Proxy Controls and Panel Data

Ben Deaner *

October 20, 2020

Abstract

We present a flexible approach to identification and estimation of causal objects in nonparametric, non-separable models using 'proxy controls': covariates that do not satisfy a standard 'unconfoundedness' assumption but are informative proxies for variables that do. Our analysis applies to cross-sectional settings but is particularly well-suited to panel models. Our identification results motivate a simple and 'well-posed' nonparametric estimator. We derive convergence rates for the estimator and construct uniform confidence bands with asymptotically correct size. In panel settings, our 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 our key identifying assumptions follow from restrictions on the serial dependence structure. We apply our methodology to two empirical settings. We estimate causal effects of grade retention on cognitive performance using cross-sectional variation and we estimate consumer demand counterfactuals using panel data.

A sizeable 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

*PhD candidate, Massachusetts Institute of Technology. Email at bdeaner@mit.edu. Previous editions of this work can be found at https://arxiv.org/abs/1810.00283 (first version September 30th 2018). Many thanks to my advisers Whitney Newey, Anna Mikusheva, and Jerry Hausman for invaluable advice. Thanks to those who attended the Econometric society session C1 at the ASSA 2020 Annual Meeting in San Diego for useful feedback and likewise to those who attended via Zoom the Treatment Effects 2 session of the 2020 Econometric Society World Congress. Thanks as well to those who attended the many internal presentations of this project at MIT.

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 Angrist & Krueger (1999) and see Griliches (1977) 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 effectst. An emerging literature considers proxy controls in non-linear and nonparametric settings.

We develop new nonparametric identification results in the context of proxy controls. Whereas existing work nonparametrically identifies the average structural function, we identify the conditional (on observed treatments) average structural function (CASF). Identification of the CASF is necessary if we wish to identify say, the effect of treatment on the treated. 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.

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. To the best of our knowledge this is the first nonparametric treatment of estimation and inference in the context of proxy controls. 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 nonparametric regression. We give conditions under which our estimator can be asymptotically approximated by a zero-mean 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.

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.

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 Fruehwirth *et al.* (2016) 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

In the biometrics literature, Miao *et al.* (2018) nonparametircally identify an average the structural function in cross-sectional settings when controls are mismeasured but do not consider estimation.[1] Our identification strategy is similar to that of Miao *et al.* (2018). However, we add to their results by identifying the conditional average structural function (CASF), establishing well-posedness of our characterization of the CASF, and using our results to identify causal objects in panel models under panel-type assumptions (i.e., assumptions on the serial dependce structure of the data).

Hu & Schennach (2008) provide identification results for nonparametric and nonseparable models with measurement error and present a related estimator. [2] Unlike Hu & Schennach (2008), 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 Hu & Schennach (2008). In our case the measurement error is only in control variables and not in treatment variables.

[1] A recent working paper Miao *et al.* (n.d.) by the same authors considers parametric estimation of the average treatment effect by GMM with proxy controls. We thank Rahul Singh for bringing this work to our attention.

[2] A working paper Rokkanen (2015) employs results from Hu & Schennach (2008) to achieve identification using proxy controls in a regression discontinuity setting.

We are uninterested in the causal effect of the latent perfect controls themselves.

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 Hausman & Taylor (1981), Holtz-Eakin *et al.* (1988), and Arellano & Bond (1991) and some more recent work that allows for nonlinearity like Arellano & Bonhomme (2016), Freyberger (2018) and Evdokimov (2009).

An extensive literature examines the effects of grade retention on cognitive and social success. For a meta-analysis see Jimerson (2001). We build on the work of Fruehwirth *et al.* (2016). 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 Blundell *et al.* (2007) and the panel approach of Chernozhukov *et al.* (2015). For a short survey see Lewbel (2008).

1 General Model and Identification

Consider the following structural model:

$$Y = y_0(X, U) \tag{1.1}$$

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) 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 $\bar{y}(x_1|x_2)$.[3]

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.[4]

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 fixed scalar, let $\tilde{Y}=1\{Y\le y\}$, and let $\tilde{y}_0(x,u)=1\{y_0(x,u)\le y\}$. 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:

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

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. 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.

[3]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.

[4]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.

When perfect controls W^* are unavailable the researcher may have access to proxy controls. In particular, we suppose that the researcher has access to two vectors of proxies V and Z. We present assumptions that imply identification when only proxy controls are available. 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. To argue persuasively that the assumptions are plausible in a given setting, a researcher will generally have to choose a particular set of unobserved perfect controls W^* and argue that the assumptions hold for those controls.

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. 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. $V \perp\!\!\!\perp (X, Z)|W^*$

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 $F_{W^*|X=x,\bar{W}=\bar{w}}$-almost sure.

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. If the remaining variation in Z is measurement error, then the measurement error must be independent of potential outcomes. Assumption 2.ii states that after controlling for W^*, any remaining variation in V is independent of remaining variation in (X, Z). V and (X, Z) could be very strongly correlated, we simply require that this correlation is explained by mutual association with W^*.

Note that the conditions are asymmetric in V and Z. V can be correlated with potential outcomes U even controlling for the perfect controls whereas Z

cannot be. On the other hand, V must be independent of X (conditional on W^*) whereas the relationship between Z and X is unrestricted; Z could codetermine X and vice-versa.

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 (Andrews (2017)). Completeness is used to achieve identification in the non-parametric instrumental variables (NPIV) models of Newey & Powell (2003) and Ai & Chen (2003).

In the NPIV context, completeness is an instrumental relevance condition analogous to the rank condition in linear IV (see Newey & Powell (2003)). 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^*.

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. [5] In our setting the relevant order condition is that each of the vectors V and Z be of a weakly larger dimension than W^*.

Assumption 3.ii differs from the corresponding assumption in Miao *et al.* (2018) (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 Miao *et al.* (2018) 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, our condition allows for the possibility that Z contains components that are entirely independent of V whereas their assumption precludes this.[6]

For discussion of Assumptions 1, 2, and 3 in the context of our empirical applications see Section 4. To provide further intuition, we show how these assumptions map into simple parametric cases in Appendix A.2.

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

[5] One such class of distributions is the conditional Gaussian class discussed in Newey & Powell (2003).

[6] 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 Miao *et al.* (2018)) holds, this weakened version of Assumption 3.ii is weaker than their condition 3 in a formal mathematical sense.

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.[7] 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:

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

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:

$$E[Y - \gamma(X, V)|X, Z] = 0 \tag{1.2}$$

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:

$$E[\varphi(x_1, x_2, Z)|X = x_1, V] = \frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(V) \tag{1.3}$$

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

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

that of Miao *et al.* (2018) 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 intution 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*:*

$$\begin{aligned}&\big(\bar{y}(x_1|x_2) - E[\tilde{\gamma}(x_1, V)|X = x_2]\big)^2 \\ \leq& C(x_1, x_2)E\left[\big(E[Y - \tilde{\gamma}(X, V)|X, Z]\big)^2 \middle| X = x_1\right]\end{aligned} \tag{1.4}$$

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*:*

$$\begin{aligned}&\big(\bar{y}(x_1|x_2) - E[Y\tilde{\varphi}(x_1, x_2, Z)|X = x_1]\big)^2 \\ \leq& D(x_1)E\left[\left(E[\tilde{\varphi}(X, x_2, Z)|X, V] - \frac{dF_{V|X=x_2}}{dF_{V|X=x_1}}(V)\right)^2 \middle| X = x_1\right]\end{aligned} \tag{1.5}$$

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

$$E[Y - \gamma(X, V)|X, Z] = 0 \tag{1.6}$$

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 (1.6). In a second step Theorems 1.1 and 1.2 suggests we estimate the CASF by $\hat{E}_V\left[\hat{\gamma}(x_1, V)|X = x_2\right]$, 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 (1.6) with small error (in a mean-squared sense), then $E_V\left[\hat{\gamma}(x_1, V)|X = x_2\right]$ is close to the CASF $\bar{y}(x_1|x_2)$. If, in addition, $E_V\left[\hat{\gamma}(x_1, V)|X = x_2\right]$ is close to the sample analogue $\hat{E}_V\left[\hat{\gamma}(x_1, V)|X = x_2\right]$, 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 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 Blundell *et al.* (2007), Chen & Pouzo (2012)).

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 (1.3) 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 (Ai & Chen (2003), Severini & Tripathi (2012), Ichimura & Newey (2017)) and to robust estimation (Deaner (2019)) 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 (1.3). 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.

2 Panel Models

The analysis in the previous section considers the model (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 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, treatements and outcomes in other periods are associatied 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) 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 $(Y_1, ..., Y_T)$. [8]

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 $V \perp\!\!\!\perp (X_t, Z)|W^*$ [9]

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

[9] 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 $U_t \perp\!\!\!\perp Z|(X_t, W^*)$.

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.

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 conditional independence restriction holds:

$$U_t \perp\!\!\!\perp (X_1, ..., X_t)|\tilde{W}^* \tag{2.1}$$

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.

Let $\lfloor t/2 \rfloor$ denote the largest natural number weakly less than half of t. Suppose that conditional on the latent variables $\tilde{W}^*$, the regressors satisfy a first-order Markov dependence structure at $\lfloor t/2 \rfloor$. Formally:

$$(X_1, ..., X_{\lfloor t/2 \rfloor -1}) \perp\!\!\!\perp (X_{\lfloor t/2 \rfloor +1}, ..., X_T)|(\tilde{W}^*, X_{\lfloor t/2 \rfloor}) \tag{2.2}$$

That is, conditional on the latent variables $\tilde{W}^*$, the treatment assignments for periods strictly prior to the given period $\lfloor t/2 \rfloor$ are only only related to treatments after $\lfloor t/2 \rfloor$ through the treatment at $\lfloor t/2 \rfloor$.

Proposition 2.1 below shows that in this setting Assumptions 1.i and 2 hold for V and Z composed of particular lagged treatments.

Proposition 2.1. *Suppose that (2.1) and (2.2) hold. Set $V = (X_1, ..., X_{\lfloor t/2 \rfloor})$, $Z = (X_{\lfloor t/2 \rfloor}, ..., X_{t-1})$, and $W^* = (\tilde{W}^*, X_{\lfloor t/2 \rfloor})$. 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 $X_{\lfloor t/2 \rfloor}$ as an observable perfect control, we therefore include it in both V and Z. Given the Markov structure, conditioning on the treatment at period $\lfloor t/2 \rfloor$ removes the dependence between V and Z.

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:

$$U_t \perp\!\!\!\perp (X_1, ..., X_t, U_1, ..., U_{t-1})|\tilde{W}^* \tag{2.3}$$

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

$$(X_1, ..., X_{\lfloor t/2 \rfloor - 1}, U_1, ..., U_{\lfloor t/2 \rfloor - 1}) \perp\!\!\!\perp (X_{\lfloor t/2 \rfloor + 1}, ..., X_t, U_{\lfloor t/2 \rfloor + 1}, ..., U_t) \tag{2.4}$$

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 (2.3), (2.4) and for all x, $y_{0,t}(x, \cdot)$ is injective. Set $W^* = (\tilde{W}^*, X_{\lfloor t/2 \rfloor}, Y_{\lfloor t/2 \rfloor})$ and set V and Z as follows:*

$$V = (X_1, ..., X_{\lfloor t/2 \rfloor}, Y_1, ..., Y_{\lfloor t/2 \rfloor})$$
$$Z = (X_{\lfloor t/2 \rfloor}, ..., X_{t-1}, Y_{\lfloor t/2 \rfloor}, ..., Y_{t-1})$$

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

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 $W^* = (\tilde{W}^*, X_{\lfloor t/2 \rfloor})$. Both Z and V contain $X_{\lfloor t/2 \rfloor}$ 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.

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^*. In the first-order Markov treatment assignment example when 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 $\lfloor T/2 \rfloor - 1$.

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 Chen & Pouzo (2012) and Chen & Pouzo (2015)). Inference is based on the multiplier bootstrap (see for example Belloni *et al.* (2015)). 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:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}_i - \hat{\pi}_{n,i}'\theta\right)^2 + \lambda_{0,n}Pen(\theta) \tag{3.1}$$

The estimate of the CASF is then given by:

$$\bar{y}(x_1|x_2) \approx \hat{\alpha}_n(x_1,x_2)'\hat{\theta} \tag{3.2}$$

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 (3.1) is minimized by setting $\hat{\theta} = \hat{\Sigma}_{\lambda_{0,n}}^{-1}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}\hat{g}_i$.

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:

$$\begin{aligned}\hat{g}_i =& \zeta_{n,i}'\hat{\Xi}_{\lambda_{1,n}}^{-1}\frac{1}{n_g}\sum_{j\in\mathcal{I}_g}\zeta_{n,j}Y_j \qquad (3.3)\\ \hat{\pi}_{n,i} =& \big(\psi_{n,i}'\hat{\Omega}_{\lambda_{2,n}}^{-1}\frac{1}{n_\pi}\sum_{j\in\mathcal{I}_\pi}\psi_{n,j}\rho_{n,j}\big)\otimes\chi_{n,i}\\ \hat{\alpha}_n(x_1,x_2) =& \big(\chi_n(x_2)'\hat{G}_{\lambda_{3,n}}^{-1}\frac{1}{n_\alpha}\sum_{j\in\mathcal{I}_\alpha}\chi_{n,j}\rho_{n,j}\big)\otimes\chi_n(x_1)\end{aligned}$$

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 $\Lambda_s^d(c)$ be the space of smooth functions defined as follows. For any vector $q = (q_1,q_2,...,q_d)\in\mathbb{N}_0^d$, 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 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 finally let us define

$\xi_{\tilde{\psi},n} = \operatorname{ess\,sup} ||\tilde{\Omega}_n^{-1/2}\tilde{\psi}_{n,i}||$.

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 $\delta \in \Lambda_s^{\dim(V)}(c)$:

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

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 $E\left[\left(\delta(X) - \chi_n(X)'\beta\right)^2\right]$ then:

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

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 $E\left[\left(\delta(X,Z) - \psi_n(X,Z)'\beta\right)^2\right]$:

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

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 $E\left[\left(\delta(X,Z) - \zeta_n(X,Z)'\beta\right)^2\right]$:

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

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. $g(\cdot,z) \in \Lambda_{s_3}^{\dim(X)}(c_3)$, 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. $\frac{\xi_{\tilde{\psi},n}^2 \log(m(n)l(n))}{n_g} \prec 1$

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 DeVore & Lorentz (1993)). [10]

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 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 (Rudelson (1999)).

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 (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*

[10] 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 Chernozhukov *et al.* (2014)).

coefficients:

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

Where $R_{\theta,n} = \ell_{\rho,n}(s_1)$ *if* X *has finite discrete support and otherwise we have* $R_{\theta,n} = \ell_{\rho,n}(s_1) + \big(\xi_{\rho,n}\ell_{\chi,n}(1)\big)^{\tilde{s}}$, *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 (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 nonparaemtric regression are attainable. Suppose $\lambda_{0,n}$ is chosen to optimize the rate in Theorem 3.1, then rate simplifies to:[11]

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

For many commonly used bases $\frac{\bar{\mu}_n}{\underline{\mu}_n}$ is bounded above and away from zero under very weak conditions (Belloni *et al.* (2015)), and in the discrete treatment case $\xi_{\tilde{\Omega},n}(x_1)$ is bounded above. Then estimator converges as quickly as either the slowest first-stage nonparametric regression or $R_{\theta,n}$.

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

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

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

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.[12] 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 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 $E_V[|\phi_n(x_1,V)'\hat{\theta}|^2]^{1/2}$.

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.

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

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 $\tilde{\alpha}_n(x_1,x_2) = E\big[\varphi(x_1,x_2,Z_i)\hat{\pi}_{n,i}\big|\mathcal{I}_\pi, X_i = x_1\big]$ 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:

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

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 (3) and $\hat{g}$, $\hat{\pi}_n$, and $\hat{\alpha}_n$ as in (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:*

$$\begin{aligned}
&\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{||s_n(x_1,x_2)||} - \frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||}\mathcal{N}_n \\
&\precsim_p R_{\mathcal{N},n} + \xi_{\zeta,n}\frac{p(n)}{n_g} + b_n(R_{\theta,n} + R_{\pi,n} + \ell_{\zeta,n}(s_3) + \lambda_{1,n}||\Xi_n^{-1}||) \\
&+ \frac{\sqrt{n_g}}{\underline{c}_n(x_1,x_2)}R_n(x_1) \qquad (3.4)
\end{aligned}$$

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.[13] 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 $\frac{\sqrt{n_g}}{\underline{c}_n(x_1,x_2)}R_n(x_1) \to 0$.

3.3 Uniform Confidence Bands

If the RHS of (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 (3) and let $\hat{g}$, $\hat{\pi}_n$, and $\hat{\alpha}_n$ be given by the forumlas in (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.[14] 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 confidence band is an interval-valued random function $\hat{\Theta}_{1-a}$ of the form:

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

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:

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

[13] Strictly speaking $R_n(x_1)$ captures more than just bias. It captures some of the error of the Guassian 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.

[14] We follow (Belloni *et al.* (2015)) and use the standard exponential, but $Q_{b,i}$ may have any distribution with mean and variance both equal to 1 and $\max_{i\in\mathcal{I}_g}|Q_{b,i}| \precsim_p ln(n_g)$.

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, suppose that $E\big[\sup_{(x_1,x_2)\in\mathcal{X}} |\frac{s_n(x_1,x_2)}{||s_n(x_1,x_2)||}'\mathcal{N}_n|\big] \precsim \bar{r}_n^{-1}$ and $\sup_{(x_1,x_2)\in\mathcal{X}} |\frac{\hat{\sigma}(x_1,x_2)}{||s_n(x_1,x_2)||} - 1| \prec_p \bar{r}_n^2$. 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 (Chernozhukov *et al.* (2014)). In the case in which $\mathcal{X}$ is a finite set they hold trivially.

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.

4.1 Causal Impact of Grade Retention

Fruehwirth *et al.* (2016) 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 Fruehwirth *et al.* (2016), we take our outcome variables to be the tests scores in reading and math when aged approximately eleven. Also in line with Fruehwirth *et al.* (2016) 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 Fruehwirth *et al.* (2016) contains only students who are retained at most once in the sample period and no students who skip a grade.

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 in the past may be associated with poor cognitive ability in the future, with retention status held fixed. Thus a difference of mean outcomes is likely overly negative compared to the true causal effect of retention.

The ECLS-K dataset contains scores that measure a student's behavioral and social skills in kindergarten as well as the student's performance in a range of cognitive tests at different ages. To account for the confounding Fruehwirth *et al.* (2016) 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. Fruehwirth *et al.* (2016) 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 early academic achievement 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. We let the set of proxies V contain the student's scores on the behavioral and social skills tests in kindergarten and Z contain the kindergarten cognitive test scores.[15]

Assumption 3 essentially requires that both the test scores Z and the behavioral scores V would each be relevant instruments for W^*. Z directly measures academic success and so Z and W^* are likely closely associated. V measures behavioral and social skills which likely reflect brain development and impact learning, thus V and W^* are also plausibly closely associated. Each of Z and V are three dimensional, so the order condition requires W^* comprise at most three latent factors.

In this context Assumptions 1.i requires that retention status and potential age-eleven test scores are only related due to mutual dependence on underlying kindergarten academic success. Assumption 2.i requires that a student's kindergarten test scores are only related to the student's potential age-eleven test scores through the mutual dependence on classroom success and retention status. This is plausible if individual-specific noise in Z results from day of the test factors that are unlikely to be associated with potential outcomes.

Assumption 2.ii requires that the student's behavioral scores are only related to cognitive test scores and retention status through mutual association with the student's underlying academic success in kindergarten. This is plausible if say, noise in Z reflects from exogenous day-of-test factors and retention decisions depend only on teachers' and parents' assessments of the child's academci progress W^* as well as attitudes toward retention which are independent of the behavioral scores V. Note that our assumptions allow for the possibility that retention decisions X depend directly on the scores in Z.

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

[15]As Fruehwirth *et al.* (2016) note, these scores are from tests taken prior to any retention.

corresponds roughly to Table 4 in Fruehwirth *et al.* (2016).

Table 1: Effects of Grade Retention on Cognitive Performance

$n = 1998$	Observed retention status:			
Difference from non-retention:	Not retained	Retained kindergarten	Retained early	Retained late
Retained kindergarten	-0.03	0.09	0.11	-0.07
	(0.04)	(0.04)	(0.06)	(0.03)
Retained early	0.04	0.04	0.11	0.00
	(0.03)	(0.03)	(0.04)	(0.02)
Retained late	-0.06	-0.10	-0.08	-0.01
	(0.05)	(0.13)	(0.18)	(0.08)

(a) Reading Ability

$n = 1999$	Observed retention status:			
Difference from non-retention:	Not retained	Retained kindergarten	Retained early	Retained late
Retained kindergarten	0.01	0.17	0.21	0.04
	(0.05)	(0.05)	(0.07)	(0.02)
Retained early	0.03	0.09	0.18	-0.02
	(0.04)	(0.04)	(0.06)	(0.02)
Retained late	-0.11	-0.20	-0.37	-0.10
	(0.09)	(0.21)	(0.27)	(0.10)

(b) Math Ability

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 1000 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.

We estimate that retention in kindergarten and retention in first or second grade, and in third or fourth grade raises the average scores for reading and math for those students who were in fact retained at these ages. These figures are statistically significant at the 95% level. In other words, the average effect of treatment on the treated (ATT) is estimated to be positive is in these cases. This contrasts with Fruehwirth *et al.* (2016) who estimate mostly negative ATTs.

By contrast the ATT for retention in grades three and four are estimated to be negative, for both the reading and math outcomes. However, the estimates are not statistically significant.

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 mostly small positive and negative effects, however none of these figures are not statistically significant.

In all, our results paint a more positive picture of schools' retention policies than those of Fruehwirth *et al.* (2016). In nearly all cases they find those who are retained on average suffered as a result and those who were not retained would have benefited from retention. We do not observe significant positive counterfactual effect for those not retained and we find that the students retained earlier on in their academic careers (in kindergarten and in first or second grade) benefited from their retention.

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 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 healthcare.

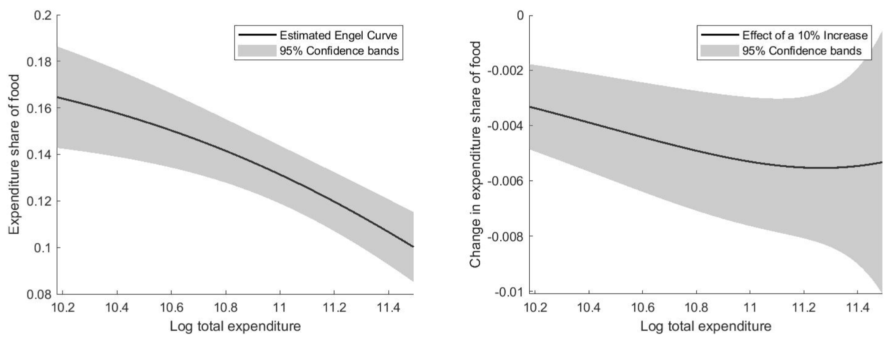
We apply the approach to identification with fixed-T panels described in Section 2, in particular the Markov treatment assignment and predetermination case. Let X_t denote the total expenditure in period t, which is the treatment in this setting. We aim to estimate the average and conditional average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

In this setting, the Markov treatment assumption requires that total expenditure in the past and in the future are only related through the expenditure today and the household's consumption preferences. Conditioning on today's expenditure X_t is important because conditioning on preferences household, expenditure may still be serially correlated due to persistence in household assets. Furthermore, total expenditure may reflect wages and labor supply which may be serially correlated even controlling for preferences.

In this context predetermination requires that time-varying preferences for spending money on food are only related to the history of total expenditure through the mutual association with the underlying time-invariant preferences of the household $\tilde{W}^*$.

Given these assumptions 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: 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 1.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 1.b we estimate $\bar{y}(1.1x, x) - \bar{y}(x, x)$ for each x. The uniform confidence bands are evaluated using 1000 replications of the multiplier boostrap as detailed in Section 3. The pointwise standard errors were set equal to pointwise standard deviations over the bootstrap replications.

Sub-Figure 1.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 1.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.

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.

References

Ai, Chunrong, & Chen, Xiaohong. 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71**(6), 1795–1843.

Andrews, Donald WK. 2017. Examples of L2-complete and boundedly-complete distributions. *Journal of Econometrics*, **199**(2), 213–220.

Angrist, Joshua D., & Krueger, Alan B. 1999. *Empirical Strategies in Labor Economics, Chapter 23 Handbook of Labor Economics, Volume 3.*

Arellano, Manuel, & Bond, Stephen. 1991. 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.

Arellano, Manuel, & Bonhomme, StÃ©phane. 2016. Nonlinear panel data estimation via quantile regressions. *The Econometrics Journal*, **19**, C61–C94.

Belloni, Alexandre, Chernozhukov, Victor, Chetverikov, Denis, & Kato, Kengo. 2015. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, **186**, 345–366.

Blundell, Richard, Chen, Xiaohong, & Kristensen, Dennis. 2007. Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica. Journal of the Econometric Society*, **75**(6), 1613–1669.

Chen, Xiaohong, & Pouzo, Demian. 2012. Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals. *Econometrica.*

Chen, Xiaohong, & Pouzo, Demian. 2015. Sieve Wald and QLR Inferences on Semi/Nonparametric Conditional Moment Models. *Econometrica*, **83**, 1013–1079.

Chernozhukov, Victor, Chetverikov, Denis, & Kato, Kengo. 2014. Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, **42**, 1787–1818.

Chernozhukov, Victor, FernÃ¡ndez-Val, Ivn, Hoderlein, Stefan, Holzmann, Hajo, & Newey, Whitney. 2015. Nonparametric identification in panels using quantiles. *Journal of Econometrics*, **188**, 378–392.

Darolles, Serge, Fan, Yanqin, Florens, Jean-Pierre, & Renault, Eric. 2011. Nonparametric Instrumental Regression. *Econometrica.*

Deaner, Ben. 2019. *Nonparametric Instrumental Variables Estimation Under Misspecification.* W.

DeVore, Ronald A., & Lorentz, George G. 1993. *Constructive Approximation.* Springer-Verlag.

Evdokimov, Kirill. 2009. *Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity.*

Freyberger, Joachim. 2018. Non-parametric panel data models with interactive fixed effects. *Review of Economic Studies*, **85**(3), 1824–1851.

Fruehwirth, Jane Cooley, Navarro, Salvador, & Takahashi, Yuya. 2016. How the Timing of Grade Retention Affects Outcomes: Identification and Estimation of Time-Varying Treatment Effects. *Journal of Labor Economics*, **34**, 979–1021.

Griliches, Zvi. 1977. Estimating the Returns to Schooling: Some Econometric Problems. *Econometrica*, **45**, 1.

Hall, Peter, & Horowitz, Joel L. 2005. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, **33**, 2904–2929.

Hausman, Jerry A., & Taylor, William E. 1981. Panel Data and Unobservable Individual Effects. *Econometrica*, **49**, 1377.

Holtz-Eakin, Douglas, Newey, Whitney, & Rosen, Harvey S. 1988. Estimating Vector Autoregressions with Panel Data. *Econometrica*, **56**, 1371.

Hu, Yingyao, & Schennach, Susanne M. 2008. Instrumental Variable Treatment of Nonclassical Measurement Error Models. *E*, **76**, 195–216.

Ichimura, Hidehiko, & Newey, Whitney K. 2017. *The influence function of semiparametric estimators.*

Jimerson, Shane R. 2001. Meta-analysis of grade retention research: Implications for practice in the 21st century. *School psychology review*, **30**(3), 420–437.

Kress, Rainer. 1999. *Linear Integral Equations.* Springer.

Lewbel, Arthur. 2008. Engel Curves. *In: The New Palgrave Dictionary of Economics, 2nd Edition.* Palgrave.

Miao, Wang, Shi, Xu, & Tchetgen, Eric Tchetgen. *A Confounding Bridge Approach for Double Negative Control Inference on Causal Effects.*

Miao, Wang, Geng, Zhi, & Tchetgen, Eric J. Tchetgen. 2018. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, **105**, 987–993.

Newey, Whitney K., & Powell, James L. 2003. Instrumental Variable Estimation of Nonparametric Models. *Econometrica*, **71**, 1565–1578.

Pollard, David. 2001. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

Rokkanen, Miikka AT. 2015. *Exam schools, ability, and the effects of affirmative action: Latent factor extrapolation in the regression discontinuity design.* Working paper.

Rudelson, M. 1999. Random Vectors in the Isotropic Position. *Journal of Functional Analysis*, **164**, 60–72.

Severini, Thomas A., & Tripathi, Gautam. 2012. Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors. *Journal of Econometrics*, **170**, 491–498.

Sunder, Viakalathur Shankar. 2016. *Operators on Hilbert space.* Vol. 71. Springer.

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 $A_{x,\bar{w}} : L_2(F_{\tilde{V}|X=x,\bar{W}=\bar{w}}) \to L_2(F_{\tilde{Z}|X=x,\bar{W}=\bar{w}})$ 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 $(X, \bar{W}) = (x, \bar{w})$.[16] 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 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

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

from the support of $\tilde{V}$ conditonal on $(X, \bar{W}) = (x, \bar{w})$. $v_k^{(x,\bar{w})}$ is a scalar valued function that maps from the support of $\tilde{Z}$ conditonal 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 $A_{x,\bar{w}}$.[17]

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 $E[\tilde{C}(x_1, x_2, \bar{W})|X = x_1] = C(x_1, x_2) < \infty$:

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

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

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

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 Hall & Horowitz (2005)). Conditions of the same form are used elsewhere in the literature, for example in Darolles *et al.* (2011) and Miao *et al.* (2018).

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 Griliches (1977).

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. Identification of the CASF is equivalent to identification of β. We also assume

[17] See, e.g., Kress (1999) Theorem 15.16 and associated discussion.

linear specifications for the heterogeneity U, and proxy controls V in terms of the perfect controls W^* and that Z is linear in W^* and treatments X. That is $U = \omega' W^* + \varepsilon$, $V = BW^* + \upsilon$, and $Z = CW^* + DX + \eta$. Where A, B, C, and D are matrices of parameters and (by linearity) $E[\varepsilon|W^*] = 0$, $E[\upsilon|W^*] = 0$, and $E[\eta|W^*, X] = 0$. Assumption 3.i implies $B'B$ and $C'C$ are both non-singular, in which case the model above implies the linear IV model:

$$\begin{aligned} Y &= X'\beta + V'\delta + \epsilon \\ V &= FZ + GX + e \end{aligned}$$

In the first equation $\delta = B(B'B)^{-1}\omega$ and $\epsilon = \varepsilon - \omega'(B'B)^{-1}B'\upsilon$. In the seciond equation $F = B(C'C)^{-1}C'$, $G = -B(C'C)^{-1}C'D$ and the residual satisfies $e = \upsilon - B(C'C)^{-1}C'\eta$.

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 $E[\upsilon|X, Z] = 0$ and $E[\eta|X, Z] = 0$ by the linear specification. Therefore $E[\epsilon|X, Z] = 0$ and $E[e|X, Z] = 0$. One can show that for β (and not necessarily δ) 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 $M = -D$.

Thus β is identified in the linear IV model above and can be recovered by standard IV methods. This is the approach of Griliches (1977). 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 Griliches (1977). 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 $\gamma(x, v) = x'\beta + v'\delta$.

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 U_1, U_2, and V in W^* and of Z in W^* and X. We have $Y = X'U_1 + U_2$, $U_1 = \omega_1 W^* + \varepsilon_1$, $U_2 = \omega_2 W^* + \varepsilon_2$, $V = BW^* + \upsilon$, and $Z = CW^* + DX + \eta$. Where $E[\varepsilon_1|W^*] = E[\varepsilon_2|W^*] = 0$, $E[\upsilon|W^*] = 0$, 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:

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

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 $\big(x', (x \otimes x)'\big)$.

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:

$$E\big[\frac{dF_{W^*|X=x_2}}{dF_{W^*|X=x_1}}(W^*)\big|X = x_1, V = v\big] = E\big[\frac{dF_{V|W^*,X=x_1}}{dF_{V|X=x_1}}(v, W^*)\big|X = x_2\big] \quad (.1)$$

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:

$$E[\varphi(x_1, x_2, Z)|W^*, X = x_1] = \frac{dF_{W^*|X=x_2}}{dF_{W^*|X=x_1}}(W^*) \quad (.2)$$

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

$$\begin{aligned}E[y_0(x_1,U)|X=x_2] &= E\big[E[y_0(x_1,U)|X=x_2,W^*]\big|X=x_2\big] \\ &= E\big[E[y_0(X,U)|X=x_1,W^*]\big|X=x_2\big] \end{aligned} \tag{.3}$$

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 (.3) and for the Radon-Nikodym derivative on the RHS by (.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 expecations and the final by the definition of Y. Combining we get:

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

□

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:

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

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

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

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:

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

Applying Cauchy-Schwartz:

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

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 $\tilde{\varphi}$ that satisfies $E[\tilde{\varphi}(x_1, x_2, Z)^2|X = x_1] \leq \infty$ 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:

$$\begin{aligned}&E[\tilde{\varphi}(x_1, x_2, Z)Y|X = x_1] \\ =&E\big[\tilde{\varphi}(x_1, x_2, Z)E[\gamma(X, V)|X, Z]\big|X = x_1\big] \\ =&E\big[E[\tilde{\varphi}(x_1, x_2, Z)|X, V]\gamma(X, V)\big|X = x_1\big] \end{aligned} \tag{.4}$$

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

$$\bar{y}(x_1|x_2) = E[\gamma(x_1, V)|X = x_2] \tag{.5}$$

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 (.4) and (.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 $E\big[\gamma(x_1, V)^2|X = x_1\big] \leq D(x_1)$. □

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:

$$W_1 \perp\!\!\!\perp W_2|(W_3, W_4) \implies W_1 \perp\!\!\!\perp (W_2, W_3)|(W_3, W_4) \tag{.6}$$
$$W_1 \perp\!\!\!\perp (W_2, W_3)|W_4 \implies W_1 \perp\!\!\!\perp (W_2, W_3)|(W_3, W_4) \tag{.7}$$
$$W_1 \perp\!\!\!\perp (W_2, W_3)|W_4 \implies W_1 \perp\!\!\!\perp W_2|W_4 \tag{.8}$$

Proof of Proposition 2.1. By supposition:

$$(X_1, ..., X_{\lfloor t/2 \rfloor -1}) \perp\!\!\!\perp (X_{\lfloor t/2 \rfloor +1}, ..., X_T)|(\tilde{W}^*, X_{\lfloor t/2 \rfloor}) \tag{.9}$$

$$U_t \perp\!\!\!\perp (X_1, ..., X_t)|\tilde{W}^* \tag{.10}$$

Using (.7), (.10) implies $U_t \perp\!\!\!\perp (X_1, ..., X_t)|(\tilde{W}^*, X_{\lfloor t/2 \rfloor})$ Applying (.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 (.6) to (.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:

$$U_t \perp\!\!\!\perp (X_1, ..., X_t, U_1, ..., U_{t-1})|\tilde{W}^* \tag{.11}$$

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

$$(X_1, ..., X_{\lfloor t/2 \rfloor -1}, U_1, ..., U_{\lfloor t/2 \rfloor -1}) \perp\!\!\!\perp (X_{\lfloor t/2 \rfloor +1}, ..., X_t, U_{\lfloor t/2 \rfloor +1}, ..., U_t) \tag{.12}$$

For all x and $u_1 \neq u_2$, $y_{0,t}(x, u_1) \neq y_{0,t}(x, u_2)$. So from (.11):

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

Applying (.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 (.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 (.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 (.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:

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

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

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

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 $E\big[|\phi_n(V,X)'\hat{\theta}|^2\big]^{1/2}\precsim\frac{\bar{\mu}_n}{\underline{\mu}_n}+\frac{\bar{\mu}_nR_n}{\lambda_{0,n}^{1/2}}$ and likewise for $E\big[|\phi_n(V,X)'(\theta_n-\hat{\theta})|^2\big]^{1/2}$, 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:

$$\begin{aligned}
&\big|\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big) - \tilde{\alpha}_n(x_1,x_2)'(\theta_n - \hat{\theta})\big| \\
\precsim_p & R_{\theta,n} + \big(\frac{\bar{\mu}_n}{\underline{\mu}_n} + \frac{\bar{\mu}_n(R_{\theta,n} + R_{g,n} + R_{\pi,n})}{\lambda_{0,n}^{1/2}}\big)\big(R_{\pi,n}(x_1) + R_{\alpha,n}(x_1,x_2)\big) \qquad (.13)
\end{aligned}$$

Uniformly when $C(x_1,x_2)$ is uniformly bounded. We now derive a rate for $\tilde{\alpha}_n(x_1,x_2)'(\theta_n - \hat{\theta})$. 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:

$$\begin{aligned}
\tilde{\alpha}_n(x_1,x_2)'(\hat{\theta} - \theta_n) =& \tilde{\alpha}_n(x_1,x_2)'\hat{\Sigma}^{-1}_{\lambda_{0,n}} \frac{1}{n_g}\sum_{i\in\mathcal{I}_g} \hat{\pi}_{n,i}(\hat{g}_i - g_i) \\
&+ \tilde{\alpha}_n(x_1,x_2)'\hat{\Sigma}^{-1}_{\lambda_{0,n}}\hat{\Sigma}_n^{1/2}\tilde{r}_n - \tilde{\alpha}_n(x_1,x_2)'\lambda_{0,n}\hat{\Sigma}^{-1}_{\lambda_{0,n}}\theta_n \qquad (.14)
\end{aligned}$$

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:

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

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

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

Substituting the rates derived above for $||\tilde{r}_n||$ and $\lambda_{0,n}||\hat{\Sigma}_{\lambda_{0,n}}^{-1/2}\theta_n||$ into (.14) and applying Cauchy-Schwartz:

$$\tilde{\alpha}_n(x_1,x_2)'(\hat{\theta}-\theta_n) - \tilde{\alpha}_n(x_1,x_2)'\hat{\Sigma}_{\lambda_{0,n}}^{-1}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}(\hat{g}_i-g_i)$$
$$\precsim_p ||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||(R_{\theta,n}+R_{\pi,n}) + ||\hat{\Sigma}_{\lambda_{0,n}}^{-1/2}\tilde{\alpha}_n(x_1,x_2)||\lambda_{0,n}^{1/2}\underline{\mu}_n^{-1} \tag{.15}$$

Again, by the properties of least squares:

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

Combining by the triangle inequality we get:

$$\tilde{\alpha}_n(x_1,x_2)'(\hat{\theta}-\theta_n) \precsim_p ||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||(R_{\theta,n}+R_{\pi,n}+R_g) + ||\hat{\Sigma}_{\lambda_{0,n}}^{-1/2}\tilde{\alpha}_n(x_1,x_2)||\lambda_{0,n}^{1/2}\underline{\mu}_n^{-1} \tag{.16}$$

From Lemma C.8 we have $||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||^2 \precsim_p C(x_1,x_2)\xi^2_{\tilde{\Omega},n}(x_1)$ and $||\hat{\Sigma}_{\lambda_{0,n}}^{-1/2}\tilde{\alpha}_n(x_1,x_2)||^2 \precsim_p C(x_1,x_2)\xi^2_{\tilde{\Omega},n}(x_1)$. Combining with (.16) and (.13) gives the conclusion. □

Proof of Theorem 3.2. Follows immediately from Lemmas C.5, C.6, and C.7. □

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 (.13) and (.15):

$$\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big) - \tilde{\alpha}(x_1,x_2)'\hat{\Sigma}_{\lambda_{0,n}}^{-1}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\hat{\pi}_{n,i}(\hat{g}_i-g_i)$$
$$\precsim_p R_{\theta,n} + \Big(\frac{\bar{\mu}_n}{\underline{\mu}_n} + \frac{\bar{\mu}_n(R_{\theta,n}+R_{g,n}+R_{\pi,n})}{\lambda_{0,n}^{1/2}}\Big)\big(R_{\pi,n}(x_1)+R_{\alpha,n}(x_1,x_2)\big)$$
$$+||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||(R_{\theta,n}+R_{\pi,n}) + \xi_{\tilde{\Omega},n}(x)\lambda_{0,n}^{1/2}\underline{\mu}_n^{-1} \tag{.17}$$

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:

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

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

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

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

$$
\begin{aligned}
&\tilde{\alpha}(x_1,x_2)'\hat{\Sigma}^{-1}_{\lambda_{0,n}} \frac{1}{n_g} \sum_{i\in\mathcal{I}_g} \hat{\pi}_{n,i}(\hat{g}_i - g_i) \\
-&\tilde{\alpha}(x_1,x_2)'\bar{\Sigma}^{-1}_{\lambda_{0,n}} E\big[\hat{\pi}_{n,i}\zeta_{n,i}'\big|\mathcal{I}_\pi\big]\Xi_n^{-1} \sum_{i\in\mathcal{I}_g} \zeta_{n,i}(r_{n,i}+\epsilon_i) \\
\precsim_p&||\hat{\Sigma}_n^{1/2}\hat{\Sigma}^{-1}_{\lambda_{0,n}}\tilde{\alpha}(x_1,x_2)||\big(||\hat{r}_{1,n}|| + ||\hat{r}_{2,n}||\big) \\
&+||\bar{\Sigma}_n^{-1/2}\tilde{\alpha}(x_1,x_2)|| \cdot ||\hat{r}_{3,n}|| \cdot ||\Xi_n^{1/2}(\tilde{\beta}-\beta_n)|| \\
&+||\Xi_n^{-1/2} E[\zeta_{n,i}\hat{\pi}_{n,i}'|\mathcal{I}_\pi]\bar{\Sigma}^{-1}_{\lambda_{0,n}}\tilde{\alpha}(x_1,x_2)|| \cdot ||\hat{r}_{4,n}|| \qquad (.18)
\end{aligned}
$$

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

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

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

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:

$$
\begin{aligned}
&||\Xi_n^{1/2}(\tilde{\beta}-\beta_n)|| \\
\leq&||\Xi_n^{1/2}\hat{\Xi}_n^{-1}\Xi_n^{1/2}|| \cdot ||\Xi_n^{-1/2} \frac{1}{n_g} \sum_{i\in\mathcal{I}_g} \zeta_{n,i}(r_{n,i}+\epsilon_i)|| \precsim_p \sqrt{\frac{p(n)}{n_g}} \qquad (.19)
\end{aligned}
$$

Where the final line above follows because $||\Xi_n^{1/2}\hat{\Xi}_n^{-1}\Xi_n^{1/2}|| \precsim_p 1$ from Assumption 5.4.iii and Lemma C.1 and $||\Xi_n^{-1/2}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}(r_{n,i}+\epsilon_i)|| \precsim_p \sqrt{\frac{p(n)}{n_g}}$ 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, $|\zeta_{n,i}'\beta_n - g_i| \precsim_p \ell_{\zeta,n}(s_3)$ almost surely. And so:

$$||\hat{r}_{1,n}|| \leq \Big(\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}(\zeta_{n,i}'\beta_n - g_i)^2\Big)^{1/2} \precsim_p \ell_{\zeta,n}(s_3) \tag{.20}$$

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:

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

$||\hat{\Xi}_n^{1/2}\hat{\Xi}_{\lambda_{1,n}}^{-1}\hat{\Xi}_n^{1/2}|| \leq 1$, 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:

$$||\hat{r}_{2,n}|| \precsim_p \lambda_{1,n}||\Xi_n^{-1}|| \tag{.21}$$

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:

$$||\hat{r}_{3,n}|| \precsim_p \big(\sqrt{\xi_{\zeta,n}^2 k(n)l(n)/n_g} + \big(\xi_{\tilde{\psi},n}^2 log\big(l(n)m(n)\big)/n_g\big)^{1/4}\big) \tag{.22}$$

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 already showed 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:

$$||\hat{r}_{4,n}|| \precsim_p \xi_{\zeta,n}\frac{p(n)}{n_g} \tag{.23}$$

Combine previous steps: Together, (4.2), (.17),(.18), (4.2), (.20), (.21), (.22), and (.23) give:

$$\begin{aligned}
&\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)\\
&-\tilde{\alpha}(x_1,x_2)'\bar{\Sigma}^{-1}_{\lambda_{0,n}} E\big[\hat{\pi}_{n,i}\zeta'_{n,i}\big|\mathcal{I}_\pi\big]\Xi_n^{-1}\sum_{i\in\mathcal{I}_g}\zeta_{n,i}(r_{n,i}+\epsilon_i)\\
\precsim_p & ||\bar{\Sigma}_n^{1/2}\bar{\Sigma}^{-1}_{\lambda_{0,n}}\tilde{\alpha}_n(x_1,x_2)||\\
&\times(R_{\theta,n}+R_{\pi,n}+\ell_{\zeta,n}(s_3)+\lambda_{1,n}||\Xi_n^{-1}||)\\
&+\xi_{\zeta,n}\frac{p(n)}{n_g}||\Xi_n^{-1/2}E\big[\zeta_{n,i}\hat{\pi}'_{n,i}\big|\mathcal{I}_\pi\big]\bar{\Sigma}^{-1}_{\lambda_{0,n}}\tilde{\alpha}(x_1,x_2)||\\
&+R_n(x_1)
\end{aligned} \tag{.24}$$

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 Pollard (2001) or Belloni *et al.* (2015)), 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:

$$P\big[||\frac{1}{\sqrt{n}}\sum_{i=1}^n a_{n,i} - \mathcal{N}_n|| > 3\delta\big] \le C_0 K_n q_n n^{-1/2}\delta^{-3}\big(1+\frac{|-log(K_n q_n \delta^{-3} n^{-1/2})|}{K_n}\big)$$

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:

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

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 $p(n)^2\xi_{\zeta,n}n_g^{-1/2}R^{-3}_{\mathcal{N},n}log(n_g) \to 0$ and so for sufficiently large n:

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

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 $\frac{1}{\sqrt{n_g}}\sum_{i\in\mathcal{I}_g}\Gamma_n^{-1/2}\zeta_{n,i}(r_{n,i}+\epsilon_i)$ 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 (.24) :

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

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:

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

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

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 $\hat{\beta}_b = \hat{\Xi}^{-1}_{\lambda_{1,n},b}\frac{1}{n_g}\sum_{i\in\mathcal{I}_g}\zeta_{n,i,b}Y_{i,b}$.

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, $E[(\sqrt{Q_{b,i}}\epsilon_i)^2|X_i,Z_i] \leq \bar{\sigma}^2_Y$. 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:

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

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 $||\Xi_n^{1/2}\hat{\Xi}^{-1}_{n,b}\Xi_n^{1/2} - I|| \precsim_p \sqrt{ln(n_g)\xi^2_{\zeta,n}p(n)/n_g} \prec 1$, 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:

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

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

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

Note that this differs from the rate (.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 $||\Xi_n^{1/2}\hat{\Xi}^{-1}_{n,b}\Xi_n^{1/2}-I||$. Subtracting (.24) from the above:

$$\begin{aligned}
&\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta} - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)\\
&-\tilde{\alpha}(x_1,x_2)'\bar{\Sigma}^{-1}_{\lambda_{0,n}} E\big[\hat{\pi}_{n,i}\zeta'_{n,i}\big|\mathcal{I}_\pi\big]\Xi_n^{-1}\sum_{i\in\mathcal{I}_g} (Q_{b,i}-1)\zeta_{n,i}(r_{n,i}+\epsilon_i)\\
\precsim_p &||\bar{\Sigma}_n^{1/2}\bar{\Sigma}^{-1}_{\lambda_{0,n}}\tilde{\alpha}_n(x_1,x_2)||(R_{\theta,n}+R_{\pi,n}+\ell_{\zeta,n}(s_3)+\xi_{\tilde{\Omega},n}(x_1)\lambda_{2,n}||\Xi_n^{-1}||)\\
&+ln(n_g)\xi_{\zeta,n}\frac{p(n)}{n_g}||\tilde{\alpha}(x_1,x_2)'\bar{\Sigma}^{-1/2}_{\lambda_{0,n}}E\big[\hat{\pi}_{n,i}\psi'_{n,i}\big|\mathcal{I}_\pi\big]\Xi_n^{-1/2}||+R_n(x_1) \qquad (.25)
\end{aligned}$$

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:

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

By Lemma C.4:

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

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:

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

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:

$$\frac{\sqrt{n_g}\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta} - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)}{||s_n(x_1,x_2)||} - \frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||}\tilde{\mathcal{N}}_n$$
$$\precsim_p R_{\mathcal{N},n} + \sqrt{\frac{\xi_{\zeta,n}^2 p(n)^2}{n_g}} + ln(n_g)\xi_{\zeta,n}\frac{p(n)}{n_g}$$
$$+b_n(R_{\theta,n} + R_{\pi,n} + \ell_{\zeta,n}(s_3) + \lambda_{1,n}||\Xi_n^{-1}||) + \frac{b_n\sqrt{n_g}R_n(x_1)}{\underline{c}_n(x_1,x_2)} \qquad (.26)$$

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:

$$\sup_{(x_1,x_2)\in\mathcal{X}} \Big|\frac{\sqrt{n_g}\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta} - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)}{||s_n(x_1,x_2)||_2} - \frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\tilde{\mathcal{N}}_n\Big| \prec_p \bar{r}_n \quad (.27)$$

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:

$$\sup_{(x_1,x_2)\in\mathcal{X}} \Big|\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{||s_n(x_1,x_2)||_2} - \frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n\Big| \prec_p \bar{r}_n \quad (.28)$$

Let $e\hat{r}r_n$ denote the quantity on the LHS of (.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:

$$\begin{aligned}
&\sup_{(x_1,x_2)\in\mathcal{X}} \Big|\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{\hat{\sigma}_n(x_1,x_2)} - \frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n\Big| \\
\leq &R_{\sigma,n} \sup_{(x_1,x_2)\in\mathcal{X}} \Big|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n\Big| + (1+R_{\sigma,n})e\hat{r}r_{1,n} \prec_p \bar{r}_n \quad (.29)
\end{aligned}$$

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:

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

which is $O(\bar{r}_n^{-1})$ by supposition. Call the LHS of the first inequality in (.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:

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

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:

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

So in all:

$$\big|P\big(\sup_{(x_1,x_2)\in\mathcal{X}}\big|\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{\hat{\sigma}_n(x_1,x_2)}\big| \leq c_n\big|\mathcal{I}_\pi\big)$$
$$-P\big(\sup_{(x_1,x_2)\in\mathcal{X}}\big|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||}\mathcal{N}_n\big| \leq c_n\big|\mathcal{I}_\pi\big)\big|$$
$$\precsim_p P\big(\big|\sup_{(x_1,x_2)\in\mathcal{X}}\big|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||}\mathcal{N}_n\big| - c_n\big| \leq \bar{r}_n\big) \tag{.30}$$

By similar reasoning, from .27 similar steps give:

$$\big|P\big(\sup_{(x_1,x_2)\in\mathcal{X}}\big|\frac{\sqrt{n_g}\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta} - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)}{\hat{\sigma}_n(x_1,x_2)}\big| \leq c_n\big|\mathcal{I}\big)$$
$$-P\big(\sup_{(x_1,x_2)\in\mathcal{X}}\big|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n\big| \leq c_n\big|\mathcal{I}_\pi\big)\big|$$
$$\precsim_p P\big(\big|\sup_{(x_1,x_2)\in\mathcal{X}}\big|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n\big| - c_n\big| \leq \bar{r}_n\big) \tag{.31}$$

In the above we have used that:

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

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 Chernozhukov *et al.* (2014) this implies:

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

So from (.30):

$$\big|P\big(\sup_{(x_1,x_2)\in\mathcal{X}}\big|\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2) - \hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{\hat{\sigma}_n(x_1,x_2)}\big| \leq c_n\big|\mathcal{I}_\pi\big)$$
$$-P\big(\sup_{(x_1,x_2)\in\mathcal{X}}\big|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n\big| \leq c_n\big|\mathcal{I}_\pi\big)\big| \prec_p 1 \tag{.32}$$

And also using (.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:

$$\frac{1}{B}\sum_{b=1}^{B}1\big\{\sup_{(x_1,x_2)\in\mathcal{X}}\big|\frac{\sqrt{n_g}\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta}-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)}{\hat{\sigma}(x_1,x_2)}\big|\leq c_n\big\}$$
$$-P\big(\sup_{(x_1,x_2)\in\mathcal{X}}\big|\frac{\sqrt{n_g}\big(\bar{y}(x_1|x_2)-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}\big)}{\hat{\sigma}_n(x_1,x_2)}\big|\leq c_n\big|\mathcal{I}_\pi\big)\prec_p 1 \qquad (.33)$$

(.33) and (.32) together imply:

$$\frac{1}{B}\sum_{b=1}^{B}1\big\{\sup_{(x_1,x_2)\in\mathcal{X}}\big|\frac{\sqrt{n_g}\big(\hat{\alpha}_n(x_1,x_2)'\hat{\theta}-\hat{\alpha}_n(x_1,x_2)'\hat{\theta}_b\big)}{\hat{\sigma}(x_1,x_2)}\big|\leq c_n\big\}$$
$$-P\big(\sup_{(x_1,x_2)\in\mathcal{X}}\big|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n\big|\leq c_n\big|\mathcal{I}_\pi\big)\prec_p 1 \qquad (.34)$$

$\sup_{(x_1,x_2)\in\mathcal{X}}\big|\frac{s_n(x_1,x_2)'}{||s_n(x_1,x_2)||_2}\mathcal{N}_n\big|$ is continuously distributed conditional on the subsample $\mathcal{I}_\pi$, 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 (.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 (.33) this then implies that with probability approaching 1:

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

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

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

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:

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

□

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. □

Lemma 1.1 proved below is an application of Theorems 15.16 and 15.18 (Picard) in Kress (1999).

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:

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

and $E\big[\tilde{C}(x_1,x_2,\bar{W})\big|X=x_1\big]\le C(x_1,x_2)$. Then if we define φ according to $\varphi(x_1,x_2,(\tilde{z},\bar{w}))=\tilde{\varphi}_{x_1,x_2,\bar{w}}(\tilde{z})$, 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 $E\big[\tilde{D}(x,\bar{w})\big|X=x\big]\le D(x)$.

To establish the relevant functions are in the ranges of the desired operators we apply Theorem 15.18 (Picard) in Kress (1999). 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 $c < \infty$:

$$\sum_{k=1}^{\infty} \frac{1}{(\mu_k^{(T)})^2} |\langle \delta, v_k^{(T)} \rangle|^2 \leq c \tag{.35}$$

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. (.35) 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. (.35) 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 Sunder (2016)). 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 Kress (1999) 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) 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 (.35) 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 (.35) 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 $\tilde{D}(x, \bar{w})$.

□

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 $E_Z[(\pi(x,Z)'\hat{\theta})^2|X = x] = \int (\pi(x,z)'\hat{\theta})^2 F_{Z|X=x}(dz)$.

Lemma C.1 below simply lists some consequences of Rudelson's matrix law of large numbers (Rudelson (1999)) 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 $\hat{A}_n = \frac{1}{n}\sum_{i=1}^n a_{n,i}a_{n,i}'$:*

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

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

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

Further, $||A_n^{-1/2}\hat{A}_n A_n^{-1/2}|| \precsim_p 1$, $||A_n^{1/2}\hat{A}_n^{-1}A_n^{1/2}|| \precsim_p 1$, $||A_n^{-1/2}\hat{A}_n^{1/2}|| \precsim_p 1$, $||A_n^{1/2}\hat{A}_n^{-1/2}|| \precsim_p 1$, $||\hat{A}_n^{-1/2}A_n^{1/2}|| \precsim_p 1$, $||\hat{A}_n^{1/2}A_n^{-1/2}|| \precsim_p 1$, 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 (Rudelson (1999)). By the triangle inequality:

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

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 $||A_n^{-1/2}\hat{A}_n A_n^{-1/2} - I|| \prec_p 1$, then $||A_n^{1/2}\hat{A}_n^{-1}A_n^{1/2} - I|| \precsim_p ||A_n^{-1/2}\hat{A}_n A_n^{-1/2} - I||$. 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 $||(A+B)^{-1}|| \leq ||A^{-1}||$. □

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 $\theta \in \mathbb{R}^{k(n)l(n)}$:

$$\begin{aligned} E\big[|\phi_n(X,V)'\theta|^2\big] &= E\big[\theta'\phi_n(X,V)\phi_n(X,V)'\theta\big] \\ &= E\big[\theta'\big[\big(\rho_n(V)\rho_n(V)'\big) \otimes \big(\chi_n(X)\chi_n(X)'\big)\big]\theta\big] \end{aligned} \tag{.36}$$

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 (.36) 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:

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

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:

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

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:*

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

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:

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

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

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

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:

$$\begin{aligned}
E\big[||Q_n^{-1/2}\epsilon_{n,i}||^2\big] \leq & 2E\big[||Q_n^{-1/2}\rho_{n,i}||^2\big] + 2E\big[||E\big[Q_n^{-1/2}\rho_n(V)\big|X_i, Z_i\big]||^2\big] \\
\leq & 4E\big[||Q_n^{-1/2}\rho_{n,i}||^2\big] = 4k(n)
\end{aligned} \quad (.37)$$

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:

$$\begin{aligned}||Q_n^{-1/2}\epsilon_{n,i}|| &\leq ||Q_n^{-1/2}\rho_{n,i}|| + ||E\big[Q_n^{-1/2}\rho_n(V)\big|X_i,Z_i\big]|| \\ &\leq \operatorname{ess\,sup}||Q_n^{-1/2}\rho_n(V)|| + \operatorname{ess\,sup}||E\big[Q_n^{-1/2}\rho_n(V)\big|X,Z\big]|| \\ &\leq 2\operatorname{ess\,sup}||Q_n^{-1/2}\rho_n(V)|| \leq 2\xi_{\rho,n} \qquad (.38)\end{aligned}$$

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 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:

$$\begin{aligned}&\big|D_q\big[E\big[\rho_n(V)'\theta|X=\cdot,Z=\cdot\big]\big](x,z)\big| \\ =&\big|E\big[\rho_n(V)'\theta D_q\big[\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot,\cdot,V)\big](x,z)\big]\big| \\ \leq& E\big[|\rho_n(V)'\theta|^2\big]^{1/2}\operatorname{ess\,sup}|D_q\big[\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot,\cdot,V)\big](x,z)| \\ \leq& c_4||Q_n^{1/2}\theta|| \qquad (.39)\end{aligned}$$

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:

$$\begin{aligned}
&\big|D_q\big[E\big[\rho_n(V)'\theta|X=x_1,Z=z_1\big]\big] \\
&\quad - D_q\big[E\big[\rho_n(V)'\theta|X=x_2,Z=z_2\big]\big]\big| \\
=&\big|E\big[\rho_n(V)'\theta\big(D_q\big[\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot,\cdot,V)\big](x_1,z_1) \\
&\quad - D_q\big[\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot,\cdot,V)\big](x_2,z_2)\big)\big]\big| \\
\leq&E\big[|\rho_n(V)'\theta|^2\big]^{1/2}\operatorname{ess\,sup}\big|D_q\big[\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot,\cdot,V)\big](x_1,z_1) \\
&\quad - D_q\big[\frac{dF_{(X,Z,V)}}{dF_{(X,Z)}\otimes F_V}(\cdot,\cdot,V)\big](x_2,z_2)\big| \\
\leq&c_4||Q_n^{1/2}\theta||\big(||x_1-x_2||^2+||z_1-z_2||^2\big)^{\frac{s_4-\lfloor s_4\rfloor}{2}}
\end{aligned} \tag{.40}$$

(.39) and (.40) 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):

$$||Q_n^{-1/2}r_n(x,z)||\precsim\ell_{\psi,n}(s_4) \tag{.41}$$

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:

$$\begin{aligned}
&E_Z\big[\big(\pi_n(x,Z)'\theta-\hat\pi_n(x,Z)'\theta\big)^2\big|X=x\big]^{1/2} \\
\leq&E[|r_n(x,Z)'\iota(\theta)\chi_n(x)|^2|X=x]^{1/2} \\
&+||\tilde\Omega_n(x)^{1/2}\hat\Omega^{-1}_{\lambda_{2,n}}\frac{1}{n_\pi}\sum_{i\in\mathcal{I}_\pi}\psi_{n,i}(\epsilon_{n,i}+r_{n,i})'\iota(\theta)\chi_n(x)|| \\
&+||\tilde\Omega_n(x)^{1/2}\hat\Omega^{-1}_{\lambda_{2,n}}\lambda_{2,n}\beta_n\iota(\theta)\chi_n(x)||
\end{aligned} \tag{.42}$$

Taking (.41) 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:

$$|r_n(x,z)'\iota(\theta)\chi_n(x_1)| \leq ||r_n(x,z)Q_n^{-1/2}|| \cdot ||Q_n^{1/2}\iota(\theta)\chi_n(x)||$$
$$\precsim E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}\ell_{\psi,n}(s_4) \quad (.43)$$

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

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

(.43) 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 $\xi_{\Omega,n}(x)$ equals $||\tilde{\Omega}_n(x)^{1/2}\Omega_n^{1/2}||$ we get:

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

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:

$$||\frac{1}{\sqrt{n_\pi}}\sum_{i\in\mathcal{I}_\pi}(\Omega_n^{-1/2}\psi_{n,i})(\epsilon_{n,i}+r_{n,i})'Q_n^{-1/2}||$$
$$\precsim_p E\big[||\Omega_n^{-1/2}\psi_{n,i}||^2||Q_n^{-1/2}(\epsilon_{n,i}+r_{n,i})||^2\big]^{1/2}$$
$$\precsim_p E\big[||\Omega_n^{-1/2}\psi_{n,i}||^2||Q_n^{-1/2}\epsilon_{n,i}||^2\big]^{1/2}$$
$$+2E\big[||\Omega_n^{-1/2}\psi_{n,i}||^2||Q_n^{-1/2}r_{n,i}||^2\big]^{1/2}$$
$$\precsim_p \min\big\{\xi_{\psi,n}E\big[||Q_n^{-1/2}\epsilon_{n,i}||\big], \xi_{\rho,n}E\big[||\Omega_n^{-1/2}\psi_{n,i}||\big]$$
$$+\operatorname{ess\,sup}||Q_n^{-1/2}r_n(X,Z)||E\big[||\Omega_n^{-1/2}\psi_{n,i}||\big]$$
$$\precsim_p \min\{\sqrt{\xi_{\psi,n}^2 k(n)}, \sqrt{\xi_{\rho,n}^2 m(n)}\} + \sqrt{m(n)}\ell_{\psi,n}(s_4)$$
$$\precsim_p \min\{\sqrt{\xi_{\psi,n}^2 k(n)}, \sqrt{\xi_{\rho,n}^2 m(n)}\} \quad (.44)$$

The second inequality follows from Young's inequality, the third by the Holder inequality, the fourth by (.37), (.38), and (.41). 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 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 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:

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

Substituting the results above into (.42) we get that uniformly over $\theta\in\mathbb{R}^{k(n)l(n)}$ and F_X-almost all x:

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

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:

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

By (.43):

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

By Lemma C.2 we have $||\hat{\Omega}_n^{1/2}\Omega_n^{-1/2}||\leq||\Omega_n^{-1/2}\hat{\Omega}_n\Omega_n^{-1/2}||^{1/2}\precsim_p 1$ 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 (.44) we get:

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

□

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*:*

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

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:

$$\begin{aligned}E\big[||Q_n^{-1/2}\epsilon_{n,i}||^2\big] &\leq 2E\big[||Q_n^{-1/2}\rho_{n,i}||^2\big] + 2E\big[||E\big[Q_n^{-1/2}\rho_n(V)\big|X_i\big]||^2\big] \\ &\leq 4E\big[||Q_n^{-1/2}\rho_{n,i}||^2\big] \\ &= 4k(n) \end{aligned} \tag{.45}$$

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

$$\begin{aligned}||Q_n^{-1/2}\epsilon_{n,i}|| &\leq ||Q_n^{-1/2}\rho_{n,i}|| + ||E\big[Q_n^{-1/2}\rho_n(V)\big|X_i\big]|| \\ &\leq \operatorname{ess\,sup}||Q_n^{-1/2}\rho_n(V)|| + \operatorname{ess\,sup}||E\big[Q_n^{-1/2}\rho_n(V)\big|X\big]|| \\ &\leq 2\xi_{\rho,n} \end{aligned} \tag{.46}$$

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:

$$||r_n(x)Q_n^{-1/2}|| \precsim \ell_{\chi,n}(s_5) \tag{.47}$$

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:

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

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

$$\begin{aligned}
&\big|\big(\alpha_n(x_1,x_2) - \hat{\alpha}_n(x_1,x_2)\big)'\theta\big| \\
\leq &|r_n(x_2)'\iota(\theta)\chi_n(x_1)| \\
+&\big(||G_n^{-1/2}\chi_n(x_2)|| \cdot ||G_n^{1/2}\hat{G}_{\lambda_{3,n}}^{-1}G_n^{1/2}|| \\
\times &||\frac{1}{n_\alpha}\sum_{i\in\mathcal{I}_\alpha} G_n^{-1/2}\chi_{n,i}(\epsilon_{n,i} + r_{n,i})'\iota(\theta)\chi_n(x_1)||\big) \\
+&|\chi_n(x_2)'\hat{G}_{\lambda_{3,n}}^{-1}\lambda_{3,n}\mu_n\iota(\theta)\chi_n(x_1)| \qquad (.48)
\end{aligned}$$

From Lemma C.3 implies $||Q_n^{1/2}\iota(\theta)\chi_n(x)||^2 \precsim \xi_{\chi,n}^2 E\big[|\phi_n(X,V)'\theta|^2\big]$ uniformly. Combining with (.47) and applying Cauchy-Schwartz:

$$\begin{aligned}
|r_n(x_2)'\iota(\theta)\chi_n(x_1)| &\leq ||r_n(x_2)Q_n^{-1/2}|| \cdot ||Q_n^{1/2}\iota(\theta)\chi_n(x)|| \\
&\precsim E\big[|\phi_n(X,V)'\theta|^2\big]^{1/2}\xi_{\chi,n}\ell_{\chi,n}(s_5) \qquad (.49)
\end{aligned}$$

Uniformly over $\theta \in \mathbb{R}^{k(n)l(n)}$ and F_X-almost all x. From (.49), 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 (.45), (.46), and (.47). So uniformly over $\theta \in \mathbb{R}^{k(n)l(n)}$ and

F_X-almost all x:

$$||\frac{1}{n}\sum_{i=1}^{n} G_n^{-1/2}\chi_{n,i}(\epsilon_{n,i}+r_{n,i})'\iota(\theta)\chi_n(x_1)|| \tag{.50}$$
$$\precsim_p ||Q_n^{1/2}\iota(\theta)\chi_n(x)|| \min\{\sqrt{\xi_{\chi,n}^2 k(n)/n_\alpha}, \sqrt{\xi_{\rho,n}^2 l(n)/n_\alpha}\}$$
$$+||Q_n^{1/2}\iota(\theta)\chi_n(x)||\ell_{\chi,n}(s_5)\sqrt{l(n)/n_\alpha}$$

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:

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

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

$$\begin{aligned}
||G_n^{1/2}\mu_n\iota(\theta)\chi_n(x)||^2 &= E\big[|\chi_n(X)'\mu_n\iota(\theta)\chi_n(x)|^2\big] \\
&\leq E\big[|E[\rho_n(V)|X]'\iota(\theta)\chi_n(x)|^2\big] \\
&\leq E\big[|\rho_n(V)'\iota(\theta)\chi_n(x)|^2\big] \\
&= ||Q_n^{1/2}\iota(\theta)\chi_n(x)||^2
\end{aligned} \tag{.51}$$

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 (.49), (.50), and (.51) into (.48) we get that uniformly over $\theta \in \mathbb{R}^{k(n)l(n)}$ and F_X-almost all x:

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

□

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:

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

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

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

By Lemma C.1 and Assumption 5.4.iii we have that $||\hat{\Xi}_n^{1/2}\Xi_n^{-1/2}|| \precsim_p 1$ and $||\Xi_n^{1/2}\hat{\Xi}^{-1}_{\lambda_{1,n}}\Xi_n^{1/2}|| \precsim_p 1$. 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 Assumption 5.2.iv we have $\operatorname{ess\,sup} E[\epsilon_i^2|X_i,Z_i] \leq E[Y_i^2|X_i,Z_i] \leq \bar{\sigma}_Y^2$. 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:

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

□

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:*

$$||\bar{\Sigma}_n^{-1/2}\hat{\Sigma}_n\bar{\Sigma}_n^{-1/2} - I||^2 \precsim_p \sqrt{\xi_{\tilde{\psi},n}^2 log\big(l(n)m(n)\big)/n_g} \tag{.52}$$

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

$$||\bar{\Sigma}_n^{1/2}\bar{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2} - \bar{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2}||^2 \precsim_p \sqrt{\xi_{\tilde{\psi},n}^2 log\big(l(n)m(n)\big)/n_g} \tag{.53}$$

Furthermore:

$$||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)|| \precsim_p ||\bar{\Sigma}_n^{1/2}\bar{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)|| \tag{.54}$$

In addition, we have $||\bar{\Sigma}_n^{-1/2}\tilde{\alpha}_n(x_1,x_2)||^2 \precsim C(x_1,x_2)\xi_{\tilde{\Omega},n}^2(x_1)$*, we have that* $||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\tilde{\alpha}_n(x_1,x_2)||^2 \precsim_p C(x_1,x_2)\xi_{\tilde{\Omega},n}^2(x_1)$*, and finally* $||\hat{\Sigma}_{\lambda_{0,n}}^{-1/2}\tilde{\alpha}_n(x_1,x_2)||^2 \precsim_p C(x_1,x_2)\xi_{\tilde{\Omega},n}^2(x_1)$*. 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 (.52).** 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:

$$\begin{aligned}
||\hat{\Sigma}_n^{1/2}\bar{\Sigma}_n^{-1/2}\theta|| &= \frac{1}{n_g}\sum_{i\in\mathcal{I}_g}|\hat{\pi}_{n,i}'\bar{\Sigma}_n^{-1/2}\theta|^2 \\
&= \frac{1}{n_g}\sum_{i\in\mathcal{I}_g}|\big(\hat{\omega}_n'\psi_n(x,z)\big)'\iota(\bar{\Sigma}_n^{-1/2}\theta)\chi_n(x)|^2 \\
&= \frac{1}{n_g}\sum_{i\in\mathcal{I}_g}|\big(\chi_n(x)\otimes\psi_n(x,z)\big)'vec\big(\hat{\omega}_n\iota(\bar{\Sigma}_n^{-1/2}\theta)\big)|^2 \\
&= \frac{1}{n_g}\sum_{i\in\mathcal{I}_g}|\tilde{\psi}_n(x,z)'vec\big(\hat{\omega}_n\iota(\bar{\Sigma}_n^{-1/2}\theta)\big)|^2 \\
&= ||\hat{\tilde{\Omega}}_n^{1/2}vec\big(\hat{\omega}_n\iota(\bar{\Sigma}_n^{-1/2}\theta)\big)|| \\
&\leq ||\hat{\tilde{\Omega}}_n^{1/2}\tilde{\Omega}_n^{-1/2}||\cdot||\tilde{\Omega}_n^{1/2}vec\big(\hat{\omega}_n\iota(\bar{\Sigma}_n^{-1/2}\theta)\big)|| \\
&= ||\hat{\tilde{\Omega}}_n^{1/2}\tilde{\Omega}_n^{-1/2}||\cdot||\theta||
\end{aligned} \tag{.55}$$

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:

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

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:

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

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 (.53). 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 (.52):

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

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(.54). 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 (.52) and (.53), 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 $\theta \in \mathbb{R}^{k(n)l(n)}$:

$$|\tilde{\alpha}_n(x_1,x_2)'\theta|^2 \leq C(x_1,x_2)E_Z\left[\left(\hat{\pi}_n(x_1,Z)'\theta\right)^2\big|X=x_1\right]$$

Using similar steps to (.55):

$$\begin{aligned}&E_Z\left[\left(\hat{\pi}_n(x_1,Z)'\theta\right)^2\big|X=x_1\right] \\ =&||\tilde{\Omega}_n(x_1)^{1/2}vec\big(\hat{\omega}_n\iota(\theta)\big)||^2 \\ \leq&||\tilde{\Omega}_n(x_1)^{1/2}\tilde{\Omega}_n^{-1/2}||E_{(X,Z)}\left[|\tilde{\psi}_n(X,Z)'vec\big(\hat{\omega}_n\iota(\theta)\big)|^2\right] \\ \leq&\xi^2_{\tilde{\Omega},n}(x_1)E_{(X,Z)}\left[\left(\hat{\pi}_n(X,Z)'\theta\right)^2\right] \\ =&\xi^2_{\tilde{\Omega},n}(x_1)||\bar{\Sigma}_n^{1/2}\theta||^2\end{aligned}$$

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:

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

Where the final line follows by $||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\bar{\Sigma}_n^{1/2}||^2 \precsim_p 1$ 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:

$$\begin{aligned}||\hat{\Sigma}_{\lambda_{0,n}}^{-1/2}\tilde{\alpha}_n(x_1,x_2)||^2 &\leq C(x_1,x_2)\xi_{\hat{\Omega},n}^2(x_1)||\hat{\Sigma}_n^{1/2}\hat{\Sigma}_{\lambda_{0,n}}^{-1}\hat{\Sigma}_n^{1/2}||||\hat{\Sigma}_n^{-1/2}\bar{\Sigma}_n^{1/2}||^2 \\ &\precsim_p C(x_1,x_2)\xi_{\hat{\Omega},n}^2(x_1)\end{aligned}$$

□

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):

$$\frac{dF_{(X,Z,V)}}{d(F_{(X,Z)}\otimes F_V)}(x,z,v)=\rho_n(v)'B_n(x,z)+r_n(x,z,v) \tag{.56}$$

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 (.56), the RHS above becomes:

$$\begin{aligned}
&\big(E\big[\gamma(x,V)\rho_n(V)'\big] - \beta_n(x)'E\big[\rho_n(V)\rho_n(V)'\big]\big)B_n(x,z) \\
+&E\big[\big(\gamma(x,V) - \beta_n(x)'\rho_n(V)\big)r_n(x,z,V)\big]
\end{aligned} \tag{.57}$$

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 (.57) 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:

$$E\big[\big(\beta_n(x)'\rho_n(V)\big)^2\big]^{1/2} \leq \bar{D} \tag{.58}$$

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:

$$E\big[\big(\beta_n(X)'\rho_n(V)\big)^2\big]^{1/2} \leq \frac{\bar{D}}{\sqrt{\underline{c}}} \precsim 1 \tag{.59}$$

In all, we get that uniformly over $F_{(X,Z)}$-almost all (x,z):

$$\left|g(x,z) - \beta_n(x)' E\left[\rho_n(V)\middle|X = x, Z = z\right]\right| \precsim \ell_{\rho,n}(s_1) \tag{.60}$$

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 (.60) we immediately get:

$$E\left[\left(g(X,Z) - \chi_n(x)'\theta_n E\left[\rho_n(V)\middle|X,Z\right]\right)^2\middle|X = x\right]^{1/2} \precsim \ell_{\rho,n}(s_1)$$

Moreover, from (.59) we get $E\left[\left(\chi_n(X)'\theta_n\rho_n(V)\right)^2\right]^{1/2} \precsim 1$, 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:

$$\sup_{x\in\mathcal{X}} ||E\left[\rho_n(V)\rho_n(V)'\right]^{1/2}\tilde{\beta}_n(x)|| \leq \sup_{x\in\mathcal{X}} ||E\left[\rho_n(V)\rho_n(V)'\right]^{1/2}\beta_n(x)||^2$$

We will show $E\left[|\tilde{\beta}_n(X)'\rho_n(V)|^2\right]$ is bounded uniformly over n. Note that:

$$\begin{aligned}
\int_{\mathcal{X}} |\tilde{\beta}_n(x)'\rho_n(v)|^2 dx &= \int_{\mathcal{X}} \left|M_n[\beta_n(X)'\rho_n(v)](x)\right|^2 dx \\
&\leq \int_{\mathcal{X}}\int_{\mathcal{X}} \frac{1\{||x'-x|| \leq b_n\}}{\int_{\mathcal{X}\cap B_{x,b_n}} dx'} |\beta_n(x')'\rho_n(v)|^2 dx'dx \\
&\leq \frac{1}{\underline{r}}\int_{\mathcal{X}}\int_{\mathcal{X}} \frac{1\{||x'-x|| \leq b_n\}}{\int_{B_{0,b_n}} dx'} dx |\beta_n(x')'\rho_n(v)|^2 dx' \\
&\leq \frac{1}{\underline{r}}\int_{\mathcal{X}} |\beta_n(x)'\rho_n(v)|^2 dx
\end{aligned}$$

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:

$$E\big[|\tilde{\beta}_n(X)'\rho_n(V)|^2\big] \leq \frac{\bar{c}\bar{f}}{\underline{c}\underline{r}\underline{f}} E\big[|\beta_n(X)'\rho_n(V)|^2\big] \leq \frac{\bar{c}\bar{f}}{\underline{c}^2\underline{r}\underline{f}} \bar{D}^2 \tag{.61}$$

Where the final inequality follows by (.59). 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 $x \in \mathcal{X}$:

$$|M_n[\delta](x_1) - M_n[\delta](x_2)| \leq \frac{2}{\underline{r}} c \frac{\dim(X)}{b_n} ||x_1 - x_2|| \tag{.62}$$

Now, we will upper bound the function $\beta_n(\cdot)'E\big[\rho_n(V)\big|X = x, Z = z\big]$. Note that (.58) 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 (.62) 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:

$$
\begin{aligned}
&|g(x,Z) - \tilde{\beta}_n(x)' E\big[\rho_n(V)\big|X = x, Z\big]| \\
\leq& \Big|\frac{\int_{\mathcal{X}\cup B_{x,b_n}} \big(g(x',Z) - \beta_n(x')' E\big[\rho_n(V)\big|X = x', Z\big]\big) dx'}{\int_{\mathcal{X}\cup B_{x,b_n}} dx'}\Big| \\
&+\big|g(x,Z) - M_n\big[g(\cdot,Z)\big]\big| \\
&+\Big|\frac{\int_{\mathcal{X}\cup B_{x,b_n}} \beta_n(x')'\big(E\big[\rho_n(V)\big|X = x, Z\big] - E\big[\rho_n(V)\big|X = x', Z\big] dx'}{\int_{\mathcal{X}\cup B_{x,b_n}} dx'}\Big|
\end{aligned}
$$

For the first term on the RHS of the inequality note that:

$$
\begin{aligned}
&\big|\Big(\int_{\mathcal{X}\cup B_{x,b_n}} dx'\Big)^{-1} \int_{\mathcal{X}\cup B_{x,b_n}} g(x',Z) - \beta_n(x')' E\big[\rho_n(V)\big|X = x', Z\big] dx'\big| \\
&\leq \sup_{x'\in\mathcal{X}} \big|g(x',Z) - \beta_n(x')' E\big[\rho_n(V)\big|X = x', Z\big]\big| \\
&\lesssim \ell_{\rho,n}(s_1)
\end{aligned}
$$

Where, for the last step we have used (.60). Next, it is easy to see that for any $\delta \in \Lambda_s^{\dim(X)}(c)$ we have $|\delta(x) - M_n[\delta](x)| \leq c b_n^{\min\{s,1\}}$. And so, using Assumption 5.3.ii, the second term satisfies:

$$
\big|g(x,z) - M_n\big[g(z,\cdot)\big]\big| \leq c_3 b_n^{\min\{s_3,1\}}
$$

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:

$$
\begin{aligned}
&\Big|\int_{\mathcal{X}\cap B_{x,b_n}} \beta_n(x')'\big(E\big[\rho_n(V)\big|X = x, Z = z\big] \\
&\quad - E\big[\rho_n(V)\big|X = x', Z = z\big] dx'\Big| \\
=&\Big|\int_{\mathcal{X}\cap B_{x,b_n}} E\big[\beta_n(x')'\rho_n(V) q(V,x,x',z)\big] dx'\Big| \\
\leq& \int_{\mathcal{X}\cap B_{x,b_n}} \big|E\big[\beta_n(x')'\rho_n(V) q(V,x,x',z)\big]\big| dx' \\
\leq& \int_{\mathcal{X}\cap B_{x,b_n}} E\big[\big(\beta_n(x')'\rho_n(V)\big)^2\big]^{1/2} E\big[q(V,x,x',z)^2\big]^{1/2} dx' \\
\leq& \bar{D} \int_{\mathcal{X}\cap B_{x,b_n}} E\big[q(V,x,x',z)^2\big]^{1/2} dx'
\end{aligned}
$$

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 (.58). 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:

$$\begin{aligned}
&\Big|\frac{\int_{\mathcal{X}\cap B_{x,b_n}} \beta_n(x')'\big(E\big[\rho_n(V)\big|X=x,Z\big] - E\big[\rho_n(V)\big|X=x',Z\big]dx'}{\int_{\mathcal{X}\cap B_{x,b_n}} dx'}\Big| \\
\leq & c_2 b_n^{\min\{s_2,1\}}
\end{aligned}$$

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:

$$E\big[|\chi_n(X)'\theta_n\rho_n(V)|^2\big] \leq \frac{\bar{c}}{\underline{c}} E\big[|\tilde{\beta}_n(X)'\rho_n(V)|^2\big] \leq \frac{\bar{c}^2\bar{f}}{\underline{c}^3\underline{r}\,\underline{f}}\bar{D}^2 \precsim 1$$

Where the final inequality follows by (.61). So $E\big[|\chi_n(X)'\theta_n\rho_n(V)|^2\big]$ is bounded above uniformly over n, and we are done. □