}^{-1}[g_0] \in \mathcal{H}$ and there is a closed ball in $\mathcal{H}$ of radius r centered at $\mathbb{A}^{-1}[g_0]$. Then if $h \in \mathcal{B}_X$ and $||h - \mathbb{A}^{-1}[g_0]||_{\mathcal{B}_X} \leq r$ we must have $h \in \mathcal{H}$. Under Assumptions 1.3 and 1.4, $\mathbb{A}^{-1}$ is an unbounded linear operator, so for any $C > 0$ we can find $h \in R(A)$ with $||h||_{\mathcal{B}_X} = 1$ and $||\mathbb{A}[h]||_{\mathcal{B}_Z} \leq C$. Thus there exists $h' \in \mathcal{B}_X$ with $||h'||_{\mathcal{B}_X} = 1$ and $||\mathbb{A}[h']||_{\mathcal{B}_Z} \leq \frac{b}{r}$. So let $h'' = \mathbb{A}^{-1}[g_0] + rh'$, then $||\mathbb{A}^{-1}[g_0] - h''||_{\mathcal{B}_X} = r$ so $h'' \in \mathcal{H}$, and $||g_0 - \mathbb{A}[h'']||_{\mathcal{B}_X} \leq b$, so $h'' \in \Theta$. Setting $h''' = \mathbb{A}^{-1}[g_0] - rh'$ we can follow the same reasoning to see $h''' \in \Theta$, and note that $||h'' - h'''||_{\mathcal{B}_X} = 2r$ so the diameter of the identified set is at least $2r$. Since this holds for every r that is the radius of a closed ball centered at $\mathbb{A}^{-1}[g_0]$, we see that the diameter of Θ must be at least twice the radius of the largest open ball in $\mathcal{H}$ centered at $\mathbb{A}^{-1}[g_0]$. $\square$

Proof of Theorem 1.3. It is well-known that a continuous and injective function defined on a compact set has a continuous inverse. So by compactness of $\mathcal{H}$, $\mathbb{A}_{\mathcal{H}}^{-1}$ is continuous. Since $\mathbb{A}_{\mathcal{H}}^{-1}$ is linear it is also uniformly continuous. P_Z is uniformly continuous by assumption, and the composition of two uniformly continuous operators is uniformly continuous, hence $\mathbb{A}_{\mathcal{H}}^{-1} P_Z$ is uniformly continuous. By uniform continuity, for any $g \in R(A)$:

$$\lim_{b \to 0} \sup_{g' \in R(A) : \, ||g' - g|| \leq b} ||\mathbb{A}_{\mathcal{H}}^{-1} P_Z[g'] - \mathbb{A}_{\mathcal{H}}^{-1} P_Z[g]||_{\mathcal{B}_X} = 0$$

Set $g = \mathbb{A}[h_0]$ in the above. Since $h_0 \in \mathcal{H}$ we have $\mathbb{A}_{\mathcal{H}}^{-1} P_Z[\mathbb{A}[h_0]] = h_0$. Since $R(A)$ is a linear space, for any $g_0 \in R(A)$ there exists $u_0 \in R(A)$ so that $g_0 = u_0 + \mathbb{A}[h_0]$. So we get:

$$\lim_{b\to 0} \sup_{u_0 \in R(A):\, ||u_0|| \le b} ||\mathbb{A}_{\mathcal{H}}^{-1} P_Z[g_0] - h_0||_{\mathcal{B}_X} = 0 \tag{2.5.1}$$

The triangle inequality implies:

$$||\hat{h}_n - h_0||_{\mathcal{B}_X} \le ||\mathbb{A}_{\mathcal{H}}^{-1} P_Z[g_0] - h_0||_{\mathcal{B}_X} + ||\hat{h}_n - \mathbb{A}_{\mathcal{H}}^{-1} P_Z[g_0]||_{\mathcal{B}_X}$$

Since $||\hat{h}_n - \mathbb{A}_{\mathcal{H}}^{-1} P_Z[g_0]||_{\mathcal{B}_X} \to^p 0$ it follows that:

$$\operatorname*{plim}_{n\to\infty} ||\hat{h}_n - h_0||_{\mathcal{B}_X} = ||\mathbb{A}_{\mathcal{H}}^{-1} P_Z[g_0] - h_0||_{\mathcal{B}_X}$$

Substituting the above into (2.5.1) and applying the definition of the worst-case asymptotic bias gives $\lim_{b\to 0} bias_{\hat{h}_n}(b) = 0$. $\square$

Lemma 1.3. *Suppose $\mathbb{A}$ satisfies Assumptions 1.3 and 1.4 and Suppose that $\mathcal{H} \subset \mathcal{B}_X$ is compact, convex, symmetric, and infinite dimensional and there exists $\alpha \in (0,1)$ so that $\frac{1}{\alpha} h_0 \in \mathcal{H}$. Then: $\lim_{b\to 0} \frac{\omega(b,h_0,\mathcal{H})}{b} = \infty$.*

Proof. Assume the contrary, then for some $b > 0$ there exists a scalar C so that for any $h \in \mathcal{H}$:

$$||\mathbb{A}[h] - \mathbb{A}[h_0]||_{\mathcal{B}_Z} \le b \implies ||h - h_0||_{\mathcal{B}_X} \le C||\mathbb{A}[h] - \mathbb{A}[h_0]||_{\mathcal{B}_Z}$$

We have $\frac{1}{\alpha} h_0 \in \mathcal{H}$ for some $\alpha \in (0,1)$ and $\mathcal{H}$ is convex. So for any $h \in \mathcal{H}$ there exists $h' \in \mathcal{H}$ so that $(1-\alpha)h = h' - h_0 \in \mathcal{H}$. Therefore the above implies that for any $h \in \mathcal{H}$:

$$||\mathbb{A}[h]||_{\mathcal{B}_Z} \le \frac{1}{1-\alpha} b \implies ||h||_{\mathcal{B}_X} \le C||\mathbb{A}[h]||_{\mathcal{B}_Z}$$

By definition of the operator norm of $\mathbb{A}$, $||\mathbb{A}||_{op}$, $||\mathbb{A}[h]||_{\mathcal{B}_Z} \le ||\mathbb{A}||_{op} ||h||_{\mathcal{B}_X}$. Assumption 1.3 implies that $||\mathbb{A}||_{op} > 0$. And so for any $h \in \mathcal{H}$:

$$||h||_{\mathcal{B}_X} \le \frac{1}{||\mathbb{A}||_{op}(1-\alpha)} b \implies ||h||_{\mathcal{B}_X} \le C||\mathbb{A}[h]||_{\mathcal{B}_Z}$$

Let R be the closed ball in $\mathcal{B}_X$ of radius $\frac{1}{||\mathbb{A}||_{op}(1-\alpha)} b$. Let $\mathcal{C}$ be the cone of $R \cap \mathcal{H}$ (that is, $\mathcal{C} = [\gamma h : \ h \in R \cap \mathcal{H}, \gamma \in \mathbb{R}_+])$.

We have already shown that for any $h \in R \cap \mathcal{H}$, $||h||_{\mathcal{B}_X} \leq C||\mathbb{A}[h]||_{\mathcal{B}_Z}$. By linearity of $\mathbb{A}$ and properties of norms, for any $h \in \mathcal{C}$ we have that $||h||_{\mathcal{B}_X} \leq C||\mathbb{A}[h]||_{\mathcal{B}_Z}$.

Let $\bar{\mathcal{C}}$ be the closure of $\mathcal{C}$. For any $h \in \bar{\mathcal{C}}$ there is a sequence h_k in $\mathcal{C}$ so that $||h - h_k||_{\mathcal{B}_X} \to 0$. For all k, $||h_k||_{\mathcal{B}_X} \leq C||\mathbb{A}[h_k]||_{\mathcal{B}_Z}$, so by the triangle inequality and the definition of the operator norm:

$$||h||_{\mathcal{B}_X} \leq C||\mathbb{A}[h]||_{\mathcal{B}_Z} + (1 + C||\mathbb{A}||_{op})||h - h_k||_{\mathcal{B}_X}$$

$\mathbb{A}$ is continuous so $||\mathbb{A}||_{op} < \infty$, and so since $||h - h_k||_{\mathcal{B}_X} \to 0$ we get $||h||_{\mathcal{B}_X} \leq C||\mathbb{A}[h]||_{\mathcal{B}_Z}$. Thus $\mathbb{A}^{-1}$ exists and is bounded on $\bar{\mathcal{C}}$.

Now, $R \cap \mathcal{H}$ is convex and symmetric which implies $\mathcal{C}$ is a linear space and therefore so is $\bar{\mathcal{C}}$. Because $\mathcal{H}$ is infinite-dimensional, convex, and symmetric (and therefore contains zero), $\mathcal{C}$ is infinite-dimensional and likewise $\bar{\mathcal{C}}$. Finally, $\bar{\mathcal{C}}$ is a closed subset of $\mathcal{B}_X$ by construction, and therefore it is complete in the norm $|| \cdot ||_{\mathcal{B}_X}$. Thus $\bar{\mathcal{C}}$ is an infinite-dimensional Banach space with norm $|| \cdot ||_{\mathcal{B}_X}$. But the inverse of a compact injective operator on an infinite-dimensional Banach space cannot be bounded, and so we have a contradiction. □

Proof of Theorem 1.4. Follows immediately from Lemmas 1.1 and 1.3. □

Proof of Theorem 1.5. Let us introduce notation. Let the $L_2(\mu_X)$ and $L_2(\mu_Z)$ inner products be $\langle \cdot, \cdot \rangle_{L_2(\mu_X)}$ and $\langle \cdot, \cdot \rangle_{L_2(\mu_Z)}$ respectively. The linear functional of interest $\mathbb{L}[h_0]$ can then be written as $\mathbb{L}[h_0] = \langle w, h_0 \rangle_{L_2(\mu_X)}$. The adjoint of the operator $\mathbb{A}$, denoted $\mathbb{A}^*$ is given by $\mathbb{A}^*[g](X) = E[g(Z)|X]$.

First we prove the following:

$$bias_{\hat{l}_n}(b) = \sup_{u_0 \in R(A):\, ||u_0||_{L_2(\mu_Z)} \leq b} |E[w(X)\mathbb{A}^{-1}[u_0](X)]| \tag{2.5.2}$$

If instruments are valid then $h_0 = \mathbb{A}^{-1}[g_0]$ and $g_0 \in R(A)$. So consistency under instrumental validity, implies that if $g_0 \in R(A)$ then:

$$|\hat{l}_n - E[w(X)\mathbb{A}^{-1}[g_0](X)]| \to^p 0$$

Let $u_0 \in R(A)$. Using $g_0 = \mathbb{A}[h_0] + u_0$:

$$|E[w(X)\mathbb{A}^{-1}[g_0](X)] - \mathbb{L}[h_0]| = |E[w(X)\mathbb{A}^{-1}[u_0](X)]|$$

And so:

$$\operatorname*{plim}_{n\to\infty} |\hat{l}_n - \mathbb{L}[h_0]| = |E[w(X)\mathbb{A}^{-1}[u_0](X)]|$$

Applying the definition of the worst-case asymptotic bias then gives (2.5.2).

Let us now prove claim a. Suppose that for some α, $w(X) = E[\alpha(Z)|X]$, equivalently $w = \mathbb{A}^*[\alpha]$. Then for any $g \in R(A)$:

$$\langle w, \mathbb{A}^{-1}[g]\rangle_{L_2(\mu_X)} = \langle \mathbb{A}^*[\alpha], \mathbb{A}^{-1}[g]\rangle_{L_2(\mu_X)} = \langle \alpha, \mathbb{A}\mathbb{A}^{-1}[g]\rangle_{L_2(\mu_Z)}$$

Which simply equals $\langle \alpha, g\rangle_{L_2(\mu_Z)}$. Suppose $||g||_{L_2(\mu_Z)} \leq b$, by Cauchy-Schwartz:

$$|\langle w, \mathbb{A}^{-1}[g]\rangle_{L_2(\mu_X)}| \leq ||g||_{L_2(\mu_Z)}||\alpha||_{L_2(\mu_Z)} \leq b||\alpha||_{L_2(\mu_Z)}$$

Combining the above with (2.5.2) gives the result.

We now prove claim b. Suppose that for some $\bar{b} > 0$ the worst-case asymptotic bias is finite. That is, there exists a scalar $c < \infty$ so that:

$$\sup_{u_0 \in R(A):\, ||u_0||_{L_2(\mu_Z)} \leq \bar{b}} |\langle w, \mathbb{A}^{-1}[u_0]\rangle_{L_2(\mu_X)}| \leq c$$

Linearity of $\mathbb{A}^{-1}$ and the inner-product then implies that for all $b > 0$:

$$\sup_{u_0 \in R(A):\, ||u_0||_{L_2(\mu_Z)} \leq b} |\langle w, \mathbb{A}^{-1}[u_0]\rangle_{L_2(\mu_X)}| \leq \frac{c}{\bar{b}} b$$

Also by linearity, the LHS above equals:

$$\sup_{u_0 \in R(A):\, ||u_0||_{L_2(\mu_Z)} \leq b} \langle w, \mathbb{A}^{-1}[u_0] \rangle_{L_2(\mu_X)}$$

Define the function $D : R(A) \to \mathbb{R}$ by $D[g] = \langle w, \mathbb{A}^{-1}[g] \rangle_{L_2(\mu_X)}$. Then:

$$\sup_{u_0 \in R(A):\, ||u_0||_{L_2(\mu_Z)} \leq b} |D[u_0]| \leq \frac{c}{\bar{b}} b$$

By the Hahn-Banach theorem we can extend D to a bounded linear function $\bar{D}$ defined on the whole space $L_2(\mu_Z)$ which then satisfies:

$$\sup_{u_0 \in L_2(\mu_Z):\, ||u_0||_{L_2(\mu_Z)} \leq b} |\bar{D}[u_0]| \leq \frac{c}{\bar{b}} b$$

Since $\bar{D}$ is a bounded linear functional defined on a Hilbert space, by the Reisz representation theorem there exists an element $\alpha \in L_2(\mu_Z)$ so that for all $g \in L_2(\mu_Z)$, $\bar{D}[g] = \langle \alpha, g \rangle_{L_2(\mu_Z)}$. And so for any $u_0 \in R(A)$:

$$D[u_0] = \langle \alpha, u_0 \rangle_{L_2(\mu_Z)} = \langle \mathbb{A}^*[\alpha], \mathbb{A}^{-1}[u_0] \rangle_{L_2(\mu_X)} = \langle w, \mathbb{A}^{-1}[u_0] \rangle_{L_2(\mu_X)}$$

Where the final equality follows by the definition of D.

Since the equality above holds for all $u_0 \in R(A)$, for all $h \in L_2(\mu_X)$ we have (using bi-linearity of the inner product) $\langle \mathbb{A}^*[\alpha] - w, h \rangle_{L_2(\mu_X)} = 0$. But we can set $h = \mathbb{A}^*[\alpha] - w$ and the above implies that the norm of $\mathbb{A}^*[\alpha] - w$ equals zero and so we have that $\mathbb{A}^*[\alpha] - w = 0$. Or equivalently $w(X) = E[\alpha(Z)|X]$. $\square$

2.6 Appendix B: Proofs of Results in Section 2

Proposition 2.1. *Under Assumption 2.1, for any linear operator* $\mathbb{L} : \mathcal{B}_X \to \mathbb{R}$, $\Theta_{\mathbb{L}} = [\mathbb{L}[h] \in \mathbb{R} : h \in \Theta]$ *is an interval.*

Proof. The constraints (2.1) and (2.2) are clearly convex and therefore so is Θ. Suppose $\theta', \theta'' \in \Theta_{\mathbb{L}}$, then $\exists h', h'' \in \mathcal{B}_X$ with $\mathbb{L}[h'] = \theta'$, $\mathbb{L}[h''] = \theta''$ and $h', h'' \in \Theta$. Since

Θ is convex $h''' = \alpha h' + (1-\alpha)h'' \in \Theta$ for any $\alpha \in [0,1]$. $\mathbb{L}$ is linear, so we then have $\alpha\theta' + (1-\alpha)\theta'' = \mathbb{L}[h'''] \in \Theta_{\mathbb{L}}$. So $\Theta_{\mathbb{L}}$ is a convex subset of $\mathbb{R}$, i.e., an interval. $\square$

Lemma 2.1. *For each $j \in [J]$ let $\mathbb{M}_j$ be a linear operator from a vector space $\mathcal{S}$ to the space of functions from $\mathcal{V}_j$ to $\mathbb{R}$ and let b_j be a function from $\mathcal{V}_j$ to $\mathbb{R}$.*[16] *Let $\mathbb{L}$ be a linear operator that maps from $\mathcal{S}$ to $\mathbb{R}$. Define:*

$$\bar{\theta}_{\mathbb{L}} = \sup_{s \in \mathcal{S}} \mathbb{L}[s] \ \ s.t. \ \ \mathbb{M}_j[s](v) \leq b_j(v), \forall v \in \mathcal{V}_j, \forall j \in [J]$$

And $\underline{\theta}_{\mathbb{L}}$ is the infimum subject the same constraints.

Consider some $\hat{s} \in \mathcal{S}$ so that for some $r > 0$ and all $j \in [J]$, $\mathbb{M}_j[\hat{s}] \leq b_j(v) + r, \forall v \in \mathcal{V}_j$. Suppose there exists $\epsilon > 0$ and $\tilde{s} \in \mathcal{S}$ so that for each $j \in [J]$, $\mathbb{M}_j[\tilde{s}](v) \leq b_j(v) - \epsilon$, $\forall v \in \mathcal{V}_j$. Then $\mathbb{L}[\hat{s}] - \bar{\theta}_{\mathbb{L}} \leq \frac{r}{\epsilon}(\bar{\theta}_{\mathbb{L}} - \mathbb{L}[\tilde{s}]) \leq \frac{r}{\epsilon}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}})$.

Proof. Define $s^* \in \mathcal{S}$ by $s^* = \frac{\epsilon}{r+\epsilon}\hat{s} + \frac{r}{r+\epsilon}\tilde{s}$. By the linearity of $\mathbb{M}_j$, for $j \in [J]$:

$$\begin{aligned}\mathbb{M}_j[s^*](v) &= \frac{\epsilon}{r+\epsilon}\mathbb{M}_j[\hat{s}](v) + \frac{r}{r+\epsilon}\mathbb{M}_j[\tilde{s}](v) \\ &\leq \frac{\epsilon}{r+\epsilon}(r + b_j(v)) - \frac{r}{r+\epsilon}(\epsilon - b_j(v)) = b_j(v)\end{aligned}$$

So s^* satisfies the constraints in the problems for $\bar{\theta}_{\mathbb{L}}$ and $\underline{\theta}_{\mathbb{L}}$, so we must have $\mathbb{L}[s^*] \leq \bar{\theta}_{\mathbb{L}}$. Substituting the definition of s^* and using linearity of $\mathbb{L}$ we get:

$$\mathbb{L}[\hat{s}] - \overline{y} \leq \frac{r}{r+\epsilon}\mathbb{L}[\hat{s} - \tilde{s}] = \frac{r}{r+\epsilon}(\bar{\theta}_{\mathbb{L}} - \mathbb{L}[\tilde{s}]) + \frac{r}{r+\epsilon}(\mathbb{L}[\hat{s}] - \bar{\theta}_{\mathbb{L}})$$

Subtracting $\frac{r}{r+\epsilon}(\mathbb{L}[\hat{s}] - \bar{\theta}_{\mathbb{L}})$ from both sides and then dividing by $\frac{\epsilon}{r+\epsilon}$ we get $\mathbb{L}[\hat{s}] - \bar{\theta}_{\mathbb{L}} \leq \frac{r}{\epsilon}(\bar{\theta}_{\mathbb{L}} - \mathbb{L}[\tilde{s}])$. Noting $\underline{\theta}_{\mathbb{L}} \leq \mathbb{L}[\tilde{s}]$ gives the conclusion. $\square$

Lemma 2.2. *For each $j \in [J]$ let $\mathbb{M}_j$ be a linear operator from a vector space $\mathcal{S}$ with norm $||\cdot||_{\mathcal{S}}$ to the space of functions from $\mathcal{V}_j$ to $\mathbb{R}$, and let b_j be a function from $\mathcal{V}_j$ to $\mathbb{R}$. Let $\mathbb{L}$ be a linear operator from $\mathcal{S}$ to $\mathbb{R}$. Define:*

$$\bar{\theta}_{\mathbb{L}} = \sup_{s \in \mathcal{S}} \mathbb{L}[s] \ \ s.t. \ \ \mathbb{M}_j[s](v) \leq b_j(v), \forall v \in \mathcal{V}_j, \forall j \in [J]$$

[16] $[J]$ denote the set of natural numbers from 1 to J.

And let $\underline{\theta}_{\mathbb{L}}$ be the infimum subject to the same constraints.

Consider a subset $\tilde{\mathcal{S}} \subseteq \mathcal{S}$. Suppose: i. There exists $\mathcal{J} \subseteq [J]$ so that for any $s \in \mathcal{S}$, if $\mathbb{M}_j[s](v) \leq b_j(v), \forall v \in \mathcal{V}_j, j \in \mathcal{J}$, then there exists $s' \in \tilde{\mathcal{S}}$ so that $||s - s'||_{\mathcal{S}} \leq \omega$ and $\mathbb{M}_j[s'](v) \leq \mathbb{M}_j[s](v), \forall v \in \mathcal{V}_j, j \in \mathcal{J}$. ii. For each $j \in \mathcal{J}^c = [J] \setminus \mathcal{J}$ the linear operators of the form $s \mapsto \mathbb{M}_j[s](v)$ for each $v \in \mathcal{V}_j$ have operator norm less than c'. iii. There exists $\epsilon > 0$ and $\tilde{s} \in \mathcal{S}$ so that for each $j \in [J]$, $\mathbb{M}_j[\tilde{s}](v) \leq b_j(v) - \epsilon$, $\forall v \in \mathcal{V}_j$. iv. $\mathbb{L}$ is a continuous linear operator with operator norm c''.

Under conditions i., ii., and iv., there exists $\tilde{s} \in \tilde{\mathcal{S}}$ so that for each $j \in [J]$, $\mathbb{M}_j[\tilde{s}](v) \leq b_j(v) + c'\omega - \epsilon$, $\forall v \in \mathcal{V}_j$.

Define $\tilde{\bar{\theta}}_{\mathbb{L}}$ by:

$$\tilde{\bar{\theta}}_{\mathbb{L}} = \sup_{s \in \tilde{\mathcal{S}}} \mathbb{L}[s] \ \ s.t. \ \ \mathbb{M}_j[s](v) \leq b_j(v), \forall v \in \mathcal{V}_j, \forall j \in [J]$$

Under conditions i. ii., iii., and iv., $|\bar{\theta}_{\mathbb{L}} - \tilde{\bar{\theta}}_{\mathbb{L}}| \leq \omega\big(\frac{c'}{\epsilon}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}}) + c''\big)$.

Proof. By conditions i. and iii. there exists an $\tilde{s}' \in \tilde{\mathcal{S}}$ so that $||\tilde{s}' - \tilde{s}||_{\mathcal{S}} \leq \omega$ and for $j \in \mathcal{J}$, $\mathbb{M}_j[\tilde{s}'](v) \leq b_j(v) - \epsilon$ for each $v \in \mathcal{V}_j$. By ii., for any $s_1, s_2 \in \mathcal{S}$ and all $v \in \mathcal{V}_j$, $|\mathbb{M}_j[s_1 - s_2](v)| \leq c'||s_1 - s_2||_{\mathcal{S}}$. Therefore for $j \in \mathcal{J}^c$, $|\mathbb{M}_j[\tilde{s} - \tilde{s}'](v)| \leq c'\omega$. By the triangle inequality and $\mathbb{M}_j[\tilde{s}](v) \leq b_j(v) - \epsilon$:

$$\mathbb{M}_j[\tilde{s}'](v) \leq b_j(v) - \epsilon + c'\omega$$

So the first claim of the lemma holds, now we prove the second claim. By definition of the supremum, for any $r > 0$ there must be some $s' \in \mathcal{S}$ that satisfies the constraints in the problem for $\bar{\theta}_{\mathbb{L}}$ and achieves $\bar{\theta}_{\mathbb{L}} - \mathbb{L}[s'] \leq r$. By the same reasoning used to establish the first claim there must exist a $s'' \in \tilde{\mathcal{S}}$ with $||s'' - s'||_{\mathcal{S}} \leq \omega$ and for $j \in \mathcal{J}$, $\mathbb{M}_j[s''](v) \leq b_j(v)$ and for $j \in \mathcal{J}^c$, $\mathbb{M}_j[s''](v) \leq b_j(v) + c'\omega$. Let $s^* = \frac{\epsilon - c'\omega}{\epsilon} s'' + \frac{c'\omega}{\epsilon}\tilde{s}'$. Since $\tilde{\mathcal{S}}$ is linear $s^* \in \tilde{S}$. Moreover, by linearity of M_j we have get that for $j \in \mathcal{J}$ $\mathbb{M}_j[s^*](v) \leq b_j(v)$ and for $j \in \mathcal{J}^c$:

$$\mathbb{M}_j[s^*](v) \leq \frac{\epsilon - c'\omega}{\epsilon}(b_j(v) + c'\omega) + \frac{c'\omega}{\epsilon}(b_j(v) - \epsilon + c'\omega) = b_j(v)$$

So s^* satisfies the constraints of the problems for $\bar{\theta}_{\mathbb{L}}$ and $\tilde{\bar{\theta}}_{\mathbb{L}}$ so $\bar{\theta}_{\mathbb{L}} \geq \tilde{\bar{\theta}}_{\mathbb{L}} \geq \mathbb{L}[s^*]$. Now, by iv., for any $s_1, s_2 \in \mathcal{S}$, $|\mathbb{L}[s_1 - s_2]| \leq c''||s_1 - s_2||_{\mathcal{S}}$. Therefore using the definition of s^*:

$$\mathbb{L}[s' - s^*] = \mathbb{L}[s' - s''] + \mathbb{L}[s' - s^*] \leq c''\omega + \frac{c'\omega}{\epsilon}\mathbb{L}[s' - \tilde{s}']$$

And so $\bar{\theta}_{\mathbb{L}} - \mathbb{L}[s^*] \leq r + c''\omega + \frac{c'\omega}{\epsilon}\mathbb{L}[s' - \tilde{s}']$. Since s' and $\tilde{s}'$ satisfy the constraints of the problems for $\bar{\theta}_{\mathbb{L}}$ and $\underline{\theta}_{\mathbb{L}}$, $\mathbb{L}[s'] \leq \bar{\theta}_{\mathbb{L}}$ and $\mathbb{L}[\tilde{s}'] \geq \underline{\theta}_{\mathbb{L}}$, and recall $\tilde{\bar{\theta}}_{\mathbb{L}} \geq \mathbb{L}[s^*]$ so (using linearity of $\mathbb{L}$) we get $\bar{\theta}_{\mathbb{L}} - \tilde{\bar{\theta}}_{\mathbb{L}} \leq r + c''\omega + \frac{c'\omega}{\epsilon}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}})$. Since this holds for any $r > 0$ we get $\bar{\theta}_{\mathbb{L}} - \tilde{\bar{\theta}}_{\mathbb{L}} \leq c''\omega + \frac{c'\omega}{\epsilon}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}})$. Finally, note that the constraints of the problem for $\tilde{\bar{\theta}}_{\mathbb{L}}$ are stronger than those for $\bar{\theta}_{\mathbb{L}}$ (because $\tilde{\mathcal{S}} \subseteq \mathcal{S}$) so $\bar{\theta}_{\mathbb{L}} - \tilde{\bar{\theta}}_{\mathbb{L}} \geq 0$. □

Lemma 2.3. *For each $j \in [J]$ let $\mathbb{M}_j$ be a linear operator from a vector space $\mathcal{S}$ to the space of functions from $\mathcal{V}_j$ to $\mathbb{R}$. Let $\mathbb{L}$ be a linear operator that maps from $\mathcal{S}$ to $\mathbb{R}$. Define:*

$$\bar{\theta}_{\mathbb{L}} = \sup_{s \in \mathcal{S}} \mathbb{L}[s] \text{ s.t. } \mathbb{M}_j[s](v) \leq b_j(v), \forall v \in \mathcal{V}_j, \forall j \in [J]$$

And let $\underline{\theta}_{\mathbb{L}}$ be the infimum subject to the same constraints. For each j let $\tilde{\mathcal{V}}_j$ be a finite subset of $\mathcal{V}_j$ and define:

$$\tilde{\bar{\theta}}_{\mathbb{L}} = \sup_{s \in \mathcal{S}} \mathbb{L}[s] \text{ s.t. } \mathbb{M}_j[s](v) \leq b_j(v), \forall v \in \tilde{\mathcal{V}}_j, \forall j \in [J]$$

Suppose there exists $\tilde{s} \in \mathcal{S}$ and $\epsilon > 0$ so that for each $j \in [J]$ we have $\mathbb{M}_j[\tilde{s}](v) \leq b_j(v) - \epsilon$, $\forall v \in \mathcal{V}_j$. Suppose for each j, $\mathcal{V}_j$ is a subset of Euclidean space, $b_j(\cdot)$ is Lipschitz continuous with constant at most ξ_j, and that if $s \in \mathcal{S}$ satisfies the constraints in the problem for $\tilde{\bar{\theta}}_{\mathbb{L}}$ then $\mathbb{M}_j[s](\cdot)$ is Lipschitz continuous with constant at most ξ_j. Then $|\tilde{\bar{\theta}}_{\mathbb{L}} - \bar{\theta}_{\mathbb{L}}| \leq \frac{2\sum_{j=1}^{J}\xi_j D_j}{\epsilon}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}})$, where $D_j = \sup_{v \in \mathcal{V}_j} \min_{v' \in \tilde{\mathcal{V}}_j} ||v - v'||_2$.

Proof. By definition of the supremum, for any $r > 0$ there must be some $s' \in \mathcal{S}$ that satisfies the constraints in the problem for $\tilde{\bar{\theta}}_{\mathbb{L}}$ and achieves $\tilde{\bar{\theta}}_{\mathbb{L}} - \mathbb{L}[s'] \leq r$. By supposition $\mathbb{M}_j[s']$ is Lipschitz continuous with constant at most ξ_j and likewise for

b_j. Thus $\mathbb{M}_j[s'] - b_j$ is Lipschitz with constant at most $2\xi_j$. This implies:

$$\begin{aligned}
&\sup_{v \in \mathcal{V}_j} \big(\mathbb{M}_j[s'](v) - b_j(v)\big) - \max_{v \in \tilde{\mathcal{V}_j}} \big(\mathbb{M}_j[s'](v) - b_j(v)\big) \\
\leq & 2\xi_j \sup_{v \in \mathcal{V}_j} \min_{v' \in \tilde{\mathcal{V}_j}} ||v - v'||_2 = 2\xi_j D_j \leq 2 \sum_{j=1}^{J} \xi_j D_j
\end{aligned}$$

We know that $\mathbb{M}_j[s'](v) \leq b_j(v)$ for all $v \in \mathcal{V}_j$, so from the above we get that $\mathbb{M}_j[s'](v) \leq b_j(v) + 2\sum_{j=1}^{J} \xi_j D_j$, $\forall v \in \mathcal{V}_j$. Applying Lemma 2.1 we then get $\mathbb{L}[s'] - \bar{\theta}_{\mathbb{L}} \leq \frac{2\sum_{j=1}^{J}\xi_j D_j}{\epsilon}(\bar{\theta}_{\mathbb{L}} - \mathbb{L}[\tilde{s}])$ which implies $\tilde{\bar{\theta}}_{\mathbb{L}} - \bar{\theta}_{\mathbb{L}} \leq r + \frac{2\sum_{j=1}^{J}\xi_j D_j}{\epsilon}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}})$. Since this holds for any r we get $\tilde{\bar{\theta}}_{\mathbb{L}} - \bar{\theta}_{\mathbb{L}} \leq \frac{2\sum_{j=1}^{J}\xi_j D_j}{\epsilon}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}})$. Finally, the constraints in the problem for $\tilde{\bar{\theta}}_{\mathbb{L}}$ are weaker than for $\bar{\theta}_{\mathbb{L}}$, so $\tilde{\bar{\theta}}_{\mathbb{L}} - \bar{\theta}_{\mathbb{L}} \geq 0$. □

Theorem 2.0. *Let $\mathcal{B}_X$ be the space of real valued functions on the support of X equipped with the supremum norm. Define $\bar{\theta}_{\mathbb{L}}$ and $\underline{\theta}_{\mathbb{L}}$ as follows:*

$$\begin{aligned}
&\bar{\theta}_{\mathbb{L}} = \sup_{h \in \mathcal{B}_X} \mathbb{L}[h],\ \underline{\theta}_{\mathbb{L}} = \inf_{h \in \mathcal{B}_X} \mathbb{L}[h] \ \ s.t.\ \ \mathbb{T}[h](x) \leq C(x), \forall x \in \mathcal{X} \\
&and\ b_{1,k}(Z_k) \leq E[Y - h(X)|Z_k] \leq b_{2,k}(Z_k), \forall k \in [K]
\end{aligned}$$

Define estimates $\hat{\bar{\theta}}_{\mathbb{L}}$ and $\hat{\underline{\theta}}_{\mathbb{L}}$ by:

$$\begin{aligned}
&\hat{\bar{\theta}}_{\mathbb{L}} = \max_{\beta \in \mathbb{R}^{K_n}} \mathbb{L}[\Phi_n']\beta,\ \hat{\underline{\theta}}_{\mathbb{L}} = \min_{\beta \in \mathbb{R}^{K_n}} \mathbb{L}[\Phi_n']\beta \\
&s.t.\ \ \mathbb{T}[\Phi_n'](x)\beta \leq C(x), \forall x \in \mathcal{X}_n \\
&and\ b_{1,k}(z) \leq \hat{g}_{n,k}(z) - \hat{\Pi}_{n,k}(z)'\beta \leq b_{2,k}(z), \forall z \in \mathcal{Z}_{n,k}, k \in [K]
\end{aligned}$$

Suppose Assumptions 2.1.i, 2.2, and 2.4 hold, and for each $k \in [K]$ Assumption 2.3 holds with $g_0(Z)$ replaced by $g_{0,k}(Z_k) = E[Y|Z_k]$, $\Pi_n(Z)$ by $\Pi_{n,k}(Z) = E[\Phi_n(X)|Z_k]$, $\hat{g}_n$ by $\hat{g}_{n,k}$, and $\hat{\Pi}_n$ by $\hat{\Pi}_{n,k}$. Suppose for each $k \in [K]$, $b_{1,k}$ and $b_{2,k}$ are Lipschitz continuous with constant at most ξ_n. Finally suppose there exists $\tilde{h} \in \mathcal{B}_X$ that satisfies the constraints of the problem for $\bar{\theta}_{\mathbb{L}}$ with slack ϵ, that is $\mathbb{T}[h](x) \leq C(x) - \epsilon$ for all $x \in \mathcal{X}$ and $b_{1,k}(Z_k) + \epsilon \leq E[Y - h(X)|Z_k] \leq b_{2,k}(Z_k) - \epsilon$ for all $k \in [K]$. Then *uniformly over all linear functionals* $\mathbb{L}$ *with operator norm less than* $c_{\mathbb{L}}$*:*

$$|\bar{\theta}_{\mathbb{L}} - \hat{\bar{\theta}}_{\mathbb{L}}| = O_p(\kappa_n + a_n + C_n \xi_n D_{1,n} + C_n G_n D_{2,n})$$

And likewise for $|\underline{\theta}_{\mathbb{L}} - \underline{\bar{\theta}_{\mathbb{L}}}|$.

Proof. Define $\tilde{a}_n$ by:

$$\tilde{a}_n = \max_k \{|g_k - g_{n,k}|_\infty + \sup_{\beta:\Phi_n'\beta\in\mathcal{H}} |(\hat{\Pi}_{n,k} - \Pi_{n,k})'\beta|_\infty\} \tag{2.6.1}$$

By Assumption 2.3, $\tilde{a}_n = O_p(a_n) = o_p(1)$. We suppose, until stated otherwise, that $\kappa_n + 2\tilde{a}_n \leq \epsilon/2$, by Assumptions 2.3 and 2.4 this holds with probability approaching 1. We also suppose that n is sufficiently large that $C_n \xi_n D_{1,n} \leq 1/2$, this must be true for sufficiently large n by Assumption 2.5.iii.

Define $\bar{\theta}^*_{\mathbb{L}}$, and $\bar{\theta}^\circ_{\mathbb{L}}$ as follows:

$$\bar{\theta}^*_{\mathbb{L}} = \sup_{\beta\in\mathbb{R}^{K_n}} \mathbb{L}[\Phi_n']\beta \text{ s.t. } \mathbb{T}[\Phi_n'](x)\beta \leq C(x), \forall x \in \mathcal{X}$$
$$\text{and } b_{1,k}(Z_k) \leq E[Y - \Phi_{n,k}(X)'\beta|Z_k] \leq b_{2,k}(Z_k), \forall k \in [K]$$

$$\bar{\theta}^\circ_{\mathbb{L}} = \sup_{\beta\in\mathbb{R}^{K_n}} \mathbb{L}[\Phi_n']\beta$$
$$\text{s.t. } \mathbb{T}[\Phi_n'](x)\beta \leq C(x), \forall x \in \mathcal{X}$$
$$\text{and } b_{1,k}(Z_k) \leq \hat{g}_{n,k}(Z_j) - \hat{\Pi}_{n,k}(Z_k)'\beta \leq b_{2,k}(Z_k), \forall k \in [K]$$

And let $\underline{\theta}^*_{\mathbb{L}}$ and $\underline{\theta}^\circ_{\mathbb{L}}$ be the respective infima subject to the same constraints. By the triangle inequality:

$$|\bar{\theta}_{\mathbb{L}} - \tilde{\bar{\theta}}_{\mathbb{L}}| \leq |\bar{\theta}_{\mathbb{L}} - \bar{\theta}^*_{\mathbb{L}}| + |\bar{\theta}^*_{\mathbb{L}} - \bar{\theta}^\circ_{\mathbb{L}}| + |\bar{\theta}^\circ_{\mathbb{L}} - \tilde{\bar{\theta}}_{\mathbb{L}}| \tag{2.6.2}$$

We use Lemma 2.2 to bound $|\bar{\theta}_{\mathbb{L}} - \bar{\theta}^*_{\mathbb{L}}|$. To apply Lemma 2.2 let $\mathcal{S} = \mathcal{B}_X$ equipped with the supremum norm and we take $\tilde{\mathcal{S}}$ to be the functions of the form $\Phi_n'\beta$. For

$j \in \mathcal{J} = [d]$ let $\mathcal{V}_j = \mathcal{X}$, let $\mathbb{M}_j[h](x) = [\mathbb{T}[h](x)]_j$, and $b_j[h](x) = [C(x)]_j$, where $[v]_j$ denotes the j^{th} component of a vector v. Then Assumption 2.4 implies condition i. of Lemma 2.2 with $\omega = \kappa_n$. To establish condition ii., of Lemma 2.2, note that the linear operators of the form $h \mapsto E[h(X)|Z_k = z]$ and $h \mapsto -E[h(X)|Z_k = z]$ have operator norm of unity because $|E[h(X)|Z_k]| \leq \sup_{x\in\mathcal{X}} |h(x)|$. So condition ii., of the lemma holds with $c' = 1$. Conditions iii. of Lemma 2.2 holds be supposition and iv. with $c'' = c_{\mathbb{L}}$. The second statement of Lemma 2.2 gives:

$$|\bar{\theta}_{\mathbb{L}} - \bar{\theta}^*_{\mathbb{L}}| \leq \kappa_n\big(\frac{1}{\epsilon}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}}) + c_{\mathbb{L}}\big) \tag{2.6.3}$$

Moreover, the first statement in Lemma 2.2 tells us there exists $\tilde{\beta}$ so that for each $j \in [J]$, $\mathbb{M}_j[\Phi'_n\tilde{\beta}](v) \leq b_j(v) - \epsilon + \kappa_n$ for all $v \in \mathcal{V}_j$. We will use this when we employ Lemma 2.1 below.

We now use Lemma 2.1 to bound $|\bar{\theta}^*_{\mathbb{L}} - \bar{\theta}^\circ_{\mathbb{L}}|$. By definition of the supremum we can find a β^* and β° that respectively satisfy the constraints of the problems for $\bar{\theta}^*_{\mathbb{L}}$ and $\bar{\theta}^\circ_{\mathbb{L}}$ so that $\bar{\theta}^*_{\mathbb{L}} - \mathbb{L}[\Phi'_n]\beta^* \leq r$ and $\bar{\theta}^\circ_{\mathbb{L}} - \mathbb{L}[\Phi'_n]\beta^\circ \leq r$. Because they satisfy the constraints we have that $\Phi'_n\beta^*, \Phi'_n\beta^\circ \in \mathcal{H}$. Recall $g_k(Z) = E[Y|Z_k]$ and $\Pi_{n,k}(Z_k) = E[\Phi_n(X)|Z_k]$. Since $\hat{\Pi}_{n,k}(Z_k)'\beta^\circ \leq \hat{g}_{n,k}(Z_k) - b_{1,k}(Z_k)$ we get from (2.6.1) that:

$$E[\Phi_{n,k}(X)|Z_k]'\beta^\circ \leq E[Y|Z_k] - b_{1,k}(Z_k) + \tilde{a}_n$$

And similarly:

$$-E[\Phi_{n,k}(X)|Z_k]'\beta^\circ \leq -E[Y|Z_k] + b_{2,k}(Z_k) + \tilde{a}_n$$

We established earlier that there exists $\tilde{\beta}$ with $\mathbb{M}_j[\Phi'_n\tilde{\beta}](v) \leq b_j(v) - \epsilon + \kappa_n$ for each $v \in \mathcal{V}_j$ and $j \in [J]$ (where $\mathbb{M}$ and b_j were defined above). Because $\kappa_n + 2\tilde{a}_n \leq \epsilon/2$ we have $\epsilon > \kappa_n$, so we can apply Lemma 2.1 to get $\mathbb{L}[\Phi_n\beta^\circ] - \bar{\theta}^*_{\mathbb{L}} \leq \frac{\tilde{a}_n}{\epsilon-\kappa_n}(\bar{\theta}^*_{\mathbb{L}} - \underline{\theta}^*_{\mathbb{L}})$ and so $\bar{\theta}^\circ_{\mathbb{L}} - \bar{\theta}^*_{\mathbb{L}} \leq r + \frac{\tilde{a}_n}{\epsilon-\kappa_n}(\bar{\theta}^*_{\mathbb{L}} - \underline{\theta}^*_{\mathbb{L}})$. Because this holds for each $r > 0$:

$$\bar{\theta}^\circ_{\mathbb{L}} - \bar{\theta}^*_{\mathbb{L}} \leq \frac{\tilde{a}_n}{\epsilon - \kappa_n}(\bar{\theta}^*_{\mathbb{L}} - \underline{\theta}^*_{\mathbb{L}}) \tag{2.6.4}$$

Now, again using (2.6.1) we get see that $\tilde{\beta}$ satisfies the constraints of the problem for θ° with slack $\epsilon - \kappa_n - \tilde{a}_n$ for each $v \in \mathcal{V}_j$ and $j \in [J]$. Moreover, since $E[\Phi_k(X)|Z_k]'\beta^* \leq E[Y|Z_k] - b_{1,k}(Z_k)$ we get from (2.6.1) that:

$$\hat{\Pi}_{n,k}(Z_k)'\beta^* \leq \hat{g}_{n,k}(Z_k) - b_{1,k}(Z_k) + \tilde{a}_n$$

$$-\hat{\Pi}_{n,k}(Z_k)'\beta^* \leq -\hat{g}_{n,k}(Z_k) + b_{2,k}(Z_k) + \tilde{a}_n$$

Recall $\kappa_n + 2\tilde{a}_n \leq \epsilon/2$ and so $\epsilon > \kappa_n + \tilde{a}_n$, then by Lemma 2.1 $\mathbb{L}[\Phi_n\beta^*] - \bar{\theta}^{\circ}_{\mathbb{L}} \leq \frac{\tilde{a}_n}{\epsilon-\kappa_n-\tilde{a}_n}(\bar{\theta}^{\circ}_{\mathbb{L}} - \mathbb{L}[\Phi'_n\tilde{\beta}])$. Using $\mathbb{L}[\Phi'_n\tilde{\beta}] \geq \underline{\theta}^*_{\mathbb{L}}$ and $\bar{\theta}^*_{\mathbb{L}} - \mathbb{L}[\Phi'_n]\beta^* \leq r$ we get:

$$\bar{\theta}^*_{\mathbb{L}} - \bar{\theta}^{\circ}_{\mathbb{L}} \leq r + \frac{\tilde{a}_n}{\epsilon - \kappa_n - \tilde{a}_n}(\bar{\theta}^{\circ}_{\mathbb{L}} - \underline{\theta}^*_{\mathbb{L}})$$

Again, this holds for each $r > 0$, and so $\bar{\theta}^*_{\mathbb{L}} - \bar{\theta}^{\circ}_{\mathbb{L}} \leq \frac{\tilde{a}_n}{\epsilon-\kappa_n-\tilde{a}_n}(\bar{\theta}^{\circ}_{\mathbb{L}} - \underline{\theta}^*_{\mathbb{L}})$. Since $\epsilon > \kappa_n + 2\tilde{a}_n$, adding $\frac{\tilde{a}_n}{\epsilon-\kappa_n-\tilde{a}_n}(\bar{\theta}^*_{\mathbb{L}} - \bar{\theta}^{\circ}_{\mathbb{L}})$ to both sides and then dividing both sides by $1 - \frac{a_n}{\epsilon-\kappa_n-\tilde{a}_n}$ we get: $\bar{\theta}^*_{\mathbb{L}} - \bar{\theta}^{\circ}_{\mathbb{L}} \leq \frac{\tilde{a}_n}{\epsilon-\kappa_n-2\tilde{a}_n}(\bar{\theta}^*_{\mathbb{L}} - \underline{\theta}^*_{\mathbb{L}})$. Combining with (2.6.4) and using that $\kappa_n + 2\tilde{a}_n \leq \epsilon/2$ we get:

$$|\bar{\theta}^*_{\mathbb{L}} - \bar{\theta}^{\circ}_{\mathbb{L}}| \leq \frac{2\tilde{a}_n}{\epsilon}(\bar{\theta}^*_{\mathbb{L}} - \underline{\theta}^*_{\mathbb{L}}) \leq \frac{2\tilde{a}_n}{\epsilon}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}}) \tag{2.6.5}$$

Where the final inequality holds because the constraints of the problems for $\bar{\theta}^*_{\mathbb{L}}$ and $\underline{\theta}^*_{\mathbb{L}}$ are stronger than those for $\bar{\theta}_{\mathbb{L}}$ and $\underline{\theta}_{\mathbb{L}}$.

Next, we apply Lemma 2.3 to bound $|\bar{\theta}^{\circ}_{\mathbb{L}} - \hat{\bar{\theta}}_{\mathbb{L}}|$. By Assumption 2.5.i, Φ_n is Lipschitz continuous with constant ξ_n, and so the function $\Phi'_n\beta$ is Lipschitz continuous with constant at most $||\beta||_2\xi_n$. Therefore:

$$|\Phi'_n\beta|_{\infty} - \max_{x\in\mathcal{X}_n}|\Phi_n(x)'\beta| \leq ||\beta||_2\xi_n D_{1,n}$$

It follows that:

$$\max_{x\in\mathcal{X}_n}|\Phi_n(x)'\beta| \leq \bar{c} \implies |\Phi'_n\beta|_{\infty} \leq \bar{c} + ||\beta||_2\xi_n D_{1,n}$$

By the definition of C_n:

$$|\Phi_n'\beta|_\infty \leq \bar{c} + ||\beta||_2 \xi_n D_{1,n} \implies ||\beta||_2 \leq C_n(\bar{c} + ||\beta||_2 \xi_n D_{1,n})$$

So if $C_n \xi_n D_{1,n} < 1$ we must have:

$$\max_{x \in \mathcal{X}_n} |\Phi_n(x)'\beta| \leq \bar{c} \implies ||\beta||_2 \leq \frac{C_n}{1 - C_n \xi_n D_{1,n}} \bar{c} \leq 2C_n \bar{c}$$

Where we used that $C_n \xi_n D_{1,n} \leq 1/2$. By Assumption 2.2, for any β that satisfies the constraints in the problem for $\hat{\bar{\theta}}_{\mathbb{L}}$ we have $\max_{x \in \mathcal{X}_n} |\Phi_n(x)'\beta| \leq \bar{c}$, and so for any such β, $||\hat{\beta}||_2 \leq 2C_n \bar{c}$. By Assumption 2.5.i, each row of $\mathbb{T}[\Phi_n']$ is Lipschitz with constant at most ξ_n and so $\mathbb{T}[\Phi_n']\beta$ is Lipschitz with constant at most $\xi_n ||\beta||$. In all, if β satisfies the constraints for the problem for $\hat{\bar{\theta}}_{\mathbb{L}}$ then $\mathbb{T}[\Phi_n']\beta$ is Lipschitz with constant at most $2\xi_n C_n \bar{c}$. Further, by Assumption 2.5.i C is Lipschitz continuous with constant at most ξ_n. Note that both $2\xi_n C_n \bar{c}$ and ξ are less than $2\xi_n(1 + C_n \bar{c})$

Now, by Assumption 2.5.ii, $b_{1,k}$, $b_{2,k}$, $\hat{g}_{n,k}$, and $\hat{\Pi}_{n,k}$ are Lipschitz with constant at most G_n. So if β satisfies the constraints for the problem for $\hat{\bar{\theta}}_{\mathbb{L}}$ then $\hat{\Pi}_{n,k}'\beta$ and $-\hat{\Pi}_{n,k}'\beta$ are both Lipschitz with constant at most $2G_n C_n \bar{c}$, and both $\hat{g}_{n,k} - b_{1,j}$ and $-\hat{g}_{n,k} + b_{2,j}$ are Lipschitz with constant at most $2G_n$. Note that $2G_n C_n \bar{c}$ and $2G_n$ are less than $2G_n(1 + C_n \bar{c})$.

So applying Lemma 2.3 we get:

$$|\bar{\theta}_{\mathbb{L}}^{\circ} - \hat{\bar{\theta}}_{\mathbb{L}}| \leq \frac{4(1 + C_n \bar{c})(d\xi_n D_{1,n} + K G_n D_{2,n})}{\epsilon - \kappa_n - \tilde{a}_n} (\bar{\theta}_{\mathbb{L}}^{\circ} - \underline{\theta}_{\mathbb{L}}^{\circ})$$

Now, recall that $|\bar{\theta}_{\mathbb{L}}^{*} - \bar{\theta}_{\mathbb{L}}^{\circ}| \leq \frac{\tilde{a}_n}{\epsilon - \kappa_n - 2\tilde{a}_n}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}})$. We can apply the same reasoning to get $|\underline{\theta}_{\mathbb{L}}^{*} - \underline{\theta}_{\mathbb{L}}^{\circ}| \leq \frac{\tilde{a}_n}{\epsilon - \kappa_n - 2\tilde{a}_n}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}})$. To see this note that we can just replace $\mathbb{L}$ with $-\mathbb{L}$ in the problems for $\bar{\theta}_{\mathbb{L}}$, $\bar{\theta}_{\mathbb{L}}^{*}$, and $\bar{\theta}_{\mathbb{L}}^{\circ}$ and then $-\underline{\theta}_{\mathbb{L}}$, $-\underline{\theta}_{\mathbb{L}}^{*}$, and $-\underline{\theta}_{\mathbb{L}}^{\circ}$ are the respective suprema of the new problems. Then by the triangle inequality:

$$\bar{\theta}_{\mathbb{L}}^{\circ} - \underline{\theta}_{\mathbb{L}}^{\circ} \leq \bar{\theta}_{\mathbb{L}}^{*} - \underline{\theta}_{\mathbb{L}}^{*} + \frac{2\tilde{a}_n}{\epsilon - \kappa_n - 2\tilde{a}_n}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}}) \leq \frac{\epsilon - \kappa_n}{\epsilon - \kappa_n - 2\tilde{a}_n}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}})$$

Where the second inequality follows because the problems for $\bar{\theta}_{\mathbb{L}}$ and $\underline{\theta}_{\mathbb{L}}$ have weaker constraints than for $\bar{\theta}^*_{\mathbb{L}}$ and $\underline{\theta}^*_{\mathbb{L}}$ so $\bar{\theta}^*_{\mathbb{L}} - \underline{\theta}^*_{\mathbb{L}} \leq \bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}}$. Combining:

$$
\begin{aligned}
|\bar{\theta}^{\circ}_{\mathbb{L}} - \hat{\bar{\theta}}_{\mathbb{L}}| &\leq \frac{8(\epsilon - \kappa_n)(1 + C_n \bar{c})(d\xi_n D_{1,n} + KG_n D_{2,n})}{(\epsilon - \kappa_n - \tilde{a}_n)(\epsilon - \kappa_n - 2\tilde{a}_n)}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}}) \\
&\leq \frac{32}{\epsilon}(1 + C_n \bar{c})(d\xi_n D_{1,n} + KG_n D_{2,n})(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}})
\end{aligned}
$$

Where the second inequality uses $\kappa_n + 2\tilde{a}_n \leq \epsilon/2$. Combining the above with (2.6.3), (2.6.5), and (2.6.2) we get:

$$
\begin{aligned}
|\bar{\theta}_{\mathbb{L}} - \hat{\bar{\theta}}_{\mathbb{L}}| \leq& \kappa_n\Big(\frac{1}{\epsilon}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}}) + c_{\mathbb{L}}\Big) + \frac{2\tilde{a}_n}{\epsilon}(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}}) \\
&+ \frac{32}{\epsilon}(1 + C_n \bar{c})(d\xi_n D_{1,n} + KG_n D_{2,n})(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}})
\end{aligned}
$$

Note that linearity and boundedness of $\mathbb{L}$ and Assumption 2.2 imply that $(\bar{\theta}_{\mathbb{L}} - \underline{\theta}_{\mathbb{L}}) \leq 2c_{\mathbb{L}}\bar{c}$ so we get:

$$
\begin{aligned}
|\bar{\theta}_{\mathbb{L}} - \hat{\bar{\theta}}_{\mathbb{L}}| \leq& \kappa_n c_{\mathbb{L}}\Big(\frac{2\bar{c}}{\epsilon} + 1\Big) + c_{\mathbb{L}}\frac{4\tilde{a}_n \bar{c}}{\epsilon} \\
&+ c_{\mathbb{L}}\frac{64\bar{c}}{\epsilon}(1 + C_n \bar{c})(d\xi_n D_{1,n} + KG_n D_{2,n})
\end{aligned}
$$

Now, $-\underline{\theta}_{\mathbb{L}}$ and $-\hat{\underline{\theta}}_{\mathbb{L}}$ are the respective suprema of the problems for $\bar{\theta}_{\mathbb{L}}$ and $\hat{\bar{\theta}}_{\mathbb{L}}$ with $\mathbb{L}$ replaced by $-\mathbb{L}$, and so we can repeat the same steps that bound $|\bar{\theta}_{\mathbb{L}} - \hat{\bar{\theta}}_{\mathbb{L}}|$ to get:

$$
\begin{aligned}
|\underline{\theta}_{\mathbb{L}} - \hat{\underline{\theta}}_{\mathbb{L}}| \leq& \kappa_n c_{\mathbb{L}}\Big(\frac{2\bar{c}}{\epsilon} + 1\Big) + c_{\mathbb{L}}\frac{4\tilde{a}_n \bar{c}}{\epsilon} \\
&+ c_{\mathbb{L}}\frac{64\bar{c}}{\epsilon}(1 + C_n \bar{c})(d\xi_n D_{1,n} + KG_n D_{2,n})
\end{aligned}
$$

Finally, recall we derived the above under the assumption that $\kappa_n + 2\tilde{a}_n \leq \epsilon/2$. By Assumptions 2.3 and 2.4 this holds with probability approaching 1. Further, by Assumption 2.3 $\tilde{a}_n = O_p(a_n)$ and by construction $C_n \geq 1$. So in all we get:

$$
|\bar{\theta}_{\mathbb{L}} - \hat{\bar{\theta}}_{\mathbb{L}}| = O_p(\kappa_n + a_n + C_n\xi_n D_{1,n} + C_n G_n D_{2,n})
$$

And likewise for $|\underline{\theta}_{\mathbb{L}} - \underline{\bar{\theta}_{\mathbb{L}}}|$. Since the above is based on an inequality that depends on $\mathbb{L}$ only through $c_{\mathbb{L}}$, the convergence in probability above must is uniform over all $\mathbb{L}$ that have the same constant $c_{\mathbb{L}}$. □

Theorem 2.1. In this case $-b_1(Z) = b_2(Z) = b$ so both are Lipschitz continuous with constant 0. The first statement of the theorem then follows immediately from Theorem 2.0. For the final statement of the theorem note that the evaluation functionals $h \mapsto h(x)$ all have operator norm of unity, because $|h(x)| \leq |h|_\infty$. □

For the proofs below, we refer to a functional γ defined as follows. Let $\hat{Q}_n = \frac{1}{n}\sum_{i=1}^n \Psi_n(Z_i)\Psi_n(Z_i)'$, then for each $h \in \mathcal{H}$ let $\hat{\gamma}[h]$ be the least squares estimator defined by $\hat{\gamma}[h] = \hat{Q}_n^{-1}\frac{1}{n}\sum_{i=1}^n \Psi_n(Z_i)h(X_i)$. If $\hat{Q}_n$ is singular we take $\hat{\gamma}[h]$ to be zero (under Assumption 2.5.i this event happens with probability approaching zero).

Lemma 2.3. *Suppose Assumptions 2.1.ii, 2.2, 2.6 and 2.7 hold. Then:*

$$\sup_{h\in\mathcal{H}} |\Psi_n'\hat{\gamma}[h] - \mathbb{A}[h]|_\infty = O_p\left(\frac{\bar{\xi}_n}{\sqrt{n}}[1 + \sqrt{log(\ell_n)} + \sqrt{L_n}R_n(s)] + R_n(s)\right)$$
$$= o_p(1)$$

Proof. The proof follows some steps of [17] Lemma 4.2 with alterations to achieve uniformity over $\mathcal{H}$. Recall $Q_n = E[\Psi_n(Z_i)\Psi_n(Z_i)']$. By Assumption 2.6.i we can normalize $E[\Psi_n(Z)\Psi_n(Z)'] = I$ and the assumptions still hold with sequences $\bar{\xi}_n$ and ℓ_n (that satisfy Assumption 2.6) changed only by a factor independent of n. We maintain this normalization throughout. Define γ by $\gamma[h] = E[\Psi_n(Z_i)h(X_i)]$, $\epsilon_i[h] = h(X_i) - E[h(X_i)|Z_i]$, and $r_i[h] = E[h(X_i)|Z_i] - \gamma[h]'\Psi_n(Z_i)$. Then:

$$h(X_i) = \Psi_n(Z_i)'\gamma[h] + \epsilon_i[h] + r_i[h]$$

By Assumption 2.2, $\mathcal{H}$ contains functions bounded by $\bar{c}$, so:

$$|\epsilon_i[h]| \leq |h(X_i)| + |E[h(X_i)|Z_i]| \leq 2\bar{c}$$

Note as well that:

$$
\begin{aligned}
|\epsilon_i[h_1] - \epsilon_i[h_2]| &\leq |h_1(X_i) - h_2(X_i)| + |E[h_1(X_i) - h_2(X_i)|Z_i]| \\
&\leq 2|h_1 - h_2|_\infty
\end{aligned} \tag{2.6.6}
$$

To bound $|r_i[h]|$ we first show that if $|h|_\infty < \infty$ then Assumption 2.6.ii implies $\mathbb{A}[h]$ is smooth. For some $m \leq s$ (where s is the smoothness in Assumption 2.6.ii) let $\{q_j\}_{j=1}^{\dim(Z)}$ be a sequence of positive integers with $\sum_{j=1}^{\dim(Z)} q_j = m$. Let D_q then be the partial derivative operator given by $D_q[f](z) = \frac{\partial^m}{\partial^{q_1}\partial^{q_2}\ldots\partial^{q_{\dim(Z)}}} f(z)$ for any sufficiently differentiable function $f : \mathcal{Z} \to \mathbb{R}$. Now, from Assumption 2.6.ii, $D_q[f_{X|Z}(x, \cdot)](z)$ is bounded uniformly over x and z by some constant $\bar{\ell}$. Then it follows by the dominated convergence theorem that:

$$
\begin{aligned}
|D_q\mathbb{A}[h](z)| &= |D_q\Big[\int h(x) f_{X|Z}(x, z) dx\Big](z)| \\
&= |\int h(x) D_q\big[f_{X|Z}(x, \cdot)\big](z) dx| \leq |h|_\infty \bar{\ell}
\end{aligned}
$$

And for $m = s$, note that:

$$
\begin{aligned}
&|D_q\mathbb{A}[h](z_1) - D_q\mathbb{A}[h](z_2)| \\
=&|\int h(x) D_q\big[f_{X|Z}(x, \cdot)\big](z_1) dx - \int h(x) D_q\big[f_{X|Z}(x, \cdot)\big](z_2) dx| \\
\leq&|h|_\infty \sup_{x \in \mathcal{X}} |D_q\big[f_{X|Z}(x, \cdot)\big](z_1) - D_q\big[f_{X|Z}(x, \cdot)\big](z_2)| \\
\leq&|h|_\infty \bar{\ell} ||z_1 - z_2||_2^s
\end{aligned}
$$

Where the last inequality again follows because Assumption 2.6.ii states $f_{X|Z}(x, \cdot)$ is of Hölder smoothness class $s$with constant $\bar{\ell}$ for all x. From the above we see that $\mathbb{A}[h]$ is of Hölder smoothness class $s$with constant at most $|h|_\infty \bar{\ell}$. So we can apply Assumption 2.7.i and get that $|r_i[h]| \leq |h|_\infty \bar{\ell} R_n(s)$. Since $|h|_\infty \leq \bar{c}$ for all $h \in \mathcal{H}$ (from Assumption 2.6.iii), we get for any $h \in \mathcal{H}$, $|r_i[h]| \leq \bar{c}\bar{\ell} R_n(s)$. Further, note that by linearity of r_i:

$$|r_i[h_1] - r_i[h_2]| = |r_i[h_1 - h_2]| \leq |h_1 - h_2|_\infty \bar{\ell} R_n(s) \tag{2.6.7}$$

Using Assumption 2.7.iv we can apply Rudelson's matrix LLN ([70], [17] Lemma 6.2) to get:

$$E||\hat{Q}_n - I||_{op} = O(\sqrt{\frac{\bar{\xi}_n^2 log(L_n)}{n}}) = o(1)$$

Which implies $||\hat{Q}_n^{-\frac{1}{2}}||_{op} = O_p(1)$, $||\hat{Q}_n^{\frac{1}{2}}||_{op} = O_p(1)$ and $||\hat{Q}_n^{-1}||_{op} = O_p(1)$. And also by Rudelson's LLN , $E||\hat{Q}_n^{\frac{1}{2}}||_{op} = O(1)$.

It follows that $\hat{Q}_n$ is non-singular with probability approaching 1 and so below we treat it as non-singular. Recall we define $\alpha_n : \mathcal{Z} \to \mathbb{R}^{L_n}$ by $\alpha_n(z) = \frac{\Psi_n(z)}{||\Psi_n(z)||}$. And note that:

$$\begin{aligned} \sqrt{n}\alpha_n(z)'(\hat{\gamma}[h] - \gamma[h]) = \alpha_n(z)'\hat{Q}_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_n(Z_i)\epsilon_i[h] \\ + \alpha_n(z)'\hat{Q}_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_n(Z_i)r_i[h] \end{aligned} \tag{2.6.8}$$

Let us bound the first term on the RHS above. Note that:

$$E\left[\alpha_n(z)'\hat{Q}_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_n(Z_i)\epsilon_i[h]\middle|Z_1, ..., Z_n\right] = 0$$

Let $(\eta_1, ..., \eta_n)$ be a sample of iid Rademachers, independent of the data. By the symmetrization inequality:

$$\begin{aligned} &E\left[\sup_{z\in\mathcal{Z}, h\in\mathcal{H}} |\alpha_n(z)'\hat{Q}_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_n(Z_i)\epsilon_i[h]|\,\middle|Z_1, ..., Z_n\right] \\ \leq &2E\left[E_\eta\big[\sup_{z\in\mathcal{Z}, h\in\mathcal{H}} |\alpha_n(z)'\hat{Q}_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_n(Z_i)\eta_i\epsilon_i[h]|\big]\middle|Z_1, ..., Z_n\right] \end{aligned} \tag{2.6.9}$$

Where the inner expectation on the RHS above is over the Rademachers with ϵ_i and

Z_i for $i = 1, ..., n$ treated as fixed.

Define the set $\mathcal{T} \subseteq \mathbb{R}^n$ by:

$$\mathcal{T} = [t = (t_1, ..., t_n) \in \mathbb{R}^n : \, t_i = \alpha_n(z)'\hat{Q}_n^{-1}\Psi_n(Z_i)\epsilon_i[h], z \in \mathcal{Z}, h \in \mathcal{H}]$$

Define a norm $||\cdot||_{n,2}$ on $\mathbb{R}^n$ by $||t||_{n,2}^2 = \frac{1}{n}\sum_{i=1}^n t_i^2$. By [36] there exists a universal constant D so that:

$$\begin{aligned} &E_\eta\left[\sup_{z\in\mathcal{Z},h\in\mathcal{H}} |\alpha_n(z)'\hat{Q}_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_n(Z_i)(\eta_i\epsilon_i[h])|\right] \\ \leq &D\int_0^\theta \sqrt{log\mathcal{N}(\mathcal{T}, ||\cdot||_{n,2}, \delta)}d\delta \end{aligned}$$

Where $\mathcal{N}(\mathcal{T}, ||\cdot||_{n,2}, \delta)$ is the smallest number of radius-δ $||\cdot||_{n,2}$ balls needed to cover $\mathcal{T}$ and θ is the smallest upper bound on the $||\cdot||_{n,2}$-distance between any two points in $\mathcal{T}$. Using our bounds on $|\epsilon_i[h]|$:

$$\theta = 2\sup_{t\in\mathcal{T}} ||t||_{n,2} \leq 4\bar{c}||\hat{Q}_n^{-\frac{1}{2}}||_{op}$$

Let $\tilde{A}_i[h] = \hat{Q}_n^{-1}\Psi_n(Z_i)\epsilon_i[h]$. Using the Lipschitz constant for α_n given in Assumption 2.7.ii, and using 2.6.6 we get (for $h_1, h_2 \in \mathcal{H}$):

$$\begin{aligned} &\left(\frac{1}{n}\sum_{i=1}^n |\alpha_n(z_1)'\tilde{A}_i[h_1] - \alpha_n(z_2)'\tilde{A}_i[h_2]|^2\right)^{\frac{1}{2}} \\ \leq &\left((\bar{c}\ell_n)^2||z_1 - z_2||_2^2 + |h_1 - h_2|_\infty^2\right)^{\frac{1}{2}} 4\sqrt{2}||\hat{Q}_n^{-\frac{1}{2}}||_{op} \end{aligned} \tag{2.6.10}$$

Define b_n by $b_n = 4||\hat{Q}_n^{-\frac{1}{2}}||_{op}$. Note that $||\hat{Q}_n^{-\frac{1}{2}}||_{op} = O_p(1)$ and so $b_n = O_p(1)$. Then from 2.6.10:

$$\mathcal{N}(\mathcal{T}, ||\cdot||_{n,2}, \delta) \leq \mathcal{N}(\mathcal{Z}, ||\cdot||_2, \frac{\delta}{\bar{c}\ell_n b_n})\mathcal{N}(\mathcal{H}, |\cdot|_\infty, \frac{\delta}{b_n})$$

And so (using sub-additivity of the square root):

$$\int_0^{\theta} \sqrt{log\mathcal{N}(\mathcal{T}, ||\cdot||_{n,2}, \delta)}d\delta \leq \int_0^{\bar{c}b_n} \sqrt{log\mathcal{N}(\mathcal{Z}, ||\cdot||_2, \frac{\delta}{\bar{c}\ell_n b_n})}d\delta + \int_0^{\bar{c}b_n} \sqrt{log\mathcal{N}(\mathcal{H}, |\cdot|_{\infty}, \frac{\delta}{b_n})}d\delta$$

Making the substitution $u = \frac{\delta}{\bar{c}b_n}$ into the first integral and $u = \frac{\delta}{b_n}$ into the second:

$$\int_0^{\theta} \sqrt{log\mathcal{N}(\mathcal{T}, ||\cdot||_{n,2}, \delta)}d\delta \leq \bar{c}b_n \int_0^{1} \sqrt{log\mathcal{N}(\mathcal{Z}, ||\cdot||_2, \frac{u}{\ell_n})}du + b_n \int_0^{\bar{c}} \sqrt{log\mathcal{N}(\mathcal{H}, |\cdot|_{\infty}, u)}du$$

Because the integrand is decreasing in u:

$$\int_0^{\bar{c}} \sqrt{log\mathcal{N}(\mathcal{H}, |\cdot|_{\infty}, u)}du \leq \max\{\bar{c}, 1\} \int_0^{1} \sqrt{log\mathcal{N}(\mathcal{H}, |\cdot|_{\infty}, u)}du$$

By Assumption 2.7.iii the integral on the RHS above is finite. So let ω_1 denote the finite constant on the RHS above.

By Assumption 2.6.ii $\mathcal{Z}$ is bounded and it has dimension $\dim(Z) < \infty$ therefore for some constant ω_2, $\mathcal{N}(\mathcal{Z}, ||\cdot||_2, \delta) \leq \omega_2 \left(\frac{1}{\delta}\right)^{\dim(Z)}$. So we have:

$$\int_0^{\theta} \sqrt{log\mathcal{N}(\mathcal{T}, ||\cdot||_{n,2}, \delta)}d\delta \leq \bar{c}b_n \int_0^{1} \sqrt{log\omega_2 + \dim(Z)log\left(\frac{\ell_n}{u}\right)}du + b_n\omega_1$$

Using sub-additivity of the square root we get from the above:

$$\begin{aligned}&\int_0^\theta \sqrt{log\mathcal{N}(\mathcal{T}, ||\cdot||_{n,2}, \delta)}d\delta \\ \leq&\bar{c}b_n\sqrt{log\omega_2} + b_n\omega_1 \\ +&\bar{c}b_n\sqrt{\dim(Z)}\Bigg(\sqrt{log(\ell_n)} + \int_0^1 \sqrt{log\bigg(\frac{1}{u}\bigg)}du\Bigg) \\ =&O_p\Bigg(1+\sqrt{log(\ell_n)}\Bigg)\end{aligned}$$

And so from 2.6.9 and Markov's inequality:

$$\sup_{z\in\mathcal{Z},h\in\mathcal{H}} |\alpha_n(z)'\hat{Q}_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_n(Z_i)\epsilon_i[h]| = O_p\Bigg(1+\sqrt{log(\ell_n)}\Bigg)$$

Now we bound the second term on the RHS in 2.6.8. Define $\mathcal{S}^{L_n-1} = [\beta \in \mathbb{R}^{L_n} : ||\beta||_2 \leq 1]$. Note that:

$$\begin{aligned}&|\alpha_n(z)'\hat{Q}_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_n(Z_i)r_i[h]| \\ \leq&||\hat{Q}_n^{-1}||_{op}\sup_{\beta\in\mathcal{S}^{L_n-1}}|\beta'\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_n(Z_i)r_i[h]|\end{aligned}$$

And we already have by the matrix LLN that $||\hat{Q}_n^{-1}||_{op} = O_p(1)$. Note as well that: $E\bigg[\beta'\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_n(Z_i)r_i[h]\bigg] = 0$. Again, let $(\eta_1, ..., \eta_n)$ be a sample of iid Rademachers, independent of the data. By the symmetrization inequality:

$$\begin{aligned}&E\bigg[\sup_{h\in\mathcal{H},\beta\in\mathcal{S}^{L_n-1}}|\beta'\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_n(Z_i)r_i[h]|\bigg] \\ \leq&2E\bigg[E_\eta\big[\sup_{h\in\mathcal{H},\beta\in\mathcal{S}^{L_n-1}}|\beta'\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_n(Z_i)r_i[h]|\big]\bigg]\end{aligned} \tag{2.6.11}$$

Where the inner expectation on the RHS above is over the Rademachers with r_i and Z_i for $i = 1, ..., n$ treated as fixed. Define a new set $\mathcal{T} \subseteq \mathbb{R}^n$ by:

$$\mathcal{T} = [t = (t_1, ..., t_n) \in \mathbb{R}^n : t_i = \beta'\psi_{n,l}(Z_i)r_i[h], h\in\mathcal{H}, \beta\in\mathcal{S}^{L_n-1}]$$

Again, by [36]:

$$
\begin{aligned}
&E_\eta\Big[\sup_{h\in\mathcal{H},\beta\in\mathcal{S}^{L_n-1}}|\beta'\frac{1}{\sqrt{n}}\sum_{i=1}^n\Psi_n(Z_i)(\eta_i r_i[h])|\Big]\\
\leq &D\int_0^\theta\sqrt{log\mathcal{N}(\mathcal{T},||\cdot||_{n,2},\delta)}d\delta
\end{aligned}
$$

Using our bound on $|r_i[h]|$: $\theta = 2\sup_{t\in\mathcal{T}}||t||_{n,2} \leq 2\bar{c}\bar{\ell}R_n(s)||\hat{Q}_n^{\frac{1}{2}}||_{op}$. Using 2.6.7 we get (for $h_1, h_2 \in \mathcal{H}$):

$$
\begin{aligned}
&\bigg(\frac{1}{n}\sum_{i=1}^n|\Psi_n(Z_i)r_i[h_1]-\Psi_n(Z_i)r_i[h_2]|^2\bigg)^{\frac{1}{2}}\\
\leq &\sqrt{2}\bar{\ell}R_n(s)||\hat{Q}_n^{\frac{1}{2}}||_{op}\bigg(\bar{c}^2||\beta_1-\beta_2||_2^2+|h_1-h_2|_\infty^2\bigg)^{\frac{1}{2}}
\end{aligned}
$$

Define c_n by $c_n = 2\bar{\ell}R_n(s)||\hat{Q}_n^{\frac{1}{2}}||_{op}$. Then from 2.6.10:

$$
\mathcal{N}(\mathcal{T},||\cdot||_{n,2},\delta)\leq\mathcal{N}(\mathcal{S}^{L_n-1},||\cdot||_2,\frac{\delta}{\bar{c}c_n})\mathcal{N}(\mathcal{H},|\cdot|_\infty,\frac{\delta}{b_n})
$$

And so (using sub-additivity of the square root) and making substitutions, much as before:

$$
\begin{aligned}
\int_0^\theta\sqrt{log\mathcal{N}(\mathcal{T},||\cdot||_{n,2},\delta)}d\delta \leq{}& \bar{c}c_n\int_0^1\sqrt{log\mathcal{N}(\mathcal{S}^{L_n-1},||\cdot||_2,u)}du\\
&+c_n\int_0^{\bar{c}}\sqrt{log\mathcal{N}(\mathcal{H},|\cdot|_\infty,u)}du
\end{aligned}
$$

We have already shown that that the second integral on the RHS above is bounded by a finite constant ω_1. The covering number of a unit ball in $\mathbb{R}^{L_n}$ satisfies for some universal constant $\omega_3 > 0$:

$$
\mathcal{N}(\mathcal{S}^{L_n-1},||\cdot||_2,\delta)=\omega_3\bigg(\frac{1}{\delta}\bigg)^{L_n}
$$

Substituting this and using sub-additivity of the square root we get:

$$\int_0^\theta \sqrt{log\mathcal{N}(\mathcal{T}, ||\cdot||_{n,2}, \delta)}d\delta \leq \bar{c}c_n\sqrt{log\omega_3} + c_n\omega_1 + \bar{c}c_n\sqrt{L_n}\int_0^1 \sqrt{log\left(\frac{1}{u}\right)}du$$

And so:

$$\begin{aligned} &E\left[\sup_{h\in\mathcal{H},\beta\in\mathcal{S}^{L_n-1}} |\beta'\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_n(Z_i)r_i[h]|\right] \\ \leq &\bar{c}E[c_n]\sqrt{L_n}\int_0^1 \sqrt{log\left(\frac{1}{u}\right)}du + \bar{c}E[c_n]\sqrt{L_n}\int_0^1 \sqrt{log\left(\frac{1}{u}\right)}du \\ = &O\big(R_n(s)\sqrt{L_n}\big) \end{aligned}$$

And so, by Markov's inequality the second term in the RHS of 2.6.8 is $O_p\big(R_n(s)\sqrt{L_n}\big)$. So in all:

$$\begin{aligned} \sup_{h\in\mathcal{H}} |\Psi_n'\hat{\gamma}[h] - \mathbb{A}[h]|_\infty \leq &\frac{\bar{\xi}_n}{\sqrt{n}} \sup_{z\in\mathcal{Z},h\in\mathcal{H}} |\sqrt{n}\alpha_n(z)'(\hat{\gamma}[h]-\gamma[h])| \\ &+ \sup_{h\in\mathcal{H}} |\Psi_n'\gamma[h] - \mathbb{A}[h]|_\infty \\ &= O_p\left(\frac{\bar{\xi}_n}{\sqrt{n}}[1+\sqrt{log(\ell_n)} + \sqrt{L_n}R_n(s)] + R_n(s)\right) \end{aligned}$$

□

Theorem 2.2. From Lemma 2.3:

$$\sup_{h\in\mathcal{H}} |\Psi_n'\hat{\gamma}[h] - \mathbb{A}[h]|_\infty = O_p\left(\frac{\bar{\xi}_n}{\sqrt{n}}[1+\sqrt{log(\ell_n)} + \sqrt{L_n}R_n(s)] + R_n(s)\right)$$

Note that if $\Phi_n'\beta \in \mathcal{H}$ then $\hat{\Pi}_n'\beta = \Psi_n'\hat{\gamma}[\Phi_n'\beta]$ and by definition $\Pi_n'\beta = \mathbb{A}[\Phi_n'\beta]$, and so:

$$\sup_{\beta\in\mathbb{R}^{K_n}:\,\Phi_n'\beta\in\mathcal{H}} |(\hat{\Pi}_n - \Pi_n)'\beta|_\infty \leq \sup_{h\in\mathcal{H}} |\Psi_n'\hat{\gamma}[h] - \mathbb{A}[h]|_\infty$$

Applying the rates in Assumption 2.8.ii:

$$\frac{\bar{\xi}_n}{\sqrt{n}}[1+\sqrt{log(\ell_n)}+\sqrt{L_n}R_n(s)]+R_n = O\left(\frac{\sqrt{L_n}}{\sqrt{n}}\sqrt{log(L_n)}+L_n^{-s_0(s)/\dim(Z)}]\right)$$

Note that Assumption 2.6.i is identical to Assumption A.2 in [17], Assumption 2.7.i implies Assumption A.3 in [17] (with '$\ell_k c_k$' in their notation equal to $R_n(s)$), Assumption 2.8.i implies Assumption A.4 in [17] and Assumption 2.8.i and 2.7.ii imply Assumption A.5. Assumption A.1 in [17], that the data are iid, is assumed in our paper throughout. Therefore we can apply [17] Theorem 4.3 for $\hat{g}_n$, and under our other assumptions the rate simplifies to:

$$|\hat{g}_n - g_0|_\infty = O\left(\frac{\sqrt{L_n}}{\sqrt{n}}\sqrt{log(L_n)}+L_n^{-s_0(s)/\dim(Z)}]\right)$$

Then the triangle inequality gives the result. $\square$

Theorem 2.3. We note two facts about B-spline basis functions on an interval. Firstly, if $K_n \geq 2$ then we can apply a linear transformation to Φ_n so that the first two entries are 1 and x. We will assume without loss of generality that Φ_n has been transformed in this way. Secondly, because the basis functions are at least third order, the vector of functions Φ_n is at least twice continuously differentiable at all but a finite set of points in its domain. For any z at which the second derivatives are defined, let $\frac{\partial^2}{\partial x^2}\Phi_n(z)$ denote the vector of second derivatives of each component of Φ_n at z. We can define this function elsewhere by right-continuity, that is if the second derivatives are not defined for some z then let $\frac{\partial^2}{\partial x^2}\Phi_n(z) = \lim_{z'\downarrow z}\frac{\partial^2}{\partial x^2}\Phi_n(z')$. $\frac{\partial^2}{\partial x^2}\Phi_n$ is an invertible linear transformation of the vector of $(s_0-2)^{th}$-order B-spline basis functions with the same knot points as the original spline basis. It then follows from the approximation properties of splines that there is a sequence $\tilde{\kappa}_n = O(K_n^{-1})$ so that for any Lipschitz continuous function h defined on $[a,b]$ with Lipschitz constant L, there exists $\beta \in \mathbb{R}^{K_n}$ with $|\frac{\partial^2}{\partial x^2}\Phi'_n\beta - h|_\infty \leq L\tilde{\kappa}_n$. See for example [34].

First we show that any $h \in \mathcal{H}$ is Lipschitz continuous with constant at most $C = \frac{2}{(b-a)} + c(b-a)$. Consider some $h \in \mathcal{H}$, since h is twice differentiable its first derivative also exists and so for any $x_1, x_2 \in [a,b]$:

$$h(x_2) - h(x_1) = \int_0^1 (x_2 - x_1)\frac{\partial}{\partial x}h(x_1 + t(x_2 - x_1))dt$$

Because $|\frac{\partial^2}{\partial x^2}h|_\infty \leq c$ the above implies:

$$|h(x_2) - h(x_1)| + c|x_2 - x_1|^2 \geq |(x_2 - x_1)\frac{\partial}{\partial x}h(x_1)|$$

Substituting $x_1 = \frac{1}{2}(a+b)$ and $x_2 = b$ into the above we get:

$$|h(x_2) - h(x_1)| + c\frac{1}{4}(b-a)^2 \geq \frac{1}{2}(b-a)|\frac{\partial}{\partial x}h(x_1)|$$

Since $h(x) \in [0,1]$, $|h(x_1) - h(x_2)| \leq 1$ and so we get:

$$|\frac{\partial}{\partial x}h(x_1)| \leq \frac{2}{(b-a)} + c\frac{1}{2}(b-a)$$

And again, since $|\frac{\partial^2}{\partial x^2}h|_\infty \leq c$ we then have that for any $x \in [a,b]$:

$$|\frac{\partial}{\partial x}h(x)| \leq \frac{2}{(b-a)} + c(b-a)$$

Denote the RHS by C. We thus have that any $h \in \mathcal{H}$ is Lipschitz continuous with constant at most C.

Let the functional P extend a function $h \in \mathcal{H}$ to a function defined on $\mathbb{R}$ as follows: $P[h](x) = h(x)$ if $x \in [a,b]$, $P[h](x) = h(a)$ if $x < a$, $P[h](x) = h(b)$ if $x > b$. For $r \in [0, b-a]$ define the linear operator $M_r : \mathcal{B}_X \to \mathcal{B}_X$ by (for all $x \in [a,b]$) $M_r[h](x) = \frac{\int_{x-r}^{x+r} P[h](y)dy}{2r}$. If $|h|_\infty < \infty$ then it is easy to see that $|M_r[h]|_\infty \leq |h|_\infty$. With a substitution we get $M_r[h](x) = \frac{\int_{-r}^{r} P[h](y+x)dy}{2r}$. For any $h \in \mathcal{H}$, $\frac{\partial^2}{\partial x^2}h$ is uniformly bounded by c. So we can use the dominated convergence theorem to get that for any $h \in \mathcal{H}$:

$$\frac{\partial^2}{\partial x^2}M_r[h](x) = \frac{1}{2r}\int_{-r}^{r}\frac{\partial^2}{\partial x^2}P[h](y+x)dy = \frac{1}{2r}\int_{\max\{a,x-r\}}^{\min\{b,x+r\}}\frac{\partial^2}{\partial y^2}h(y)dy$$

Where we have used that the second derivative of $P[h]$ is zero outside of $[a, b]$. So for any $h \in \mathcal{H}$, $|\frac{\partial^2}{\partial x^2} M_r[h](x)| \leq c$.

Assuming without loss of generality that $x_1 \geq x_2$, we see for any $h \in \mathcal{H}$:

$$
\begin{aligned}
&|\frac{\partial^2}{\partial x^2} M_r[h](x_1) - \frac{\partial^2}{\partial x^2} M_r[h](x_2)| \\
=&\frac{1}{2r}\left| \int_{\max\{x_1 - r, x_2 + r\}}^{x_1 + r} \frac{\partial^2}{\partial y^2} P[h](y)dy - \int_{x_2 - r}^{\min\{x_2 + r, x_1 - r\}} \frac{\partial^2}{\partial y^2} P[h](y)dy \right| \\
\leq&\frac{c}{r}|x_1 - x_2|
\end{aligned}
$$

Where the last inequality follows because $\frac{\partial^2}{\partial x^2} P[h]$ is uniformly bounded by c for $h \in \mathcal{H}$. So we have established that for any $h \in \mathcal{H}$ the function $M_r[h]$ has second derivatives that are Lipschitz continuous with Lipschitz constant at most $\frac{c}{r}$.

And now note that, because $h \in \mathcal{H}$ is Lipschitz continuous with constant C for any $x \in [a, b]$, $|M_r[h](x) - h(x)| \leq rC$.

Now, recall the properties of B-splines discussed at the beginning of this proof. Because $\frac{\partial^2}{\partial x^2} M_r[h]$ is Lipschitz continuous with constant $\frac{c}{r}$, there is some $\beta \in \mathbb{R}^{K_n}$ with $|\frac{\partial^2}{\partial x^2} \Phi'_n \beta - \frac{\partial^2}{\partial x^2} M_r[h]|_\infty \leq \frac{c}{r}\tilde{\kappa}_n$. In which case $|\frac{\partial^2}{\partial x^2} \Phi'_n \beta|_\infty \leq c(1 + \frac{\tilde{\kappa}_n}{r})$. And so letting $\tilde{\beta} = (1 + \frac{\tilde{\kappa}_n}{r})^{-1}\beta$ we get $|\frac{\partial^2}{\partial x^2} \Phi'_n \tilde{\beta}|_\infty \leq c$, and from the triangle inequality:

$$
|\frac{\partial^2}{\partial x^2} \Phi'_n \tilde{\beta} - \frac{\partial^2}{\partial x^2} M_r[h]|_\infty \leq \frac{c}{r}\tilde{\kappa}_n + \left(\frac{\tilde{\kappa}_n / r}{1 + (\tilde{\kappa}_n / r)}\right)|\frac{\partial^2}{\partial x^2} \Phi'_n \beta|_\infty
$$

Which is less than $2\frac{c}{r}\tilde{\kappa}_n$, and so:

$$
\begin{aligned}
|\frac{\partial}{\partial x} \Phi'_n \tilde{\beta} - \frac{\partial}{\partial x} M_r[h]|_\infty \leq &\left| \frac{\partial}{\partial x} \Phi_n(\frac{a + b}{2})' \tilde{\beta} - \frac{\partial}{\partial x} M_r[h](\frac{a + b}{2}) \right| \\
&+ (b - a)\frac{c}{r}\tilde{\kappa}_n
\end{aligned}
$$

Where again $\frac{\partial}{\partial x} \Phi_n$ is defined at points at which the derivative is undefined by right continuity. The inequality above then implies:

$$\begin{aligned}|\Phi_n'\tilde{\beta} - M_r[h]|_\infty \leq &\, |\Phi_n(\frac{a+b}{2})'\tilde{\beta} - M_r[h](\frac{a+b}{2})| \\ &+ \frac{b-a}{2}\left|\frac{\partial}{\partial x}\Phi_n(\frac{a+b}{2})'\tilde{\beta} - \frac{\partial}{\partial x}M_r[h](\frac{a+b}{2})\right| \\ &+ \left(\frac{b-a}{2}\right)^2 2\frac{c}{r}\tilde{\kappa}_n\end{aligned} \tag{2.6.12}$$

But the first two entries of $\Phi_n(x)$ are 1 and x, so let β^* be identical to $\tilde{\beta}$ aside from its first two entries β_1^* and β_2^* which are given by:

$$\beta_2^* = \frac{\partial}{\partial x}M_r[h](\frac{a+b}{2}) - \frac{\partial}{\partial x}\Phi_n(\frac{a+b}{2})'\tilde{\beta}$$

And:

$$\beta_1^* = M_r[h](\frac{a+b}{2}) - \Phi_n(\frac{a+b}{2})'\tilde{\beta} - (\frac{a+b}{2})\beta_2^*$$

Then $\frac{\partial^2}{\partial x^2}\Phi_n'\beta^* = \frac{\partial^2}{\partial x^2}\Phi_n'\tilde{\beta}$, and so $|\frac{\partial^2}{\partial x^2}\Phi_n'\beta^*|_\infty \leq c$. Moreover, repeating the same steps used to get (2.6.12) above we see:

$$|\Phi_n'\beta^* - M_r[h]|_\infty \leq \left(\frac{b-a}{2}\right)^2 2\frac{c}{r}\tilde{\kappa}_n$$

We already showed $|M_r[h](x) - h(x)| \leq rC$ and so

$$|\Phi_n'\beta^* - h|_\infty \leq \left(\frac{b-a}{2}\right)^2 2\frac{c}{r}\tilde{\kappa}_n + rC$$

And so setting $r = \sqrt{\tilde{\kappa}_n}$:

$$|\Phi_n'\beta^* - h|_\infty \leq \left[\left(\frac{b-a}{2}\right)^2 2c + C\right]\sqrt{\tilde{\kappa}_n}$$

Since we found such a β^* for any h the result follows. □

Chapter 3

Approximation-Robust Inference in Dynamic Discrete Choice

Models of dynamic decision making have been applied to a wide range of important problems in empirical economics (see [56] for a survey). In these models agents face a dynamic decision problem that can often be solved only numerically and at substantial computational cost. Estimation of these models generally requires that the empirical practitioner solve the agents' dynamic programming problem for each candidate set of parameter values as the likelihood is maximized or GMM objective minimized.[1] For estimation to be computationally feasible, the researcher may need to use approximate dynamic programming methods. The most common approach in empirical economics is approximation of the 'value function' or of related functions like the expected value function or the choice-specific expected value function. The function is iteratively interpolated between a small subset of states using say, polynomials, splines, or kernel regression.

Approximation of the solution to the dynamic programming problem has a long history both in operations research and in econometrics. [16] suggests a method based on polynomial approximation of the expected value function. Approximate dynamic programming methods were introduced to empirical economics by [57] who

[1]Exceptions include the constrained optimization approach of [75] and the conditional choice probability approaches of [48] and [3].

developed an interpolation method designed for use in finite-horizon dynamic discrete choice models. More recent work includes that of [65] who approximates the expected value function using artificial neural networks in the dynamic discrete choice setting.

In many cases value function approximation occurs implicitly. When the state space is continuous some kind of interpolation of the value function is unavoidable, and in many cases models with a discrete state-space implicitly approximate a true model with a continuous state-space.

Unfortunately, the use of approximation leads to inconsistent estimates, and confidence intervals that do not account for the approximation error are not asymptotically valid. One employs approximation methods hoping that the resulting bias is not of too great a magnitude. Without explicit bounds on the approximation error it is unclear how close the approximated value function is to the true value function and it is also unclear how sensitive the estimates are to the approximation error.

Monte Carlo simulation can be used to assess the biases due to approximation (see [57]) but simulation studies are inherently limited. Firstly, the findings of simulation studies apply only to the specific model that is simulated. Secondly, one must be able to exactly solve the dynamic programming problem in order to simulate the data, which means that simulation studies can only be carried out on relatively simple models without a very large state space.

Ideally estimation and inference would explicitly account for the use of approximation. We show how this can be achieved. Using the contraction mapping property of the Bellman operator as well as some other features of the problem, we derive bounds on the differences between choices of the choice-specific expected value function. These bounds can be calculated at relatively little computational cost. Using these bounds one can attain a set estimate that must contain the infeasible estimate that could be evaluated if the researcher were able to fully solve the dynamic programming problem. The bounds can be used to achieve confidence sets that are valid regardless of the closeness of the approximation.

Bounds on the approximation error in approximate dynamic programming are considered in for example [10], but these bounds are generally on the value function, expected value function, or choice-specific expected value function itself. The value function enters the likelihood only through the differences between the the expected values at different choices. We directly bound the latter and then, using these bounds, derive error bounds on the likelihood.

We develop set estimates and confidence sets that are asymptotically valid regardless of the distance between the exact value function. Our methodology is valid regardless of the method used to approximate the value function.

In short, we provide a methodology that allows researchers to conduct asymptotically valid inference and consistent set estimation, regardless of the accuracy of their value function approximation. Of course, if the approximation is poor then the resulting set estimates and confidence sets will be large, but if the approximation is close they will be small.

Our methodology allows a researcher to directly assess the closeness of their approximation, and balance the trade-off between ease of computation and accuracy of estimation. If the researcher observes that the set estimates and robust confidence intervals are very large compared to standard non-robust confidence sets, then they can decrease the coarseness of their discretization of the state space, increase the order of the approximating polynomial, etc. The tools presented in this paper also apply to novel methods of approximate dynamic programming like Q-learning. Economic researchers might be reluctant to try less well-established methods because they fear that the strength of the approximation may be poor. Our methodology can allay these fears because it allows researchers to explicitly assess the biases due to approximation error and perform estimation and inference that is robust to this type of error.

The paper is organized as follows: Section 1 describes the general model in which the results apply and provides background to the problem. Section 2 provides the derivation of bounds on the value function differences. Section 3 explains how the bounds can be used for estimation and inference that is robust to approximation. Section 4 provides Monte Carlo evidence of the efficacy of the method, and Section 5 concludes.

3.1 Background

3.1.1 Modeling Dynamic Discrete Choices

Consider an agent who at a each time t is at some state $S_t \in \mathcal{S}$ and makes choice $D_t \in \mathcal{D}$. The 'choice set' $\mathcal{D}$ is assumed to be discrete, but we make no such assumption about the 'state space' $\mathcal{S}$. The state evolves according to a first-order Markov process in which the probability distribution of next period's state may depend on both the current state S_t and choice D_t, and is otherwise unrelated to the history of states and choices. The static utility of the agent in each period is given by $u(S_t, D_t) + \epsilon_t(D_t)$ where ϵ_t is an additive iid random function from $\mathcal{D}$ to $\mathbb{R}$. We refer to ϵ_t as 'shocks' to the utility because they are not known to the agent prior to period t. We assume throughout that $E[\max_{d\in\mathcal{D}} \epsilon_t(d)] < \infty$. We assume the deterministic part of the utility u is bounded, i.e., there exists a scalar $b < \infty$ so that $|u(s,d)| \leq b$ for all $s \in \mathcal{S}$ and $d \in \mathcal{D}$.

The agent sets $D_t = d(S_t, \epsilon_t)$, where $d(\cdot,\cdot)$ is a static decision rule that maximizes the expected discounted (by discount factor β) sum of utilities:

$$d \in \operatorname*{arg\,sup}_{d: S\times E\to D} E\bigg[\sum_{t=1}^{\infty} \beta^t u\big(S_t, d(S_t,\epsilon_t)\big) + \epsilon_t\big(d(S_t,\epsilon_t)\big)\bigg]$$

There are potentially two sources of uncertainty in the agent's problem. Firstly, ϵ_t is a random function. Secondly, the transition between states may be stochastic. We denote by $F_{s,d}$ the probability distribution of the random variable S_{t+1} for an agent who at time t is at state $S_t = s$ and chooses $D_t = d$. $F_{s,d}$ is referred to as the 'transition distribution' at (s,d) and we refer to F, which maps from (s,d) to $F_{s,d}$ as the 'transition rule'. That the state in the next period depends on the decision in the current period is what makes the agent's problem 'dynamic': the decision today effects not only today's utility, but the utility tomorrow because it partially determines tomorrow's state.

If the dynamic optimization problem is sufficiently regular, then the decision rule

$d(s, \epsilon)$ is the unique solution to the following static optimization problem:

$$d(s, \epsilon) = \arg\max_{d \in \mathcal{D}} [u(s, d) + \epsilon(d) + \beta V(s, d)]$$

Where $V : \mathcal{S} \times \mathcal{D} \to \mathbb{R}$ is the unique fixed point of the non-linear operator $T : \mathbb{R}^{(\mathcal{S} \times \mathcal{D})} \to \mathbb{R}^{(\mathcal{S} \times \mathcal{D})}$ which is defined below:

$$T[V](s, d) = E\big[E[\max_{d'} u(S_{t+1}, d) + \beta V(S_{t+1}, d) + \epsilon_{t+1}(d) | S_{t+1}]\big| S_t = s, D_t = d\big] \tag{3.1.1}$$

And so V solves the non-linear operator equation:

$$V = T[V] \tag{3.1.2}$$

This characterization of the optimal decision rule is due to the work of Bellman (see [15]), and is therefore known as 'Bellman's principle of optimality'. We refer to V as the 'choice-specific expected value function' or just as the 'value function' for short. Note that in the literature the 'value function' usually refers to instead to the quantity $\max_d u(s, d) + \epsilon(d) + \beta V(s, d)$ which is a function of both s and ϵ. $\beta V(s, d)$ is the expected discounted future returns from time $t + 1$ onwards for an agent who, at time t, is at state s and makes choice d, and acts optimally thereafter. We refer to (3.1.2) as the 'Bellman equation' and T as the 'Bellman operator' although these terms are often used slightly differently (see [56]).

3.1.2 Point Estimation

To create a statistical model within the dynamic discrete choice framework let us introduce a vector of parameters of interest θ. θ may parameterize the utility function, the distribution of ϵ_t, and the state transition rule F. For simplicity we focus on the case in which θ parameterizes the utility and distribution of ϵ_t. Both the utility function and distribution of ϵ_t are treated as known up to the parameters θ. To emphasize the dependence of the utility function on θ we let $u(s, d; \theta)$ be the deterministic part of the utility at state s and from choice d under the parameters θ.

Note that the Bellman operator (3.1.1) depends on the utility function, thus the solution V to the Bellman equation (3.1.2) depends on the parameters θ and discount factor β. In addition, the Bellman operator depends on the transition rule F and therefore so does the solution V. To make clear this dependence we write $V_{(\theta,\beta,F)}(s, d)$ to denote the choice-specific expected value function at choice d and state s when the utility parameters are equal to θ, the discount factor equals β, and the transition rule is F.

It will be convenient to define the differences in shocks, utilities and choice-specific values between pairs of choices d and d':

$$\begin{aligned}
\Delta_{d,d'}\epsilon_t &= \epsilon_t(d) - \epsilon_t(d') \\
\Delta_{d,d'}u(s;\theta) &= u(s, d; \theta) - u(s, d'; \theta) \\
\Delta V_{(\theta,\beta,F)}(s, d, d') &= V_{(\theta,\beta,F)}(s, d) - V_{(\theta,\beta,F)}(s, d')
\end{aligned}$$

Given the model in the previous section, the probability that an agent at state $S_t = s$ makes decision $D_t = d$ is given by:

$$L_{(\theta,\beta;F)}(s, d) = P_\theta\Big[\Delta_{d,d'}\epsilon_t \geq \Delta_{d',d}u(s;\theta) + \beta\Delta V_{(\theta,\beta,F)}(s, d', d), \forall d' \in \mathcal{D}\Big] \qquad (3.1.3)$$

Where P_θ is the probability given the distribution of ϵ_t which possibly depends on parameters θ. If the transition rule F were known then $L_{(\theta,\beta;F)}(S_t, D_t)$ is the likelihood of the observation (S_t, D_t) at time t. If F is unknown then it is a nuisance parameter. F can be estimated directly from the data in a first stage either parametrically or nonparametrically. Let $\hat{F}$ be some estimate of F that is evaluated in a first stage. With F replaced by the estimate $\hat{F}$, the partial log-likelihood of the data $\{S_t, D_t\}_{t=1}^T$

evaluated at parameters θ and β is then given by:

$$\mathcal{L}_{\{S_t,D_t\}_{t=1}^T}(\theta,\beta;\hat{F}) = \sum_{t=1}^{T} \ln L_{(\theta,\beta;\hat{F})}(S_t,D_t)$$

An estimate of the parameters θ and β is then given by:

$$(\hat{\theta},\hat{\beta}) = \arg\max_{\theta,\beta} \mathcal{L}_{\{S_t,D_t\}_{t=1}^\infty}(\theta;\beta,\hat{F})$$

Identification of the discount factor can be difficult (see [71], [60], [1]), and so the discount factor may be calibrated in some applications. In this case one can simply plug-in the assumed value for β and maximize the log-likelihood only over the argument θ. The maximization problem above seldom admits a closed-form solution, and so numerical optimization methods are required. Thus, for a number of candidate values of θ and possibly β one must evaluate $\mathcal{L}_{\{S_t,D_t\}_{t=1}^T}(\theta,\beta;\hat{F})$ and/or its derivatives. In order to calculate $\mathcal{L}_{\{S_t,D_t\}_{t=1}^T}(\theta,\beta;F)$ one must evaluate the value function differences $\Delta V_{(\theta,\beta,F)}(s,d',d)$ at all choices in $\mathcal{D}$ and at all states observed in the data. This procedure is known as the 'nested fixed point' or NFXP algorithm because nested within each step of a numerical maximum likelihood procedure one applies an algorithm to find a fixed point of the Bellman operator ([80], [71]).

The standard method for evaluating V at all realized states up to close approximation is 'Bellman iteration'. To apply the method one begins with some guess of the solution $V_{(\theta,\beta,F)}(\cdot,\cdot)$ and then iteratively applies the Bellman operator until there is little change between iterations. Let us be more precise. Let $T_{(\theta,\beta,F)}$ be the Bellman operator (defined in (3.1.1)) when the parameters are set to values θ, β and F. For any $n \in \mathbb{N}$ and $\tilde{V}$, let $T^n_{(\theta,\beta,F)}[\tilde{V}] = \left[T^{n-1}_{(\theta,\beta,F)}[\tilde{V}]\right]$ and let $T^1_{(\theta,\beta,F)}[\tilde{V}] = T_{(\theta,\beta,F)}[\tilde{V}]$. Then for some initial guess $\tilde{V}_0$ one approximates $V_{(\theta,\beta,F)}(\cdot,\cdot)$ by $T^n_{(\theta,\beta,F)}[\tilde{V}_0](s,d)$ where n is large enough that $T^n_{(\theta,\beta,F)}[\tilde{V}_0]$ and $T^{n-1}_{(\theta,\beta,F)}[\tilde{V}_0]$ sufficiently close in some sense. Because $T_{(\theta,\beta,F)}$ is a contraction mapping, one can show that for any bounded function $\tilde{V}_0$, $T^n_{(\theta,\beta,F)}[\tilde{V}_0]$ converges in the supremum norm to the unique fixed point of $T_{(\theta,\beta,F)}$ as $n \to \infty$. [2]

[2] In practice it may be more computationally efficient to combine this procedure with a faster but

Unfortunately, each iteration of the fixed point procedure above requires that one evaluate the right hand side of (3.1.1) at every state s and choice d, not just those states and decisions observed in the data.[3] If the state space is discrete but large then this can be computationally burdensome and if the state space is not discrete and finite it is generally impossible. This same problem applies for many alternative methods of finding a fixed point to the Bellman equation like Newton-Kantorovich.

An alternative to the NXFP algorithm, Mathematical Program with Equilibrium Constraints (MPEC) avoids the need for Bellman iteration ([75]). MPEC treats the estimation problem as a constrained optimization problem in which both θ and (possibly) β and the value function itself are parameters and the the Bellman equation (3.1.2) is treated as a constraint. Thus in MPEC the number of scalar constraints is proportional to the cardinality of the state space. As such, when the state space is large MPEC may still be computationally burdensome.

In many cases, the RHS of (3.1.1) cannot be evaluated exactly even at a single state and choice. The expectation over the future states on the RHS of (3.1.1) generally does not have an analytical form and so must be evaluated by numerical methods. The inner expectation on the RHS of (3.1.1) may also be analytically intractable. For instance, this is the case when ϵ_t (understood as a vector with each entry corresponding to $\epsilon_t(d)$ for some $d \in \mathcal{D}$) is multivariate normal with a non-diagonal covariance matrix. This further compounds the computational burden of applying the Bellman operator at a large number of states.

To avoid the need to evaluate the RHS of the Bellman equation at a large and possibly infinite number of states, some kind of approximation is needed. For example, one may apply the Bellman operator at only some subset of states and then for other states approximate the value function as needed by interpolation. Let us denote by $\tilde{T}_{(\theta,\beta,F)}$ an approximate Bellman operator so that $\tilde{T}_{(\theta,\beta,F)}[\tilde{V}](s,d) = T_{(\theta,\beta,F)}[\tilde{V}](s,d)$ for all s in some small, finite subset $\tilde{\mathcal{S}} \subset \mathcal{S}$ and all choices $d \in \mathcal{D}$ and for all other

less stable method like Newton-Kontorovich ([71]).

[3] Strictly speaking one can restrict the state-space to only those states for which there exists a decision rule $d(\cdot,\cdot)$ under which the state is reached with positive probability starting at the states observed in the data.

$s \in \mathcal{S}$ and $d \in \mathcal{D}$ takes some value given by interpolation. We refer to $\tilde{\mathcal{S}}$ as a 'discrete grid'. It is important to note that in general, $\tilde{T}$ is not a contraction. The term 'interpolation' is used here as a catch-all for a range of different approaches that include fitting a parametric function by least squares, kernel methods and modern machine-learning regression techniques like random forests.

If one replaces $\Delta_{(\theta,\beta,F)}V(s; d', d)$ in the expression for the likelihood (3.1.3) with an approximation, then one evaluates the log-likelihood with error. Thus the resulting maximum likelihood estimates will generally be inconsistent. When applying approximation methods one hopes that the approximated value function differences are close to the exact differences $\Delta_{(\theta,\beta,F)}V(s, d', d)$, that the approximate log-likelihood is then close to the exact log-likelihood, and that the estimates are then close to the exact maximum likelihood estimates and the degree of inconsistency is therefore small. The need to bound the estimation error that results from value function approximation motivates the method described in the next section.

Finally, we note that a class of methods referred to as Conditional Choice Probability (CCP) methods side-step the need for dynamic programming entirely. This approach was introduced by [48], and the literature contains a number of different methods based on their approach, for example [3] and [14]. These methods are based on the observation that under the structure of a given dynamic discrete choice model, the value function is often equal to an analytic (or easy to numerically compute) function of the choice probabilities of the agent at each state. Thus in a first stage one nonparametrically estimates from the data the probability of each discrete choice at each stage and plugs this in to get an estimate of the value function. However, these methods have certain drawbacks. The first-stage nonparametric choice probability estimation contributes to the variance and bias of the parameter estimates, and so the parameter estimates are generally inefficient compared to estimates based on dynamic programming.[4] This is particularly problematic because the choice probabilities may be very imprecisely estimates at parts of the state space that are visited with very low probability (for further discussion see [3]).

[4]The method of [2], which lies somewhere between CCP and NFXP is efficient up to first order.

Our method can be used regardless of the manner in which the value function is approximated or estimated and can therefore complement CCP methods. Firstly, our method allows for valid inference and consistent set estimation even when the conditional choice probabilities cannot be consistently estimated. Secondly, our inference method does not need to account for any first stage nonparametric estimation of the conditional choice probabilities and as such confidence sets based on our method could be smaller than those based on CCP point estimates.

3.2 Bounding the Approximation Error

In the previous section we discussed the need for approximation methods in the evaluation of the value function differences $\Delta_{(\theta,\beta,F)}V$ at the states observed in the data. In this section we describe how one can bound the difference between the exact value function differences $\Delta_{(\theta,\beta,F)}V$ and the differences from some approximate value function $\tilde{V}$. In Section 3 we show how these bounds can be used to achieve set estimation and inference that is valid regardless of the approximation error.

For ease of notation let us again suppress the arguments θ, β, and F in the value function V and Bellman operator T. Theorem 1 below provides an upper bound on the distance between the exact value function differences $V(s,d) - V(s,d')$ and the differences $\tilde{V}(s,d) - \tilde{V}(s,d')$ for some approximate value function $\tilde{V}$. The theorem refers to a function $b(s,d,d')$ defined as follows:

$$b(s,d,d') = \frac{1 - \beta[\sup_{s,s' \in S, d,d' \in D} \delta(F_{s,d}, F_{s',d'})] + \beta\delta(F_{s,d}, F_{s,d'})}{1 - \beta[\sup_{s,s' \in S, d,d' \in D} \delta(F_{s,d}, F_{s',d'})]}$$

Where $\delta(F_{s,d}, F_{s',d'})$ is the total variation distance between $F_{s,d}$ and $F_{s',d'}$.

Theorem 1. *Suppose u is bounded and $E[\max_{d \in \mathcal{D}} \epsilon_t(d)] < \infty$. Let V solve the Bellman equation (3.1.2) for transition rule F and discount factor $\beta < 1$. For any*

$d, d' \in \mathcal{D}$ and $s \in \mathcal{S}$ and bounded function $\tilde{V}$ on $(\mathcal{S} \times \mathcal{D})$:

$$\begin{aligned}
& \left|\left(\tilde{V}(s,d) - \tilde{V}(s,d')\right) - \Delta V(s,d,d')\right| \\
\leq \; & b(s,d,d') \sup_{s,s' \in S, d, d' \in D} |[T[\tilde{V}](s',d') - T[\tilde{V}](s,d)] - [\tilde{V}(s',d') - \tilde{V}(s,d)]]|
\end{aligned}$$

The upper bound given in Theorem 1 does not involve the exact value function V. Instead it involves only the discount factor β, the transition law F, and the approximate value function $\tilde{V}$. The upper bound is small when the discount factor β is small, when the transition distribution $F_{s,d}$ does not change much between states and choices, and when $\tilde{V}$ comes close to satisfying the Bellman equation. Note that in the case where the approximation is exact, i.e. $\tilde{V} = V$ the upper bound is equal to zero.

The total variation distance $\delta(F_{s,d}, F_{s',d'})$, could be computed analytically or numerically depending on the transition law F or first-stage estimate thereof. Note that the total variation distance must be bounded above by 1 and so:

$$b(s,d,d') \leq \frac{1}{1-\beta}$$

The inequality above allows us to avoid calculating the total variation distance but results in a looser bound than that given in Theorem 1. In order to evaluate the upper-bound in Theorem 1 one must also evaluate the supremum over states s and s' and choices d and d' of the following quantity:

$$|T\tilde{V}(s,d) - T\tilde{V}(s',d') - [\tilde{V}(s,d) - \tilde{V}(s',d')]|$$

It may be feasible to directly evaluate the supremum of the above, however if $\mathcal{S}$ is large then evaluating the supremum is still computationally expensive. If $\mathcal{S}$ is continuous one could discretize the state space before finding the supremum, this of course will still result in approximation error, but this error may be acceptable because it is not compounded by repeated Bellman iteration. For the case where the supremum cannot be effectively computed we describe an algorithm that upper bounds the supremum and that can be run until the exact supremum and upper bound are arbitrarily close.

Define $\bar{b}_{\tilde{V}}$ by:

$$\bar{b}_{\tilde{V}} = \sup_{s,s' \in S, d \in D} |u(s,d) + \beta\tilde{V}(s,d) - u(s',d) - \beta\tilde{V}(s',d)|$$

Note that $\bar{b}_{\tilde{V}}$ can be evaluated without applying the Bellman operator. Further, because $\bar{b}_{\tilde{V}}$ only involves u and $\tilde{V}$ which are generally parametric functions, it may be possible to evaluate (or at least upper bound) $\bar{b}_{\tilde{V}}$ analytically.

Theorem 2 below motivates our algorithm.

Theorem 2. *Let $\mathcal{A}$ be a finite subset of $\mathcal{S} \times \mathcal{D}$, then:*

$$\begin{aligned} & \left|\left(\tilde{V}(s,d) - \tilde{V}(s,d')\right) - \Delta V(s,d,d')\right| \\ \leq \; & b(s,d,d') \sup_{s,s' \in S, d,d' \in D} \Bigg[\min_{(s^*,d^*) \in \mathcal{A}} \{T[\tilde{V}](s^*,d^*) + \delta(F_{s,d}, F_{s^*,d^*})\bar{b}_{\tilde{V}}\} \\ & - \max_{(s^*,d^*) \in \mathcal{A}} \{T[\tilde{V}](s^*,d^*) - \delta(F_{s',d'}, F_{s^*,d^*})\bar{b}_{\tilde{V}}\} - [\tilde{V}(s,d) - \tilde{V}(s',d')]\Bigg] \end{aligned}$$

The upper bound in Theorem 2 only requires that one apply the Bellman operator T at those states and choices in the finite set $\mathcal{A}$. The supremum in the upper bound may have an analytical solution for some parametric choices of $\tilde{V}$.

If $\mathcal{A}$ is 'dense' in a certain sense, and the functions involved are uniformly continuous, then the bound in Theorem 2 is not very conservative compared to the bound in Theorem 1. We formalize this in Theorem 3 below. Theorem 2 suggests the following procedure to upper bound $\sup_{s,s',d,d'} |T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - [\tilde{V}(s,d) - \tilde{V}(s',d')]|$. The algorithm returns a quantity $\bar{B}$ that is guaranteed (by Theorem 2) to upper bound $\sup_{s,s',d,d'} |T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - [\tilde{V}(s,d) - \tilde{V}(s',d')]|$ and that is guaranteed to exceed this quantity by less than a desired tolerance τ.

1. Choose some finite set of states and choices $\mathcal{A} \subset \mathcal{S} \times \mathcal{D}$. Evaluate $T[\tilde{V}](s,d)$ for each $(s,d) \in \mathcal{A}$. Evaluate $\bar{b}_{\tilde{V}}$.

2. Calculate $\bar{B}$ given by:

$$\bar{B} = \sup_{s,s' \in S, d,d' \in D} \Bigg[\min_{(s^*,d^*) \in \mathcal{A}} \{T[\tilde{V}](s^*,d^*) + \delta(F_{s,d}, F_{s^*,d^*})\bar{b}_{\tilde{V}}\} \\ - \max_{(s^*,d^*) \in \mathcal{A}} \{T[\tilde{V}](s^*,d^*) - \delta(F_{s',d'}, F_{s^*,d^*})\bar{b}_{\tilde{V}}\} - [\tilde{V}(s,d) - \tilde{V}(s',d')]\Bigg]$$

Evaluate $\underline{B}$ by:

$$\underline{B} = \max_{(s,d),(s',d') \in \mathcal{A}} \{|T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - [\tilde{V}(s,d) - \tilde{V}(s',d')]|\}$$

3. If $\bar{B} - \underline{B} \leq \tau$ then we are done and $\bar{B}$ is the upper bound. If $\bar{B} - \underline{B} > \tau$ then increase the density of $\mathcal{A}$ and return to step 2 else continue to step 4.

4. Take as the bound $b(s,d,d')\bar{B}$.

Note that at each iteration in the algorithm above, the upper bound $\bar{B}$ exceeds $\sup_{s,s',d,d'} |T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - [\tilde{V}(s,d) - \tilde{V}(s',d')]|$. It is not too difficult to see that $\underline{B}$ lower bounds this quantity. Hence τ bounds the conservativeness of the upper bound $\bar{B}$. In Theorem 3 we provide simple conditions that ensure the algorithm above halts and therefore guarantees we find a bound $\bar{B}$ that is within τ of the quantity we wish to bound.

We define the density of $\mathcal{A}$ as the reciprocal of $\sup_{s \in \mathcal{S}, d \in \mathcal{D}} \min_{(s',d) \in \mathcal{A}} ||s - s'||$ (if the denominator is zero we take the density to be infinity).

Theorem 3. *If $\mathcal{S} \subset \mathbb{R}^k$ is compact and the functions $s \mapsto T[\tilde{V}](s,d)$, $s \mapsto \tilde{V}(s,d)$, and $(s,s') \mapsto \delta(F_{s,d}, F_{s',d})$ are each uniformly continuous for all $d \in \mathcal{D}$, then as the density of $\mathcal{A}$ increases to infinity $\bar{B} - \underline{B} \to 0$.*

Theorem 3 implies that the algorithm described above must eventually halt and thus ensure that $\bar{B}$ exceeds the bound in Theorem 1 by no more than the pre-specified tolerance τ.

3.3 Subset Estimation and Robust Inference

Let us reintroduce parameters θ, β, and F into our notation for the value function V and Bellman operator. For any bounded function $\tilde{V}$, The results in the previous section imply an upper-bound of the form:

$$\begin{aligned} & \left|\big(\tilde{V}(s,d) - \tilde{V}(s,d')\big) - \Delta_{(\theta,\beta,F)}V(s,d,d')\big)\right| \\ \leq \;& Q_{(\theta,\beta,F)}(s,d,d';\tilde{V}) \end{aligned}$$

Where $Q_{(\theta,\beta,F)}$ is the bound given in Theorem 2 (or possibly Theorem 1 if this can be feasibly calculated) for the particular choice of θ, β, F. Note then that for any $\tilde{V}$:

$$\begin{aligned} & \Delta_{(\theta,\beta,F)}\tilde{V}(s;d',d) - Q_{(\theta,\beta,F)}(s,d,d';\tilde{V}) \\ \leq \;& \Delta_{(\theta,\beta,F)}V(s;d',d) \\ \leq \;& \Delta_{(\theta,\beta,F)}\tilde{V}(s;d',d) + Q_{(\theta,\beta,F)}(s,d,d';\tilde{V}) \end{aligned}$$

For any bounded function $\hat{V} : \mathcal{D} \to \mathbb{R}$ let:

$$L_{(\theta,\beta,F)}(s,d;\hat{V}) = P_\theta\Big[\Delta_{d,d'}\epsilon_t \geq \Delta_{d',d}u(s;\theta) + \beta\hat{V}(d'), \forall d' \in \mathcal{D}, d' \neq d\Big]$$

Note that $L_{(\theta,\beta,F)}(s,d) = L_{(\theta,\beta,F)}\big(s,d;\Delta V_{(\theta,\beta,F)}(s,d,\cdot)\big)$. It is easy to see that $L_{(\theta,\beta,F)}(s,d;\hat{V})$ is monotonically decreasing in $\hat{V}(d')$ for each $d' \neq d$. As such, for any bounded $\tilde{V} : \mathcal{S} \times \mathcal{D} \to \mathbb{R}$:

$$\begin{aligned} & L_{(\theta,\beta,F)}\big(s,d;\tilde{V}(s,\cdot) - \tilde{V}(s,d) + Q_{(\theta,\beta,F)}(s,d,\cdot;\tilde{V})\big) \\ \leq & L_{(\theta,\beta,F)}(s,d) \\ \leq & L_{(\theta,\beta,F)}\big(s,d;\tilde{V}(s,\cdot) - \tilde{V}(s,d) - Q_{(\theta,\beta,F)}(s,d,\cdot;\tilde{V})\big) \end{aligned}$$

So we have found upper and lower bounds the likelihood. Thus we can define an upper- bound on the sample log likelihood of the data by:

$$\begin{aligned}&\mathcal{L}^{U}_{\{S_t,D_t\}_{t=1}^{T}}(\theta,\beta;F,\tilde{V})\\ =&\sum_{t=1}^{T}\ln L_{(\theta,\beta,F)}\big(S_t,D_t;\tilde{V}(s,\cdot)-\tilde{V}(s,d)-Q_{(\theta,\beta,F)}(S_t,D_t,\cdot;\tilde{V})\big)\end{aligned}$$

And a lower bound by:

$$\begin{aligned}&\mathcal{L}^{L}_{\{S_t,D_t\}_{t=1}^{T}}(\theta,\beta;F,\tilde{V})\\ =&\sum_{t=1}^{T}\ln L_{(\theta,\beta,F)}\big(S_t,D_t;\tilde{V}(s,\cdot)-\tilde{V}(s,d)+Q_{(\theta,\beta,F)}(S_t,D_t,\cdot;\tilde{V})\big)\end{aligned}$$

Now suppose that $\hat{\theta}^*$ and $\hat{\beta}^*$ maximize the exact log likelihood $\mathcal{L}_{\{S_t,D_t\}_{t=1}^{T}}(\theta,\beta;\hat{F})$ (the log-likelihood obtained if the full dynamic programming solution were available). Suppose we have some approximate dynamic programming algorithm that, for each θ, β and F, provides an approximate value function $\tilde{V}_{(\theta,\beta,F)}$. It follows that:

$$\mathcal{L}^{U}_{\{S_t,D_t\}_{t=1}^{T}}(\hat{\theta}^*,\hat{\beta}^*;\hat{F},\tilde{V}_{(\hat{\theta}^*,\hat{\beta}^*,\hat{F})})\geq\sup_{\theta,\beta}\mathcal{L}^{L}_{\{S_t,D_t\}_{t=1}^{T}}(\theta,\beta;\hat{F},\tilde{V}_{(\theta,\beta,\hat{F})})$$

Note that the inequality above applies regardless of the algorithm used to generate $\tilde{V}_{(\theta,\beta,F)}$. The above implies that $(\hat{\theta}^*,\hat{\beta}^*)$ lies in the following set:

$$\left[(\theta,\beta):\mathcal{L}^{U}_{\{S_t,D_t\}_{t=1}^{T}}(\theta,\beta;\hat{F},\tilde{V}_{(\theta,\beta,\hat{F})})\geq\sup_{\theta',\beta'}\mathcal{L}^{L}_{\{S_t,D_t\}_{t=1}^{T}}(\theta',\beta';\hat{F},\tilde{V}_{(\theta',\beta',\hat{F})})\right]$$

The set above can be used as a set estimator for the true parameters θ and β. By construction the set estimator must contain the infeasible maximum likelihood estimates, that is, the estimates obtained if the full dynamic programming solution were available. If the infeasible maximum likelihood estimates $\hat{\theta}^*$ and $\hat{\beta}^*$ are consistent estimates, then set estimator is consistent (in the Hausdorff metric).

We can also use the upper and lower bounds on the likelihood to derive confidence sets that are valid regardless of the accuracy of the approximate value function $\tilde{V}_{(\theta,\beta,\hat{F})}$.

The standard likelihood-ratio confidence set for the parameters is given by:

$$[(\theta,\beta):\ 2\big(\sup_{\theta',\beta'}\mathcal{L}_{\{S_t,D_t\}_{t=1}^T}(\theta',\beta';\hat{F})-\mathcal{L}_{\{S_t,D_t\}_{t=1}^T}(\theta,\beta;\hat{F})\big)\le c_{1-\alpha}]$$

where $c_{1-\alpha}$ is the $1-\alpha$-level critical value, usually the α-quantile of the chi-squared distribution with degrees of freedom equal to the dimension of θ and β. The above is infeasible, it cannot be evaluated because $\mathcal{L}_{\{S_t,D_t\}_{t=1}^T}$ incorporates the exact solution to the dynamic programming problem. However the confidence set above is necessarily a subset of the following feasible set:

$$[(\theta,\beta):\ 2\big(\sup_{\theta',\beta'}\mathcal{L}^U_{\{S_t,D_t\}_{t=1}^T}(\theta',\beta';\hat{F},\tilde{V}_{(\theta,\beta,\hat{F})})-\mathcal{L}^L_{\{S_t,D_t\}_{t=1}^T}(\theta,\beta;\hat{F},\tilde{V}_{(\theta,\beta,\hat{F})})\big)\le c_{1-\alpha}]$$

Thus, if the exact, infeasible likelihood ratio-based confidence set has asymptotically correct size, then so too must the feasible one based on approximate dynamic programming above.

3.4 Monte Carlo Exercise

As a practical demonstration of our methods we present a Monte Carlo simulation. We simulate data from a modified Rust model ([71]). We apply our methods to account for approximation error in maximum likelihood estimates that employ approximate dynamic programming. We evaluate set estimates and robust confidence sets when the approximate dynamic programming is applied with differing degrees of coarseness. We compare our set estimates and robust confidence sets to conventional point estimates and confidence sets that do not account for the error due to approximation in the value function.

In the modified Rust model, at each period a decision maker chooses whether or not to send a bus engine for repairs. The choice set is binary $\mathcal{D}=\{0,1\}$ with $d=0$ representing the decision not to repair the engine and $d=1$ the decision to repair the engine. The state space is given by $\mathcal{S}=[0,20]$, each $s\in\mathcal{S}$ represents a possible value for the engine's mileage given in thousands of miles. We treat the mileage as a metaphor for the general health the engine and thus it is possible for the mileage to decrease between periods. The cost of running (and not repairing) an engine increases with the mileage because with high mileage the engine becomes unreliable and fuel-hungry. At state s running the engine incurs a cost of $-\theta_1 s$ where θ_1 is a scalar parameter. Repairing an engine incurs a fixed cost of $-\theta_2$ regardless of the mileage. Thus the deterministic parts of the utilities from the two choices at state s are given by $u(s, 0) = \theta_1 s$ and $u(s, 1) = \theta_2$.

In each period t there is an additive, stochastic shock to both the cost of repair and non-repair. The cost shock to non-repair and repair in period t are respectively $\epsilon_t(0)$ and $\epsilon_t(1)$. The shocks are iid and independent between the two choices, they are observed by the decision-maker before making a decision at time t but are not known prior to t. We let the shocks be standard type-1 extreme value distributed. The bus's route is subject to unexpected changes due to road closures and driver error, as such the mileage evolves stochastically. If the decision at time t is D_t and the engine has mileage S_t, then the mileage in period $t + 1$ is distributed according to:

$$S_{t+1}|S_t, D_t \sim \min\{\max\{0, S_t + 1\{D_t = 0\}\gamma_1 + 1\{D_t = 1\}\gamma_2 + U[-\gamma_3, \gamma_3]\}, 50\}$$

Where $U[-\gamma_3, \gamma_3]$ is a uniform random variable between $-\gamma_3$ and γ_3. γ_1, γ_2, and γ_3 are model parameters. In our simulations γ_2 is negative, thus repair of the engine reduces the mileage.

The modified replacement model differs from Rust's original formulation in two key ways. Firstly, we treat the mileage as a continuous variable, In the original model estimated by Rust the mileage is discretized. Secondly, in the modified model the decision maker chooses whether or not to repair the engine, reducing its effective mileage by some fixed amount. In the original model the decision maker chooses whether or not to replace the engine, bringing the mileage down to zero. The use of a continuous state-space allows us to examine our method when the value function is approximated with different degrees of coarseness. Repairing rather than replacing the engine ensures that the total variation distance $\delta(F_{s,0}, F_{s,1})$ is strictly less than unity even for large values of s. Without this feature the factor $b(s,d,d')$ in the upper bound in Theorem 1 would be equal to $\frac{1}{1-\beta}$ when the mileage s is large. Thus this modeling decision allows us to demonstrate the improvement our specific bound provides over the more crude bound with a factor of $\frac{1}{1-\beta}$ when the distribution of future states from different choices overlaps.

Note that in the model above:

$$b(s,d,d') \leq \frac{1-\beta+\beta\min\{|\frac{\gamma_1-\gamma_2}{2\gamma_3}|,1\}}{1-\beta}$$

Note also that because $\epsilon_t(d)$ is independently type-1 extreme value distributed for each $d \in \mathcal{D}$ we have for any $\tilde{V}$:

$$T[\tilde{V}](s,d) = E\big[ln\big(\sum_{d\in\mathcal{D}}\exp(u(S_{t+1},d)+\beta\tilde{V}(S_{t+1},d)\big)\big|S_t=s,D_t=d\big]+c$$

Where c is the Euler–Mascheroni constant. In order to generate data from the model we must solve for the value function. An exact solution is infeasible and so we apply approximate value function iteration. We employ a dense grid $\tilde{\mathcal{S}}$ that consists of 1001 evenly spaced points between 0 and 20. We replace the outer expectation in the RHS above by an empirical expectation (over 100 draws of S_{t+1}):

$$T[\tilde{V}](s,d) \approx \frac{1}{n}\sum_{i=1}^{n} ln\big(\sum_{d\in\mathcal{D}}\exp(u(S_{s,d,i},d)+\beta\tilde{V}(S_{s,d,i},d)\big)+c$$

Where $S_{s,d,i}$ is the i^{th} draw from the distribution of S_{t+1} given $S_t=s$ and $D_t=d$. We begin with an initial guess V_0 of the value function (our initial guess is that the function equals zero everywhere), and at the j^{th} iteration we evaluate $T[V_j](s,d)$ for each s in the dense grid $\tilde{\mathcal{S}}$ and each $d \in \mathcal{D}$. We then linearly interpolate the resulting function between the points in $\tilde{\mathcal{S}}$ in order to evaluate $V_{j+1}(S_{s,d,i},d)$ for each $d \in \mathcal{D}$, $s \in \tilde{\mathcal{S}}$ and $i \in \{1,...,n\}$. Note that the samples $\{\{\{S_{s,d,i}\}_{i=1}^{n}\}_{d\in\mathcal{D}}\}_{s\in\tilde{\mathcal{S}}}$ are drawn before the first iteration and do not change between iterations. We continue until the change in the value function between iterations is smaller than a pre-specified tolerance at all $s \in \tilde{\mathcal{S}}$ and $d \in \mathcal{D}$.

Each simulated dataset consists of 5000 periods of data generated from the model above. Table 1 summarizes the data generating process and states the parameter values used in the simulation.

Table 3.1: DGP summary and Parameter Values

Model Parameters	DGP Summary
β =0.8	$\big(\epsilon_t(0), \epsilon_t(1)\big) \sim$ independent type-1 extreme value
$\theta_1 = -0.6$	$u(s,0) = \theta_1 s$
$\theta_2 = -4$	$u(s,1) = \theta_2$
$\gamma_1 = 1$	$\mathcal{S} = [0,20]$, $\mathcal{D} = \{0,1\}$
$\gamma_2 = -1$	$S_{t+1}\|S_t, D_t = 0 \sim \min\{\max\{S_t + \gamma_1 + U[-\gamma_3, \gamma_3], 20\}, 0\}$
$\gamma_3 = 5$	$S_{t+1}\|S_t, D_t = 1 \sim \min\{\max\{S_t + \gamma_2 + U[-\gamma_3, \gamma_3], 20\}, 0\}$
$T = 1000$	$\tilde{\mathcal{S}} = \{0, \frac{1}{50}, \frac{2}{50}, ..., \frac{999}{50}, 20\}$

On each simulated dataset we apply maximum likelihood to estimate the parameters θ_1 and θ_2. We treat β, γ_1, γ_2, and γ_3 as known. For each candidate set of parameter values in the maximization routine we approximately solve the dynamic programming problem using Bellman iteration on a grid $\hat{\mathcal{S}}$ of evenly spaced points in the interval $[0,20]$ and interpolating as described in Section 1. The interpolation method used is linear interpolation. Let $\tilde{V}$ be an approximate value function. In order to evaluate the bound on the RHS of the inequality in Theorem 1 we maximize the following quantity over all $d, d' \in \mathcal{D}$ and all $s, s' \in \tilde{\mathcal{S}}$ where $\tilde{\mathcal{S}}$ is the dense grid used in the data generating process:

$$|[T[\tilde{V}](s', d') - T[\tilde{V}](s,d)] - [\tilde{V}(s',d') - \tilde{V}(s,d)]]|$$

We then multiply the above by the factor $\frac{1-\beta+\beta\min\{|\frac{\gamma_1-\gamma_2}{2\gamma_3}|,1\}}{1-\beta}$ to get the bound in Theorem 1.

To examine the coverage of our robust confidence set and set estimator we numerically maximize $\mathcal{L}^L_{\{S_t,D_t\}_{t=1}^T}(\theta, \beta; F, \tilde{V}_{(\theta,\beta,F)})$ over θ. The true parameters (θ_1, θ_2) lie in the set estimator if:

$$\sup_{\theta} \mathcal{L}^{L}_{\{S_t,D_t\}_{t=1}^{T}}(\theta,\beta;F,\tilde{V}_{(\theta,\beta,F)}) - \mathcal{L}^{U}_{\{S_t,D_t\}_{t=1}^{T}}\big((\theta_1,\theta_2),\beta;F,\tilde{V}_{(\theta,\beta,F)}\big) \leq 0$$

and the true parameters are contained in the robust confidence set if:

$$\sup_{\theta} \mathcal{L}^{L}_{\{S_t,D_t\}_{t=1}^{T}}(\theta,\beta;F,\tilde{V}_{(\theta,\beta,F)}) - \mathcal{L}^{U}_{\{S_t,D_t\}_{t=1}^{T}}\big((\theta_1,\theta_2),\beta;F,\tilde{V}_{(\theta,\beta,F)}\big) \leq \frac{1}{2}c_{1-\alpha}$$

Where $c_{1-\alpha}$ is the $1-\alpha$-quantile of the chi-squared distribution with two degrees of freedom. The true parameters (θ_1,θ_2) are in the non-robust likelihood-ratio confidence set if:

$$\sup_{\theta} \mathcal{L}_{\{S_t,D_t\}_{t=1}^{T}}(\theta,\beta;F,\tilde{V}_{(\theta,\beta,F)}) - \mathcal{L}_{\{S_t,D_t\}_{t=1}^{T}}\big((\theta_1,\theta_2),\beta;F,\tilde{V}_{(\theta,\beta,F)}\big) \leq \frac{1}{2}c_{1-\alpha}$$

We use numerical methods to evaluate the supremum on the LHS above.

We apply our methodology with different densities for the grid $\hat{\mathcal{S}}$ that is used in the approximate dynamic programming routine. Since the grid density controls the coarseness of the approximation, this allows us to examine how our set estimation and inference procedures perform as the quality of the approximation changes.

Table 1 provides results from our simulations. Each column corresponds to different choice for $\hat{\mathcal{S}}$, the grid used to perform approximate dynamic programming in estimation. The figures in the first row are the mean (over 500 simulations) squared errors of the maximum likelihood point estimates. In the second row are the frequencies with which the true parameters lie within the set estimate. In the third row are the frequencies with which the robust confidence sets contain the true parameters. Finally, in the fourth row are the frequencies with which the standard likelihood ratio-based confidence sets contain the true parameters.

Table 3.2: Simulation Results

	Number of points in $\hat{\mathcal{S}}$:		
	10	100	500
Mean Squared Error	0.023	0.023	0.022
Set Estimator Coverage	1	1	0.908
Robust Confidence Set Coverage	1	1	0.996
Standard Confidence Set Coverage	0.702	0.942	0.938

We see that the standard non-robust confidence set has coverage much lower than 95% when $\hat{\mathcal{S}}$ contains only 10 points. This is not surprising, when the grid used for the approximate dynamic programming is sparse, the quality of approximation is low leading to biased point estimates and confidence sets with low coverage. When the number of points in $\hat{\mathcal{S}}$ is larger the coverage is close to the stated level of 95%. By contrast, the robust confidence set has coverage of at least 95% regardless of the density of the grid used in the approximate dynamic programming. The robust confidence sets are based on worst-case scenarios for the approximation error and as such they are generally conservative, having coverage greater than the stated level. As the error from approximation is reduced and statistical noise dominates, the robust confidence sets should have coverage approaching the stated 95% level, and indeed we do see that the coverage is slightly lower in the case of $\hat{\mathcal{S}}$ with 500 points.

The set estimator is not required to cover the true parameters with any particular frequency, but is nonetheless interesting to observe its coverage in our simulations. We note that the set estimates cover the true parameters in 100% of our simulations when $\hat{\mathcal{S}}$ has either 10 or 100 points. When the grid is made dense the set estimator should shrink around the maximum likelihood point estimates and the coverage should fall, and indeed when $\hat{\mathcal{S}}$ contains 500 points the coverage of the set estimator is reduced to 91%.

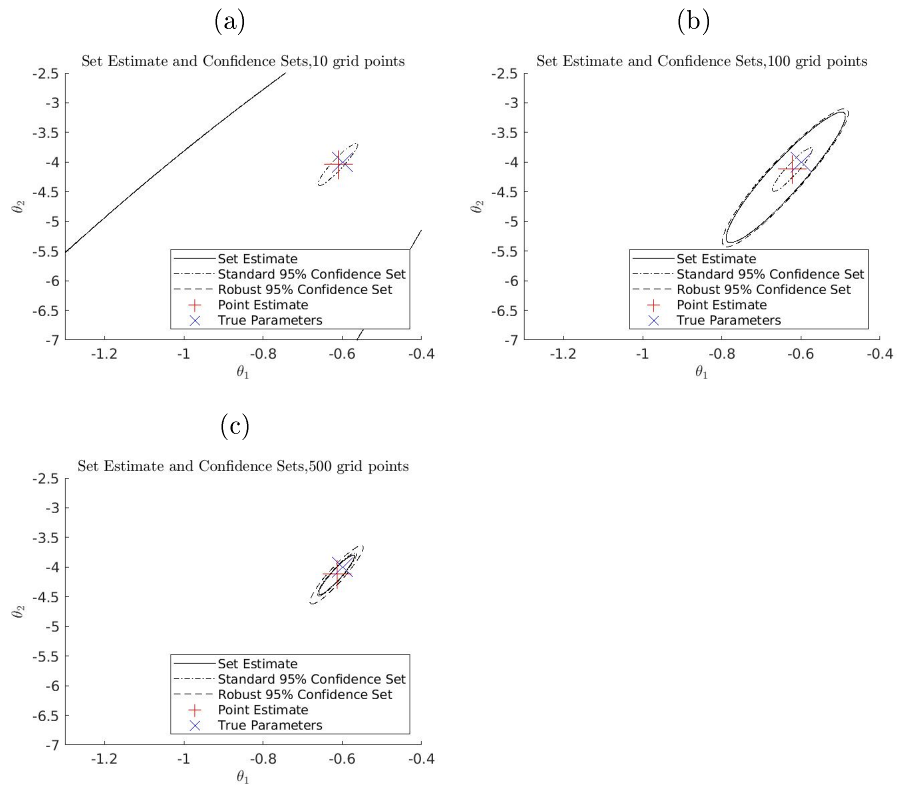
To get a sense of what the set estimates and confidence sets look like, for each choice of $\hat{\mathcal{S}}$ we plot the sets for one simulation draw. We evaluate $\mathcal{L}^U_{\{S_t,D_t\}_{t=1}^T}$ on a grid of values for θ_1 and θ_2. We then interpolate the function over a denser grid and for each set of parameters in this grid we compare $\mathcal{L}^U_{\{S_t,D_t\}_{t=1}^T}$ and $\sup_\theta \mathcal{L}^L_{\{S_t,D_t\}_{t=1}^T}(\theta, \beta; F, \tilde{V}_{(\theta,\beta,F)})$ to determine whether the parameters tlie inside the set estimate and robust confidence set. Similarly, we evaluate $\mathcal{L}_{\{S_t,D_t\}_{t=1}^T}$ for each pair of parameter values on the same grid and interpolate to assess which points lie inside the standard non-robust confidence set.

We plot these sets in Figure 1 below. Also indicated in each sub-figure are the maximum likelihood point estimates and the true values of parameters θ_1 and θ_2. As one might expect, as $\hat{\mathcal{S}}$ grows dense, the robust confidence set shrinks and becomes close to the standard confidence set, and the set estimate shrinks around the MLE estimator.

In the case of $\hat{\mathcal{S}}$ with 10 points, the approximation is weak and there is substantial undercoverage of the non-robust confidence set. We see from Sub-Figure 1.a that in this case the set estimates and robust confidence sets are very large, reflecting the poor strength of approximation. Note that because the worst-case approximation error dominates the statistical noise in this case, the robust confidence set and the set estimates are visibly indistinguishable. In the case of $\hat{\mathcal{S}}$ with 100 points the approximation quality is improved and we see from Sub-Figure 1.b that the set estimates and confidence sets are much smaller, although still significantly larger than the non-robust confidence set, which (according to Table 1) has nearly correct coverage in this case. When $\hat{\mathcal{S}}$ contains 500 points the approximation is good and we see from Sub-Figure 1.c that the robust confidence set is only a little larger than the non-robust, moreover in this case the set estimate is contained within the non-robust confidence set.

An empirical researcher using our methodology would see that in the case of $\hat{\mathcal{S}}$ with 10 points, the robust confidence set and set estimate are impractically large, and should conclude that the quality of approximation is poor. The researcher can then decide to increase the density of the grid in order to achieve a better approximation. Thus our methodology allows the researcher to make a more informed decision about the approximation method and to better balance the trade-off between computational expedience and the bias due to approximate dynamic programming.

Figure 3-1: Simulated Set Estimates



3.5 Conclusion

We develop an approach to estimation and inference in dynamic discrete choice models that directly accounts for error due to the use of approximate dynamic programming. We prove the validity of our approach and provide simulation evidence that our methodology is of practical use. Of course, our simulation results pertain to one particular model, and it is possible that the robust confidence sets are overly conservative in other settings. To better assess the practical efficacy of our methods we hope to implement the procedure on real data in future work.

The methodology detailed in the paper is very flexible in that it applies regardless of the approximation method used. It may be possible to refine our approach taking into account features of the specific approximate dynamic programming method used. It may also be possible to tighten our error bounds by incorporating information about the value function that can be derived analytically, for example, in some cases one can prove the value function is convex. This avenue may also be worth exploring in future research.

3.6 Appendix: Proofs

Proof Theorem 1. By the definition of T for any bounded functions $\tilde{V}$ and $\hat{V}$ that map from $\mathcal{S} \times \mathcal{D}$ to $\mathbb{R}$:

$$\begin{aligned}
&= T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - \big(T[\hat{V}](s',d') - T[\hat{V}](s',d')\big) \\
&= E[E[\max_{d'' \in \mathcal{D}} u(S_{t+1},d'') + \beta\tilde{V}(S_{t+1},d'') + \epsilon_{t+1}(d'')]|S_t = s, D_t = d] \\
&- E[E[\max_{d'' \in \mathcal{D}} u(s',d'') + \beta\tilde{V}(s',d'') + \epsilon_{t+1}(d'')]|S_t = s', D_t = d'] \\
&+ E[E[\max_{d'' \in \mathcal{D}} u(S_{t+1},d'') + \beta\hat{V}(S_{t+1},d'') + \epsilon_{t+1}(d'')]|S_t = s, D_t = d] \\
&- E[E[\max_{d'' \in \mathcal{D}} u(S_{t+1},d'') + \beta\hat{V}(S_{t+1},d'') + \epsilon_{t+1}(d'')]|S_t = s', D_t = d']
\end{aligned}$$

Because ϵ_{t+1} and S_{t+1} are independent we can rewrite the right-hand side of the final equality above as an integral:

$$\begin{aligned}
&\int \Bigg(E[\max_{d'' \in \mathcal{D}} u(s'',d'') + \beta\tilde{V}(s'',d'') + \epsilon_t(d'')] \\
&- E[\max_{d'' \in \mathcal{D}} u(s'',d'') + \beta\hat{V}(s'',d'') + \epsilon_t(d'')]\Bigg) d(F_{s,d} - F_{s',d'})
\end{aligned}$$

It is an elementary property of the total variation distance that for any bounded function $g : \mathcal{S} \to \mathbb{R}$:

$$\int g(s'')d(F_{s,d} - F_{s',d'}) \quad \leq \quad \delta(F_{s,d}, F_{s',d'})\big(\sup_{s \in \mathcal{S}} g(s) - \inf_{s \in \mathcal{S}} g(s)\big)$$

Since u, $\tilde{V}$, and $\hat{V}$ are bounded and $E[\max_{d \in \mathcal{D}} \epsilon_t(d)] < \infty$, $E[\max_{d \in \mathcal{D}} u(s,d) + \beta\tilde{V}(s,d) + \epsilon_t(d)]$ and $E[\max_{d \in \mathcal{D}} u(s,d) + \beta\hat{V}(s,d) + \epsilon_t(d)]$ are also uniformly bounded over $s \in \mathcal{S}$ and so applying the inequality above:

$$
\begin{aligned}
& \int \Bigg(E[\max_{d''\in\mathcal{D}} u(s'',d'') + \beta \tilde{V}(s'',d'') + \epsilon_t(d'')] \\
& -E[\max_{d''\in\mathcal{D}} u(s'',d'') + \beta \hat{V}(s'',d'') + \epsilon_t(d'')]\Bigg) d(F_{s,d} - F_{s',d'}) \\
\leq \; & \delta(F_{s,d}, F_{s',d'}) \sup_{s''\in\mathcal{S}} \Bigg(E[\max_{d''\in\mathcal{D}} u(s'',d'') + \beta \tilde{V}(s'',d'') + \epsilon_t(d'')] \\
& -E[\max_{d''\in\mathcal{D}} u(s'',d'') + \beta \hat{V}(s'',d'') + \epsilon_t(d'')]\Bigg) \\
- \; & \delta(F_{s,d}, F_{s',d'}) \inf_{s''\in\mathcal{S}} \Bigg(E[\max_{d''\in\mathcal{D}} u(s'',d'') + \beta \tilde{V}(s'',d'') + \epsilon_t(d'')] \\
& -E[\max_{d''\in\mathcal{D}} u(s'',d'') + \beta \hat{V}(s'',d'') + \epsilon_t(d'')]\Bigg)
\end{aligned}
$$

It is easy to see that:

$$
\begin{aligned}
& \sup_{s\in\mathcal{S}} \Bigg(E[\max_{d\in\mathcal{D}} u(s,d) + \beta \tilde{V}(s,d) + \epsilon_t(d)] \\
& - E[\max_{d\in\mathcal{D}} u(s,d) + \beta \hat{V}(s,d) + \epsilon_t(d)]\Bigg) \\
\leq & \beta \sup_{s\in\mathcal{S}} \max_{d\in\mathcal{D}} \big(\tilde{V}(s,d) - \hat{V}(s,d)\big)
\end{aligned}
$$

And:

$$
\begin{aligned}
& \inf_{s\in\mathcal{S}} \Bigg(E[\max_{d} u(s,d) + \beta \tilde{V}(s,d) + \epsilon_t(d)] \\
& - E[\max_{d''} u(s,d) + \beta \hat{V}(s,d) + \epsilon_t(d)]\Bigg) \\
\geq & \beta \inf_{s} \min_{d\in\mathcal{D}} \big(\tilde{V}(s,d) - \hat{V}(s,d)\big)
\end{aligned}
$$

Combining we get:

$$
\begin{aligned}
& T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - \big(T[\hat{V}](s',d') - T[\hat{V}](s',d')\big) \\
\leq \; & \delta(F_{s,d}, F_{s',d'})\beta \Bigg(\sup_{s\in\mathcal{S}} \max_{d\in\mathcal{D}} \big(\tilde{V}(s,d) - \hat{V}(s,d)\big) - \inf_{s} \min_{d\in\mathcal{D}} \big(\tilde{V}(s,d) - \hat{V}(s,d)\big)\Bigg) \\
= \; & \delta(F_{s,d}, F_{s',d'})\beta \sup_{s,s'\in\mathcal{S},d,d'\in\mathcal{D}} \big|\tilde{V}(s,d) - \tilde{V}(s',d') - \big(\hat{V}(s,d) - \hat{V}(s',d')\big)\big|
\end{aligned}
$$

Since the same reasoning holds with $\tilde{V}$ and $\hat{V}$ switched we get:

$$\begin{aligned}
& \left|T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - \left(T[\hat{V}](s',d') - T[\hat{V}](s',d')\right)\right| \\
\leq \; & \delta(F_{s,d}, F_{s',d'})\beta \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \left|\tilde{V}(s,d) - \tilde{V}(s',d') - \left(\hat{V}(s,d) - \hat{V}(s',d')\right)\right|
\end{aligned}$$

Now, V is a fixed point of T and it must be bounded because u is bounded and $E[\max_{d \in \mathcal{D}} \epsilon_t(d)] < \infty$. Thus the above implies:

$$\begin{aligned}
& \left|T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - \left(V(s',d') - V(s',d')\right)\right| \\
\leq \; & \delta(F_{s,d}, F_{s',d'})\beta \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \left|\tilde{V}(s,d) - \tilde{V}(s',d') - \left(V(s,d) - V(s',d')\right)\right| \quad (3.6.1)
\end{aligned}$$

Taking the supremum over both sides:

$$\begin{aligned}
& \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \left|T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - \left(V(s',d') - V(s',d')\right)\right| \\
\leq \; & \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \delta(F_{s,d}, F_{s',d'})\beta \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \left|\tilde{V}(s,d) - \tilde{V}(s',d') - \left(V(s,d) - V(s',d')\right)\right|
\end{aligned}$$

By the reverse triangle inequality:

$$\begin{aligned}
& \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \left|\tilde{V}(s,d) - \tilde{V}(s',d') - \left(V(s',d') - V(s',d')\right)\right| \\
- \; & \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \left|T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - \left(\tilde{V}(s',d') - \tilde{V}(s',d')\right)\right| \\
\leq \; & \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \delta(F_{s,d}, F_{s',d'})\beta \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \left|\tilde{V}(s,d) - \tilde{V}(s',d') - \left(V(s,d) - V(s',d')\right)\right|
\end{aligned}$$

Since $\delta(F_{s,d}, F_{s',d'})\beta \leq \beta < 1$ we can solve to get:

$$\begin{aligned}
& \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \left|\tilde{V}(s,d) - \tilde{V}(s',d') - \left(V(s',d') - V(s',d')\right)\right| \\
\leq \; & \frac{1}{1 - \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \delta(F_{s,d}, F_{s',d'})\beta} \\
\times \; & \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \left|T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - \left(\tilde{V}(s',d') - \tilde{V}(s',d')\right)\right| \quad (3.6.2)
\end{aligned}$$

Also by the triangle inequality:

$$\begin{aligned}
& |\tilde{V}(s,d) - \tilde{V}(s',d') - [V(s,d) - V(s',d')]| \\
\leq \ & |T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - [\tilde{V}(s,d) - \tilde{V}(s',d')]| \\
+ \ & |T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - [V(s,d) - V(s',d')]|
\end{aligned}$$

Substituting using (3.6.1) we get:

$$\begin{aligned}
& |\tilde{V}(s,d) - \tilde{V}(s',d') - [V(s,d) - V(s',d')]| \\
\leq \ & |T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - [\tilde{V}(s,d') - \tilde{V}(s',d')]| \\
+ \ & \delta(F_{s,d}, F_{s',d'})\beta \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \left| \tilde{V}(s,d) - \tilde{V}(s',d') - \big(V(s,d) - V(s',d')\big) \right|
\end{aligned}$$

And so from (3.6.2):

$$\begin{aligned}
& |\tilde{V}(s,d) - \tilde{V}(s',d') - [V(s,d) - V(s',d')]| \\
\leq \ & |T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - [\tilde{V}(s,d') - \tilde{V}(s',d')]| \\
+ \ & \Bigg(\frac{\delta(F_{s,d}, F_{s',d'})\beta}{1 - \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \delta(F_{s,d}, F_{s',d'})\beta} \\
\times \ & \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \left| T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - \big(\tilde{V}(s',d') - \tilde{V}(s',d')\big) \right| \Bigg)
\end{aligned}$$

Taking the supremum over the RHS we get:

$$\begin{aligned}
& |\tilde{V}(s,d) - \tilde{V}(s',d') - [V(s,d) - V(s',d')]| \\
\leq \ & \Bigg(\frac{1 - \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \delta(F_{s,d}, F_{s',d'})\beta + \delta(F_{s,d}, F_{s',d'})\beta}{1 - \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \delta(F_{s,d}, F_{s',d'})\beta} \\
\times \ & \sup_{s,s' \in \mathcal{S}, d,d' \in \mathcal{D}} \left| T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - \big(\tilde{V}(s',d') - \tilde{V}(s',d')\big) \right| \Bigg)
\end{aligned}$$

The conclusion then follows immediately. □

Proof Theorem 2. First we show that for any $s, s' \in \mathcal{S}$ and $d, d' \in \mathcal{D}$:

$$|T\tilde{V}(s,d) - T\tilde{V}(s',d')| \leq \delta(F_{s,d}, F_{s',d'})\bar{b}_{\tilde{V}}$$

By the definition of T:

$$\begin{aligned}
& T[\tilde{V}](s,d) - T[\tilde{V}](s',d') \\
= \quad & E[E[\max_{d''\in\mathcal{D}} u(S_{t+1},d'') + \beta\tilde{V}(S_{t+1},d'') + \epsilon_{t+1}(d'')]|S_t = s, D_t = d] \\
- \quad & E[E[\max_{d''\in\mathcal{D}} u(s',d'') + \beta\tilde{V}(s',d'') + \epsilon_{t+1}(d'')]|S_t = s', D_t = d']
\end{aligned}$$

Using independence of ϵ_{t+1} and S_{t+1} the right-hand side above is equal to:

$$\int E[\max_{d''\in\mathcal{D}} u(s'',d'') + \beta\tilde{V}(s'',d'') + \epsilon_t(d'')]d(F_{s,d} - F_{s',d'})$$

As in the proof of Theorem 1 we can apply elementary properties of the total variation distance to get:

$$\int g(x)d(F_{s,d} - F_{s',d'}) \leq \delta(F_{s,d}, F_{s',d'}) \sup_{x,x'\in\mathcal{S}(s,d)\cup\mathcal{S}(s',d')} g(x) - g(x')$$

$$\begin{aligned}
& \int E[\max_{d''\in\mathcal{D}} u(s'',d'') + \beta\tilde{V}(s'',d'') + \epsilon_t(d'')]d(F_{s,d} - F_{s',d'}) \\
\leq \quad & \delta(F_{s,d}, F_{s',d'})\bigg(\sup_{s'\in\mathcal{S}} E[\max_{d''\in\mathcal{D}} u(s'',d'') + \beta\tilde{V}(s'',d'') + \epsilon_t(d'')] \\
& - \inf_{s'\in\mathcal{S}} E[\max_{d''\in\mathcal{D}} u(s'',d'') + \beta\tilde{V}(s'',d'') + \epsilon_t(d'')]\bigg) \\
= \quad & \delta(F_{s,d}, F_{s',d'})\bigg(\sup_{s',s\in\mathcal{S}} E[\max_{d''\in\mathcal{D}} u(s'',d'') + \beta\tilde{V}(s'',d'') + \epsilon_t(d'')] \\
& - E[\max_{d''\in\mathcal{D}} u(s'',d'') + \beta\tilde{V}(s'',d'') + \epsilon_t(d'')]\bigg)
\end{aligned}$$

It is easy to see that for any $s' \in \mathcal{S}$:

$$E[\max_{d\in\mathcal{D}} u(s,d)+\beta\tilde{V}(s,d)+\epsilon_t(d)] - E[\max_{d'\in\mathcal{D}} u(s',d)+\beta\tilde{V}(s',d)+\epsilon_t(d)]\Big)$$
$$\leq \max_{d\in\mathcal{D}} \big(u(s,d)+\beta\tilde{V}(s,d)-u(s',d)+\beta\tilde{V}(s',d)\big)$$

And so in all:

$$\begin{aligned}
&T[\tilde{V}](s,d)-T[\tilde{V}](s',d')\\
\leq&\delta(F_{s,d},F_{s',d'})\sup_{s,s'\in\mathcal{S},d\in\mathcal{D},}\big(u(s,d)+\beta\tilde{V}(s,d)-u(s',d)+\beta\tilde{V}(s',d)\big)\\
=&\delta(F_{s,d},F_{s',d'})\Big|\sup_{s,s'\in\mathcal{S},d\in\mathcal{D},}\big(u(s,d)+\beta\tilde{V}(s,d)-u(s',d)+\beta\tilde{V}(s',d)\big)\Big|
\end{aligned}$$

By symmetry:

$$\begin{aligned}
&|T[\tilde{V}](s,d)-T[\tilde{V}](s',d')|\\
\leq&\delta(F_{s,d},F_{s',d'})\Big|\sup_{s,s'\in\mathcal{S},d\in\mathcal{D},}\big(u(s,d)+\beta\tilde{V}(s,d)-u(s',d)+\beta\tilde{V}(s',d)\big)\Big|\\
=&\delta(F_{s,d},F_{s',d'})\bar{b}_{\tilde{V}}
\end{aligned}$$

Now adding and subtracting terms, for any $(s^*,d^*),(s^{**},d^{**})\in\mathcal{A}$:

$$\begin{aligned}
& T[\tilde{V}](s,d)-T[\tilde{V}](s',d')-[\tilde{V}(s,d)-\tilde{V}(s',d')]\\
= \;& T[\tilde{V}](s^*,d^*)+T[\tilde{V}](s,d)-T[\tilde{V}](s^*,d^*)\\
- \;& T[\tilde{V}](s^{**},d^{**})+T[\tilde{V}](s^{**},d^{**})-T[\tilde{V}](s',d')\\
- \;& [\tilde{V}(s,d)-\tilde{V}(s',d')]\\
\leq \;& T[\tilde{V}](s^*,d^*)+|T[\tilde{V}](s,d)-T[\tilde{V}](s^*,d^*)|\\
- \;& T[\tilde{V}](s^{**},d^{**})+|T[\tilde{V}](s',d')-T[\tilde{V}](s^{**},d^{**})|\\
- \;& [\tilde{V}(s,d)-\tilde{V}(s',d')]
\end{aligned}$$

Using the bound we derived earlier we get from the above:

$$\begin{aligned}
& T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - [\tilde{V}(s,d) - \tilde{V}(s',d')] \\
\leq \;& T[\tilde{V}](s^*,d^*) + \delta(F_{s,d}, F_{s^*,d^*})\bar{b}_{\tilde{V}} \\
- \;& T[\tilde{V}](s^{**},d^{**}) + \delta(F_{s',d'}, F_{s^{**},d^{**}})\bar{b}_{\tilde{V}} \\
- \;& [\tilde{V}(s,d) - \tilde{V}(s',d')]
\end{aligned}$$

Since the above holds for any $(s^*,d^*),(s^{**},d^{**}) \in \mathcal{A}$ we can choose these to minimize it:

$$\begin{aligned}
& T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - [\tilde{V}(s,d) - \tilde{V}(s',d')] \\
\leq \;& \min_{(s^*,d^*)\in\mathcal{A}} \{T[\tilde{V}](s^*,d^*) + \delta(F_{s,d}, F_{s^*,d^*})\bar{b}_{\tilde{V}}\} \\
- \;& \max_{(s^*,d^*)\in\mathcal{A}} \{T[\tilde{V}](s^*,d^*) - \delta(F_{s',d'}, F_{s^*,d^*})\bar{b}_{\tilde{V}}\} \\
- \;& [\tilde{V}(s,d) - \tilde{V}(s',d')]
\end{aligned}$$

Taking the supremum of both sides above over $s, s' \in \mathcal{S}$ and $d, d' \in \mathcal{D}$ and noting that by symmetry:

$$\sup_{s,s'\in\mathcal{S},d,d'\in\mathcal{D}} T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - [\tilde{V}(s,d) - \tilde{V}(s',d')] > 0$$

We get the result. $\square$

Proof Theorem 3. Let B be defined by:

$$B = \sup_{s,s'\in S,d,d'\in D} |[T[\tilde{V}](s',d') - T[\tilde{V}](s,d)] - [\tilde{V}(s',d') - \tilde{V}(s,d)]]|$$

Fix some $\tau > 0$. Recall $\underline{B}$ is defined by:

$$\begin{aligned}
\underline{B} &= \max_{(s,d),(s',d')\in\mathcal{A}} \{|T[\tilde{V}](s,d) - T[\tilde{V}](s',d') - [\tilde{V}(s,d) - \tilde{V}(s',d')]|\} \\
&= \max_{(s,d)\in\mathcal{A}} \big(T[\tilde{V}](s,d) - \tilde{V}(s,d)\big) - \min_{(s',d')\in\mathcal{A}} \big(T[\tilde{V}](s',d') - \tilde{V}(s',d')\big)
\end{aligned}$$

A linear combination of two uniformly continuous functions is also uniformly continuous, and so $s \mapsto T[\tilde{V}](s,d) - \tilde{V}(s,d)$ is uniformly continuous for each $d \in \mathcal{D}$, there must exist some $\eta_1 > 0$ so that for every $d \in \mathcal{D}$:

$$\sup_{s,s' \in \mathcal{S}:||s-s'|| \leq \eta} |\big(T[\tilde{V}](s,d) - \tilde{V}(s,d)\big) - \big(T[\tilde{V}](s',d) - \tilde{V}(s',d)\big)| \leq \frac{1}{4}\tau$$

So suppose that for any $d \in \mathcal{D}$ and $s \in \mathcal{S}$ there is an s' with $||s - s'|| \leq \eta_1$ and $(s', d) \in \mathcal{A}$, then:

$$\sup_{s \in \mathcal{S}, d \in \mathcal{D}} \min_{s' \in \mathcal{S}:(s',d) \in \mathcal{A}} |\big(T[\tilde{V}](s,d) - \tilde{V}(s,d)\big) - \big(T[\tilde{V}](s',d) - \tilde{V}(s',d)\big)| \leq \frac{1}{4}\tau$$

And so:

$$\sup_{s \in \mathcal{S}, d \in \mathcal{D}} \big(T[\tilde{V}](s,d) - \tilde{V}(s,d)\big) - \max_{(s,d) \in \mathcal{A}} \big(T[\tilde{V}](s,d) - \tilde{V}(s,d)\big) \leq \frac{1}{4}\tau$$

And:

$$\inf_{s \in \mathcal{S}, d \in \mathcal{D}} \big(T[\tilde{V}](s,d) - \tilde{V}(s,d)\big) - \min_{(s,d) \in \mathcal{A}} \big(T[\tilde{V}](s,d) - \tilde{V}(s,d)\big) \leq \frac{1}{4}\tau$$

And so:

$$B - \underline{B} \leq \frac{1}{2}\tau$$

Next, recall the definition of $\bar{B}$:

$$\bar{B} = \sup_{s,s' \in S, d, d' \in D} \Bigg[\min_{(s^*,d^*) \in \mathcal{A}} \{T[\tilde{V}](s^*,d^*) + \delta(F_{s,d}, F_{s^*,d^*})\bar{b}_{\tilde{V}}\}$$

$$- \max_{(s^*,d^*) \in \mathcal{A}} \{T[\tilde{V}](s^*,d^*) - \delta(F_{s',d'}, F_{s^*,d^*})\bar{b}_{\tilde{V}}\} - [\tilde{V}(s,d) - \tilde{V}(s',d')]\Bigg]$$

$$= \sup_{s,s' \in S, d, d' \in D} \Bigg[\min_{(s^*,d^*) \in \mathcal{A}} \big\{T[\tilde{V}](s^*,d^*) - T[\tilde{V}](s,d)$$

$$+ \big(\delta(F_{s,d}, F_{s^*,d^*}) - \delta(F_{s,d}, F_{s,d})\big)\bar{b}_{\tilde{V}}\big\}$$

$$+ \min_{(s^*,d^*) \in \mathcal{A}} \big\{T[\tilde{V}](s',d') - T[\tilde{V}](s^*,d^*)$$

$$+ \big(\delta(F_{s',d'}, F_{s^*,d^*}) - \delta(F_{s',d'}, F_{s',d'})\big)\bar{b}_{\tilde{V}}\big\}$$

$$- [\big(\tilde{V}(s,d) - T[\tilde{V}](s,d)\big) + \big(T[\tilde{V}](s',d') - \tilde{V}(s',d')\big)]\Bigg]$$

Where we have used that $\delta(F_{s,d}, F_{s,d}) = \delta(F_{s',d'}, F_{s',d'}) = 0$. By the triangle inequality, the RHS of the final equality above is bounded by:

$$2 \sup_{s \in S, d \in D} \min_{(s^*,d^*) \in \mathcal{A}} |T[\tilde{V}](s^*,d^*) - T[\tilde{V}](s,d) + \big(\delta(F_{s,d}, F_{s^*,d^*}) - \delta(F_{s,d}, F_{s,d})\big)\bar{b}_{\tilde{V}}|$$

$$+B$$

A linear combination of two uniformly continuous functions is uniformly continuous and so $(s, s') \mapsto T[\tilde{V}](s,d) + \delta(F_{s',d}, F_{s,d})\bar{b}_{\tilde{V}}$ is uniformly continuous for all $d \in \mathcal{D}$. It follows that for some $\eta_2 > 0$:

$$\sup_{s,s' \in S: ||s-s'|| \leq \eta} |T[\tilde{V}](s',d) - T[\tilde{V}](s,d) + \big(\delta(F_{s,d}, F_{s',d}) - \delta(F_{s,d}, F_{s,d})\big)\bar{b}_{\tilde{V}}| \leq \frac{1}{4}\tau$$

For all $d \in \mathcal{D}$. So suppose that for any $d \in \mathcal{D}$ and $s \in \mathcal{S}$ there is an s' with $||s - s'|| \leq \eta_2$ and $(s', d) \in \mathcal{A}$, then:

$$\sup_{s \in S, d \in D} \min_{(s^*,d^*) \in \mathcal{A}} |T[\tilde{V}](s^*,d^*) - T[\tilde{V}](s,d) + \big(\delta(F_{s,d}, F_{s^*,d^*}) - \delta(F_{s,d}, F_{s,d})\big)\bar{b}_{\tilde{V}}| \leq \frac{1}{2}\tau$$

And so:

$$\bar{B} - B \leq \frac{1}{2}\tau$$

Finally, for dense enough $\mathcal{A}$, for any $d \in \mathcal{D}$ and $s \in \mathcal{S}$ there is an s' with $||s - s'|| \leq \min\{\eta_1, \eta_2\}$ and $(s', d) \in \mathcal{A}$ and hence:

$$\bar{B} - \underline{B} \leq \tau$$

τ was chosen arbitrarily and so as the density of $\mathcal{A}$ grows to infinity $\bar{B} - \underline{B} \to 0$. $\square$

Bibliography

[1] Abbring and ystein Daljord. Identifying the discount factor in dynamic discrete choice models. *Quantitive Economics*, 2020.

[2] Victor Aguirregabiria and Pedro Mira. Swapping the nested fixed point algorithm: A class of estimators for discrete markov decision models. *Econometrica*, 70:1519–1543, 2002.

[3] Victor Aguirregabiria and Pedro Mira. Sequential estimation of dynamic discrete games. *Econometrica*, 75:1–53, 2007.

[4] Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

[5] Chunrong Ai and Xiaohong Chen. Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141:5–43, 2007.

[6] JosephÂ G. Altonji, ToddÂ E. Elder, and ChristopherÂ R. Taber. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, 113:151–184, 2005.

[7] Donald WK Andrews. Examples of l2-complete and boundedly-complete distributions. *Journal of Econometrics*, 199(2):213–220, 2017.

[8] Isaiah Andrews, Matthew Gentzkow, and Jesse M. Shapiro. Measuring the sensitivity of parameter estimates to estimation moments. *The Quarterly Journal of Economics*, 132:1553–1592, 2017.

[9] Joshua D. Angrist and Alan B. Krueger. Empirical strategies in labor economics, chapter 23 handbook of labor economics, volume 3, 1999.

[10] Peter Arcidiacono, Patrick Bayer, Federico A. Bugni, and Jonathan James. Approximating high-dimensional dynamic models: Sieve value function iteration, 2013.

[11] Manuel Arellano and Stephen Bond. Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297, 1991.

[12] Manuel Arellano and Stephane Bonhomme. Nonlinear panel data estimation via quantile regressions. *The Econometrics Journal*, 19:C61–C94, 2016.

[13] Timothy B. Armstrong and Michal Kolesar. Sensitivity analysis using approximate moment condition models. *Cowles Foundation Discussion Papers*, 2018.

[14] Patrick Bajari, C. Lanier Benkard, and Jonathan Levin. Estimating dynamic models of imperfect competition. *Econometrica*, 75:1331–1370, 2007.

[15] R. Bellman. *Dynamic Programming.* Princeton University Press, 1957.

[16] Richard Bellman, Robert Kalaba, and Bella Kotkin. Polynomial approximation–a new computational technique in dynamic programming: Allocation processes. *Mathematics of Computation*, 17:155, 1963.

[17] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186:345–366, 2015.

[18] Richard Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica. Journal of the Econometric Society*, 75(6):1613–1669, 2007.

[19] StÃ©phane Bonhomme and Martin Weidner. Minimizing sensitivity to model misspecification. 2020.

[20] Ivan A. Canay, Andres Santos, and Azeem M. Shaikh. On the testability of identification in some nonparametric models with endogeneity. *Econometrica. Journal of the Econometric Society*, 81(6):2535–2559, 2013.

[21] Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. the handbook of econometrics, jj heckman and ee leamer (eds.), 6b, 2007.

[22] Xiaohong Chen, Victor Chernozhukov, Sokbae Lee, and Whitney K. Newey. Local identification of nonparametric and semiparametric models. *Econometrica*, 2014.

[23] Xiaohong Chen and Timothy M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9:39–84, 2018.

[24] Xiaohong Chen and Demian Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 2012.

[25] Xiaohong Chen and Demian Pouzo. Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, 83:1013–1079, 2015.

[26] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42:1787–1818, 2014.

[27] Victor Chernozhukov, Ivn FernÃ¡ndez-Val, Stefan Hoderlein, Hajo Holzmann, and Whitney Newey. Nonparametric identification in panels using quantiles. *Journal of Econometrics*, 188:378–392, 2015.

[28] Timothy G. Conley, Christian B. Hansen, Robert E. McCulloch, and Peter E. Rossi. A semi-parametric Bayesian approach to the instrumental variable problem. *Journal of Econometrics*, 144(1):276–305, 2008.

[29] Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference, November 2020.

[30] Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 2011.

[31] Carl de Boor. (b)asic-spline basics, 2014.

[32] Ben Deaner. Nonparametric estimation and identification in non-separable models using panel data. September 2018.

[33] Ben Deaner. Nonparametric instrumental variables estimation under misspecification. W, January 2019.

[34] Ronald A. DeVore and George G. Lorentz. *Constructive Approximation*. Springer-Verlag, 1993.

[35] Xavier D'Haultfoeuille. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27(3):460–471, 2011.

[36] R. M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1:290–330, 1967.

[37] Kirill Evdokimov. Identification and estimation of a nonparametric panel data model with unobserved heterogeneity, 2009.

[38] Jean-Pierre Florens. Non-parametric models with instrumental variables, 2011.

[39] Joachim Freyberger. On completeness and consistency in nonparametric instrumental variable models. *Econometrica. Journal of the Econometric Society*, 85(5):1629–1644, 2017.

[40] Joachim Freyberger. Non-parametric panel data models with interactive fixed effects. *Review of Economic Studies*, 85(3):1824–1851, 2018.

[41] Joachim Freyberger and Matthew A Masten. A practical guide to compact infinite dimensional parameter spaces. *Econometric Reviews*, 2019.

[42] Jane Cooley Fruehwirth, Salvador Navarro, and Yuya Takahashi. How the timing of grade retention affects outcomes: Identification and estimation of time-varying treatment effects. *Journal of Labor Economics*, 34:979–1021, 2016.

[43] Zvi Griliches. Estimating the returns to schooling: Some econometric problems. *Econometrica*, 45:1, 1977.

[44] Peter Hall and Joel L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33:2904–2929, 2005.

[45] Jerry A. Hausman and William E. Taylor. Panel data and unobservable individual effects. *Econometrica*, 49:1377, 1981.

[46] Douglas Holtz-Eakin, Whitney Newey, and Harvey S. Rosen. Estimating vector autoregressions with panel data. *Econometrica*, 56:1371, 1988.

[47] Joel L. Horowitz. Applied nonparametric instrumental variables estimation. *Econometrica. Journal of the Econometric Society*, 79(2):347–394, 2011.

[48] V. Joseph Hotz and Robert A. Miller. Conditional choice probabilities and the estimation of dynamic models. *Review of Economic Studies*, 60:497, 1993.

[49] Yingyao Hu and Susanne M. Schennach. Instrumental variable treatment of nonclassical measurement error models. *E*, 76:195–216, 2008.

[50] Yingyao Hu and Ji-Liang Shiu. Nonparametric identification using instrumental variables: sufficient conditions for completeness. *Econometric Theory*, 34(3):659–693, 2018.

[51] Yingyao Hu and Matthew Shum. Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics*, 171:32–44, 2012.

[52] Peter J. Huber. Robust statistics, 2011.

[53] Hidehiko Ichimura and Whitney K. Newey. The influence function of semiparametric estimators. 2017.

[54] Guido W. Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economics Review, P&P*, 93:126–132, 2003.

[55] Shane R Jimerson. Meta-analysis of grade retention research: Implications for practice in the 21st century. *School psychology review*, 30(3):420–437, 2001.

[56] Michael P. Keane, Petra E. Todd, and Kenneth I. Wolpin. The structural estimation of behavioral models: Discrete choice dynamic programming methods and applications, chapter 4, handbook of labor economics volume 4a, 2011.

[57] Michael P. Keane and Kenneth I. Wolpin. The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte carlo evidence. *The Review of Economics and Statistics*, 76(4):684–672, 1992.

[58] Rainer Kress. *Linear Integral Equations*. Springer, 1999.

[59] Arthur Lewbel. Engel curves. In *The New Palgrave Dictionary of Economics, 2nd Edition*. Palgrave, 2008.

[60] Thierry Magnac and David Thesmar. Identifying dynamic discrete decision processes. *Econometrica. Journal of the Econometric Society*, 70(2):801–816, 2002.

[61] Matthew A. Masten and Alexandre Poirier. Identification of treatment effects under conditional partial independence. *Econometrica*, July 2018.

[62] Wang Miao, Zhi Geng, and Eric J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105:987–993, 2018.

[63] Wang Miao, Xu Shi, and Eric Tchetgen Tchetgen. A confounding bridge approach for double negative control inference on causal effects, August 2018. First version appeared on Arxiv on 15th August 2018.

[64] Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71:1565–1578, 2003.

[65] Andriy Norets. Estimation of dynamic discrete choice models using artificial neural network approximations. *Econometric Reviews*, 31:84–106, 2012.

[66] Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37:187–204, 2019.

[67] Judea Pearl. *Causality: Models, Reasoning, and Inference (Second Edition)*. Cambridge University Press, 2009.

[68] David Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, 2001.

[69] Miikka AT Rokkanen. Exam schools, ability, and the effects of affirmative action: Latent factor extrapolation in the regression discontinuity design. Working paper, 2015.

[70] M. Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164:60–72, 1999.

[71] John Rust. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica*, 55:999, 1987.

[72] Andres Santos. Inference in nonparametric instrumental variables with partial identification. *Econometrica. Journal of the Econometric Society*, 80(1):213–275, 2012.

[73] Thomas A. Severini and Gautam Tripathi. Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors. *Journal of Econometrics*, 170:491–498, 2012.

[74] Xu Shi, Wang Miao, Jennifer C. Nelson, and Eric J. Tchetgen Tchetgen. Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society*, 82:521–540, 2020.

[75] Che-Lin Su and Kenneth L. Judd. Constrained optimization approaches to estimation of structural models. *Econometrica*, 2012.

[76] Viakalathur Shankar Sunder. *Operators on Hilbert space*, volume 71. Springer, 2016.

[77] Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning, September 2020. Appeared on Arxiv 23 Sep 2020.

[78] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[79] Daniel Wilhelm. Identification and estimation of nonparametric panel data regressions with measurement error. CENMAP working paper, June 2015.

[80] Kenneth I. Wolpin. An estimable dynamic stochastic model of fertility and child mortality. *Journal of Political Economy*, 92:852–874, 1984.