Discussion Papers
Department of Economics
University of Copenhagen

No. 20-04

Testing a Class of Semi- or Nonparametric Conditional Moment Restriction Models using Series Methods

by

Jesper R.-V. Sørensen

Øster Farimagsgade 5, Building 26, DK-1353 Copenhagen K., Denmark
Tel.: +45 35 32 30 01 – Fax: +45 35 32 30 00
http://www.econ.ku.dk

ISSN: 1601-2461 (E)

Testing a Class of Semi- or Nonparametric Conditional Moment Restriction Models using Series Methods

Jesper R.-V. Sørensen*

Abstract

This paper proposes a new test for a class of conditional moment restrictions whose parameterization involves unknown, unrestricted conditional expectation functions. Examples of such conditional moment restrictions are conditional mean independence (leading to a nonparametric significance test) and conditional homoskedasticity (with an otherwise unrestricted conditional mean) and also arise from models of single-agent discrete choice under uncertainty and static games of incomplete information. The proposed test may be viewed as a semi-/nonparametric extension of the Bierens (1982) goodness-of-fit test of a parametric model for the conditional mean. Estimating conditional expectations using series methods and employing a Gaussian multiplier bootstrap to obtain critical values, the resulting test is shown to be asymptotically correctly sized and consistent. A simulation study applies the procedure to test the specification of a two-player, binary-action static game of incomplete information, treating equilibrium beliefs as nonparametric conditional expectations.

Keywords: Omnibus specification testing; Semiparametric; Conditional moment restrictions; Conditional expectation; Series estimation; Bootstrap; Cramér-von Mises distance.

JEL classification: C01, C14.

Dates: First version: November 2017. This version: July 2020. *Address:* Department of Economics, University of Copenhagen, Øster Farimagsgade 5, building 26, 1353 Copenhagen K, Denmark. *Email:* jrvs@econ.ku.dk. *Website:* https://sites.google.com/site/jesperrvs/.

1 Introduction

In this paper I propose a general method for constructing omnibus specification tests for a wide class of semi- or nonparametric conditional moment restriction models. The paper aims to test the validity of the model assertion that there exists a finite-dimensional parameter β such that

$$\mathrm{E}\left[\rho_\ell\left(Z, \beta, \mathrm{E}\left[Y_\ell \middle| W_\ell\right]\right) \middle| X_\ell\right] = 0 \text{ almost surely } X_\ell \text{ for all } \ell \in \{1, \ldots, L\}, \tag{1.1}$$

where the ρ_ℓ's are known functions—which may be thought of as model residuals—each $\mathrm{E}\left[Y_\ell \middle| W_\ell\right]$ is an unrestricted, possibly vector, conditional expectation, and each W_ℓ is a subvector of the conditioning variables X_ℓ. I use Z for all model observables, i.e., the union of distinct elements of the X_ℓ's and Y_ℓ's. The alternative hypothesis is that (1.1) is violated. Allowing for unknown, unrestricted conditional expectation functions (CEFs) as part of the model parameterization constitutes a main novelty of this paper.

The conditional moment restrictions (CMRs) framework studied in this paper encompasses several semi- or nonparametric models encountered in empirical work. An example of a *non*parametric hypothesis leading to an expression of the form in (1.1) is that of *conditional mean independence*, which states that the conditional mean of an outcome variable Y depends only on a subset of the candidate conditioning variables, i.e., that $\mathrm{E}\left[Y \middle| X\right] = \mathrm{E}\left[Y \middle| W\right]$, with W being a given strict subvector of X. Conditional mean independence may be rephrased as $\mathrm{E}\left[Y - \mathrm{E}\left[Y \middle| W\right] \middle| X\right] = 0$. Given that irrelevant regressors ought to be dropped from the regression analysis, a test of conditional mean independence is often referred to as a *nonparametric significance test*. The nonparametric hypothesis of *conditional variance independence,* $\mathrm{var}\left(Y \middle| X\right) = \mathrm{var}\left(Y \middle| W\right)$, is similarly nested in the (1.1) framework. An example of a *semi*parametric hypothesis is that of *conditional homoskedasticity*. This hypothesis states that $\mathrm{var}\left(Y \middle| X\right) = \sigma^2$ for some constant σ^2, which may be expressed as $\mathrm{E}[Y^2 - \left(\mathrm{E}\left[Y \middle| X\right]\right)^2 - \sigma^2 \,|X] = 0$ for some σ^2. More structural examples of semiparametric models leading to expressions of the form (1.1) are (single-agent) *discrete choice under uncertainty* [as in Manski (1991); Ahn and Manski (1993)] and *static games of incomplete information* (see, e.g., Bajari, Hong, Krainer, and Nekipelov, 2010). In a model of discrete choice under uncertainty, CEFs may be introduced via the model assumption that the agents' beliefs are correct in the aggregate—a *ratio-*

nal expectations hypothesis. In static games of incomplete information, CEFs appear under the model assumption that beliefs are correct in a *Bayesian Nash equilibrium*.

The test I propose is an extension of the one given by Bierens (1982) in the context of parametric mean regression. The idea of Bierens' method is to recast a *conditional* moment restriction as a collection of testable *un*conditional moment restrictions, which are then suitably integrated (or otherwise aggregated). Within the context of parametric mean regression, a test of correct specification may be obtained by checking whether the least-squares residuals correlate with any member of a suitably rich family of transformations of the regressors. Bierens' idea carries over to any setting where one may speak of model residuals, including (1.1).

An alternative approach estimates the model under both the null and alternative and contrasts the estimates according to some notion of distance [see, e.g., Härdle and Mammen (1993) and Zheng (1996)]. The Bierens approach is convenient in that it only requires estimation of the (potentially substantially) simpler null model. However, the two approaches have different (local) power properties and should be viewed as complementary.[1]

The suggested test statistic is a Cramér-von Mises-type measure of distance between the collection of residual-to-transformation correlations and zero. One rejects the null hypothesis that the semi-/nonparametric model in (1.1) is correctly specified whenever said distance is "unreasonably" large. Under the null hypothesis, the proposed test statistic has a nonpivotal limiting distribution and can therefore not be tabulated. I propose and formally justify the use of a multiplier bootstrap procedure for obtaining critical values. Calculation of the test statistic and critical values requires estimation of CEFs. These are here estimated using series methods and therefore boil down to linear regressions.

The resulting test is shown to have attractive theoretical properties: it is both asymptotically of correct size and consistent against any fixed alternative. To illustrate these properties, I implement my procedure in a comprehensive simulation study testing the specification of a static discrete game of incomplete information. The simulations by and large reproduce the asymptotic properties in small samples.

Static discrete-choice models with social or strategic interactions have been applied in numerous contexts including firm entry (Seim, 2006), the timing of radio commercials (Sweeting, 2009), labor force participation (Bjorn and Vuong, 1984), and teen

[1] For a formal comparison of their power properties, see Fan and Li (2000).

sex (Card and Giuliano, 2013). These models may be conveniently estimated in two steps. In the first step, the conditional choice probabilities (CCPs) are estimated in a nonparametric manner. The estimated CCPs are then employed in a second step to estimate the structural parameters of the model (see, e.g., Bajari, Hong, Krainer, and Nekipelov, 2010). Construction of my test follows along the same lines.[2]

There exists a vast literature on *omnibus* (i.e., consistent against any model violating the null hypothesis) *specification testing*[3] of *parametric* models of i.i.d. data for the conditional *mean* [see, e.g., Bierens (1982; 1990); Härdle and Mammen (1993); de Jong and Bierens (1994); Hong and White (1995); Zheng (1996); Bierens and Ploberger (1997); Stute (1997); Whang (2000); Horowitz and Spokoiny (2001); Stengos and Sun (2001); Guerre and Lavergne (2005); Stute and Zhu (2005); Escanciano (2006); Sun and Li (2006): Hsiao, Li, and Racine (2007)] as well as for a (single) conditional *quantile* [see, e.g., Zheng (1998); Bierens and Ginther (2001); Horowitz and Spokoiny (2002); He and Zhu (2003); Whang (2006)]. Whang (2001), Donald, Imbens, and Newey (2003), Tripathi and Kitamura (2003) and Delgado, Domínguez, and Lavergne (2006) propose tests for a *class* of *parametric* CMRs nesting parametric specifications of the conditional mean as a special case. There also exists a sizeable literature on the topic of consistent *nonparametric significance testing* [see, e.g., Fan and Li (1996); Lavergne and Vuong (2000); Aït-Sahalia, Bickel, and Stoker (2001); Delgado and Manteiga (2001); Racine, Hart, and Li (2006); Lavergne, Maistre, and Patilea (2015)].

The results of this paper complement those obtained by Song (2010) and Bravo (2012), both of whom develop test statistics for a class of semiparametric CMRs similar to (1.1). Song (2010) confines interest to the case where the nonparametric part of the parameterization takes a composite-index form. His treatment of the nonparametric part rules out unrestricted CEFs but does allow for single-index models not nested in (1.1). Song's framework is thus neither more nor less general. Unlike the nonpivotal test statistic proposed in this paper, Song (2010) uses a conditional martingale transform to obtain an asymptotically distribution-free test statistic, thus

[2]Implicit in this two-step estimation strategy is an assumption of equilibrium uniqueness. See Hahn, Moon, and Snider (2017) for a test aiming at detecting neglected heterogeneity, which may be used to test for equilibrium multiplicity.

[3]Given the extraordinary number of papers that have appeared over the past three decades or so, my referencing will necessarily be incomplete. For a fairly recent review of methods for specification testing, see Davidson and Zinde-Walsh (2017).

allowing for tabulation of critical values. However, since the martingale transform is generally unknown, pivotality comes at the cost of additional steps of nonparametric estimation. In addition, as indicated by Song's simulation studies (and remarked by Song himself), the martingale transform approach appears more sensitive to the choice of tuning parametes than the bootstrap—the latter approach being the one taken in this paper.

Bravo (2012) uses a generalized empirical likelihood approach to obtain specification tests similar in spirit to classical Kolmogorov-Sminov and Cramér-von Mises goodness-of-fit statistics. As in this paper, Bravo's tests statistic has a nonpivotal limit distribution and a multiplier bootstrap procedure is used to obtain critical values. His framework is broader than (1.1) in that his residual function may depend on arbitrary nonparametric components. Morover, this dependence is allowed to be functional. Naturally, Bravo's greater generality comes at the cost of relatively "high level" (i.e., abstract) conditions. Specifically, his level of generality makes it difficult to analytically derive the adjustments terms necessary to account for nonparametric estimation. These adjustments must be estimated in order to obtain valid critical values and therefore constitute crucial elements of the implementation of the test. In contrast, by restricting attention to the nonparametric CEFs (a type of mean-square projection), I may obtain these necessary adjustments in closed form under fairly primitive conditions, such as (ordinary) differentiability. The added CEF structure also allows me to tailor my assumptions to the nonparametric method of estimation, here chosen to be *series estimation* [see, e.g., Newey (1994; 1995; 1997); Belloni, Chernozhukov, Chetverikov, and Kato (2015)].

The problem studied in this paper also relates to the task of testing the validity of a model for the entire conditional *distribution* [see, e.g., Andrews (1997); Delgado and Stute (2008); Escanciano and Velasco (2010); Bierens and Wang (2012); Rothe and Wied (2013); Escanciano and Goh (2014)], which corresponds to testing *infinitely* many CMRs (in fact, a continuum) of a particular form. In addition, Fan and Li (1996), Aït-Sahalia, Bickel, and Stoker (2001), Li, Hsiao, and Zinn (2003) and Korolev (2018) also develop tests for particular *semi-* or *non*parametric specifications of the conditional mean, not all of which are nested in (1.1). This paper should therefore also be viewed as complementary to the work of these authors.[4]

[4]Given that I take W_ℓ in (1.1) to be a subvector of X_ℓ, settings with "outside" exogenous (or instrumental) variables are not subsumed by the CMR framework of this paper. Breunig (2015)

The remainder of this paper is organized as follows. I define the testing problem and test statistic in Section 2. Section 3.1 analyzes the limiting behavior of the test statistic. Motivated by this limiting behavior, in Section 3.2 I construct critical values based on a bootstrap procedure and establishes their asymptotic validity. The limiting properties of the resulting test are given in Section 3.3. I investigate the small-sample properties of the test in a simulation study in Section 4. Section 5 concludes and discusses possible directions for future research. Proofs of formal statements can be found in the appendices.

Notation

I use $\|f\|_{\mathcal{D}} \coloneqq \sup_{x\in\mathcal{D}} |f(x)|$ to denote the supremum norm of $f : \mathcal{D} \to \mathbf{R}$, and write $L^{\infty}(\mathcal{D}) = \{f : \mathcal{D} \to \mathbf{R}; \|f\|_{\mathcal{D}} < \infty\}$ for the collection of bounded real-valued functions on $\mathcal{D}$.

2 Testing Semi-/Nonparametric CMRs

2.1 Testing Problem

Let $\{Z_i\}_1^n$ be n independent copies of Z, such that Z_i is random element of $\mathbf{R}^{d_z}$ composed by the distinct elements elements of $X_{\ell i}$, thus subsuming $W_{\ell i}$,[5] and $Y_{\ell i}, \ell \in \{1,\dots,L\}$. The support of Z is denoted by $\mathcal{Z}$ and that of X_ℓ by $\mathcal{X}_\ell \subseteq \mathbf{R}^{d_{x,\ell}}$. Let $\mathcal{B} \subseteq \mathbf{R}^d$ be a parameter space. The *null hypothesis* we wish to test is

$$\mathrm{H}_0 : \text{For some } \beta \in \mathcal{B},\ \mathrm{E}\left[\rho_\ell\left(Z,\beta,h_\ell^*\left(W_\ell\right)\right)\middle| X_\ell\right] = 0 \text{ a.s. } X_\ell \text{ for all } \ell \in \{1,\dots,L\}, \tag{2.1}$$

where, for notational convenience, I have abbreviated the CEFs by

$$h_\ell^*\left(W_\ell\right) \coloneqq \mathrm{E}\left[Y_\ell \middle| W_\ell\right], \quad \ell \in \{1,\dots,L\}.$$

develops goodness-of-fit tests (also based on series estimators) for the nonparametric instrumental-variables (NPIV) model. For inference in NPIV more broadly, see Santos (2012), who also allows partial identification.

[5] Here W_ℓ need not be a literal subvector of X_ℓ; only X_ℓ-measurability is required.

(Henceforth "a.s." connotes "almost surely.") The null is tested against the general *alternative hypothesis*

$$\mathrm{H}_1 : \text{For all } \beta \in \mathcal{B},\ \mathrm{P}\left(\mathrm{E}\left[\rho_\ell\left(Z, \beta, h_\ell^*\left(W_\ell\right)\right)\middle| X_\ell\right] = 0\right) < 1 \text{ for some } \ell \in \{1, \ldots, L\} \tag{2.2}$$

under a collection of regularity conditions presented below. In this paper, I propose a procedure for testing (2.1) versus (2.2) assuming the existence of some $\beta_0 \in \mathcal{B}$ such that (a) β_0 may be consistently estimated (irrespective of the null or alternative being true), and (b) Equation (1.1) is satisfied at β_0 under the null. Due to property (b), β_0 will be referred to as *pseudo true*.

Example 1 (**Pseudo Truths Defined via Moment Conditions**)**.** Given the CMR setting of this paper, a natural definition of a pseudo true parameter is as the assumed unique solution to

$$\mathrm{E}\left[m\left(Z, \beta, h^*\left(W\right)\right)\right] = 0, \tag{2.3}$$

where $h^*\left(W\right)$ denotes the vector of all unique elements of the $h_\ell^*\left(W_\ell\right), \ell \in \{1, \ldots, L\}$, and $m\left(Z, \beta, h^*\left(W\right)\right)$ denotes a vector of moment functions arising from interacting one or more of the residuals $\rho_\ell\left(Z, \beta, h_\ell^*\left(W_\ell\right)\right)$ with transformations of the corresponding conditioning variables X_ℓ and stacking the results. Via iterated expectations, this implicit product form in $m\left(Z, \beta, h^*\left(W\right)\right)$ ensures that any parameter satisfying (2.1) must also satisfy (2.3). Hence, the solution of (2.3) must be pseudo true.

Example 2 (**Pseudo Truth in Testing Conditional Homoskedasticity**)**.** Consider the *conditional homoskedasticity* hypothesis mentioned in the introduction. Then, under the null, one has $\mathrm{E}\left[Y^2 - h^*\left(X\right)^2 \middle| X\right] = \sigma^2$ for some constant σ^2 and $h^*\left(X\right) = \mathrm{E}\left[Y \middle| X\right]$. Irrespective of the null or alternative being true, one may define $\sigma_0^2 \coloneqq \mathrm{E}\left[Y^2 - h^*\left(X\right)^2\right]$. It follows by iterated expectations that $\mathrm{E}\left[Y^2 - h^*\left(X\right)^2 \middle| X\right] = \sigma^2$ implies $\sigma^2 = \sigma_0^2$, so σ_0^2 is pseudo true. Given an estimator $\widehat{h}$ of h^*, a natural estimator of σ_0^2 solves the sample moment condition $n^{-1}\sum_{i=1}^n\{Y_i^2 - \widehat{h}\left(X_i\right)^2 - \sigma^2\} = 0$. In this example, the estimator $\widehat{\sigma}^2 \coloneqq n^{-1}\sum_{i=1}^n\{Y_i^2 - \widehat{h}\left(X_i\right)^2\}$ is available in closed form and is one example of a two-step generalized method of moments (two-step GMM) estimator based on a nonparametric first step.

The simulation design of Section 4 gives an example of (2.3) in the context of a two-player, binary-action static game of incomplete information for which the parameters may be estimated using two-step (pseudo) maximum likelihood. While defining a

pseudo truth via a moment condition may seem natural in the present context, in order to allow other methods of estimation, I do not force the pseudo truth to satisfy a moment condition of the form (2.3).

2.2 Recasting the Problem

To motivate a test statistic for testing (2.1) against (2.2), note that the presence of a pseudo truth $\beta_0 \in \mathcal{B}$ implies that the null hypothesis may be equivalently stated as

$$\mathrm{H}_0 : \mathrm{E}\left[\rho_\ell\left(Z, \beta_0, h_\ell^*\left(W_\ell\right)\right) \middle| X_\ell\right] = 0 \text{ a.s. } X_\ell \text{ for all } \ell \in \{1, \ldots, L\}. \tag{2.4}$$

Suppose for the moment that there is only one CMR to be tested and let $U \coloneqq \rho_1\left(Z, \beta_0, h_1^*\left(W_1\right)\right)$, $X = X_1$ and $d_x = d_{x,1}$ abbreviate the model residual and the conditioning variables, respectively. Then we may write the null hypothesis as,

$$\mathrm{E}\left[U \middle| X\right] = 0 \text{ a.s.} \tag{2.5}$$

Note that $\mathrm{E}\left[U \middle| X\right] = 0$ a.s. if and only if $\mathrm{E}\left[Ug\left(X\right)\right] = 0$ for all bounded "test functions" $g : \mathcal{X} \to \mathbf{R}$. Following Bierens and Ploberger (1997), Stinchcombe and White (1998) and Stute (1997), among others, I construct a test of the *conditional* moment restriction in (2.4) by testing the *unconditional* moment restrictions (UMRs)

$$\mathrm{E}\left[U\omega\left(t, X\right)\right] = 0 \text{ for almost every } t \in \mathcal{X}, \tag{2.6}$$

where $\mathcal{X} \coloneqq \mathcal{X}_1$ and ω denotes a proper weight function chosen so as to make (2.5) and (2.6) equivalent even though (2.6) only employs the subset $\{g; g = \omega\left(t, \cdot\right), t \in \mathcal{X}\}$ of possible test functions. (See Assumption 2 below for formal requirements of this choice.)

Bierens and Ploberger (1997) [with its addendum in Bierens (2017)] and Stinchcombe and White (1998) give detailed discussions on how to choose weight functions ensuring the equivalence between a CMR and a family of UMRs. Example weight functions from these references are the *exponential* $\omega\left(t, x\right) = \exp\left(t^\top x\right)$, *logistic* $\omega\left(t, x\right) = 1/[1 + \exp(c - t^\top x)]$ with $c \neq 0$, and *cosine-sine* $\omega\left(t, x\right) = \cos\left(t^\top x\right) + \sin\left(t^\top x\right)$.[6]

[6]Strictly speaking, exponential weighting requires X bounded in order to ensure $\omega\left(t, \cdot\right)$ bounded. However, for unbounded X, one may replace X with any bounded, one-to-one transformation thereof.

In general, $L \geqslant 1$ and one must choose a proper weight function for each CMR. Having settled on such proper weight functions $\{\omega_\ell\}_1^L$, the null hypothesis (2.1) may be rephrased as

$$\mathrm{H}_0 : \mathrm{E}\left[\rho_\ell\left(Z, \beta_0, h_\ell^*\left(W_\ell\right)\right) \omega_\ell\left(t_\ell, X_\ell\right)\right] = 0 \text{ almost every } t_\ell \in \mathcal{X}_\ell \text{ and all } \ell \in \{1, \ldots, L\}. \tag{2.7}$$

To further motivate the test statistic, define functions $M_\ell : \mathcal{X}_\ell \to \mathbf{R}, \ell \in \{1, \ldots, L\}$, by

$$M_\ell\left(t_\ell\right) \coloneqq \mathrm{E}\left[\rho_\ell\left(Z, \beta_0, h_\ell^*\left(W_\ell\right)\right) \omega_\ell\left(t_\ell, X_\ell\right)\right],$$

and let F_{X_ℓ} denote the distribution of the conditioning variables X_ℓ from the ℓth CMR.[7] Then squaring and integrating the collection of UMRs in (2.7), we may equivalently express the testing problem as

$$\mathrm{H}_0 : \sum_{\ell=1}^{L} \int_{\mathcal{X}_\ell} M_\ell\left(t_\ell\right)^2 \mathrm{d}F_{X_\ell}\left(t_\ell\right) = 0 \tag{2.8}$$

versus

$$\mathrm{H}_1 : \sum_{\ell=1}^{L} \int_{\mathcal{X}_\ell} M_\ell\left(t_\ell\right)^2 \mathrm{d}F_{X_\ell}\left(t_\ell\right) > 0. \tag{2.9}$$

The left-hand side expresses the null hypothesis as the sum of mean-square deviations of each M_ℓ from the zero function (on $\mathcal{X}_\ell$). Based on this representation of the testing problem, I propose to construct a test of the CMRs in (1.1) using a Cramér-von Mises-type (CM-type) measure of distance. Concretely, I define my test statistic as

$$T_n \coloneqq n \sum_{\ell=1}^{L} \int_{\mathcal{X}_\ell} \widehat{M}_\ell\left(t_\ell\right)^2 \mathrm{d}\widehat{F}_{X_\ell}\left(t_\ell\right) = \sum_{\ell=1}^{L} \sum_{i=1}^{n} \widehat{M}_\ell\left(X_{\ell i}\right)^2, \tag{2.10}$$

where $\widehat{F}_{X_\ell}$ is the empirical distribution,

$$\widehat{F}_{X_\ell}\left(t_\ell\right) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(X_{\ell i} \leqslant t_\ell\right), \quad \ell \in \{1, \ldots, L\}, \tag{2.11}$$

See also the discussion following Assumption 2.

[7] Throughout this paper, for a random variable U, I use F_U to denote both its distribution and CDF.

and I have estimated each M_ℓ by the plug-in method

$$\widehat{M}_\ell(t_\ell) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \rho_\ell(Z_i, \widehat{\beta}, \widehat{h}_\ell(W_{\ell i}))\omega_\ell(t_\ell, X_{\ell i}), \quad \ell \in \{1, \ldots, L\}, \tag{2.12}$$

with $\widehat{\beta}$ being an estimate of β_0 and each $\widehat{h}_\ell$ a nonparametric estimate of the corresponding h_ℓ^*. The formal requirements of the $\widehat{\beta}$ estimates are given in Assumption 1 below. I estimate the h_ℓ's using *series methods* (see Section 2.3).

Under general conditions presented in Section 3, the stochastic processes $\{\widehat{M}_\ell\}_1^L$ all converge to the zero function in probability under the null hypothesis, while at least one of them converges to a nonzero probability limit under the alternative. A "large" realization of T_n thus telegraphs a violation of the null.

As shown in Section 3.1 (see Theorem 3), the asymptotic distribution of T_n under the null is generally nonpivotal and its dependence on the data-generating process involved. In Section 3.2, I propose to obtain critical values via a *multiplier bootstrap* procedure and establish its asymptotic validity.

Remark 1 (Conditional vs. Unconditional). An advantage of recasting one or more CMRs as a collection of unconditional ones is that one avoids estimating the model under the alternative, thus partially circumventing the "curse of dimensionality" associated with nonparametric estimation. The potential gain may be illustrated in the nonparametric significance test described in the introduction. A direct test of $\mathrm{E}[Y|X] = \mathrm{E}[Y|W]$ a.s. may be construed by estimating both sides of the equality and calculating the distance between the two. However, when the list of candidate regressors X is moderately long, nonparametric estimation of $\mathrm{E}[Y|X]$ may be imprecise. In contrast, testing $\mathrm{E}[Y|X] = \mathrm{E}[Y|W]$ a.s. indirectly via a test of $\mathrm{E}[(Y - \mathrm{E}[Y|W])\,\omega(t, X)] = 0$ for almost every $t \in \mathcal{X}$ only involves nonparametric estimation of the CEF of Y as a function of the regressors W relevant under the null, which is an easier problem.

2.3 Series Estimation

In constructing the test statistic (2.10), I take a series approach to estimating the h_ℓ^*'s.[8] To keep notation at a minimum, suppose for the moment that the CMRs

[8] Detailed accounts of the properties of least-squares series estimators may be found in Newey (1995; 1997), Chen (2007), and Belloni, Chernozhukov, Chetverikov, and Kato (2015).

involve only a single CEF and denote it $h^*(W) = \mathrm{E}[Y|W]$. For any nonnegative integer k, let

$$w \mapsto p^k(w) \coloneqq (p_1(w), \ldots, p_k(w))^\top$$

be a k-vector of known approximating functions $\{p_j\}_1^k$.[9] Then *a series estimator* of $h^*(w)$ is given by

$$\widehat{h}(w) \coloneqq \widehat{h}_{k_n}(w) \coloneqq p^{k_n}(w)^\top \widehat{\pi}, \tag{2.13}$$

$$\widehat{\pi} \coloneqq \widehat{\pi}_{k_n} \coloneqq \left(\frac{1}{n} \sum_{i=1}^n p^{k_n}(W_i) p^{k_n}(W_i)^\top \right)^- \frac{1}{n} \sum_{i=1}^n p^{k_n}(W_i) Y_i. \tag{2.14}$$

Here $\widehat{\pi}$ denotes the vector of regression coefficients from a regression of Y_i on $p^{k_n}(W_i)$ using observations $i \in \{1, \ldots, n\}$, with k_n being a sequence of positive integers growing without bound as $n \to \infty$, and $(\cdot)^-$ is short for the Moore-Penrose generalized inverse of a matrix.[10]

In the event that more than one CEF is at play, I construct a series estimate of the form (2.13) for each distinct element of the $h_\ell^*(W_\ell)$'s. Given that the subvectors W_ℓ may differ in both content and dimension, the approximating functions p_ℓ^k must be indexed by ℓ, in general. However, to avoid further cluttering notation, I assume the same approximating functions are used for each entry of $h_\ell^*(W_\ell)$.[11]

3 Theoretical Properties

3.1 Limiting Behavior of Test Statistic

Some regularity is required to derive the limiting behavior of the test statistic. To control the influence of estimation of β_0 I assume that:

Assumption 1 (Parametric Estimation). *The pseudo truth β_0 is interior to $\mathcal{B} \subseteq \mathbf{R}^{d_\beta}$. For each $n \in \mathbf{N}$, $\widehat{\beta}$ is a random element of $\mathcal{B}$. Moreover, there exists $s : \mathcal{Z} \to \mathbf{R}^{d_\beta}$*

[9] These approximating functions may in principle change with k, which is not reflected in my notation.

[10] Under the conditions stated below, the matrix $n^{-1} \sum_{i=1}^n p^k(W_i) p^k(W_i)^\top$ is asymptotically non-singular. The particular choice of a generalized inverse is therefore asymptotically irrelevant.

[11] Use of different approximating functions for different entries is in principle allowed, provided Assumptions 5–8 are suitably modified.

such that

$$\sqrt{n}(\widehat{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s(Z_i) + o_{\mathrm{P}}(1), \tag{3.1}$$

where $s(Z)$ is centered and square integrable.

Assumption 1 requires that the centered and scaled parametric estimator is asymptotically linear with influence function s. Asymptotic linearity is admittedly a "high-level" condition.[12] That being said, Example 3 illustrates that for certain classes of two-step GMM estimators it is possible to obtain asymptotic linearity through more primitive assumptions.

Example 3 (Asymptotic Linearity in Two-Step GMM). Similar to Example 1, let β_0 be the unique solution to

$$\mathrm{E}\left[m\left(Z, \beta, h^*(W)\right)\right] = 0_{d_m \times 1},$$

with number of moments $d_m \geqslant d_\beta$. For now, take $h^*(W) = \mathrm{E}[Y|W]$ to be scalar, and define $\widehat{\beta}$ as the minimizer of

$$\beta \mapsto \widehat{m}(\beta)^\top \widehat{W} \widehat{m}(\beta) \text{ where } \widehat{m}(\beta) \coloneqq \frac{1}{n} \sum_{i=1}^{n} m(Z_i, \beta, \widehat{h}(W_i)), \quad \widehat{W} \xrightarrow{\mathrm{P}} W,$$

where $\widehat{h}$ is some nonparametric estimator of h^*, and W is a positive definite, non-stochastic $d_m \times d_m$ matrix. Then $\widehat{\beta}$ is a *two-step GMM estimator* based on a nonparametric first step. Newey (1994, Lemma 5.3) provides conditions under which such a two-step GMM estimator based on a nonparametric first step is $\sqrt{n}$-asymptotically normal and provide tools for calculating its asymptotic variance.[13] Inspection of Newey's argument reveals that the same set of conditions actually yields the slightly

[12] Assumption 1 also implies that the pseudo-truth β_0 is root-n estimable, which is not an innocuous requirement in conditional moment models. (see, for instance, Chen and Pouzo, 2015). It may be possible to this assumption 1 to allow for slower-than-root-n estimability by rescaling appropriate quantities by the relevant rate of convergence.

[13] Newey's (1994) framework is more general than presented here. Specifically, in his setup, nonparametric component(s) h^* need not be CEF(s) and moment functions may depend on the entire function h^* rather than just their values $h^*(w)$.

stronger result of asymptotic linearity. Specifically, under Newey's conditions,

$$\sqrt{n}(\widehat{\beta}-\beta_0) = -\left(M^\top W M\right)^{-1} M^\top W \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{m\left(Z_i, \beta_0, h^*\left(W_i\right)\right) + \alpha\left(Z_i\right)\right\} + o_{\mathrm{P}}\left(1\right), \tag{3.2}$$

where $M \coloneqq \mathrm{E}\left[\left(\partial/\partial\beta^\top\right) m\left(Z, \beta_0, h^*\left(W\right)\right)\right]$ is a Jacobian term, and α is an adjustment to the moment function due to estimation of h^*. Because h^* is a CEF, Newey (1994, Proposition 4) shows that, irrespective of the choice of nonparametric estimator, under some conditions, the adjustment is of the form

$$\alpha\left(z\right) = \delta\left(w\right)\left\{y - h^*\left(w\right)\right\}, \quad \delta\left(W\right) \coloneqq \mathrm{E}\left[\frac{\partial}{\partial h} m\left(Z, \beta_0, h^*\left(W\right)\right) \middle| W\right], \tag{3.3}$$

where $\left(\partial/\partial h\right) m\left(z, \beta_0, h^*\left(w\right)\right)$ denotes the (ordinary) derivative $\left(\partial/\partial h\right) m(z, \beta_0, h)$ with respect to the third argument evaluated at $h = h^*\left(w\right)$. The influence function is therefore given by

$$s\left(z\right) = -\left(M^\top W M\right)^{-1} M^\top W \left(m\left(z, \beta_0, h^*\left(w\right)\right) + \delta\left(w\right)\left\{y - h\left(w\right)\right\}\right),$$

with δ defined in (3.3).[14]

When the moment function $m\left(z, \beta, h^*\left(w\right)\right)$ depends on a *vector* $h^*\left(W\right)$ [abbreviating the distinct elements of the $h_\ell^*\left(W_\ell\right)$'s], the total adjustment to the moment function is given by adding up the individual adjustment terms (Newey, 1994, p. 1357) . That is, the adjustment in (3.2) becomes

$$\begin{aligned} \alpha\left(z\right) &= \sum_{\ell=1}^{L} \alpha_\ell\left(z\right) = \sum_{\ell=1}^{L} \delta_\ell\left(w_\ell\right)\left\{y_\ell - h_\ell^*\left(w_\ell\right)\right\}, \\ \delta_\ell\left(W_\ell\right) &\coloneqq \mathrm{E}\left[\frac{\partial}{\partial h_\ell^\top} m\left(Z, \beta_0, h^*\left(W\right)\right) \middle| W_\ell\right], \quad \ell \in \{1, \ldots, L\}, \end{aligned} \tag{3.4}$$

where $\partial/\partial h_\ell^\top$ denotes (ordinary) differentiation with respect to the arguments corresponding to $h_\ell^*\left(W_\ell\right)$.

While primitive, easy-to-verify conditions are desirable, Assumption 1 leaves free-

[14] See also Chen, Linton, and Van Keilegom (2003), who extend Newey's (1994) more general results on two-step GMM estimation to more general Z-estimation with a possibly nonsmooth criterion function.

dom in choice beyond the two-step GMM estimation outlined in Example 3. For example, (3.1) allows for other or more general two- or multi-step estimation procedures, such as two-step extremum estimation. Such procedures typically estimate the nonparametric components in a first step, use their estimates to contruct a criterion function, and maximize or minimize over β in order to produce a second-step estimator $\widehat{\beta}$. For example, one may let $\widehat{\beta}$ be a sieve minimum distance (SMD) estimator (Ai and Chen, 2003) or a penalized sieve minimum distance (PSMD) estimator (Chen and Pouzo, 2009; 2012).

The calculations outlined in Example 3 needed to verify Assumption 1 can be made explicit in the case of testing conditional homoskedasticity.

Example 4 (**Asymptotic Linearity in Testing Conditional Homoskedasticity**). Recall that σ_0^2 is identified by $\mathrm{E}\left[Y^2 - h^*(X)^2 - \sigma^2\right] = 0$ with $h^*(X) = \mathrm{E}[Y|X]$, and may be estimated by $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n \{Y_i^2 - \widehat{h}(X_i)^2\}$ with $\widehat{h}$ being a nonparametric estimator of h^*. Differentiation of the moment function given by $m(z, \sigma^2, h) := y^2 - h^2 - \sigma^2$ shows that $M := \mathrm{E}[(\partial/\partial\sigma^2) m(Z, \sigma_0^2, h^*(X))] = -1$ and $\delta(X) := \mathrm{E}[(\partial/\partial h) m(Z, \sigma_0^2, h^*(X))|X] = -2h^*(X)$. It follows that $\alpha(z) = \delta(x)\{y - h^*(x)\} = -2h^*(x)\{y - h^*(x)\}$ and thus from (3.2) that (under some conditions),

$$\sqrt{n}(\widehat{\sigma}^2 - \sigma_0^2) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\{Y_i^2 - h^*(X_i)^2 - \sigma_0^2\} - 2h^*(X_i)\{Y_i - h^*(X_i)\}\right) + o_{\mathrm{P}}(1).$$

Hence, in this example $s(z) = \{y^2 - h^*(x)^2 - \sigma_0^2\} - 2h^*(x)\{y - h^*(x)\}$.

I impose the following conditions on the choice of weight functions used in converting CMRs into an equivalent collection of UMRs.

Assumption 2 (**Weight Function**). *Each $\mathcal{X}_\ell \subset \mathbf{R}^{d_{x,\ell}}$ is compact. Each weight function $\omega_\ell : \mathcal{X}_\ell \times \mathcal{X}_\ell \to \mathbf{R}$ is continuous, has the property that (2.5) if and only if (2.6), and satisfies the Lipschitz condition: for all $t_1, t_2, x_\ell \in \mathcal{X}_\ell$ and some finite constant C_ℓ, $|\omega_\ell(t_1, x_\ell) - \omega_\ell(t_2, x_\ell)| \leqslant C_\ell \|t_1 - t_2\|$.*

Examples of weight functions satisfying Assumption 2 and references giving detailed discussion of the equivalence between (2.5) and (2.6) were provided in Section 2.2.

At first glance, the compactness of each $\mathcal{X}_\ell$ appears to rule out unbounded conditioning variables. However, if X_ℓ is not bounded, one may replace it with $\widetilde{X}_\ell \coloneqq \Phi(X_\ell)$ for any $\Phi : \mathbf{R}^{d_{x,\ell}} \to \mathbf{R}^{d_{x,\ell}}$ bounded. Provided Φ is also one-to-one, such a transformation entails no loss in generality in the sense that $\mathrm{E}[U|X] = \mathrm{E}[U|\Phi(X)]$ a.s. The compactness "assumption" thus only acts as a reminder to conduct such a preliminary transformation, if necessary. In the simulation study (Section 4) I use an elementwise arctan transform to reduce otherwise unbounded conditioning variables to a bounded set prior to calculating weights.

I next impose conditions on the residual functions. For this purpose, let d_ℓ be the number of elements in Y_ℓ [hence $h_\ell^*(W_\ell)$], and let $\mathcal{W}_\ell$ be the support of $W_\ell, \ell \in \{1, \ldots, L\}$.

Assumption 3 (Residual). *For each $\ell \in \{1, \ldots, L\}$, the following holds:*

1. *For each $z \in \mathcal{Z}, v_\ell \in \mathbf{R}^{d_\ell}, \beta \mapsto \rho_\ell(z, \beta, v)$ is continuous on $\mathcal{B}$ and continuously differentiable on an open neighborhood $\mathcal{N}_\ell$ of β_0. Moreover, there exist $c_\ell \in (0, 1]$ and $a_\ell : \mathcal{Z} \to \mathbf{R}_+$ integrable such that for each $z \in \mathcal{Z}, \beta \in \mathcal{N}_\ell, v_\ell \in \mathbf{R}^{d_\ell}$,*

$$\left\| \frac{\partial}{\partial \beta} \rho_\ell(z, \beta, v_\ell) - \frac{\partial}{\partial \beta} \rho_\ell(z, \beta, h_\ell^*(w_\ell)) \right\| \leqslant a_\ell(z) \left\| v_\ell - h_\ell^*(w_\ell) \right\|^{c_\ell}.$$

2. *For each $z \in \mathcal{Z}, v_\ell \mapsto \rho_\ell(z, \beta_0, v_\ell)$ is continuously differentiable on $\mathbf{R}^{d_\ell}$. Moreover, there exists $\gamma_\ell \in (0, 1]$ and $R_\ell : \mathcal{Z} \to \mathbf{R}_+$, such that for each $z \in \mathcal{Z}, v_\ell \in \mathbf{R}^{d_\ell}$,*

$$\left\| \frac{\partial}{\partial h_\ell} \rho_\ell(z, \beta_0, v_\ell) - \frac{\partial}{\partial h_\ell} \rho_\ell(z, \beta_0, h_\ell^*(w_\ell)) \right\| \leqslant R_\ell(z) \left\| v_\ell - h_\ell^*(w_\ell) \right\|^{\gamma_\ell}, \tag{3.5}$$

   and $\mathrm{E}[R_\ell(Z)] \sqrt{n} \max_{1 \leqslant m \leqslant d_\ell} \|\widehat{h}_{\ell m} - h_{\ell m}^*\|_{\mathcal{W}_\ell}^{1+\gamma_\ell} \to_{\mathrm{P}} 0$.

3. *The following are integrable: $|\rho_\ell(Z, \beta_0, h_\ell^*(W_\ell))|$, $\|(\partial/\partial h)\rho_\ell(Z, \beta_0, h_\ell^*(W_\ell))\|^2$ and $\sup_{\beta \in \mathcal{N}_\ell} \|(\partial/\partial \beta)\rho_\ell(Z, \beta, h_\ell^*(W_\ell))\|$.*

Assumptions 3.1 and 3.2 are smoothness conditions facilitating a linearization around (β_0, h^*) in order to extract the dominant component of the processes $\{\widehat{M}_\ell\}_1^L$ used in constructing the test statistic. The differentiability assumptions may likely be relaxed at the expense of longer proofs. I leave such extensions for future research.

Assumption 3.2 generally requires each element of $\widehat{h}_\ell$ to converge to the corresponding element of h_ℓ^* uniformly over $\mathcal{W}_\ell$ at a sufficiently fast rate.[15] Such a rate requirement often boils down to assuming that the estimand is sufficiently smooth relative to its number of arguments.

While the previous assumptions in principle allow for general nonparametric estimation methods, the following three conditions are tailored to series estimators. The first assumption is prevalent in the series estimation literature [see, e.g., Stone (1985); Newey (1994; 1997); and Belloni et al. (2015)].

Assumption 4 (Variance). $\operatorname{var}(Y_{\ell m}|\, W_\ell)$ *is bounded for all* $m \in \{1, \ldots, d_\ell\}, \ell \in \{1, \ldots, L\}$.

The second assumption imposes regularity conditions on the approximating functions.

Assumption 5 (Eigenvalues). *The eigenvalues of* $\mathrm{E}[p_\ell^k(W_\ell)\, p_\ell^k(W_\ell)^\top]$ *are bounded from above and away from zero uniformly over* $k \in \mathbf{N}$ *for all* $\ell \in \{1, \ldots, L\}$.

Loosely speaking, Assumption 5 requires that the technical regressors $p_\ell^k(W_\ell)$ are not too co-linear. See, e.g., Belloni et al. (2015, Proposition 2.1) for more primitive sufficient conditions.

Assumptions 4 and 5 are used to control the variance of the series estimators, but do not provide control over the bias arising from approximating the estimands by a linear form. The following assumption restricts the bias—or, approximation error—provided by the approximating functions $w_\ell \mapsto p_\ell^k(w_\ell) = (p_{\ell 1}(w_\ell), \ldots, p_{\ell k}(w_\ell))^\top, \ell \in \{1, \ldots, L\}$, relative to the supremum metric.

Assumption 6 (Approximation). $h_{\ell m}^*$ *is bounded,* $m \in \{1, \ldots, d_\ell\}, \ell \in \{1, \ldots, L\}$. *Moreover, for each* $\ell \in \{1, \ldots, L\}, m \in \{1, \ldots, d_\ell\}$ *and each* $k \in \mathbf{N}$, *there exists constants* $\alpha_{\ell m} \in (0, 1), C_{\ell m} \in (0, \infty)$, *and* $\widetilde{\pi}_{\ell m} \in \mathbf{R}^k$ *such that* $\|p_\ell^{k\top} \widetilde{\pi}_{\ell m} - h_{\ell m}^*\|_{\mathcal{W}_\ell} \leqslant C_{\ell m} k^{-\alpha_{\ell m}}$.

Assumption 6 is a high-level asssumption, but it is satisfied in many cases. The exponent $\alpha_{\ell m}$ usually depends on the smoothness of the estimand $h_{\ell m}^*$ and its number

[15] A notable exception occurs when the residual is *linear* in $h_\ell^*(w)$. In this case, R_ℓ may be taken as the zero function, and the requirement $\mathrm{E}[R_\ell(Z)]\sqrt{n}\max_{1\leqslant m\leqslant d_\ell}\|\widehat{h}_{\ell m} - h_{\ell m}^*\|_{\mathcal{W}_\ell}^{1+\gamma_\ell} \to_{\mathrm{P}} 0$ becomes vacuous.

of arguments d_ℓ. When the estimand can be viewed as a member of some smooth class of functions, this exponent will typically be available from the approximation theory literature. For example, if $h^*_{\ell m}$ belongs to a Hölder ball with Hölder exponent $s_{\ell m}$ (often referred to as $h^*_{\ell m}$ being "$s_{\ell m}$-smooth"), then Assumption 6 holds with $\alpha_{\ell m} = s_{\ell m}/d_\ell$, provided $p^k_{\ell m}$ is constructed using either power series [see, e.g., Timan, 1963, Section 5.3.2; Lorentz, 1966, Theorem 8] or splines [see, e.g., Schumaker, 2007; DeVore and Lorentz, 1993].

Assumption 5 is a normalization that restricts the magnitude of the series terms. The theory to follow will also require that the size of each p^k_ℓ does not grow too fast relative to the sample size, where "size" is quantified by

$$\zeta_{\ell,k} \coloneqq \sup_{w \in \mathcal{W}_\ell} \left\| p^k_\ell(w_\ell) \right\|. \tag{3.6}$$

For specific choices of approximating functions p^k_ℓ, bounds on the corresponding $\zeta_{\ell,k}$ are readily available. For example, for p^k_ℓ power series $\zeta_{\ell,k} \leqslant Ck$, and for regression splines $\zeta_{\ell,k} \leqslant C\sqrt{k}$ (cf. Newey, 1997). See Belloni et al. (2015, Section 3) for a comprehensive list.

Remark 2 (Smallest Size of Approximating Functions). Let $L = 1$, $d_1 = 1$, $W = W_1$ and $p^k = p^k_1$. When the eigenvalues of $Q_k \coloneqq \mathrm{E}[p^k(W)p^k(W)^\top]$ are bounded away from zero, Q_k^{-1} exists and has eigenvalues bounded from above, such that $\mathrm{E}[p^k(W)^\top Q_k^{-1} p^k(W)] \leqslant C\mathrm{E}[\|p^k(W)\|^2]$. Given that

$$\mathrm{E}[p^k(W)^\top Q_k^{-1} p^k(W)] = \mathrm{tr}\left(Q_k^{-1}\mathrm{E}[p^k(W)\,p^k(W)^\top]\right) = \mathrm{tr}(I_k) = k,$$

we must have

$$\zeta_k^2 \geqslant \mathrm{E}[\|p^k(W)\|^2] \geqslant (1/C)\,\mathrm{E}[p^k(W)^\top Q_k^{-1} p^k(W)] = (1/C)\,k,$$

Thus, under Assumption 5, one necessarily has $\sqrt{k} \leqslant C\zeta_k$, i.e., $\sqrt{k}$ is the smallest order of size for p^k.

The probabilistic behavior of the test statistic T_n defined in (2.10) depends crucially on the probabilistic behavior of the stochastic processes $\{\sqrt{n}\widehat{M}_\ell\}_1^L$ given in (2.12). In fact, a linearization argument (cf. Lemma 1) shows that each $\sqrt{n}\widehat{M}_\ell$ is asymptotically equivalent to a stochastic process $t_\ell \mapsto n^{-1/2}\sum_{i=1}^n [g_\ell(t_\ell, Z_i)]$, $t_\ell \in \mathcal{X}_\ell$,

defined by

$$g_\ell(t_\ell, z) := \rho_\ell(z, \beta_0, h_\ell^*(w_\ell))\,\omega_\ell(t_\ell, x_\ell) + b_\ell(t_\ell)^\top s(z) + \delta_\ell(t_\ell, w_\ell)^\top \{y_\ell - h_\ell^*(w_\ell)\}, \tag{3.7}$$

$$b_\ell(t_\ell) := \mathrm{E}\left[\omega_\ell(t_\ell, X_\ell)\frac{\partial}{\partial\beta}\rho_\ell(Z, \beta_0, h_\ell^*(W_\ell))\right], \tag{3.8}$$

$$\delta_\ell(t_\ell, W_\ell) := \mathrm{E}\left[\omega_\ell(t_\ell, X_\ell)\frac{\partial}{\partial h_\ell}\rho_\ell(Z, \beta_0, h_\ell^*(W_\ell))\middle| W_\ell\right], \tag{3.9}$$

with s provided by Assumption 1. Here $b_\ell(t_\ell)^\top s(z)$ and $\delta_\ell(t_\ell, w_\ell)^\top \{y_\ell - h_\ell^*(w_\ell)\}$ are adjustments to the moment function $z \mapsto \rho_\ell(z, \beta_0, h_\ell^*(w_\ell))\,\omega_\ell(t_\ell, x_\ell)$ due to estimation of β_0 and h_ℓ^*, respectively. The form of the β-adjustment follows from a mean-value expansion with $b_\ell(t_\ell)$ being a Jacobian term. The form of the h-adjustment is akin to the adjustment to the influence function in two-step GMM estimation with a nonparametric first step as summarized in Example 3, in particular, (3.2) and (3.3). The main difference is that, while two-step semiparametric GMM estimation requires adjustment of a finite number of moments used in defining the GMM criterion function, I here need to adjust a possibly infinite collection of moment functions $\{z \mapsto \rho_\ell(z, \beta_0, h_\ell^*(w_\ell))\,\omega_\ell(t_\ell, x_\ell); t_\ell \in \mathcal{X}_\ell\}$ for estimation of h_ℓ^*.

For the purpose of stating the following assumption, define the mean-square projection coefficients

$$\pi_{h_{\ell m},k} := \operatorname*{argmin}_{\pi\in\mathbf{R}^k} \mathrm{E}\left[\left\{p_\ell^k(W_\ell)^\top \pi - h_{\ell m}^*(W_\ell)\right\}^2\right], \tag{3.10}$$

$$\pi_{\delta_{\ell m},k}(t_\ell) := \operatorname*{argmin}_{\pi\in\mathbf{R}^k} \mathrm{E}\left[\left\{p_\ell^k(W_\ell)^\top \pi - \delta_{\ell m}(t_\ell, W_\ell)\right\}^2\right], \tag{3.11}$$

and their induced mean-square errors

$$r_{h_{\ell m},k}^2 := \min_{\pi\in\mathbf{R}^k} \mathrm{E}\left[\left\{p_\ell^k(W_\ell)^\top \pi - h_{\ell m}^*(W_\ell)\right\}^2\right], \tag{3.12}$$

$$r_{\delta_{\ell m},k}^2(t_\ell) := \mathrm{E}\left[\left\{p_\ell^k(W_\ell)^\top \pi - \delta_{\ell m}(t_\ell, W_\ell)\right\}^2\right], \tag{3.13}$$

$$R_{\delta_{\ell m},k}^2 := \mathrm{E}\left[\sup_{t_\ell\in\mathcal{X}_\ell}\left|p_\ell^k(W_\ell)^\top \pi_{\delta_{\ell m},k}(t_\ell) - \delta_{\ell m}(t_\ell, W_\ell)\right|^2\right], \tag{3.14}$$

where $\ell \in \{1, \ldots, L\}$, $t_\ell \in \mathcal{X}_\ell$ and $m \in \{1, \ldots, d_\ell\}$. Assumption 7 contains rate con-

ditions sufficient to show that the difference between $\sqrt{n}\widehat{M}_\ell$ and $n^{-1/2}\sum_{i=1}^n g_\ell(\cdot, Z_i)$ is asymptotically negligible, $\ell \in \{1, \ldots, L\}$.

Assumption 7 (**Rate Conditions**). *For all $\ell \in \{1, \ldots, L\}$ and $m \in \{1, \ldots, d_\ell\}$ and $\alpha_{\ell m}$ provided by Assumption 6,*

$$\zeta_{k_{\ell m,n}} r_{h_{\ell m},k_{\ell m,n}} \to 0, \qquad \frac{\zeta^2_{\ell m,k_{\ell m,n}} k_{\ell m,n} \ln(k_{\ell m,n})}{n} \to 0, \quad n r^2_{h_{\ell m},k_{\ell m,n}} \|r_{\delta_{\ell m},k_{\ell m,n}}\|^2_{\mathcal{X}_\ell} \to 0,$$
$$R_{\delta_{\ell m},k_{\ell m,n}} \to 0, \quad R_{\delta_{\ell m},k_{\ell m,n}} \sqrt{\ln\left(\frac{k_{\ell m,n}}{R_{\delta_{\ell m},k_{\ell m,n}}}\right)} \to 0,$$

and

$$\Big(\sum_{j=1}^{k_{\ell m,n}} \|p_{\ell,j}\|^2_{\mathcal{W}_\ell}\Big)^{1/2} \left(\sqrt{\frac{k_{\ell m,n}}{n}} + k_{\ell m,n}^{-\alpha_{\ell m}}\right) \to 0.$$

In discussing the rate conditions, consider the scalar case and drop the ℓ and m subscripts. Given that $\zeta_k \leqslant (\sum_{j=1}^k \|p_{jk}\|^2_{\mathcal{W}})^{1/2}$, the last rate condition ensures that $\zeta_{k_n}(\sqrt{k_n/n} + k_n^{-\alpha}) \to 0$, which I use to argue uniform consistency of the series estimators. Note that the presence of ζ_k in the rate conditions formally requires one to use approximating functions that are bounded on $\mathcal{W}$. However, the simulations in Section 4—where I construct approximating functions based on power series even though the conditioning variables have unbounded support—suggest that this formal requirement can be relaxed.

Observe that the mean-square error r_{h,k_n} resulting from approximating h^* by linear forms is not required to go to zero at a rate faster than $n^{-1/2}$. Such a condition would otherwise require choosing k_n larger than what would maximize its rate of convergence (sometimes referred to as "undersmoothing"). Instead Assumption 7 requires the *product* of r_{h,k_n} and the maximal approximation mean-square error $\|r_{\delta,k_n}\|_{\mathcal{X}}$ to be $o(n^{-1/2})$. This property arises from the orthogonality property of mean-square projections. Specifically, for the projections $h_k(\cdot) = p^k(\cdot)^\top \pi_{h,k}$ and $\delta_k(t,\cdot) = p^k(\cdot)^\top \pi_{\delta,k}(t)$ of h^* and $\delta(t,\cdot)$, respectively, the bias term $\mathrm{E}[\delta(t,W)\{h_k(W) - h^*(W)\}]$ is equal to $\mathrm{E}[\{\delta_k(t,W) - \delta(t,W)\}\{h_k(W) - h^*(W)\}]$ for each $t \in \mathcal{X}$. Consequently, if the family $\{\delta(t,\cdot); t \in \mathcal{X}\}$ can be sufficiently well approximated by linear forms, there is no need to "undersmooth."[16] Newey (1994) shows that a similar feature arises in the context

[16] While undersmoothing may not be necessary to achieve the claimed asymptotic approximation, it may be "optimal" in the sense of minimizing the remainder resulting from this approximation as remarked by Donald and Newey (1994) in the context of partially linear regression.

of two-step GMM estimation with a first step based on series estimation of projection functionals, such as CEFs.

Remark 3 (On Alternatives to Series Estimation). Alternative nonparametric estimators of the CEFs may not be able to exploit such built-in orthogonality properties and may therefore require undersmoothing through choice of the tuning parameter(s). One estimation method, which has recently gained significant attention, is the Lasso (Tibshirani, 1996), also known as ℓ^1-penalized least squares. Due to penalization, the Lasso does *not* give rise to an orthogonal projection. However, theoretical guidance for choosing the penalty [see, e.g., Bickel, Ritov, and Tsybakov (2009); Belloni and Chernozhukov (2011; 2013)] does not allow for undersmoothing. Nonetheless, building on ideas from the literatures on "double machine learning" (see, e.g., Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2017a) and high-dimensional central limit theorems [see Chernozhukov, Chetverikov and Kato (2013; 2017b)], in related work I develop a specification test for moment restrictions with CEFs estimated by the Lasso (see Sørensen, 2018).

The previous assumptions suffice for the asymptotic equivalence posited above.

Lemma 1 (**Asymptotic Equivalence**). *If Assumptions 1–7 hold, then for* $\{\widehat{M}_\ell\}_1^L$ *in (2.12) and* $\{g_\ell\}_1^L$ *in (3.7) we have*

$$\max_{1\leqslant\ell\leqslant L}\left\|\sqrt{n}\widehat{M}_\ell\left(\cdot\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_\ell\left(\cdot,Z_i\right)\right\|_{\mathcal{X}_\ell}\xrightarrow{\mathrm{P}}0.$$

Lemma 1 implies that the probabilistic behavior of $\|\sqrt{n}\widehat{M}_\ell\|$ may be approximated by that of $\|n^{-1/2}\sum_i g(\cdot,Z_i)\|$ for any norm $\|\cdot\|$ weaker than the supremum norm, such as the empirical L^2-norm implicit in the definition of T_n.

A class $\mathcal{F}$ of real-valued functions is called a *Donsker class* (van der Vaart and Wellner, 1996, pp. 81-82), if the sequence of empirical processes $\{n^{-1/2}\sum_{i=1}^n\{f\left(Z_i\right)-\mathrm{E}\left[f\left(Z\right)\right]\};f\in\mathcal{F}\}$ induced by $\mathcal{F}$—viewed as random elements of $L^\infty(\mathcal{F})$—converges weakly[17] to a centered Gaussian process $\{\mathbb{G}\left(f\right);f\in\mathcal{F}\}$ with covariance function

$$\mathrm{E}\left[\mathbb{G}\left(f_1\right)\mathbb{G}\left(f_2\right)\right]=\mathrm{E}\left[f_1\left(Z\right)f_2\left(Z\right)\right]-\mathrm{E}\left[f_1\left(Z\right)\right]\mathrm{E}\left[f_2\left(Z\right)\right],\quad f_1,f_2\in\mathcal{F}.$$

[17] A sequence X_n of stochastic processes taking values in a metric space $\mathbb{D}$ are said to *converge weakly* to X if $\mathrm{E}[h(X_n)]\to\mathrm{E}[h(X)]$ for all $h:\mathbb{D}\to\mathbf{R}$ continuous and bounded.

Define function classes

$$\mathcal{G}_\ell \coloneqq \{g_\ell(t_\ell, \cdot) : \mathcal{Z} \to \mathbf{R}; t_\ell \in \mathcal{X}_\ell\}, \quad \ell \in \{1, \ldots, L\}, \quad \mathcal{G} \coloneqq \times_{\ell=1}^{L} \mathcal{G}_\ell.$$

The same set of assumptions then also shows:

Lemma 2 (**Donsker Class**). *If Assumptions 1–7 hold, then $\mathcal{G}$ is Donsker.*

Note that, for each $t_\ell \in \mathcal{X}_\ell$ and $\ell \in \{1, \ldots, L\}$,

$$\mathrm{E}[g_\ell(t_\ell, Z)] = \mathrm{E}[\rho_\ell(Z, \beta_0, h_\ell^*(W_\ell))\, \omega_\ell(t_\ell, X_\ell)] = M_\ell(t_\ell), \tag{3.15}$$

which follows from Assumption 1 and (3.9). Since we may identify each $\mathcal{G}_\ell$ with the corresponding $\mathcal{X}_\ell$, Lemma 2 states that the sequence of L-variate stochastic processes

$$G_n(t) \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{g(t, Z_i) - \mathrm{E}[g(t, Z)]\}, \quad t \in \mathcal{T}, \quad \mathcal{T} \coloneqq \times_{\ell=1}^{L} \mathcal{X}_\ell,$$

converges weakly in $\times_{\ell=1}^{L} L^\infty(\mathcal{X}_\ell)$ to an L-variate zero-mean Gaussian process G_M indexed by $\mathcal{T}$ and with (matrix) covariance kernel

$$\mathbb{C}_M(t, t') \coloneqq \mathrm{E}\left[\{g(t, Z) - M(t)\}\{g(t', Z) - M(t')\}^\top\right], \quad t, t' \in \mathcal{T}, \tag{3.16}$$

where $M : \mathcal{T} \to \mathbf{R}^L$ denotes the vector function

$$M(t) = (M_1(t_1), \ldots, M_L(t_L))^\top, \quad t \in \mathcal{T}.$$

The behavior of the test statistic under the null and alternative follows.

Theorem 1 (**Asymptotic Behavior of Test Statistic**). *If Assumptions 1–7 hold, then*

$$T_n \xrightarrow{d} T_0 \coloneqq \sum_{\ell=1}^{L} \int_{\mathcal{X}_\ell} G_{0\ell}(t_\ell)^2 \, \mathrm{d}F_{X_\ell}(t_\ell) \quad \textit{under the null hypothesis (2.1)},$$

$$\frac{T_n}{n} \xrightarrow{\mathrm{P}} \sum_{\ell=1}^{L} \int_{\mathcal{X}_\ell} M_\ell(t_\ell)^2 \, \mathrm{d}F_{X_\ell}(t_\ell) > 0 \quad \textit{under the fixed alternative hypothesis (2.2)},$$

where G_0 is an L-variate centered Gaussian process indexed by $\mathcal{T}$ and with covariance kernel $\mathbb{C}_0 \coloneqq \mathbb{C}_M|_{M \equiv 0}$.

Remark 4.

1. The proof of Theorem invokes a (second-order) delta method argument to show that, under the null, $T_n = n\sum_{\ell=1}^{L}\int_{\mathcal{X}_\ell}\widehat{M}_\ell(t_\ell)^2\,\mathrm{d}F_{X_\ell}(t_\ell) + o_{\mathrm{P}}(1)$. That is, the limiting null distribution is unaffected by the use of empirical distributions in place of their (unknown) population counterparts.

2. The second claim of Theorem 1 implies that $T_n \to_{\mathrm{P}} \infty$ (at the rate n) under the alternative, which plays a key role in establishing consistency in Theorem 2.

3.2 Bootstrap Critical Values

The limit results in Theorem 1 cannot be implemented for inference without a consistent estimator for the appropriate critical values. For this purpose, I employ a *Gaussian multiplier bootstrap* procedure.

By Theorem 1, the limiting law of T_n under the null hypothesis is given by that of $T_0 = \sum_{\ell=1}^{L}\int_{\mathcal{X}_\ell} G_{0\ell}(t_\ell)^2\,\mathrm{d}F_{X_\ell}(t_\ell)$. To obtain a consistent bootstrap, it is therefore necessary to estimate the law of the Gaussian process G_0. Toward this end, let $\{\xi_i\}_1^\infty$ be a sequence of i.i.d. standard normal random variables independent of the stream of data $\{Z_i\}_1^\infty$ and let $\overline{\xi} \coloneqq n^{-1}\sum_{i=1}^{n}\xi_i$ abbreviate their average.

To fix ideas, consider first the *multiplier process* $G_n^* \coloneqq (G_{1n}^*, \ldots, G_{Ln}^*)$ defined by

$$G_n^*(t) \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\xi_i - \overline{\xi}\right) g(t, Z_i), \quad t \in \mathcal{T}. \tag{3.17}$$

By independence, the summands of G_n^* are centered even if one or more of the $g_\ell(t_\ell, Z)$'s are not, i.e., even when the null is false. The purpose of including $\overline{\xi}$ in (3.17) is to take into account that the $g_\ell(t_\ell, Z_i)$ may not be centered with respect to the empirical distribution even if the null is true.[18] The sample centering aims for less conservative critical values in finite sample by correctly accounting for sample variation.

[18] Rearranging, this connection can be made explicit:

$$G_n^*(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i\left\{g(t, Z_i) - \overline{g}(t)\right\}, \quad \overline{g}(t) \coloneqq \frac{1}{n}\sum_{i=1}^{n} g(t, Z_i).$$

The following discussion relies on the notion of *weak convergence in probability*. The multiplier process G_n^* is said to *converge weakly in probability* to G^* in $\times_{\ell=1}^L L^\infty(\mathcal{X}_\ell)$, written $G_n^* \rightsquigarrow_{\mathrm{P},\xi} G^*$, if G_n^* converges weakly to G^* conditional on the data, in probability.[19] Given that $\mathcal{G}$ is Donsker (Lemma 2), Kosorok (2008, Theorem 10.4) shows that $G_n^* \rightsquigarrow_{\mathrm{P},\xi} G_M$ in $\times_{\ell=1}^L L^\infty(\mathcal{X}_\ell)$. Under the null, $M \equiv 0$ on $\mathcal{T}$, and the covariance function of G_M coincides with that of G_0. Since both G_M and G_0 are Gaussian, under the null the two must therefore be identically distributed. This observation suggests using the $(1-\alpha)$-quantile of $\sum_{\ell=1}^L \int_{\mathcal{X}_\ell} G_{\ell n}^*(t_\ell)^2 \,\mathrm{d}F_{X_\ell}(t_\ell)$ conditional on $\{Z_i\}_1^n$ to approximate

$$c_{T_M}(\alpha) \coloneqq (1-\alpha)\text{-quantile of } T_M, \quad T_M \coloneqq \sum_{\ell=1}^L \int_{\mathcal{X}_\ell} G_{M,\ell}(t_\ell)^2 \,\mathrm{d}F_{X_\ell}(t_\ell).$$

Of course, the g_ℓ's—as well as the F_{X_ℓ}'s—are generally unknown, thus rendering the above procedure infeasible. However, endowed with an estimator $\widehat{s}$ of the influence function s from Assumption 1, one may estimate g and define the *bootstrap process* $\widehat{G} \coloneqq (\widehat{G}_1, \ldots, \widehat{G}_L)^\top$ as the feasible version of G_n^*. Specifically, I define

$$\widehat{G}(t) \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\xi_i - \overline{\xi}\right) \widehat{g}(t, Z_i), \quad t \in \mathcal{T}, \tag{3.18}$$

$$\widehat{g}(t,z) \coloneqq \left(\widehat{g}_1(t_1,z), \ldots, \widehat{g}_L(t_L,z)\right)^\top, \tag{3.19}$$

$$\widehat{g}_\ell(t_\ell,z) \coloneqq \rho_\ell(z,\widehat{\beta},\widehat{h}_\ell(w_\ell))\omega_\ell(t_\ell,x_\ell) + \widehat{b}_\ell(t_\ell)^\top \widehat{s}(z) + \widehat{\delta}_\ell(t_\ell,w_\ell)^\top \{y_\ell - \widehat{h}_\ell(w_\ell)\}, \tag{3.20}$$

$$\widehat{b}_\ell(t) \coloneqq \frac{1}{n}\sum_{i=1}^n \omega_\ell(t_\ell, X_{\ell i}) \frac{\partial}{\partial \beta} \rho_\ell(Z_i, \widehat{\beta}, \widehat{h}_\ell(W_{\ell i})), \tag{3.21}$$

$$\begin{aligned}\widehat{\delta}_{\ell m}(t_\ell, w_\ell) \coloneqq\ & p_\ell^{k_{\ell m,n}}(w_\ell)^\top \left(\frac{1}{n}\sum_{i=1}^n p_\ell^{k_{\ell m,n}}(W_{\ell i}) p_\ell^{k_{\ell m,n}}(W_{\ell i})^\top\right)^- \\ & \times \frac{1}{n}\sum_{i=1}^n p_\ell^{k_{\ell m,n}}(W_{\ell i}) \omega_\ell(t_\ell, X_{\ell i}) \frac{\partial}{\partial h_{\ell m}} \rho_\ell(Z_i, \widehat{\beta}, \widehat{h}_\ell(W_{\ell i})).\end{aligned} \tag{3.22}$$

Replacing the multiplier process G_n^* with the bootstrap process $\widehat{G}$ and the F_{X_ℓ}'s

[19] That is, if $\mathrm{E}[h(G_n^*)|\{Z_i\}_1^n] \to_{\mathrm{P}} \mathrm{E}[h(X)]$ for all $h : \times_{\ell=1}^L L^\infty(\mathcal{X}_\ell) \to \mathbf{R}$ continuous and bounded. An equivalent definition is $\sup_{h \in \mathrm{BL}_1(\times_{\ell=1}^L L^\infty(\mathcal{X}_\ell))} |\mathrm{E}[h(G_n^*)|\{Z_i\}_1^n] - \mathrm{E}[h(G^*)]| \to_{\mathrm{P}} 0$, where $\mathrm{BL}_1(\mathbb{D})$ denotes the space of functionals $h : \mathbb{D} \to \mathbf{R}$ defined on the metric space $(\mathbb{D}, d)$ whose Lipschitz norm is bounded by one, i.e., functionals satisfying $\|h\|_{\mathbb{D}} \leqslant 1$ and $|h(f) - h(g)| \leqslant d(f,g)$ for all $f, g \in \mathbb{D}$.

with their empirical analogs, we arrive at a feasible bootstrap test statistic, namely

$$\widehat{T} \coloneqq \sum_{\ell=1}^{L} \int_{\mathcal{X}_\ell} \widehat{G}_\ell\left(t_\ell\right)^2 \mathrm{d}\widehat{F}_{X_\ell}\left(t_\ell\right) = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i=1}^{n} \widehat{G}_\ell\left(X_{\ell i}\right)^2 . \tag{3.23}$$

A feasible critical value is therefore given by

$$c_{\widehat{T}}\left(\alpha\right) \coloneqq \left(1-\alpha\right)\text{-quantile of } \widehat{T} \text{ conditional on } \left\{Z_i\right\}_1^n . \tag{3.24}$$

For a given significance level $\alpha \in (0,1)$, the critical value $c_{\widehat{T}}\left(\alpha\right)$ may be obtained through simulation of the Gaussian multipliers $\left\{\xi_i\right\}_1^n$ holding the data constant.[20]

The test rejects the null hypothesis (2.1) if $T_n > c_{\widehat{T}}\left(\alpha\right)$ for some prespecified significance level $\alpha \in (0,1)$, where the test statistic is defined in (2.10) and the critical value in (3.24).

Remark 5 (Additively Separable Residuals). If a residual function is *additively separable* in the conditioning variables, in the sense that $\rho_\ell(z, \beta_0, h_\ell^*(w_\ell)) = \phi_\ell(y_\ell, \beta_0) + \varphi_\ell(x_\ell, \beta_0, h_\ell^*(w_\ell))$, then $(\partial/\partial h_\ell)\rho_\ell(z, \beta_0, h_\ell^*(w_\ell)) = (\partial/\partial h_\ell)\varphi_\ell(x_\ell, \beta_0, h_\ell^*(w_\ell))$ depends on z through x_ℓ alone. If, in addition, X_ℓ and W_ℓ coincide,[21] then the term $\omega_\ell(t_\ell, X_\ell)$ $(\partial/\partial h_\ell)\rho_\ell(Z, \beta_0, h_\ell^*(W_\ell))$ must be conditionally known given W_ℓ (up to β_0 and h_ℓ^*) and hence equal to $\delta_\ell\left(t_\ell, W_\ell\right)$. For such models, one may therefore drop the projection element of the estimator in (3.22) and replace it by the simpler $\widehat{\delta}_\ell\left(t_\ell, W_{\ell i}\right) = \omega_\ell\left(t_\ell, X_{\ell i}\right)(\partial/\partial h_\ell)\varphi_\ell(X_{\ell i}, \widehat{\beta}, \widehat{h}_\ell\left(W_{\ell i}\right))$. This simplification is utilized in Section 4.

A potentially difficult step in this bootstrap procedure is the construction of $\widehat{s}$. If s is a function $s\left(\cdot, \beta_0, h^*\right)$ known up to β_0 and h^*, a natural estimator $\widehat{s}$ is $\widehat{s}\left(\cdot\right) \coloneqq s(\cdot, \widehat{\beta}, \widehat{h})$. This structure is found in testing for conditional homoskedasticity (see Example 4). For other two-step GMM estimators, $\widehat{s}$ will typically require estimation of a Jacobian (matrix) term (cf. Example 3). In general, one may construct s estimates by first obtaining an analytic formula for s and then replacing unknown components by estimates. However, at the level of generality for the parametric component considered in this paper, it does not appear possible to give primitive conditions under which $\widehat{s}$ is consistent for s.

Assumption 8 (Bootstrap Conditions). *For each* $\ell \in \{1, \ldots, L\}$, *the following holds:*

[20] In practice, this simulation is terminated after a finite but large number of draws.

[21] Formally: if X_ℓ is measurable with respect to W_ℓ.

1. *For each* $z \in \mathcal{Z}, \beta \in \mathcal{N}_\ell, v_\ell \mapsto \rho_\ell(z, \beta, v_\ell)$ *is continuously differentiable on* $\mathbf{R}^{d_\ell}$. *Moreover, there exists* $R'_\ell : \mathcal{Z} \to \mathbf{R}_+$ *such that for each* $z \in \mathcal{Z}, \beta \in \mathcal{N}_\ell, v_\ell \in \mathbf{R}^{d_\ell}$,

$$\left\| \frac{\partial}{\partial h_\ell} \rho_\ell(z, \beta, v_\ell) - \frac{\partial}{\partial h_\ell} \rho_\ell(z, \beta_0, h^*_\ell(w_\ell)) \right\| \leqslant R'_\ell(z) \left(\|\beta - \beta_0\| + \|v_\ell - h^*_\ell(w_\ell)\| \right),$$

where $\mathrm{E}[R'_\ell(Z_i)] \sqrt{n} \max_{1 \leqslant m \leqslant d_\ell} \|\widehat{h}_{\ell m} - h^*_{\ell m}\|^2_{\mathcal{W}_\ell} \to_{\mathrm{P}} 0$;

2. $n^{-1} \sum_{i=1}^n \{\widehat{s}(Z_i) - s(Z_i)\}^2 \to_{\mathrm{P}} 0$; *and,*

3. *for all* $m \in \{1, \ldots, d_\ell\}$ *and* $\alpha_{\ell m}$*'s provided by Assumption 6,*

$$\zeta_{\ell, k_{\ell m, n}} \sqrt{k_{\ell m, n}} \left(\sqrt{k_{\ell m, n}/n} + k^{-\alpha_{\ell m}}_{\ell m, n} \right) \to 0,$$
$$\left(\sum_{j=1}^{k_{\ell m, n}} \|p_{\ell j}\|^2_{\mathcal{W}} \right)^{1/2} \max_{1 \leqslant m' \leqslant d_\ell} \left(\sqrt{k_{\ell m', n}/n} + k^{-\alpha_{\ell m'}}_{\ell m', n} \right) \to 0.$$

With the addition of Assumption 8, we obtain:

Lemma 3 (**Bootstrap Equivalence**). *If Assumptions 1–8 hold, then* $\max_{1 \leqslant \ell \leqslant L} \|\widehat{G}_\ell - G^*_{\ell n}\|_{\mathcal{X}_\ell} \to_{\mathrm{P}} 0$.

Lemma 3 establishes that the unknown character of g is asymptotically irrelevant. Given that G^*_n converges weakly in probability to G_M, by the lemma, so must its feasible analog $\widehat{G}$.

Now, the limit T_M is a nonnegative random variable arising from applying a convex functional (the sum of squares of L^2-type norms) to a Gaussian process G_M. It follows from Davydov, Lifshits, and Smorodina (1998, Theorem 11.1) that its CDF

$$F_{T_M}(u) \coloneqq \mathrm{P}(T_M \leqslant u), \quad u \in \mathbf{R}, \tag{3.25}$$

is everywhere continuous, except possibly at the separation point zero. I explicitly rule out a mass point at separation by invoking the high-level assumption:

Assumption 9 (**Continuity**). $F_{T_M}(0) = 0$.

More primitive conditions may be used to satisfy Assumption 9. For example, using the continuity of sample paths of G_M, $F_{T_M}(0) = 0$ may be obtained under the "nondegeneracy" assumption that $\mathrm{var}[g_\ell(t_\ell, Z)] > 0$ for some $t_\ell \in \mathcal{X}_\ell$ and some $\ell \in$

$\{1, \ldots, L\}$, when combined with an assumption that the corresponding distribution F_{X_ℓ} is absolutely continuous with density bounded away from zero.

Given the continuous nature of the weak in-probability limit T_M of $\widehat{T}$, convergence of quantiles essentially follows.

Lemma 4 (**Quantile Consistency**). *If Assumptions 1–9 hold, and F_{T_M} is increasing at its $(1-\alpha)$-quantile $c_{T_M}(\alpha)$, then $c_{\widehat{T}}(\alpha) \to_{\mathrm{P}} c_{T_M}(\alpha) \in (0, \infty)$.*

3.3 Limiting Behavior of Test

Theorem 2 contains the main results of this paper, namely that that the test which rejects the null hypothesis if and only if $T_n > c_{\widehat{T}}(\alpha)$ is (1) correctly sized and (2) consistent against any fixed alternative.

Theorem 2 (**Asymptotic Properties of Test**). *If Assumptions 1–9 hold, and F_{T_M} is increasing at its $(1-\alpha)$-quantile, then*

$$\mathrm{P}\left(T_n > c_{\widehat{T}}(\alpha)\right) \to \begin{cases} \alpha, & \textit{under the null hypothesis (2.1)}, \\ 1, & \textit{under the fixed alternative hypothesis (2.2)}. \end{cases}$$

4 Simulations

To demonstrate the usefulness of the proposed testing procedure and assess its finite-sample properties, I carry out a simulation experiment.

4.1 Setup: A Two-by-Two Game of Incomplete Information

One potential application of the test lies in testing for correct specification of static binary choice models with social and strategic interactions. (See the introduction for references.) I therefore use a two-player, binary-action game of incomplete information as data-generating process (DGP). The DGP considered here is a slight modification of the one in Hahn, Moon, and Snider (2017) with the addition of continuous conditioning variables.[22] Two players, indexed $j \in \{1, 2\}$, simultaneously

[22] When conditioning variables are discrete, the CEFs may be represented using a finite set of values and the estimation problem becomes parametric.

choose one of two alternatives $y_j \in \{0,1\}$. Utility of the players is parameterized as

$$u\left(y_j, y_{-j}, x_j, \varepsilon_j\left(0\right), \varepsilon_j\left(1\right); \theta\right) = \begin{cases} Ax_j + Cx_j^2 + \gamma_0 y_{-j} + \varepsilon_j\left(1\right), & y_j = 1, \\ Bx_j + Dx_j^2 + \gamma_0\left(1 - y_{-j}\right) + \varepsilon_j\left(0\right), & y_j = 0, \end{cases}$$

where y_{-j} denotes the action of the other player, x_j is a player-specific *public* payoff shock, and $\left(\varepsilon_j\left(0\right), \varepsilon_j\left(1\right)\right)$ is a vector of iid (over both players and alternatives) payoff shocks private to player j drawn from a commonly known distribution. In a Bayesian Nash equilibrium (BNE), both players maximize their expected utility given their beliefs, and their beliefs turn out correct, thus leading to the decision rule

$$Y_j = \mathbf{1}\left(\alpha_0 X_j + \delta_0 X_j^2 + \gamma_0\left(2\mathrm{E}\left[Y_{-j} \middle| X\right] - 1\right) \geqslant \varepsilon_j\right), \tag{4.1}$$

where I abbreviate $\alpha_0 \coloneqq A - B, \delta_0 \coloneqq C - D, X \coloneqq (X_1, X_2)$, and $\varepsilon_j \coloneqq \varepsilon_j\left(0\right) - \varepsilon_j\left(1\right)$. The $\varepsilon_j\left(y_j\right)$'s are here taken to be Type 1 Extreme Value distributed independent of the X_j's. Correctness of beliefs therefore leads to the CCPs

$$\mathrm{E}\left[Y_j \middle| X\right] = \Lambda\left(\alpha_0 X_j + \delta X_j^2 + \gamma_0\left(2\mathrm{E}\left[Y_{-j} \middle| X\right] - 1\right)\right), \quad j \in \{1,2\}, \tag{4.2}$$

with $\Lambda\left(t\right) = 1/(1 + \mathrm{e}^{-t})$ being the logistic CDF.

Let $\left\{\{(Y_{ij}, X_{ij})\}_{j=1}^2\right\}_{i=1}^n$ denote data from n independent games. We wish to test the null hypothesis

$$\exists \beta \coloneqq (\alpha, \gamma) \text{ s.t. } \mathrm{E}\left[Y_j - \Lambda\left(\alpha X_j + \gamma\left(2\mathrm{E}\left[Y_{-j} \middle| X\right] - 1\right)\right) \middle| X\right] = 0 \text{ a.s. for both } j \in \{1,2\}.$$

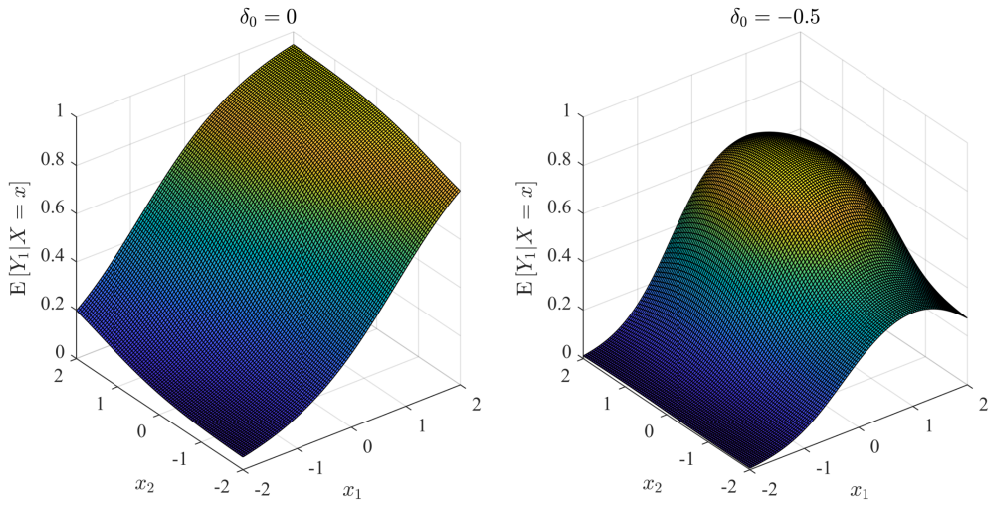
To generate data from the model, I first draw conditioning variables $X = (X_1, X_2)$, which are taken to be bivariate normal with unit variances and correlation ρ. I then solve the two equations

$$\sigma_j = \Lambda\left(\alpha_0 X_j + \delta_0 X_j^2 + \gamma_0\left(2\sigma_{-j} - 1\right)\right), \quad j \in \{1,2\}, \tag{4.3}$$

in the unknowns (σ_1, σ_2) to obtain beliefs consistent with a BNE. Outcomes are subsequently generated using the decision rules in (4.1).[23] Throughout I set $\alpha_0 =$

[23]Depending on the value of X, the nonlinear system (4.3) may in principle have multiple solutions resulting in different equilibria. The notation employed in (4.1)–(4.2) implicitly assumes uniqueness of equilibrium beliefs. The parameter values are here selected to guarantee a unique solution to this

Figure 1: Equilibrium Beliefs of Player 1 as a Function of Public Information



$\gamma_0 = 1$. To generate data consistent with the null hypothesis, I set $\delta_0 = 0$. To generate data under the alternative, I set $\delta_0 = -0.5$. The resulting equilibrium beliefs from the perspective of Player 1 as a function of the public signals are depicted in Figure 1. (The equilibrium beliefs of Player 2 mirror those of Player 1, i.e., they may be found by swapping the labels of the first and secondary axes.) In both cases, equilibrium beliefs are smooth functions in public information.

To allow for different parts of the equilibrium belief surface to be likely to be explored, in generating the public information I allow for different levels of correlation.

4.2 Construction of Test

To construct the *test statistic*, I first take a series approach to estimating the equilibrium beliefs $h_j^*(\cdot) \coloneqq \mathrm{E}[Y_j | X = \cdot]$ of both players. For both estimands I employ the power series approximating functions $p_\ell^k \coloneqq p^k$ defined as the tensor-product

$$p^k(x)^\top \coloneqq (1, x_1, \ldots, x_1^{\sqrt{k}-1}) \otimes (1, x_2, \ldots, x_2^{\sqrt{k}-1})$$

of the monomials in each argument up to the same order. The formal results of this paper are developed under the assumption that the series length $k = k_n$ grows with n at a suitably rate. However, for a given sample size, one must settle on a particular

nonlinear system of equations no matter the realization of X, thus ensuring equilibrium uniqueness. See Hahn et al. (2017) for a test of neglected heterogeneity, which may be used to detect multiplicity.

k. In order to investigate the sensitity of the test with respect to this (user) choice, I carry out my procedure for each series length $k \in \{4, 9, 16\}$.

Next, based on the logit conditional choice probabilities,

$$f\left(y_j \middle| x, (\alpha, \gamma), h\right) = \Lambda\left(\alpha x_j + \gamma(2h_{-j} - 1)\right)^{y_j} \left[1 - \Lambda\left(\alpha x_j + \gamma(2h_{-j} - 1)\right)\right]^{1-y_j},$$

I formulate a (pseudo) maximum likelihood estimator of $\beta_0 = (\alpha_0, \gamma_0)^\top$,

$$\widehat{\beta} \coloneqq \underset{\beta \in \mathbf{R}^2}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{2} \ln f(Y_{ij} | X_i, \beta, \widehat{h}(X_i)).$$

Following Bierens (1990), I use exponential weighting $\omega(\tilde{t}, \tilde{x}) = \exp(\tilde{x}^\top \tilde{t})$ combined with a preliminary arctan transformation $\widetilde{X}_j \coloneqq \tan^{-1}(X_j)$ of each (otherwise unbounded) conditioning variable. I use the same weights for both residuals.[24] The test statistic then follows from (2.10), (2.11) and (2.12) using

$$\rho_\ell(z, \beta, h) \coloneqq y_\ell - \Lambda\left(\alpha x_\ell + \gamma(2h_{-\ell} - 1)\right), \quad \ell \in \{1, 2\},$$

as residual functions,[25] and integration is understood to be against the empirical distribution of the *transformed* conditioning variables.

I obtain a *critical value* using (3.18)–(3.24). Given that the argument of the CEFs coincides with the conditioning variables and that $(\partial/\partial h_{-\ell})\rho_\ell(z, \beta_0, h^*(x)) = -2\gamma_0\Lambda'(\alpha_0 x_\ell + \gamma_0(2h^*_{-\ell}(x) - 1))$ depends on z only through x, I construct the $\widehat{\delta}_{\ell,-\ell}$'s defined in (3.22) without projections, i.e.,

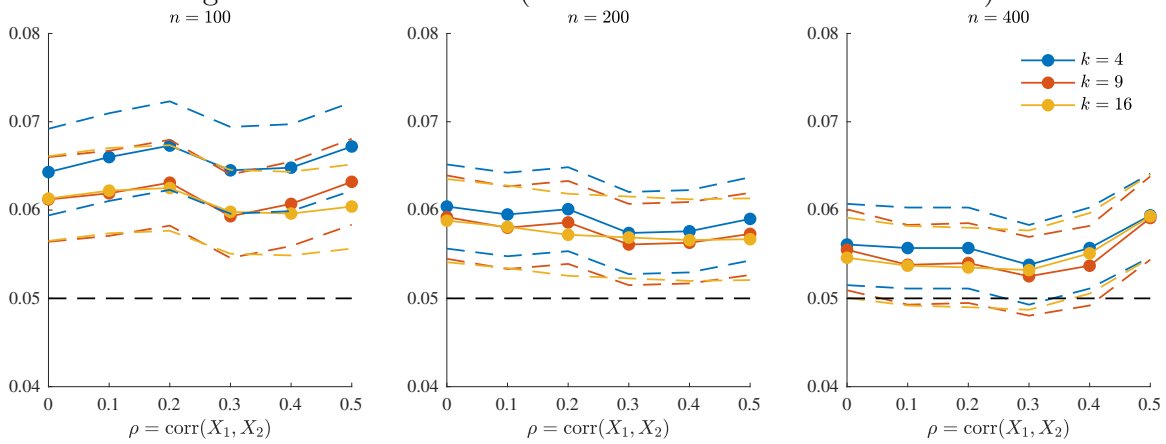
$$\widehat{\delta}_{\ell,-\ell}(\tilde{t}, X_i) \coloneqq -2\widehat{\gamma}\omega(\tilde{t}, \widetilde{X}_i)\Lambda'(\widehat{\alpha} X_{i\ell} + \widehat{\gamma}[2\widehat{h}_{-\ell}(X_i) - 1]).$$

(See also Remark 8.) The $\widehat{\delta}_{\ell,\ell}$'s are zero. To obtain the $\widehat{s}(Z_i)$ estimates needed in (3.20) to adjust for estimation of β_0, I first derive the influence function of $\sqrt{n}(\widehat{\beta} - \beta_0)$ as outlined in Example 3 for general two-step GMM estimation with a nonparametric

[24] Given that the arctan function is close to constant for values of its argument far away from zero, prior to applying a bounded one-to-one transformation, Bierens (1990) advocates centering and scaling the conditioning variables by their sample means and variances, respectively. This centering and scaling is strictly speaking not allowed for in my notation, which treats the weight function as known.

[25] To simplify derivations, the test thus constructed only makes use of the CMRs arising from the marginal distributions of the Y_j's (conditional on X). Three additional CMRs may be deduced from their joint (conditional) distribution.

Figure 2: Size Estimates (±2 Monte Carlo Standard Errors)



first step, using the (pseudo) scores $\sum_{j=1}^{2}(\partial/\partial\beta)\ln f(y_j|x,\beta,h)$ as moment functions $m(z,\beta,h)$. I then replace unknowns (including the moment Jacobian) with sample analogs.

I consider sample sizes $n \in \{100, 200, 400\}$ and levels of correlation $\rho = \text{corr}\,(X_1, X_2) \in \{0, .1, \ldots, .5\}$. The number of Monte Carlo replications is 10,000. I implement the test at a 5 percent nominal level and approximate the critical value $c_{\widehat{T}}\,(.05)$ in (3.24) using $1,000$ draws of the Gaussian multipliers within each replication.

4.3 Results

Figure 2 shows the size estimates of the test for each each sample size and series length as a function of the correlation level. The test is oversized by 1–2 percentage points for $n = 100$. For this (limited) sample size, the amount of overrejection may depend on the choice of series length by about half a percentage point. However, as the sample size increases, the size estimates appear to converge towards the nominal level across all series lengths and all correlation levels, except perhaps $\rho = 0.5$.

Figure 3 plots the power of the test when $\delta_0 = -0.5$. The power may depend on the choice of series length by upwards of 10 percentage points. As the sample size increases, the power appears to converge to one for all series lengths and all correlations. This convergence is expected, since the test is consistent against all deviations from the null.

To further investigate the size and power properties of the test, I carry out the test as if β_0 is known. For this "β-oracle" version of the test, only the equilibrium

Figure 3: Global Power Estimates (±2 Monte Carlo Standard Errors)

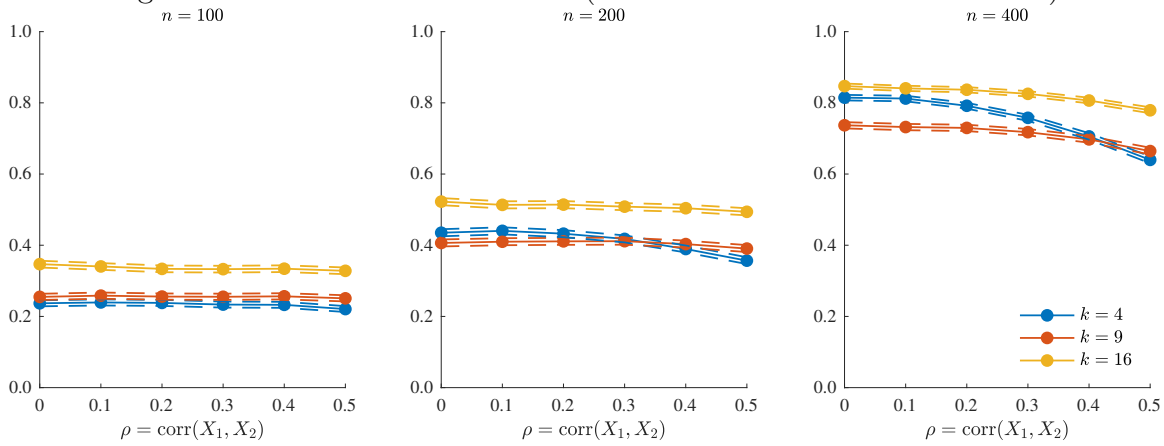
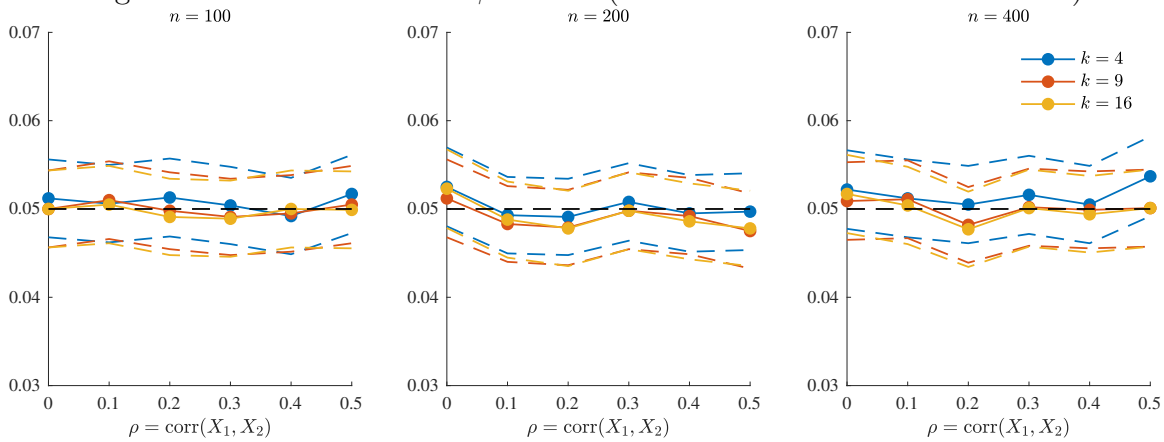


Figure 4: Size Estimates for β-Oracle (±2 Monte Carlo Standard Errors)



belief functions are estimated. The size and global power of the β-oracle test are given in Figures 4 and 5, respectively. These figures illustrate that the test delivers on its promises of asymptotic size control and consistency upon removing the need for adjustment due to (two-step) estimation of β_0.

Lastly, I briefly explore the local power of the (non-oracle) test developed in this paper. While the proposed test is not formally proven to exhibit nontrivial local power, in Figure 6 I depict estimates of its local power for the sequence of alternatives resulting from $\delta_n = -5/\sqrt{n}$. The test does have nontrivial local power, at least against this particular sequence of alternatives. Moreover, its local power appears stable across series lengths as well as correlations (at least for $n = 400$).

Figure 5: Global Power Estimates for β-Oracle (± 2 Monte Carlo Standard Errors)

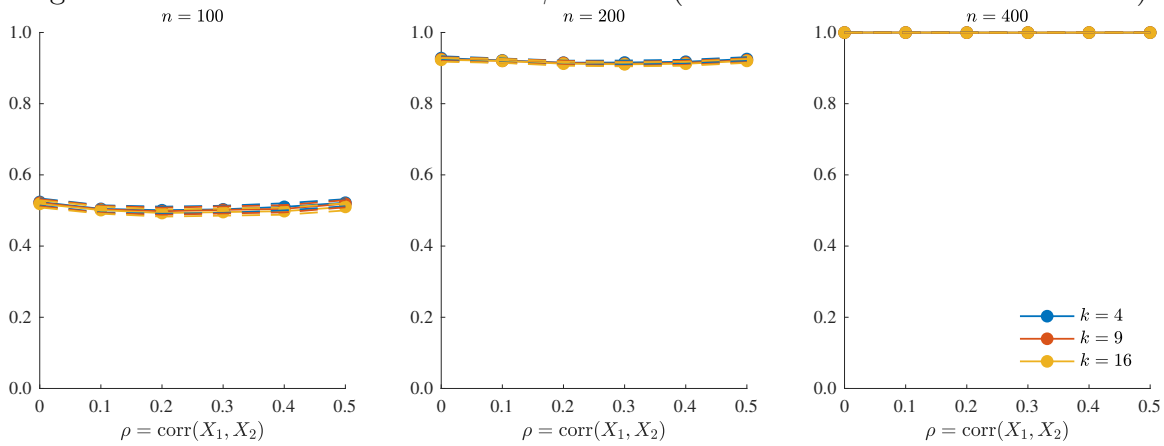
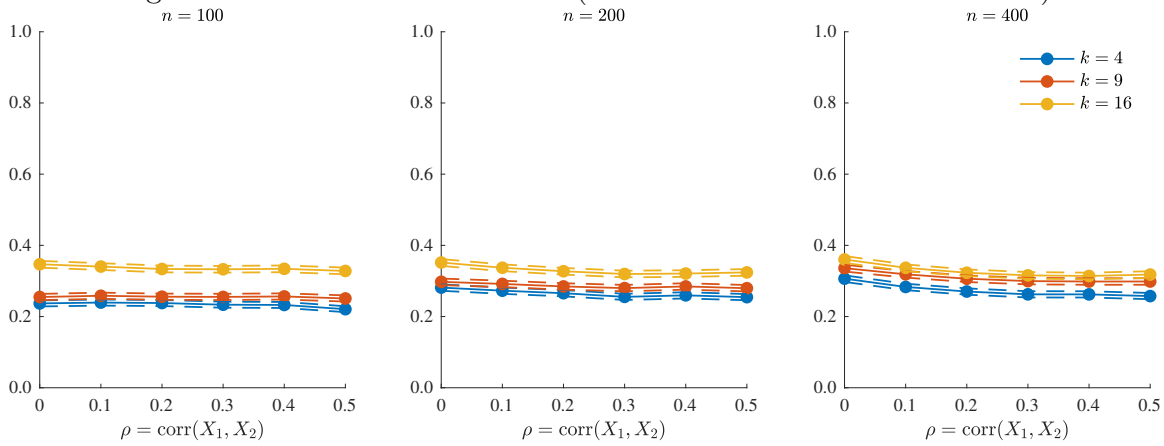


Figure 6: Local Power Estimates (± 2 Monte Carlo Standard Errors)



5 Conclusion

I this paper, I develop an omnibus specification test for a class of semi- or nonparametric conditional moment restrictions in part parameterized by conditional expectation functions. The test is a suitable semi-/nonparametric extension of the Bierens (1982) goodness-of-fit test of a parametric model for the conditional mean. Estimating conditional expectations using series methods, I construct a bootstrap-based test which is proven both asymptotically correctly sized and consistent against any fixed alternative.

I implement my procedure in a comprehensive simulation study testing the specification of a static game of incomplete information. The simulations by and large reproduce the asymptotic properties in small samples. A possible application of my test therefore lies in testing functional form assumptions used in specifying static discrete-choice models with social and strategic interactions (for references, see the introduction). These models may be conveniently estimated in two steps. In the first step, the conditional choice probabilities (CCPs) are estimated in a nonparametric manner. The estimated CCPs are then employed in a second step to estimate the structural parameters of the model. Construction of my test follows along the same lines.

My simulations also indicate that the test has nontrivial power versus root-n-local alternatives, although further effort is needed to investigate the local power properties of the test in a formal manner. Future research might also consider relaxing the assumption of root-n estimability of the parametric component, relaxing the requirement of differentiability to allow for nonsmooth residual functions, and developing formally justified data-driven methods for choosing the series length(s).

Acknowledgements

Parts of the paper derive from my doctoral dissertation, completed under the guidance and encouragement of Denis Chetverikov and Jinyong Hahn. In addition to my former advisors, I thank Andres Santos and Rasmus Søndergaard Pedersen for comments that helped greatly improve this paper.

A Appendix

Appendix Abbreviations and Notation

To conserve space I use the abbreviations CS, H, J, M and T for the Cauchy-Schwarz, Hölder, Jensen, Markov and triangle inequalities, respectively. CMT, LOIE, MVE and MVT are short for the "continuous mapping theorem," "law of iterated expectations" and "mean-value expansion" and "mean-value theorem," respectively. I also abbreviate "with probability approaching one" by wp $\to 1$. I employ empirical process notation and write $\mathbb{E}_n[f] \coloneqq \mathbb{E}_n[f(Z_i)]$ for the average $n^{-1}\sum_{i=1}^n f(Z_i)$ and $\mathbb{G}_n(f)$ for the centered-and-scaled average $\mathbb{G}_n(f) \coloneqq \mathbb{G}_n[f(Z_i)] = \sqrt{n}(\mathbb{E}_n - \mathrm{E})[f]$. If need be, I subscript the expectation operator E to highlight over which variables are integrated out (e.g., E_Z). $\|f\|_{n,2}$ is short for the empirical norm L^2-norm (i.e., $\|f\|_{n,2}^2 = \mathbb{E}_n[f^2]$). $\|A\|_{\mathrm{op}}$ is the operator norm of a matrix A induced by the ℓ_2-norm for vectors. $C, C_1, C_2, ...$ denote positive and finite constants, the meaning of which may change between appearances. For nonrandom sequences, the notation $a_n \lesssim b_n$ means that $|a_n| \leqslant Cb_n$ for C not depending on n. For potentially random sequences, the relation $X_n \lesssim_{\mathrm{P}} b_n$ means $X_n/b_n = O_{\mathrm{P}}(1)$, where $O_{\mathrm{P}}(1)$ denotes stochastic boundedness.

A.1 Proofs for Section 3.1

Lemma A1. *If Assumption 3 holds, then for any $z \in \mathcal{Z}$ and any $h_\ell : \mathcal{W}_\ell \to \mathbf{R}^{d_\ell}$*

$$\begin{aligned}&\left|\rho_\ell(z, h_\ell(w_\ell)) - \rho_\ell(z, h_\ell^*(w_\ell)) - \frac{\partial}{\partial h_\ell}\rho_\ell(z, h_\ell^*(w_\ell))[h_\ell(w_\ell) - h_\ell^*(w_\ell)]\right| \\ &\leqslant R_\ell(z)\, d_\ell^{(1+\gamma_\ell)/2} \max_{1\leqslant m\leqslant d_\ell} \|h_{\ell m} - h_{\ell m}^*\|_{\mathcal{W}_\ell}^{1+\gamma_\ell},\end{aligned}$$

where $\rho_\ell(z, v_\ell) \coloneqq \rho_\ell(z, \beta_0, v_\ell)$.

Proof. Let $z \in \mathcal{Z}, h_\ell : \mathcal{W}_\ell \to \mathbf{R}^{d_\ell}$ be arbitrary. Then $h_\ell(w) \in \mathbf{R}^{d_\ell}$, so by Assumption

3 and a MVE of $v_\ell \mapsto \rho_\ell(z, v_\ell)$ at $h_\ell(w)$ around $h_\ell^*(w_\ell)$ yields

$$
\begin{aligned}
&|\rho_\ell(z, h_\ell(w_\ell)) - \rho_\ell(z, h_\ell^*(w_\ell)) - (\partial/\partial h_\ell)\rho_\ell(z, h_\ell^*(w_\ell))[h_\ell(w_\ell) - h_\ell^*(w_\ell)]| \\
&= \left| \left[\frac{\partial}{\partial h_\ell} \rho_\ell(z, \widetilde{h}_\ell(w_\ell)) - \frac{\partial}{\partial h_\ell} \rho_\ell(z, h_\ell^*(w_\ell)) \right] [h_\ell(w_\ell) - h_\ell^*(w_\ell)] \right| \\
&\leqslant R_\ell(z) \|\widetilde{h}_\ell(w_\ell) - h_\ell^*(w_\ell)\|^\gamma \|h_\ell(w_\ell) - h_\ell^*(w_\ell)\| \leqslant R_\ell(z) \|h_\ell(w_\ell) - h_\ell^*(w_\ell)\|^{1+\gamma} \\
&\leqslant R_\ell(z) d_\ell^{(1+\gamma_\ell)/2} \max_{1 \leqslant m \leqslant d_\ell} \|h_{\ell m} - h_{\ell m}^*\|_{\mathcal{W}_\ell}^{1+\gamma_\ell},
\end{aligned}
$$

where $\widetilde{h}_\ell(w_\ell)$ lies on the line segment connecting $h_\ell(w_\ell)$ and $h_\ell^*(w_\ell)$, thus satisfying $\|\widetilde{h}_\ell(w_\ell) - h_\ell^*(w_\ell)\| \leqslant \|h_\ell(w_\ell) - h_\ell^*(w_\ell)\|$. □

The following result is the crucial step in proving Lemma 1.

Lemma A2. *If Assumptions 1–7 hold, then for each* $\ell \in \{1, \ldots, L\}$,

$$
\begin{aligned}
&\|\sqrt{n}\widehat{M}_\ell - \sqrt{n}\mathbb{E}_n[g_\ell(\cdot, Z_i)]\|_{\mathcal{X}_\ell} \\
&\lesssim_{\mathrm{P}} \max_{1 \leqslant m \leqslant d_\ell} \Big\{ \mathrm{E}[R_\ell(Z)] \sqrt{n} \|\widehat{h}_{\ell m} - h_{\ell m}^*\|_{\mathcal{W}_\ell}^{1+\gamma_\ell} \\
&\quad + \Big(\sum_{j=1}^{k_{\ell m,n}} \|p_{\ell,j}\|_{\mathcal{W}_\ell}^2 \Big)^{1/2} \Big(\sqrt{k_{\ell m,n}/n} + k_{\ell m,n}^{-\alpha_{\ell m}} \Big) \\
&\quad + \sqrt{n} r_{h_{\ell m,n}, k_{\ell m,n}} \sup_{t_\ell \in \mathcal{X}_\ell} r_{\delta_{\ell m,n}, k_{\ell m,n}}(t_\ell) + \sqrt{\zeta_{\ell,k_{\ell m,n}}^2 k_{\ell m,n} \ln(k_{\ell m,n})/n} \\
&\quad + R_{\delta_{\ell m,n}, k_{\ell m,n}} \sqrt{\ln\left(k_{\ell m,n}/R_{\delta_{\ell m,n}, k_{\ell m,n}}\right)} + \zeta_{\ell,k_{\ell m,n}} r_{h_{\ell m,n}, k_{\ell m,n}} \Big\} + o_{\mathrm{P}}(1).
\end{aligned}
$$

The proof of Lemma A2 is long and technical in nature and has therefore been relegated to the online appendix (see Section A2).

Proof of Lemma 1. The claim follows from Lemma A2 and Assumption 7. □

Proof of Lemma 2. A multivariate CLT shows joint convergence of all marginals of the sequences of processes $\{n^{-1/2} \sum_{i=1}^n g_\ell(t_\ell, Z_i); t_\ell \in \mathcal{X}_\ell\}, n \in \mathbf{N}, \ell \in \{1, \ldots, L\}$. To show $\mathcal{G}$ is Donsker, it therefore suffices to show that each

$$
\mathcal{G}_\ell := \{z \mapsto g_\ell(t_\ell, z); t_\ell \in \mathcal{X}_\ell\}, \quad \ell \in \{1, \ldots, L\},
$$

is Donsker [cf. van der Vaart and Wellner (1996, Problem 1.5.3)]. In what follows I therefore omit the subscript ℓ. Moreover, given that β_0 and h^* are held fixed through-

out the argument, I write $\rho(z) \coloneqq \rho(z, \beta_0, h^*(w))$, $(\partial/\partial\beta)\rho(z) \coloneqq (\partial/\partial\beta)\rho(z, \beta_0, h^*(w))$ and $(\partial/\partial h)\rho(z) \coloneqq (\partial/\partial h)\rho(z, \beta_0, h^*(w))$. By Assumption 2 and J we have both $\|b(t)\| \leqslant \mathrm{E}[|\omega(t, X)| \, \|(\partial/\partial\beta)\rho(Z)\|] \lesssim \mathrm{E}[\|(\partial/\partial\beta)\rho(Z)\|]$ and $\|\delta(t, w)\| \leqslant \mathrm{E}[|\omega(t, X)| \, \|(\partial/\partial h)\rho(Z)\| | W = w] \lesssim \mathrm{E}[\|(\partial/\partial h)\rho(Z)\| | W = w]$, where b and δ are defined in (3.8) and (3.9), respectively. Letting $g(t, \cdot) \in \mathcal{G}$ be arbitrary, T and CS therefore imply

$$\begin{aligned}
|g(t, z)| &\leqslant |\omega(t, x)| \, |\rho(z)| + \|b(t)\| \, \|s(z)\| + \|\delta(t, w)\| \|y - h^*(w)\| \\
&\leqslant C_1 |\rho(z)| + \mathrm{E}[\|(\partial/\partial\beta)\rho(Z)\|] \, \|s(z)\| \\
&\quad + \mathrm{E}[\|(\partial/\partial h)\rho(Z)\| | W = w] \|y - h^*(w)\| \\
&=: G_1(z),
\end{aligned}$$

with s stemming from Assumption 1. Taking the expectation and using the inequality $(a + b)^2 \leqslant 2a^2 + 2b^2$ repeatedly alongside the integrability and boundedness parts of Assumptions 1 and 3, we see that $G_1(Z)^2$ is integrable. Hence, G_1 is a square-integrable envelope for $\mathcal{G}$. Let $g_1 = g(t_1, \cdot), g_2 = g(t_2, \cdot) \in \mathcal{F}$ be arbitrary. Then by T and CS, followed by J, CS and Assumption 2,

$$\begin{aligned}
|g(t_1, z) - g(t_2, z)| &\leqslant |\omega(t_1, x) - \omega(t_2, x)| \, |\rho(z)| + \|b(t_1) - b(t_2)\| \, \|s(z)\| \\
&\quad + \|y - h^*(w)\| \|\delta^*(t_1, w) - \delta^*(t_2, w)\| \\
&\leqslant C_2 \Big(|\rho(z)| + \mathrm{E}[\|(\partial/\partial\beta)\rho(Z)\|] \, \|s(z)\| \\
&\quad + \|y - h^*(w)\| \mathrm{E}[\|(\partial/\partial h)\rho(Z)\| | W = w] \Big) \|t_1 - t_2\| \\
&=: G_2(z) \|t_1 - t_2\|
\end{aligned}$$

Defining $G \coloneqq G_1 \vee G_2$, we see that G is a square-integrable envelope for $\mathcal{G}$ satisfying

$$|g(t_1, z) - g(t_2, z)| \leqslant G(z) \|t_1 - t_2\|.$$

Given that $\mathcal{T} = \times_{\ell=1}^{L} \mathcal{X}_\ell$ is compact [Assumption 2 and Tychonoff's theorem (cf. Aliprantis and Border, 2006, Theorem 2.61)], we thus have

$$N_{[\,]}(\varepsilon \|G\|_{P,2}, \mathcal{G}, L^2(P)) \leqslant N(\varepsilon, \mathcal{T}, \|\cdot\|) \leqslant (\mathrm{diam}(\mathcal{T})/\varepsilon)^d \leqslant (C/\varepsilon)^d, \ \varepsilon \in (0, \mathrm{diam}(\mathcal{T})],$$

so using $\|G\|_{P,2} < \infty$,

$$N_{[\,]}(\varepsilon, \mathcal{G}, L^2(P)) \leqslant (C/\varepsilon)^d, \quad \varepsilon > 0.$$

The previous display implies

$$\int_0^\infty \sqrt{\ln(N_{[\,]}(\varepsilon, \mathcal{G}, L^2(P)))}\mathrm{d}\varepsilon \leqslant \sqrt{d}\int_0^\infty \sqrt{\ln(C/\varepsilon)}\mathrm{d}\varepsilon < \infty.$$

The desired conclusion now follows from van der Vaart (2000, Theorem 19.5), which uses the Ossiander (1987) sufficient condition for $\mathcal{G}$ to be Donsker. □

Proof of Theorem 1. Given that $\mathrm{E}[g(\cdot, Z)] = M$, $\mathcal{G}$ being Donsker (Lemma 2) means that $G_n = \sqrt{n}(\mathbb{E}_n[g(\cdot, Z_i)] - M_0) \rightsquigarrow G_M$ in $\times_{\ell=1}^L L^\infty(\mathcal{X}_\ell)$ to an L-variate centered Gaussian process with covariance kernel $\mathbb{C}_M$ given in (3.16). Donsker's theorem shows that $\sqrt{n}(\widehat{F}_X - F_X) \rightsquigarrow G_{F_X}$ in $D([-\infty, \infty]^{d_x})$ where d_x is the number of distinct elements of the X_ℓ's. A multivariate CLT establishes joint convergence of the marginals of the above processes from which we may deduce joint convergence in their product space [cf. van der Vaart and Wellner (1996, Problem 1.5.3)]. A CMT therefore shows weak convergence of $(\sqrt{n}(\mathbb{E}_n[g(\cdot, Z_i)] - M), \{\sqrt{n}(\widehat{F}_{X_\ell} - F_{X_\ell})\}_1^L)$ in $[\times_{\ell=1}^L L^\infty(\mathcal{X}_\ell)] \times [\times_{\ell=1}^L D([-\infty, \infty]^{d_{x,\ell}})]$ to an $2L$-variate centered Gaussian process.

Under the null, $M \equiv 0$ and asymptotic equivalence (Lemma 1) yields

$$(\sqrt{n}\widehat{M}, \{\sqrt{n}(\widehat{F}_{X_\ell} - F_{X_\ell})\}_1^L) \rightsquigarrow (G_0, \{G_{F_{X_\ell}}\}_1^L) \text{ in } [\times_{\ell=1}^L L^\infty(\mathcal{X}_\ell)] \times [\times_{\ell=1}^L D([-\infty, \infty]^{d_{x,\ell}})]$$

with G_0 centered Gaussian and covariance kernel $\mathbb{C}_0$ given by (3.16) with $M \equiv 0$. Let $BV_K(\mathcal{A})$ be the set of real-valued functions on $\mathcal{A}$ which are nondecreasing in each variable (holding the other arguments fixed) and of variation no more than $K \in \mathbf{R}_{++}$. Then the functional $\phi : [\times_{\ell=1}^L C(\mathcal{X}_\ell)] \times [\times_{\ell=1}^L BV_1(\mathcal{X}_\ell)] \subseteq [\times_{\ell=1}^L L^\infty(\mathcal{X}_\ell)] \times [\times_{\ell=1}^L D([-\infty, \infty]^{d_{x,\ell}})] \to \mathbf{R}$ defined by $\phi(\{m_\ell\}_1^L, \{f_\ell\}_1^L) \coloneqq \sum_{\ell=1}^L \int_{\mathcal{X}_\ell} m_\ell^2 \mathrm{d}f_\ell$ is second-order Hadamard differentiable at $(0, \{F_{X_\ell}\}_1^L) \in [\times_{\ell=1}^L C(\mathcal{X}_\ell)] \times [\times_{\ell=1}^L BV_1(\mathcal{X}_\ell)]$ with vanishing first-order Hadamard derivative and second-order Hadamard derivative $\phi''_{(0,\{F_{X_\ell}\}_1^L)} : [\times_{\ell=1}^L C(\mathcal{X}_\ell)] \times [\times_{\ell=1}^L BV_1(\mathcal{X}_\ell)] \to \mathbf{R}$ given by

$$\phi''_{(0,\{F_{X_\ell}\}_1^L)}(h_1, h_2) \coloneqq 2\sum_{\ell=1}^L \int_{\mathcal{X}_\ell} h_{1\ell}^2 \mathrm{d}F_{X_\ell}.$$

The functional (second-order) delta method therefore produces

$$\begin{aligned} T_n &= n\sum_{\ell=1}^{L}\int_{\mathcal{X}_\ell}\widehat{M}_\ell^2 \mathrm{d}\widehat{F}_{X_\ell} = n[\phi(\widehat{M},\{\widehat{F}_{X_\ell}\}_1^L) - \phi(0,\{F_{X_\ell}\}_1^L)] \\ &= \frac{1}{2}\phi''_{(0,\{F_{X_\ell}\}_1^L)}(\sqrt{n}\widehat{M},\{\sqrt{n}(\widehat{F}_{X_\ell}-F_{X_\ell})\}_1^L)) + o_{\mathrm{P}}(1) \\ &= \sum_{\ell=1}^{L}\int_{\mathcal{X}_\ell}(\sqrt{n}\widehat{M}_\ell)^2 \mathrm{d}F_{X_\ell} + o_{\mathrm{P}}(1), \end{aligned}$$

i.e., use of the empirical distribution has no impact on the asymptotic distribution. Again appealing to Lemmas 1 and 2, the previous display and CMT combine to yield $T_n \to_d \sum_{\ell=1}^{L}\int_{\mathcal{X}_\ell} G_{0\ell}^2 \mathrm{d}F_{X_\ell}$.

To establish the second claim, note that asymptotic equivalence and the reverse triangle inequality imply

$$\left|\sqrt{\sum_{\ell=1}^{L}\|\widehat{M}_{\ell n}\|^2_{\widehat{F}_{X_\ell},2}} - \sqrt{\sum_{\ell=1}^{L}\|\mathbb{E}_n[g(\cdot,Z_i)]\|^2_{\widehat{F}_{X_\ell},2}}\right| \xrightarrow{\mathrm{P}} 0.$$

Given that $\sqrt{n}(\mathbb{E}_n[g(\cdot,Z_i)] - M_0) \rightsquigarrow G_{M_0}$ in $\times_{\ell=1}^{L} L^\infty(\mathcal{X}_\ell)$, the CMT yields

$$\begin{aligned} \max_{1\leqslant\ell\leqslant L}\|\mathbb{E}_n[g(\cdot,Z_i)] - M_{0\ell}\|_{\mathcal{X}_\ell} &= \frac{1}{\sqrt{n}}\max_{1\leqslant\ell\leqslant L}\left\|\sqrt{n}(\mathbb{E}_n[g(\cdot,Z_i)] - M_{0\ell})\right\|_{\mathcal{X}_\ell} \\ &= o(1)\,O_{\mathrm{P}}(1) = o_{\mathrm{P}}(1). \end{aligned}$$

Another application of the reverse triangle inequality therefore shows

$$\begin{aligned} &\left|\sqrt{\sum_{\ell=1}^{L}\|\mathbb{E}_n[g(\cdot,Z_i)]\|^2_{\widehat{F}_{X_\ell},2}} - \sqrt{\sum_{\ell=1}^{L}\|M_{0\ell}\|^2_{\widehat{F}_{X_\ell},2}}\right| \\ &\leqslant \sqrt{L}\max_{1\leqslant\ell\leqslant L}\|\mathbb{E}_n[g(\cdot,Z_i)] - M_{0\ell}\|_{\mathcal{X}_\ell} \xrightarrow{\mathrm{P}} 0, \end{aligned}$$

so by the the triangle inequality and the CMT, we see that

$$\frac{T_n}{n} = \sum_{\ell=1}^{L}\|M_{0\ell}\|^2_{\widehat{F}_{X_\ell},2} + o_{\mathrm{P}}(1) = \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{L} M_{0\ell}(X_{\ell i})^2 + o_{\mathrm{P}}(1).$$

The LLN now yields $T_n/n \to_{\mathrm{P}} \sum_{\ell=1}^{L} \|M_{0\ell}\|^2_{F_{X_\ell},2}$, which is positive under the alternative (2.2) by the choice of weights (Assumption 2). □

A.2 Proofs for Section 3.2

Proof of Lemma 3 (Sketch). The proof parallels that of Lemma A2 (see Section B.1) with some added complexity due to the presence of multipliers and the error introduced from estimating the g_ℓ's and recentering at the sample values. Due to space constraints and to avoid repetition, I relegate the argument to the online supplement not intended for publication (see Section C.1). □

Proof of Lemma 4. Given that $\mathcal{G}$ is Donsker (Lemma 2), Kosorok (2008, Theorem 10.4(iv)) implies that $\mathbb{G}_n'' \rightsquigarrow_{\mathrm{P},\xi} \mathbb{G}^*$ in $\times_{\ell=1}^{L} L^*(\mathcal{G}_\ell)$, where $\mathbb{G}_n''(g) := n^{-1/2}\sum_{i=1}^{n} \xi_i\{g(Z_i) - \mathrm{E}[g(Z)]\}$ and $\mathbb{G}^*$ is an L-variate zero-mean Gaussian process with covariance kernel $\mathrm{E}[g(Z)\,g'(Z)^\top] - \mathrm{E}[g(Z)]\mathrm{E}[g'(Z)^\top], g, g' \in \times_{\ell=1}^{L}\mathcal{G}_\ell$. Since we may identify each $\mathcal{G}_\ell$ with $\mathcal{X}_\ell$, this result is equivalent to $G_n^* \rightsquigarrow_{\mathrm{P},\xi} G_M$ in $\times_{\ell=1}^{L} L^\infty(\mathcal{X}_\ell)$, where G_M is a zero-mean Gaussian process with covariance kernel $\mathbb{C}_M$ given in (3.16). Lemma B6 now implies $\widehat{G}_n \rightsquigarrow_{\mathrm{P},\xi} G_M$ in $\times_{\ell=1}^{L} L^\infty(\mathcal{X}_\ell)$. An application of the (second-order) delta method for the bootstrap now establishes that $\widehat{T}$ converges weakly in probability to T_M. Invoking continuity (Assumption 9) Kosorok (2008, Lemma 10.11) therefore shows that the $F_{\widehat{T}}$ converges in probability to F_{T_M} pointwise on $[0,\infty)$. Fix $\varepsilon > 0$ and $\alpha \in (0,1)$. Since F_{T_M} is continuous, by the hypothesis that it is also increasing at $c_{T_M}(\alpha)$, there exists $r_1 \in \mathbf{R}$ such that $c_{T_M}(\alpha) - \varepsilon < r_1 < c_{T_M}(\alpha)$ and $F_{T_M}(r_1) < 1-\alpha$. Then $F_{\widehat{T}}(r_1) < 1-\alpha$ wp $\to 1$, which implies $c_{T_M}(\alpha) - \varepsilon < r_1 \leqslant c_{\widehat{T}}(\alpha)$ wp $\to 1$. In particular, $\mathrm{P}(c_{\widehat{T}}(\alpha) \geqslant c_{T_M}(\alpha) - \varepsilon) \to 1$. Similarly, there exists $r_2 \in \mathbf{R}$ be such that $c_{T_M}(\alpha) < r_2 < c_{T_M}(\alpha) + \varepsilon$ and $1-\alpha < F_{T_M}(r_2)$. Then $1-\alpha < F_{\widehat{T}}(r_2)$ wp $\to 1$, which implies $c_{\widehat{T}}(\alpha) \leqslant r_2 < c_{T_M}(\alpha) + \varepsilon$ wp $\to 1$. In particular, $\mathrm{P}(c_{\widehat{T}}(\alpha) < c_{T_M}(\alpha) + \varepsilon) \to 1$. It follows that

$$\begin{aligned}
&\overline{\lim_{n\to\infty}}\, \mathrm{P}(|c_{\widehat{T}}(\alpha) - c_{T_M}(\alpha)| \geqslant \varepsilon) \\
&\leqslant \overline{\lim_{n\to\infty}}\, \mathrm{P}(c_{\widehat{T}}(\alpha) \geqslant c_{T_M}(\alpha) + \varepsilon) + \overline{\lim_{n\to\infty}}\, \mathrm{P}(c_{\widehat{T}}(\alpha) \leqslant c_{T_M}(\alpha) - \varepsilon) = 0.
\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, the lemma follows. □

A.3 Proofs for Section 3.3

PROOF OF THEOREM 2. Fix $\alpha \in (0,1)$. *Under the null*, $T_n \to_d T_0$ (Theorem 1). Letting F_{T_0} denote the CDF of T_0, by F_{T_0} is continuous on $\mathbf{R}$ (using Assumption 9) and increasing at $c_{T_0}(\alpha)$ (by hypothesis). It therefore follows from Lemma 4 that $c_{\widehat{T}}(\alpha) \to_{\mathrm{P}} c_{T_0}(\alpha) \in (0,\infty)$. Slutsky's theorem shows $T_n - c_{\widehat{T}}(\alpha) \to_d T_0 - c_{T_0}(\alpha)$, which establishes the first claim. *Under the alternative*, $T_n/n \to_{\mathrm{P}} \sum_{\ell=1}^{L} \int_{\mathcal{X}_\ell} M_\ell^2 \mathrm{d}F_{X_\ell} \in (0,\infty)$. Since F_{T_M} is increasing at its $c_{T_M}(\alpha)$, Lemma 4 yields $c_{\widehat{T}}(\alpha) \to_{\mathrm{P}} c_{T_M}(\alpha) \in (0,\infty)$. In particular, $c_{\widehat{T}}(\alpha) = O_{\mathrm{P}}(1)$, so for any $\varepsilon \in (0,1)$, there exists $K_\varepsilon \in (0,\infty)$ such that $\overline{\lim}_{n\to\infty} \mathrm{P}\left(c_{\widehat{T}}(\alpha) > K_\varepsilon\right) \leqslant \varepsilon$. Letting $\varepsilon \in (0,1)$ be arbitrary, we see that

$$\begin{aligned}
\mathrm{P}\left(T_n \leqslant c_{\widehat{T}}(\alpha)\right) &= \mathrm{P}\left(T_n \leqslant c_{\widehat{T}}(\alpha) \cap c_{\widehat{T}}(\alpha) \leqslant K_\varepsilon\right) + \mathrm{P}\left(T_n \leqslant c_{\widehat{T}}(\alpha) \cap c_{\widehat{T}}(\alpha) > K_\varepsilon\right) \\
&\leqslant \mathrm{P}\left(T_n \leqslant K_\varepsilon\right) + \mathrm{P}\left(c_{\widehat{T}}(\alpha) > K_\varepsilon\right) \\
&= \mathrm{P}\left(T_n/n \leqslant K_\varepsilon/n\right) + \mathrm{P}\left(c_{\widehat{T}}(\alpha) > K_\varepsilon\right),
\end{aligned}$$

which—by the preceding remarks—implies $\overline{\lim}_{n\to\infty} \mathrm{P}\left(T_n \leqslant c_{\widehat{T}}(\alpha)\right) \leqslant \varepsilon$. The second claim now follows from taking $\varepsilon \to 0_+$. □

References

AHN, H. AND C. F. MANSKI (1993): "Distribution theory for the analysis of binary choice under uncertainty with nonparametric estimation of expectations," *Journal of Econometrics*, 56, 291–321.

AI, C. AND X. CHEN (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71, 1795–1843.

AÏT-SAHALIA, Y., P. J. BICKEL, AND T. M. STOKER (2001): "Goodness-of-fit tests for kernel regression with an application to option implied volatilities," *Journal of Econometrics*, 105, 363–412.

ALIPRANTIS, C. D. AND K. BORDER (2006): *Infinite dimensional analysis: A Hitchhiker's Guide*, Springer Science & Business Media.

ANDREWS, D. W. (1997): "A conditional Kolmogorov test," *Econometrica: Journal of the Econometric Society*, 1097–1128.

Bajari, P., H. Hong, J. Krainer, and D. Nekipelov (2010): "Estimating static models of strategic interactions," *Journal of Business & Economic Statistics*, 28, 469–482.

Belloni, A. and V. Chernozhukov (2011): *High dimensional sparse econometric models: An introduction*, Springer.

——— (2013): "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19, 521–547.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015): "Some new asymptotic theory for least squares series: Pointwise and uniform results," *Journal of Econometrics*, 186, 345–366.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous analysis of Lasso and Dantzig selector," *The Annals of Statistics*, 1705–1732.

Bierens, H. J. (1982): "Consistent model specification tests," *Journal of Econometrics*, 20, 105–134.

——— (1990): "A consistent conditional moment test of functional form," *Econometrica: Journal of the Econometric Society*, 1443–1458.

——— (2017): *Econometric Model Specification: Consistent Model Specification Tests and Semi-nonparametric Modeling and Inference*, World Scientific Publishing Company Pte. Limited.

Bierens, H. J. and D. K. Ginther (2001): "Integrated conditional moment testing of quantile regression models," *Empirical Economics*, 26, 307–324.

Bierens, H. J. and W. Ploberger (1997): "Asymptotic theory of integrated conditional moment tests," *Econometrica: Journal of the Econometric Society*, 1129–1151.

Bierens, H. J. and L. Wang (2012): "Integrated conditional moment tests for parametric conditional distributions," *Econometric Theory*, 28, 328–362.

Bjorn, P. A. and Q. H. Vuong (1984): "Simultaneous equations models for dummy endogenous variables: a game theoretic formulation with an application to labor force participation," .

Bravo, F. (2012): "Generalized empirical likelihood testing in semiparametric conditional moment restrictions models," *The Econometrics Journal*, 15, 1–31.

Breunig, C. (2015): "Goodness-of-fit tests based on series estimators in nonparametric instrumental regression," *Journal of Econometrics*, 184, 328–346, publisher: Elsevier.

Card, D. and L. Giuliano (2013): "Peer effects and multiple equilibria in the risky behavior of friends," *Review of Economics and Statistics*, 95, 1130–1149.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics*, 6, 5549–5632.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 1591–1608.

Chen, X. and D. Pouzo (2009): "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals," *Journal of Econometrics*, 152, 46–60.

——— (2012): "Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals," *Econometrica*, 80, 277–321.

——— (2015): "Sieve Wald and QLR inferences on semi/nonparametric conditional moment models," *Econometrica*, 83, 1013–1079.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2017a): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013): "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors," *The Annals of Statistics*, 41, 2786–2819.

——— (2017b): "Central limit theorems and bootstrap in high dimensions," *The Annals of Probability*, 45, 2309–2352.

Davidson, R. and V. Zinde-Walsh (2017): "Advances in specification testing," *Canadian Journal of Economics/Revue canadienne d'économique*, 50, 1595–1631.

Davydov, Y. A., M. A. Lifshits, and N. V. Smorodina (1998): *Local properties of distributions of stochastic functionals*, American Mathematical Soc.

de Jong, R. M. and H. J. Bierens (1994): "On the limit behavior of a chi-square type test if the number of conditional moments tested approaches infinity," *Econometric Theory*, 10, 70–90.

Delgado, M. A., M. A. Domínguez, and P. Lavergne (2006): "Consistent tests of conditional moment restrictions," *Annales d'Economie et de Statistique*, 33–67.

Delgado, M. A. and W. G. Manteiga (2001): "Significance testing in nonparametric regression based on the bootstrap," *The Annals of Statistics*, 29, 1469–1507.

Delgado, M. A. and W. Stute (2008): "Distribution-free specification tests of conditional models," *Journal of Econometrics*, 143, 37–55.

DeVore, R. A. and G. G. Lorentz (1993): *Constructive approximation*, vol. 303, Springer Science & Business Media.

Donald, S. G., G. W. Imbens, and W. K. Newey (2003): "Empirical likelihood estimation and consistent tests with conditional moment restrictions," *Journal of Econometrics*, 117, 55–93.

Donald, S. G. and W. K. Newey (1994): "Series estimation of semilinear models," *Journal of Multivariate Analysis*, 50, 30–40.

Escanciano, J. and S. Goh (2014): "Specification analysis of linear quantile models," *Journal of econometrics*, 178, 495–507.

Escanciano, J. C. (2006): "A consistent diagnostic test for regression models using projections," *Econometric Theory*, 22, 1030–1051.

Escanciano, J. C. and C. Velasco (2010): "Specification tests of parametric dynamic conditional quantiles," *Journal of Econometrics*, 159, 209–221.

Fan, Y. and Q. Li (1996): "Consistent model specification tests: omitted variables and semiparametric functional forms," *Econometrica: Journal of the econometric society*, 865–890.

——— (2000): "Consistent model specification tests: Kernel-based tests versus Bierens' ICM tests," *Econometric Theory*, 16, 1016–1041.

Guerre, E. and P. Lavergne (2005): "Data-driven rate-optimal specification testing in regression models," *The Annals of Statistics*, 33, 840–870.

Hahn, J., H. R. Moon, and C. Snider (2017): "LM test of neglected correlated random effects and its application," *Journal of Business & Economic Statistics*, 35, 359–370.

Härdle, W. and E. Mammen (1993): "Comparing nonparametric versus parametric regression fits," *The Annals of Statistics*, 1926–1947.

He, X. and L.-X. Zhu (2003): "A lack-of-fit test for quantile regression," *Journal of the American Statistical Association*, 98, 1013–1022.

Hong, Y. and H. White (1995): "Consistent specification testing via nonparametric series regression," *Econometrica: Journal of the Econometric Society*, 1133–1159.

Horowitz, J. L. and V. G. Spokoiny (2001): "An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative," *Econometrica*, 69, 599–631.

——— (2002): "An adaptive, rate-optimal test of linearity for median regression models," *Journal of the American Statistical Association*, 97, 822–835.

Hsiao, C., Q. Li, and J. S. Racine (2007): "A consistent model specification test with mixed discrete and continuous data," *Journal of Econometrics*, 140, 802–826.

Korolev, I. (2018): "A Consistent LM Type Specification Test for Semiparametric Models," *Working Paper, Department of Economics, Binghamton University*.

Kosorok, M. R. (2008): *Introduction to empirical processes and semiparametric inference*, Springer Science & Business Media.

Lavergne, P., S. Maistre, and V. Patilea (2015): "A significance test for covariates in nonparametric regression," *Electronic journal of statistics*, 9, 643–678.

Lavergne, P. and Q. Vuong (2000): "Nonparametric significance testing," *Econometric Theory*, 16, 576–601.

Li, Q., C. Hsiao, and J. Zinn (2003): "Consistent specification tests for semiparametric/nonparametric models based on series estimation methods," *Journal of Econometrics*, 112, 295–325.

Lorentz, G. G. (1966): *Approximation of functions*, Holt, Rinehart and Winston.

Manski, C. F. (1991): "Nonparametric estimation of expectations in the analysis of discrete choice under uncertainty," in *Nonparametric and semiparametric methods in econometrics and statistics*, 259–275.

Newey, W. K. (1994): "The asymptotic variance of semiparametric estimators," *Econometrica: Journal of the Econometric Society*, 1349–1382.

——— (1995): "Convergence rates for series estimators," in *Advances in Econometrics and Quantitative Economics*, ed. by G. Maddala, P. C. Phillips, and T. Srinivasan, Blackwell.

——— (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147–168.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Ossiander, M. (1987): "A central limit theorem under metric entropy with L2 bracketing," *The Annals of Probability*, 897–919.

Pollard, D. (1990): "Empirical processes: theory and applications," in *NSF-CBMS regional conference series in probability and statistics*, JSTOR, i–86.

Racine, J. S., J. Hart, and Q. Li (2006): "Testing the significance of categorical predictor variables in nonparametric regression models," *Econometric Reviews*, 25, 523–544.

Rao, C. R. (1973): *Linear statistical inference and its applications*, vol. 2, Wiley New York.

Rothe, C. and D. Wied (2013): "Misspecification testing in a class of conditional distributional models," *Journal of the American Statistical Association*, 108, 314–324.

Rudelson, M. (1999): "Random vectors in the isotropic position," *Journal of Functional Analysis*, 164, 60–72.

Rudin, W. (1976): "Principles of Mathematical Analysis (International Series in Pure & Applied Mathematics)," .

Santos, A. (2012): "Inference in nonparametric instrumental variables with partial identification," *Econometrica*, 80, 213–275.

Schumaker, L. (2007): *Spline functions: basic theory*, Cambridge University Press.

Seim, K. (2006): "An empirical model of firm entry with endogenous product-type choices," *The RAND Journal of Economics*, 37, 619–640.

Song, K. (2010): "Testing semiparametric conditional moment restrictions using conditional martingale transforms," *Journal of Econometrics*, 154, 74–84.

Sørensen, J. R.-V. (2018): "Essays on Nonparametric and High-Dimensional Econometrics," PhD Thesis, UCLA.

Stengos, T. and Y. Sun (2001): "A consistent model specification test for a regression function based on nonparametric wavelet estimation," *Econometric Reviews*, 20, 41–60.

Stinchcombe, M. B. and H. White (1998): "Consistent specification testing with nuisance parameters present only under the alternative," *Econometric theory*, 14, 295–325.

Stone, C. J. (1985): "Additive regression and other nonparametric models," *The annals of Statistics*, 689–705.

Stute, W. (1997): "Nonparametric model checks for regression," *The Annals of Statistics*, 613–641.

Stute, W. and L.-X. Zhu (2005): "Nonparametric checks for single-index models," *The Annals of Statistics*, 33, 1048–1083.

Sun, Y. and Q. Li (2006): "An alternative series based consistent model specification test," *Economics Letters*, 93, 37–44.

Sweeting, A. (2009): "The strategic timing incentives of commercial radio stations: An empirical analysis using multiple equilibria," *The RAND Journal of Economics*, 40, 710–742.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Timan, A. F. (1963): *Theory of approximation of functions of a real variable*, vol. 34, Courier Corporation.

Tripathi, G. and Y. Kitamura (2003): "Testing conditional moment restrictions," *The Annals of Statistics*, 31, 2059–2095.

van der Vaart, A. and J. A. Wellner (2011): "A local maximal inequality under uniform entropy," *Electronic Journal of Statistics*, 5, 192.

van der Vaart, A. W. (2000): *Asymptotic statistics*, vol. 3, Cambridge university press.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

Whang, Y.-J. (2000): "Consistent bootstrap tests of parametric regression functions," *Journal of Econometrics*, 98, 27–46.

——— (2001): "Consistent specification testing for conditional moment restrictions," *Economics Letters*, 71, 299–306.

——— (2006): "Consistent specification testing for quantile regression models," *Econometric Theory and Practice: Frontiers of Analysis and Applied Research*, 288–308.

Zheng, J. X. (1996): "A consistent test of functional form via nonparametric estimation techniques," *Journal of Econometrics*, 75, 263–289.

——— (1998): "A consistent nonparametric test of parametric regression models under conditional quantile restrictions," *Econometric Theory*, 14, 123–138.

B Online Appendix (Intended for Publication)

This appendix contains a proof of Lemma A2 (which is key to establishing the asymptotic equivalence in Lemma 1) and some supporting lemmas.

B.1 Proof of Lemma A2

Proof of Lemma A2. The proof proceeds in a number of steps. Since the lemma is stated for a given ℓ, for notational convenience I drop the ℓ subscripts throughout and refer to the (ℓth) index set $\mathcal{X}_\ell$ as $\mathcal{T}$ itself.

Step 0 (Main)

Let $t \in \mathcal{T}$ be arbitrary. Assumption 1 and M implies that $\|\widehat{\beta} - \beta_0\| \lesssim_{\mathrm{P}} n^{-1/2} \to 0$. Letting $\mathcal{N}_0$ be any open neighborhood of β_0, $\widehat{\beta} \in \mathcal{N}_0$ wp $\to 1$. To simplify notation and ensure that objects are globally well defined, in what follows I will—without loss of generality—assume that $\widehat{\beta} \in \mathcal{N}_0$ with *probability one for all* n. Then by Assumption 3, for any z, v, we may conduct a mean value expansion of $\beta \mapsto \rho(z, \beta, v)$ at $\widehat{\beta}$ around β_0 to get

$$\begin{aligned}
\widehat{M}(t) &= \sqrt{n}\mathbb{E}_n[\omega(t, X_i)\,\rho(Z_i, \beta_0, \widehat{h}(W_i))] + \mathrm{I}_n(t)^\top \sqrt{n}(\widehat{\beta} - \beta_0), \\
\mathrm{I}_n(t) &\coloneqq \mathbb{E}_n\left[\omega(t, X_i)\,(\partial/\partial\beta)\,\rho(Z_i, \overline{\beta}_n, \widehat{h}(W_i))\right],
\end{aligned}$$

where $\overline{\beta}$ lies on the line segment connecting $\widehat{\beta}$ and β_0, thus satisfying $\|\overline{\beta} - \beta_0\| \leqslant \|\widehat{\beta} - \beta_0\| \to_{\mathrm{P}} 0$. Recall the definition of $b(t)$ in (3.8), which is well defined on $\mathcal{T}$ since β_0 is interior to $\mathcal{B}$ (Assumption 3). Step B.1 below shows that $\sup_{t\in\mathcal{T}}\|\mathrm{I}_n(t) - b(t)\| \to_{\mathrm{P}} 0$, and that b is bounded on $\mathcal{T}$, so Assumption 1 and the previous display combine to yield

$$\sqrt{n}\widehat{M}(t) = \sqrt{n}\mathbb{E}_n[\omega(t, X_i)\,\rho(Z_i, \beta_0, \widehat{h}(W_i))] + b(t)^\top \sqrt{n}\mathbb{E}_n[s(Z_i)] + o_{\mathrm{P}}(1), \quad \text{(B.1)}$$

uniformly on $\mathcal{T}$.

The remainder of the proof is about adjusting for estimation of h^*. Given that β_0 is held fixed throughout this argument, I will suppress the β argument and write $\rho(z, v) \coloneqq \rho(z, \beta_0, v)$. For the purpose of the adjustment, denote the first term on the

right-hand side of (B.1)

$$\sqrt{n}\widehat{M}^{*}(t) \coloneqq \sqrt{n}\mathbb{E}_{n}[\omega(t, X_i)\,\rho(Z_i, \widehat{h}(W_i))], \tag{B.2}$$

and conduct a MVE of $v \mapsto \rho(Z_i, v)$ at $\widehat{h}(W_i)$ around $h^{*}(W_i)$ to arrive at

$$\sqrt{n}\widehat{M}^{*}(t) = \sqrt{n}\mathbb{E}_{n}\left[\omega(t, X_i)\left\{\rho(Z_i, h^{*}(W_i)) + \frac{\partial}{\partial h^{\top}}\rho(Z_i, \overline{h}_{n}(W_i))[\widehat{h}(W_i) - h^{*}(W_i)]\right\}\right],$$

where $\overline{h}(W_i)$ lies on the line segment connecting $\widehat{h}(W_i)$ and $h^{*}(W_i)$. Such an expansion is justified by Assumption 3. Further decomposition of the right-hand side yields

$$\begin{aligned}
&\sqrt{n}\widehat{M}^{*}(t) \\
&= \sqrt{n}\mathbb{E}_{n}\left[\omega(t, X_i)\,\rho(Z_i, h^{*}(W_i)) + \delta(t, W_i)^{\top}\{Y_i - h^{*}(W_i)\}\right] \\
&\quad + \sqrt{n}\mathbb{E}_{n}\left[\omega(t, X_i)\left\{\frac{\partial}{\partial h^{\top}}\rho(Z_i, \overline{h}(W_{\ell i})) - \frac{\partial}{\partial h^{\top}}\rho(Z_i, h^{*}(W_i))\right\}\{\widehat{h}(W_i) - h^{*}(W_i)\}\right] \\
&\quad + \mathbb{G}_{n}\left[\omega(t, X_i)\frac{\partial}{\partial h^{\top}}\rho(Z_i, h^{*}(W_i))\right][\widehat{h}(W_i) - h^{*}(W_i)] \\
&\quad + \sqrt{n}\Big(\mathrm{E}_{Z}\left[\omega(t, X)\frac{\partial}{\partial h^{\top}}\rho(Z, h^{*}(W))\,[\widehat{h}(W) - h^{*}(W)]\right] \\
&\qquad - \mathbb{E}_{n}[\delta(t, W_i)^{\top}\{Y_i - h^{*}(W_i)\}]\Big) \\
&=: \sqrt{n}\mathbb{E}_{n}\left[\omega(t, X_i)\,\rho(Z_i, h^{*}(W_i)) + \delta(t, W_i)^{\top}\{Y_i - h^{*}(W_i)\}\right] \\
&\quad + \mathrm{II}_{n}(t) + \mathrm{III}_{n}(t) + \mathrm{IV}_{n}(t),
\end{aligned} \tag{B.3}$$

where $\mathrm{E}_{Z}[\cdot]$ denotes integration with respect to the distribution of Z, and $\delta(t, Z)$ is defined as in (3.9). The $k \times k$ design matrix $Q_k = \mathrm{E}[p^{k}(W)\,p^{k}(W)^{\top}]$ is invertible by Assumption 5. Let h_k and $\delta_k(t, \cdot)$ denote the mean-square projections of h^{*} and $\delta(t, \cdot)$, respectively, onto the span of p^{k}, i.e.,

$$h_{k,m}(\cdot) \coloneqq p^{k}(\cdot)^{\top} Q_{k}^{-1}\mathrm{E}[p^{k}(W)\,h_{m}^{*}(W)] = p^{k}(\cdot)^{\top}\pi_{h_m,k}, \tag{B.4}$$

$$\delta_{k,m}(t, \cdot) \coloneqq p^{k}(\cdot)^{\top} Q_{k}^{-1}\mathrm{E}[p^{k}(W)\,\delta_{m}(t, W)] = p^{k}(\cdot)^{\top}\pi_{\delta_m,k}(t), \tag{B.5}$$

where $\pi_{h_{m,k}}$ and $\pi_{\delta_{m,k}}$ are defined in (3.10) and (3.11), respectively. Consequently,

$$\begin{aligned}
\mathrm{E}[\{h_{m,k}(W) - h_m^*(W)\}^2] &= r_{h_{m,k}}^2, \\
\mathrm{E}[\{\delta_{m,k}(t,W) - \delta_m(t,W)\}^2] &= r_{\delta_{m,k}}^2(t), \\
\mathrm{E}\{\|\delta_{m,k}(\cdot,W) - \delta_m(\cdot,W)\|_{\mathcal{T}}^2\} &= R_{\delta_{m,k}}^2,
\end{aligned}$$

for $r_{h_{m,k}}^2$, $r_{\delta_{m,k}}^2$ and $R_{\delta_{m,k}}^2$ defined in (3.12), (3.13) and (3.14), respectively. Steps 2–4 below show that the three remainder terms in the decomposition (B.3) satisfy:

$$\begin{aligned}
\|\mathrm{II}_n\|_{\mathcal{T}} &\lesssim_{\mathrm{P}} \mathrm{E}[R(Z)]\sqrt{n}\max_{1\leqslant m\leqslant d}\|\widehat{h}_m - h_m^*\|_{\mathcal{W}}^{1+\gamma}, \\
\|\mathrm{III}_n\|_{\mathcal{T}} &\lesssim_{\mathrm{P}} \max_{1\leqslant m\leqslant d}\Big(\sum_{j=1}^{k_{m,n}}\|p_j\|_{\mathcal{W}}^2\Big)^{1/2}\Big(\sqrt{k_{m,n}/n} + k_{m,n}^{-\alpha_m}\Big), \\
\text{and } \|\mathrm{IV}_n\|_{\mathcal{T}} &\lesssim_{\mathrm{P}} \max_{1\leqslant m\leqslant d}\Big\{\sqrt{n}r_{h_{m,k_{m,n}}}\sup_{t\in\mathcal{T}} r_{\delta_{m,k_{m,n}}}(t) + \sqrt{\zeta_{k_{m,n}}^2 k_{m,n}\ln(k_{m,n})/n} \\
&\qquad + R_{\delta_{m,k_{m,n}}}\sqrt{\ln\left(k_{m,n}/R_{\delta_{m,k_{m,n}}}\right)} + \zeta_{k_{m,n}}r_{h_{m,k_{m,n}}}\Big\}
\end{aligned}$$

Plug (B.3) into (B.1), apply T and use the definition of $\widehat{M}^*$ in (B.2) to get the claimed in-probability bound.

Step 1: I_n and b_*

In this step I show that I_n defined in (B.1) and b defined (3.8) satisfy

$$\text{(a) } \sup_{t\in\mathcal{T}}\|\mathrm{I}_n(t) - b(t)\| \xrightarrow{\mathrm{P}} 0 \quad \text{and} \quad \text{(b) } \sup_{t\in\mathcal{T}}\|b(t)\| < \infty.$$

Decompose I_n as

$$\begin{aligned}
\mathrm{I}_n(t) &= \mathbb{E}_n\left[\omega(t,X_i)(\partial/\partial\beta)\rho(Z_i,\overline{\beta}_n,h_*(W))\right] \\
&\quad + \mathbb{E}_n\left[\omega(t,X_i)\left\{(\partial/\partial\beta)\rho(Z_i,\overline{\beta}_n,\widehat{h}_n(W_i)) - (\partial/\partial\beta)\rho(Z_i,\overline{\beta}_n,h_*(W_i))\right\}\right] \\
&=: \mathrm{I}_{a,n}(t) + \mathrm{I}_{b,n}(t).
\end{aligned}$$

Since $\|\overline{\beta} - \beta_0\| \leqslant \|\widehat{\beta} - \beta_0\|$ and $\widehat{\beta} \in \mathcal{N}_0$, we must have $\overline{\beta} \in \mathcal{N}_0$ wp $\to 1$, so using T, Assumptions 2 and 3 and Lemma B5.4, we get

$$\begin{aligned}\sup_{t\in\mathcal{T}} \|\mathrm{I}_{a,n}(t)\| &\leqslant \mathbb{E}_n\left[a(Z_i)\,\|\widehat{h}(W_i) - h^*(W_i)\|^c\right] \\ &\leqslant \sqrt{d}\mathbb{E}_n[a(Z_i)] \max_{1\leqslant m\leqslant d} \|\widehat{h}_m - h^*_m\|^c_{\mathcal{W}} \xrightarrow{\mathrm{P}} 0\end{aligned}$$

Given that $\beta_0 \in \mathcal{N}_0$ open, there is an $r > 0$ such that the open ball $B_r(\beta_0)$ in $\mathbf{R}^{d_\beta}$ centered at β_0 with radius r is contained in $\mathcal{N}_0$. Let $\overline{B} \coloneqq \overline{B}_{r/2}(\beta_0)$ denote the closed ball in $\mathbf{R}^{d_\beta}$ with the same center but half the radius. Given that $\overline{B}$ is a closed and bounded subset of a finite-dimensional Euclidean space, by the Heine–Borel theorem it is compact. Assumptions 2 and 3 imply that $(t,\beta) \mapsto \omega(t,x)(\partial/\partial\beta)\rho(z,\beta,h^*(w))$ is continuous on $\mathcal{T}\times\mathcal{N}_0$ for each $z \in \mathcal{Z}$, hence on the subset $\mathcal{T}\times\overline{B}$, and this function is is dominated by an integrable function depending on z only. Moreover, via Tychonoff's theorem, $\mathcal{T}$ and $\overline{B}$ compact imply that is $\mathcal{T}\times\overline{B}$ compact. Combining these observations with the fact that the data are i.i.d., Newey and McFadden (1994, Lemma 2.4) tells us that

(i) $(t,\beta) \mapsto \mathrm{E}[\omega(t,X)(\partial/\partial\beta)\rho(Z,\beta,h^*(W))]$ is continuous on $\mathcal{T}\times\overline{B}$,

(ii) $\sup_{(t,\beta)\in\mathcal{T}\times\overline{B}} \|(\mathbb{E}_n - \mathrm{E})[\omega(t,X_i)(\partial/\partial\beta)\rho(Z_i,\beta,h^*(W_i))]\| \xrightarrow{\mathrm{P}} 0.$

Given (i) and $\mathcal{T}\times\overline{B}$ compact, we must have (cf. Rudin, 1976, Theorem 4.19) that

(iii) $(t,\beta) \mapsto \mathrm{E}[\omega(t,X)(\partial/\partial\beta)\rho(Z,\beta,h^*(W))]$ is uniformly continuous on $\mathcal{T}\times\overline{B}$.

Let $\widetilde{\beta}$ be an arbitrary consistent estimator of β_0. Then $\widetilde{\beta} \in \overline{B}$ wp $\to 1$, and, on this

event,

$$
\begin{aligned}
&\sup_{t\in\mathcal{T}}\left\|\mathbb{E}_n\left[\omega\left(t,X_i\right)\left(\partial/\partial\beta\right)\rho(Z_i,\widetilde{\beta},h^*\left(W_i\right))\right]-b\left(t\right)\right\| \\
&\leqslant \sup_{t\in\mathcal{T}}\left\|\left(\mathbb{E}_n-\mathrm{E}_Z\right)\left[\omega\left(t,X_i\right)\left(\partial/\partial\beta\right)\rho(Z_i,\widetilde{\beta},h^*\left(W_i\right))\right]\right\| \\
&\quad+\sup_{t\in\mathcal{T}}\left\|\mathrm{E}_Z\left[\omega\left(t,X\right)\left(\partial/\partial\beta\right)\rho(Z,\widetilde{\beta},h^*\left(W\right))\right]-b\left(t\right)\right\| \\
&\leqslant \sup_{(t,\beta)\in\mathcal{T}\times\overline{B}}\left\|\left(\mathbb{E}_n-\mathrm{E}\right)\left[\omega\left(t,X_i\right)\left(\partial/\partial\beta\right)\rho(Z_i,\beta,h^*\left(W_i\right))\right]\right\| \\
&\quad+\sup_{t\in\mathcal{T}}\left\|\mathrm{E}_Z\left[\omega\left(t,X\right)\left(\partial/\partial\beta\right)\rho(Z,\widetilde{\beta},h^*\left(W\right))\right]-b\left(t\right)\right\|\xrightarrow{\mathrm{P}}0,
\end{aligned}
$$

where the first inequality is due to T, the second uses $\{\widetilde{\beta}\in\overline{\mathcal{N}}\}$, and we have used (ii) uniform convergence and (iii) uniform continuity. Invoking the conclusion of the previous display for the mean value $\widetilde{\beta}\coloneqq\overline{\beta}$ we see that $\sup_{t\in\mathcal{T}}\|\mathrm{I}_{a,n}\left(t\right)-b\left(t\right)\|\to_{\mathrm{P}}0$, which combined with $\sup_{t\in\mathcal{T}}\|\mathrm{I}_{b,n}\left(t\right)\|\to_{\mathrm{P}}0$ and T establishes Part (a).

Continuity and $\mathcal{T}\times\overline{B}$ compact also imply $(t,\beta)\mapsto\mathrm{E}[\omega\left(t,X\right)\left(\partial/\partial\beta\right)\rho(Z,\beta,h^*\left(W\right))]$ is *bounded* on $\mathcal{T}\times\overline{B}$ (cf. Rudin, 1976, Theorem 4.15). Part (b) then follows from $\beta_0\in\overline{B}$.

Step 2: $\|\mathbf{II}_n\|_{\mathcal{T}}$

In this step I show that II_n defined in (B.3) satisfies

$$
\|\mathrm{II}_n\|_{\mathcal{T}}\lesssim_{\mathrm{P}}\mathrm{E}\left[R\left(Z\right)\right]\sqrt{n}\max_{1\leqslant m\leqslant d}\|\widehat{h}_m-h_m^*\|_{\mathcal{W}}^{1+\gamma}
$$

for R and γ given by Assumption 3. Using T and CS, Assumptions 2 and 3 imply that

$$
\begin{aligned}
\|\mathrm{II}_n\|_{T}&\leqslant\|\omega\|_{\mathcal{T}\times\mathcal{T}}\sqrt{n}\mathbb{E}_n\left[\|\frac{\partial}{\partial h^\top}\rho(Z_i,\overline{h}\left(W_i\right))-\frac{\partial}{\partial h^\top}\rho(Z_i,h^*\left(W_i\right))\|\|\widehat{h}\left(W_i\right)-h\left(W_i\right)\|\right] \\
&\lesssim\sqrt{n}\mathbb{E}_n[R\left(Z_i\right)\|\overline{h}\left(W_i\right)-h^*\left(W_i\right)\|^{\gamma}\|\widehat{h}\left(W_i\right)-h^*\left(W_i\right)\|] \\
&\leqslant\sqrt{n}\mathbb{E}_n[R\left(Z_i\right)\|\widehat{h}\left(W_i\right)-h^*\left(W_i\right)\|^{1+\gamma}] \\
&\leqslant d^{(1+\gamma)/2}\mathbb{E}_n\left[R\left(Z_i\right)\right]\sqrt{n}\max_{1\leqslant m\leqslant d}\|\widehat{h}_m-h_m^*\|_{\mathcal{W}}^{1+\gamma} \\
&\lesssim_{\mathrm{P}}\mathrm{E}\left[R\left(Z\right)\right]\sqrt{n}\max_{1\leqslant m\leqslant d}\|\widehat{h}_m-h_m^*\|_{\mathcal{W}}^{1+\gamma},
\end{aligned}
$$

where $\overline{h}(W_i)$ is on the line segment connecting $\widehat{h}(W_i)$ and $h(W_i)$, thus satisfying $\|\overline{h}(W_i) - h^*(W_i)\| \leqslant \|\widehat{h}(W_i) - h^*(W_i)\|$, and $\mathbb{E}_n[R(Z_i)] \lesssim_{\mathrm{P}} \mathrm{E}[R(Z)]$ follows from M.

Step 3: $\|\mathbf{III}_n\|_{\mathcal{T}}$

In this step I show that III_n defined in (B.3) satisfies

$$\|\mathrm{III}_n\|_{\mathcal{T}} \lesssim_{\mathrm{P}} \max_{1\leqslant m\leqslant d} \Big(\sum_{j=1}^{k_{m,n}} \|p_j\|_{\mathcal{W}}^2\Big)^{1/2} \Big(\sqrt{k_{m,n}/n} + k_{m,n}^{-\alpha}\Big)$$

for α given by Assumption 6. For $h : \mathcal{W} \to \mathbf{R}^d$ composed by maps $\{h_m\}_1^d$ in $L^2(W)$, define the map D

$$D(t,z,h) \coloneqq \omega(t,x)(\partial/\partial h^\top)\rho(z,h^*(w))\,h(w) \tag{B.6}$$

such that $h \mapsto D(t,z,h)$ is a linear functional for given $(t,z) \in \mathcal{T}\times\mathcal{Z}$. Let Δ denote the centered version of D, i.e.,

$$\begin{aligned}\Delta(t,z,h) \coloneqq{}& \omega(t,x)(\partial/\partial h^\top)\rho(z,h^*(w))\,h(w)\\ &- \mathrm{E}_Z\left[\omega(t,X)(\partial/\partial h^\top)\rho(Z,h^*(W))\,h(W)\right]\end{aligned} \tag{B.7}$$

which is also linear in h. Letting $\widetilde{h} = p^{k\top}\widetilde{\pi}$ be as in Assumption 6, by linearity we may write

$$\begin{aligned}\mathrm{III}_n(t) &= \sqrt{n}\mathbb{E}_n\left[\Delta(t,Z_i,\widehat{h}-h^*)\right]\\ &= \sqrt{n}\mathbb{E}_n\left[\Delta(t,Z_i,\widehat{h}-\widetilde{h})\right] + \sqrt{n}\mathbb{E}_n\left[\Delta(t,Z_i,\widetilde{h}-h^*)\right]\\ &=: \mathrm{III}_{a,n}(t) + \mathrm{III}_{b,n}(t).\end{aligned} \tag{B.8}$$

Given that $\zeta_k = \sup_{w\in\mathcal{W}}[\sum_{j=1}^k p_j(w)^2]^{1/2}$ and $\zeta_{k_n} \to \infty$ (see Remark 2), we must have $\sum_{j=1}^{k_n}\|p_j\|_{\mathcal{W}}^2 \to \infty$. In particular, $\sum_{j=1}^{k_n}\|p_j\|_{\mathcal{W}}^2$ is bounded away from zero as $n\to\infty$.

By T, the desired conclusion will therefore follow from showing

$$\|\mathrm{III}_{a,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} \max_{1 \leqslant m \leqslant d} \Big(\sum_{j=1}^{k_{m,n}} \|p_j\|_{\mathcal{W}}^2 \Big)^{1/2} \Big(\sqrt{k_{m,n}/n} + k_{m,n}^{-\alpha_m} \Big),$$
$$\|\mathrm{III}_{b,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} k_n^{-\alpha}.$$

Step 3a: $\|\mathbf{III}_{a,n}\|_{\mathcal{T}}$ In this step I show that $\mathrm{III}_{a,n}$ defined in (B.8) satisfies

$$\|\mathrm{III}_{a,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} \max_{1 \leqslant m \leqslant d} \Big(\sum_{j=1}^{k_{m,n}} \|p_j\|_{\mathcal{W}}^2 \Big)^{1/2} \Big(\sqrt{k_{m,n}/n} + k_{m,n}^{-\alpha_m} \Big)$$

for α_m given by Assumption 6. Given that

$$\mathrm{III}_{a,n}(t) = \sqrt{n}\mathbb{E}_n \left[\Delta(t, Z_i, \widehat{h} - \widetilde{h}) \right] = \sum_{m=1}^{d} \sqrt{n}\mathbb{E}_n \left[\Delta_m(t, Z_i, \widehat{h}_m - \widetilde{h}_m) \right],$$
$$\begin{aligned} \Delta_m(t, Z_i, h_m) \coloneqq{}& \omega(t, x) (\partial/\partial h_m) \rho(z, h^*(w)) h_m(w) \\ & - \mathrm{E}\left[\omega(t, X) (\partial/\partial h_m) \rho(Z, h^*(W)) h_m(W) \right], \end{aligned}$$

by T, we may focus on bounding a single $\sup_{t \in \mathcal{T}} |\sqrt{n}\mathbb{E}_n[\Delta_m(t, Z_i, \widehat{h}_m - \widetilde{h}_m)]|$ in probability. For the remainder of this section I therefore drop the m subscript and write $(\partial/\partial h)\, \rho(Z, h^*(Z))$ for the scalar $(\partial/\partial h_m)\, \rho(Z, h^*(Z))$. Let

$$\Delta_i^k(t) \coloneqq (\Delta(t, Z_i, p_1), \ldots, \Delta(t, Z_i, p_k))^\top.$$

Then CS implies

$$\begin{aligned} \|\mathrm{III}_{a,n}\|_{\mathcal{T}} &= \sup_{t \in \mathcal{T}} \left| \sqrt{n}\mathbb{E}_n \left[\Delta(t, Z_i, p^{k_n\top}(\widehat{\pi} - \widetilde{\pi}) \right] \right| = \sup_{t \in \mathcal{T}} \left| \sqrt{n} \left\{ \mathbb{E}_n[\Delta_i^{k_n}(t)] \right\}^\top (\widehat{\pi} - \widetilde{\pi}) \right| \\ &\leqslant \|\widehat{\pi} - \widetilde{\pi}\| \sup_{t \in \mathcal{T}} \left\| \sqrt{n}\mathbb{E}_n[\Delta_i^{k_n}(t)] \right\|. \end{aligned}$$

Lemma B5 tells us that $\|\widehat{\pi} - \widetilde{\pi}\| \lesssim_{\mathrm{P}} \sqrt{k_n/n} + k_n^{-\alpha}$, so it remains to show that

$$\sup_{t \in \mathcal{T}} \left\| \sqrt{n}\mathbb{E}_n[\Delta_i^{k_n}(t)] \right\| \lesssim_{\mathrm{P}} \Big(\sum_{j=1}^{k_n} \|p_j\|_{\mathcal{W}}^2 \Big)^{1/2}.$$

By M it suffices to show the finite-sample moment bound, for any $k \in \mathbf{N}$,

$$\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left\|\sqrt{n}\mathbb{E}_n[\Delta_i^k(t)]\right\|^2\right] \lesssim \sum_{j=1}^{k}\|p_j\|_{\mathcal{W}}^2.$$

Given that

$$\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left\|\sqrt{n}\mathbb{E}_n[\Delta_i^k(t)]\right\|^2\right] \leqslant \sum_{j=1}^{k}\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left[\sqrt{n}\mathbb{E}_n[\Delta(t,Z_i,p_j)]\right]^2\right],$$

it suffices to show that

$$\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left[\sqrt{n}\mathbb{E}_n[\Delta(t,Z_i,p_j)]\right]^2\right] \lesssim \|p_j\|_{\mathcal{W}}^2, \quad j\in\{1,\ldots,k\}.$$

To this end, fix $j \in \{1,\ldots,k\}$, and consider the function class $\mathcal{F}_j \coloneqq \mathcal{F}_j(\mathcal{T}) \coloneqq \{f : z \mapsto \Delta(t,z,p_j); t \in \mathcal{T}\}$. For $f_1 \coloneqq f(\cdot;t_1), f_2 \coloneqq f(\cdot;t_2) \in \mathcal{F}_j$ arbitrary, by T, J and Assumptions 2 and 3,

$$\begin{aligned}
&|f_1(z) - f_2(z)| \\
&= \left|\left[\omega(t_1,x)-\omega(t_2,x)\right]\frac{\partial}{\partial h}\rho(z,h^*(w))\,p_j(w)\right. \\
&\quad \left. - \mathrm{E}\left[\left\{\omega(t_1,X)-\omega(t_2,X)\right\}\frac{\partial}{\partial h}\rho(Z,h^*(W))\,p_j(W)\right]\right| \\
&\leqslant |\omega(t_1,x)-\omega(t_2,x)|\left|\frac{\partial}{\partial h}\rho(z,h^*(w))\right||p_j(w)| \\
&\quad + \mathrm{E}\left[|\omega(t_1,X)-\omega(t_2,X)|\left|\frac{\partial}{\partial h}\rho(Z,h^*(W))\right||p_j(W)|\right] \\
&\lesssim \left(\left|\frac{\partial}{\partial h}\rho(z,h^*(w))\right||p_j(w)| + \mathrm{E}\left[\left|\frac{\partial}{\partial h}\rho(Z,h^*(W))\right||p_j(W)|\right]\right)\|t_1-t_2\| \\
&\leqslant \left(\left|\frac{\partial}{\partial h}\rho(z,h^*(w))\right| + \mathrm{E}\left[\left|\frac{\partial}{\partial h}\rho(Z,h^*(W))\right|\right]\right)\|p_j\|_{\mathcal{W}}\|t_1-t_2\| \\
&= L_1(z)\,\|p_j\|_{\mathcal{W}}\|t_1-t_2\|,
\end{aligned}$$

such that we may write

$$|f_1(z)-f_2(z)| \leqslant F_{1j}(z)\,\|t_1-t_2\|, \quad F_{1j}(z) \coloneqq C_1 L_1(z)\,\|p_j\|_{\mathcal{W}},$$

for some constant $C_1 \in \mathbf{R}_{++}$. Similarly, for $f \coloneqq f(\cdot;t) \in \mathcal{F}_j$ arbitrary, by T, J and Assumptions 2 and 3,

$$\begin{aligned} |f(z)| &= \left| \omega(t,x) \frac{\partial}{\partial h} \rho(z, h^*(w)) p_j(w) - \mathrm{E}_Z \left[\omega(t,X) \frac{\partial}{\partial h} \rho(Z, h^*(W)) p_j(W) \right] \right| \\ &\lesssim L_1(z) \|p_j\|_{\mathcal{W}}, \end{aligned}$$

such that we may write

$$|f(z)| \leqslant F_{2j}(z), \quad F_{2j}(z) \coloneqq C_2 L_1(z) \|p_j\|_{\mathcal{W}},$$

for some constant $C_2 \in (0, \infty)$. Let $C_3 \coloneqq C_1 \vee C_2$ and

$$F_j(z) \coloneqq C_3 L_1(z) \|p_j\|_{\mathcal{W}}.$$

Then $\|F_j\|_{P,2} \lesssim \|p_j\|_{\mathcal{W}}$, so F_j is an square-integrable envelope for $\mathcal{F}_j$ satisfying

$$|f_1(z) - f_2(z)| \leqslant F_j(z) \|t_1 - t_2\|.$$

Given that $\mathcal{T}$ is compact (Assumption 2), we must have $\operatorname{diam}(\mathcal{T}) < \infty$. Pollard (1990, Lemma 4.1) and the fact that covering numbers are bounded by packing numbers (cf. van der Vaart and Wellner, 1996, p. 98) therefore combine to yield $N(\varepsilon, \mathcal{T}, \|\cdot\|) \leqslant (3\operatorname{diam}(\mathcal{T})/\varepsilon)^{d_t}$ for $\varepsilon \in (0, \operatorname{diam}(\mathcal{T})]$. Hence, by van der Vaart and Wellner (1996, Theorem 2.7.11) and the previous display,

$$N_{[]}(\varepsilon \|F_j\|_{P,2}, \mathcal{F}_j, L^2(P)) \leqslant N(\varepsilon/2, \mathcal{T}, \|\cdot\|) \leqslant (6\operatorname{diam}(\mathcal{T})/\varepsilon)^d \leqslant (C/\varepsilon)^d$$

for $\varepsilon \in (0, \operatorname{diam}(\mathcal{T})]$ (and $= 1$ otherwise). The bracketing integral of $\mathcal{F}_j$ therefore satisfies the bound

$$J_{[]}\left(\delta, \mathcal{F}_j, L^2(P)\right) \leqslant \int_0^\delta \sqrt{1 + C \ln(1/\varepsilon)} \mathrm{d}\varepsilon.$$

Note that the right-hand side depends on neither j nor k. In particular, the integral $J_{[]}(1, \mathcal{F}_j, L^2(P))$ is bounded uniformly in $j \in \{1, \ldots, k\}, k \in \mathbf{N}$. By construction, $\mathrm{E}[f(Z)] = \mathrm{E}[\Delta(t, Z, p_j)] = 0$ for any $f \in \mathcal{F}_j$, so we may view the stochastic process $\{\sqrt{n}\mathbb{E}_n[\Delta(t, Z_i, p_j)]; t \in \mathcal{T}\}$ as an *empirical* process $\{\mathbb{G}_n(f); f \in \mathcal{F}_j\}$. van der Vaart

and Wellner (1996, Theorem 2.14.2) therefore implies the finite-sample bound

$$\mathrm{E}\left[\|\mathbb{G}_n\|_{\mathcal{F}_j}\right] \lesssim J_{[\,]}\left(1, \mathcal{F}_j, L^2\left(P\right)\right)\|F_j\|_{P,2} \lesssim \|F_j\|_{P,2} \lesssim \|p_j\|_{\mathcal{W}}\,.$$

van der Vaart and Wellner (1996, Theorem 2.14.5) now shows

$$\left(\mathrm{E}[\|\mathbb{G}_n\|^2_{\mathcal{F}_j}]\right)^{1/2} \lesssim \mathrm{E}\left[\|\mathbb{G}_n\|_{\mathcal{F}_j}\right] + \|F_j\|_{P,2} \lesssim \|p_j\|_{\mathcal{W}}\,,$$

which is the desired bound.

Step 3b: $\|\mathbf{III}_{b,n}\|_{\mathcal{T}}$ In this step I show that $\mathrm{III}_{b,n}$ defined in (B.8) satisfies

$$\|\mathrm{III}_{b,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} \max_{1\leqslant m\leqslant d} k_{m,n}^{-\alpha_m},$$

for α_m given by Assumption 6. Given that

$$\mathrm{III}_{b,n}\left(t\right) = \sqrt{n}\mathbb{E}_n\left[\Delta(t, Z_i, \widetilde{h} - h^*)\right] = \sum_{m=1}^{d} \sqrt{n}\mathbb{E}_n\left[\Delta_m(t, Z_i, \widetilde{h}_m - h^*_m)\right],$$

as was the case for $\|\mathrm{III}_{a,n}\|_{\mathcal{T}}$, by T we may focus on bounding each right-hand side term in probability and therefore drop the m subscript. For this purpose, fix $k \in \mathbf{N}$ and consider the function class $\mathcal{F}_k \coloneqq \mathcal{F}_k\left(\mathcal{T}\right) \coloneqq \{f\colon z \mapsto \Delta(t, z, \widetilde{h} - h^*); t \in \mathcal{T}\}$. For $f \coloneqq f(\cdot, t), f_1 \coloneqq f(\cdot, t_1), f_2 \coloneqq f(\cdot, t_2) \in \mathcal{F}_k$ arbitrary, arguments analogous to the ones applied to handle $\|\mathrm{III}_{a,n}\|_{\mathcal{T}}$ establish that

$$\begin{aligned}\left|f_1\left(z\right) - f_2\left(z\right)\right| &\leqslant C_1 L_1\left(z\right)\|\widetilde{h} - h^*\|_{\mathcal{W}}\left\|t_1 - t_2\right\|,\\ \left|f\left(z\right)\right| &\leqslant C_2 L_1\left(z\right)\|\widetilde{h} - h^*\|_{\mathcal{W}}.\end{aligned}$$

Define $C_3 \coloneqq C_1 \vee C_2$ and $F_k\left(z\right) \coloneqq C_3 L_1\left(z\right)\|\widetilde{h} - h^*\|_{\mathcal{W}}$. Then $\|F_k\|_{P,2} = C_4\|\widetilde{h} - h^*\|_{\mathcal{W}} \lesssim k^{-\alpha}$ by Assumption 6. Hence F_k is an square-integrable envelope for $\mathcal{F}_k$, and arguments analogous to the ones used for $\|\mathrm{III}_{a,n}\|_{\mathcal{T}}$ show that the resulting bracketing integral $J_{[\,]}\left(\delta, \mathcal{F}_k, L^2\left(P\right)\right)$ is bounded by a constant independent of k. van der Vaart and Wellner (1996, Theorem 2.14.2) therefore implies

$$\mathrm{E}\left[\|\mathbb{G}_n\|_{\mathcal{F}_k}\right] \lesssim J_{[\,]}\left(1, \mathcal{F}_k, L^2\left(P\right)\right)\|F_k\|_{P,2} \lesssim \|F_k\|_{P,2} \lesssim k^{-\alpha}.$$

and the claim follows from M.

Step 4: $\|\mathbf{IV}_n\|_{\mathcal{T}}$

In this step I show that IV_n defined in (B.3) satisfies

$$\|\mathrm{IV}_n\|_{\mathcal{T}} \lesssim_{\mathrm{P}} \max_{1\leqslant m\leqslant d}\Big\{\sqrt{n}r_{h_m,k_{m,n}}\sup_{t\in\mathcal{T}} r_{\delta_m,k_{m,n}}(t) + \sqrt{\zeta^2_{k_{m,n}}k_{m,n}\ln(k_{m,n})/n} \\ + R_{\delta_m,k_{m,n}}\sqrt{\ln\left(k_{m,n}/R_{\delta_m,k_{m,n}}\right)} + \zeta_{k_{m,n}}r_{h_m,k_{m,n}}\Big\},$$

where $\zeta_k, r_{h_m,k}, r_{\delta_m,k}$ and $R_{\delta_m,k}$ are defined in (3.6), (3.12), (3.13) and (3.14), respectively. Given the decomposition

$$\begin{aligned}\mathrm{IV}_n(t) &= \sqrt{n}\Big(\mathrm{E}_Z\Big[\omega(t,X)(\partial/\partial h^\top)\rho(Z,h^*(W))\{\widehat{h}(W)-h^*(W)\}\Big] \\ &\quad - \mathbb{E}_n[\delta(t,W_i)^\top\{Y_i-h^*(W_i)\}]\Big) \\ &= \sum_{m=1}^{d}\sqrt{n}\Big(\mathrm{E}_Z\Big[\omega(t,X)(\partial/\partial h_m)\rho(Z,h^*(W))\{\widehat{h}_m(W)-h^*_m(W)\}\Big] \\ &\quad - \mathbb{E}_n[\delta_m(t,W_i)\{Y_{mi}-h^*_m(W_i)\}]\Big),\end{aligned}$$

by T we may drop the m subscript and focus on bounding a single summand uniformly over $\mathcal{T}$ in probability. For this purpose, recall that h_k and $\delta_k(t,\cdot)$ are the mean-square projections of $h^*$and $\delta(t,\cdot)$, respectively, onto the linear span of p^k and $r^2_{h,k}$ and $r^2_{\delta,k}(t)$ are the mean-square errors resulting from these projections. Define

$$\psi_k(t) \coloneqq \mathrm{E}\left[\delta(t,W)p^k(W)\right] \tag{B.9}$$

By Assumption 5, the population least-square coefficients $\pi_k = Q_k^{-1}\mathrm{E}[p^k(W)Y]$ are well defined for all $k\in\mathbf{N}$. Applying Lemma B1, we see that the inverse of $\widehat{Q}_{k_n}$ exists wp $\to 1$. As a consequence, the sample least-squares coefficients take the form $\widehat{\pi} = \widehat{Q}^{-1}_{k_n}\mathbb{E}_n[p^{k_n}(W_i)Y_i]$ wp $\to 1$. Assuming—without loss of generality—that $\widehat{Q}^{-1}_{k_n}$

exists with probability one for all n,

$$\begin{aligned}\sqrt{n}\mathrm{E}_W\{\delta(t,W)[\widehat{h}(W)-h_{k_n}(W)]\} &= \sqrt{n}\mathrm{E}_W\{\delta(t,W)\,p^{k_n}(W)^\top(\widehat{\pi}-\pi_{k_n})\}\\ &= \psi_{k_n}(t)^\top\sqrt{n}(\widehat{\pi}-\pi_{k_n})\\ &= \psi_{k_n}(t)^\top\sqrt{n}\left(\widehat{Q}_{k_n}^{-1}\mathbb{E}_n\left[p^{k_n}(W_i)Y_i\right]-\pi_{k_n}\right)\\ &= \psi_{k_n}(t)^\top\widehat{Q}_{k_n}^{-1}\sqrt{n}\left(\mathbb{E}_n\left[p^{k_n}(W_i)Y_i\right]-\widehat{Q}_{k_n}\pi_{k_n}\right)\\ &= \psi_{k_n}(t)^\top\widehat{Q}_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}(W_i)\{Y_i-h_{k_n}(W_i)\}\right],\end{aligned}$$

where $\mathrm{E}_W[\cdot]$ denotes integration with respect to the distribution of W. By definition of $\delta(t,W)$ [see (3.9)] and iterated of expectations, for any measurable function h of W alone,

$$\mathrm{E}[\omega(t,X)(\partial/\partial h)\rho(Z,h^*(W))\,h(W)] = \mathrm{E}[\delta(t,W)\,h(W)].$$

Using the previous two displays and adding and subtracting

$$\begin{aligned}&\sqrt{n}\mathbb{E}_n[\delta_{k_n}(t,W_i)\{Y_i-h_{k_n}(W_i)\}]\\ &= \sqrt{n}\mathbb{E}_n\left[p^{k_n}(W_i)^\top Q_{k_n}^{-1}\mathrm{E}[p^{k_n}(W)\,\delta(t,W)]\{Y_i-h_{k_n}(W_i)\}\right]\\ &= \psi_{k_n}(t)^\top Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}(W_i)\{Y_i-h_{k_n}(W_i)\}\right],\end{aligned}$$

we may decompose $\mathrm{IV}_n(t)$ as

$$\begin{aligned}\mathrm{IV}_n(t) &= \sqrt{n}\mathrm{E}_W[\delta(t,W)\{\widehat{h}(W)-h^*(W)\}]-\sqrt{n}\mathbb{E}_n[\delta(t,W_i)\{Y_i-h^*(W_i)\}]\\ &= \sqrt{n}\mathrm{E}_W\{\delta(t,W)[h_{k_n}(W)-h^*(W)]\}+\sqrt{n}\mathrm{E}_W[\delta(t,W)\{\widehat{h}(W)-h_{k_n}(W)\}]\\ &\quad+\sqrt{n}\mathbb{E}_n[\delta(t,W_i)\{Y_i-h^*(W_i)\}]\\ &= \sqrt{n}\mathrm{E}_W[\delta(t,W)\{h_{k_n}(W)-h^*(W)\}]\\ &\quad+\psi_{k_n}(t)^\top(\widehat{Q}_{k_n}^{-1}-Q_{k_n}^{-1})\sqrt{n}\mathbb{E}_n\left[p^{k_n}(W_i)\{Y_i-h_{k_n}(W_i)\}\right]\\ &\quad+\sqrt{n}\mathbb{E}_n[\delta_{k_n}(t,W_i)\{Y_i-h_{k_n}(W_i)\}-\delta^*(t,W_i)\{Y_i-h_*(W_i)\}]\\ &=: \mathrm{IV}_{a,n}(t)+\mathrm{IV}_{b,n}(t)+\mathrm{IV}_{c,n}(t).\end{aligned}$$

By T it therefore suffices to show that

$$\begin{aligned} \|\mathrm{IV}_{a,n}\|_{\mathcal{T}} &\leqslant \sqrt{n} r_{h,k_n} \sup_{t\in\mathcal{T}} r_{\delta,k_n}(t), \\ \|\mathrm{IV}_{b,n}\|_{\mathcal{T}} &\lesssim_{\mathrm{P}} \sqrt{\zeta_{k_n}^2 k_n \ln(k_n)/n}, \\ \text{and} \quad \|\mathrm{IV}_{c,n}\|_{\mathcal{T}} &\lesssim_{\mathrm{P}} R_{\delta,k_n} \sqrt{\ln(k_n/R_{\delta,k_n})} + \zeta_{k_n} r_{h,k_n}. \end{aligned}$$

Step 4a: $\|\mathbf{IV}_{a,n}\|_{\mathcal{T}}$ In order to establish the inequality

$$\|\mathrm{IV}_{a,n}\|_{\mathcal{T}} \leqslant \sqrt{n} r_{h,k_n} \sup_{t\in\mathcal{T}} r_{\delta,k_n}(t),$$

recall that h_k defined in (B.4) is the mean-square projection of h^* onto the span of p^k, so by orthogonality of projections we have $\mathrm{E}[\delta_k(t,W)\{h_k(W) - h^*(W)\}] = 0$ for each $t \in \mathcal{T}$. Now J followed by CS yield

$$\begin{aligned} \|\mathrm{IV}_{a,n}\|_{\mathcal{T}} &= \sqrt{n} \sup_{t\in\mathcal{T}} |\mathrm{E}[\delta(t,W)\{h_{k_n}(W) - h^*(W)\}]| \\ &= \sqrt{n} \sup_{t\in\mathcal{T}} |\mathrm{E}[\{\delta_{k_n}(t,W) - \delta(t,W)\}\{h_{k_n}(W) - h^*(W)\}]| \\ &\leqslant \sqrt{n} \|h_{k_n} - h^*\|_{P,2} \sup_{t\in\mathcal{T}} \|\delta_{k_n}(t,\cdot) - \delta(t,\cdot)\|_{P,2} = \sqrt{n} r_{h,k_n} \sup_{t\in\mathcal{T}} r_{\delta,k_n}(t). \end{aligned}$$

Step 4b: $\|\mathbf{IV}_{b,n}\|_{\mathcal{T}}$ In this step I show that

$$\|\mathrm{IV}_{b,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} \sqrt{\zeta_{k_n}^2 k_n \ln(k_n)/n}.$$

Using the fact that mean-square projections are $L^2(P)$-contractions followed by Assumptions 2 and 3, we see that

$$\begin{aligned} \psi_k(t)^\top Q_k^{-1} \psi_k(t) &= \{Q_k^{-1}\mathrm{E}[p^k(W)\delta(t,W)]\}^\top Q_k \{Q_k^{-1}\mathrm{E}[p^k(W)\delta(t,W)]\} \\ &= \mathrm{E}[\delta_k(t,W)] \leqslant \mathrm{E}[\delta(t,W)^2] = \mathrm{E}[\omega(t,W)^2 (\partial/\partial h)\rho(Z,h^*(W))^2] \\ &\lesssim \mathrm{E}[(\partial/\partial h)\rho(Z,h^*(W))^2] < \infty, \end{aligned}$$

with an upper bound that depends on neither t nor k. By the Min-Max Theorem, Assumption 5, and the previous display, it follows that

$$\begin{aligned}\|\psi_k(t)\,Q_k^{-1}\|^2 &= [\psi_k(t)\,Q_k^{-1/2}]^\top Q_k^{-1}[Q_k^{-1/2}\psi_K(t)] \lesssim \|\psi_k(t)\,Q_k^{-1/2}\|^2 \\ &\leqslant \sup_{k\in\mathbf{N},t\in\mathcal{T}} |\psi_k(t)^\top Q_k^{-1}\psi_k(t)| < \infty,\end{aligned}$$

thus implying $\sup_{k\in\mathbf{N},t\in\mathcal{T}}\|\psi_k(t)\,Q_k^{-1}\| < \infty$. By Lemma B4 we have $\|\widehat{Q}_{k_n} - Q_{k_n}\|_{\text{op}} \lesssim_{\text{P}} [\zeta_{k_n}^2 \ln(k_n)/n]^{1/2} \to 0$ under Assumption 7. Moreover, Lemma B1 shows that $\|\widehat{Q}_{k_n}^{-1}\|_{\text{op}} \lesssim_{\text{P}} 1$. Using these observations and the previous display,

$$\begin{aligned}\sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1} - \psi_{k_n}(t)^\top Q_k^{-1}\| &= \sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top Q_{k_n}^{-1}(Q_{k_n} - \widehat{Q}_{k_n})\widehat{Q}_{k_n}^{-1}\| \\ &\leqslant \|(Q_{k_n} - \widehat{Q}_{k_n})\widehat{Q}_{k_n}^{-1}\|_{\text{op}} \sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top Q_{k_n}^{-1}\| \\ &\leqslant \|\widehat{Q}_{k_n} - Q_{k_n}\|_{\text{op}}\|\widehat{Q}_{k_n}^{-1}\|_{\text{op}} \sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top Q_{k_n}^{-1}\| \\ &\lesssim_{\text{P}} \sqrt{\zeta_{k_n}^2 \ln(k_n)/n} \to 0.\end{aligned}$$

From the previous display and $\sup_{k\in\mathbf{N},t\in\mathcal{T}}\|\psi_k(t)\,Q_k^{-1}\| < \infty$ it follows that

$$\sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1}\| \lesssim_{\text{P}} 1.$$

Observe also that, by the Assumption 5, the Min-Max theorem, and the fact that $\mathrm{E}[p^k(W)\{Y - h_k(W)\}] = 0$ (which follows from h_k being the mean-square projection of h^*),

$$\begin{aligned}&\mathrm{E}\left[\|Q_k^{-1}\sqrt{n}\mathbb{E}_n\left[p^k(W_i)\{Y_i - h_k(W_i)\}\right]\|^2\right] \\ &\lesssim \mathrm{E}\left[\|Q_k^{-1/2}\sqrt{n}\mathbb{E}_n\left[p^k(W_i)\{Y_i - h_k(W_i)\}\right]\|^2\right] \\ &= \mathrm{E}\left[p^k(W)^\top Q_k^{-1}p^k(W)\{Y - h_k(W)\}^2\right] \\ &= \mathrm{E}\left[U^2 p^k(W)^\top Q_k^{-1}p^k(W)\right] + \mathrm{E}\left[p^k(W)^\top Q_k^{-1}p^k(W)\{h_k(W) - h^*(W)\}^2\right],\end{aligned}$$

where I have used $U = Y - h^*(W)$. By Assumption 4, $\mathrm{E}[U^2|W]$ is bounded, so

$$\begin{aligned}\mathrm{E}[U^2 p^k(W)^\top Q_k^{-1}p^k(W)] &= \mathrm{E}[\mathrm{E}\left[U^2|W\right]p^k(W)^\top Q_k^{-1}p^k(W)] \\ &\lesssim \mathrm{E}[p^k(W)^\top Q_k^{-1}p^k(W)] = k.\end{aligned}$$

Moreover,

$$\begin{aligned}&\mathrm{E}\left[p^k(W)^\top Q_k^{-1} p^k(W)\{h_k(W)-h^*(W)\}^2\right]\\&\lesssim \mathrm{E}\left[\|p^k(W)\|^2\{h_k(W)-h^*(W)\}^2\right]\leqslant \zeta_k^2 r_{h,k}^2.\end{aligned}$$

Given Assumption 7, $\zeta_k^2 r_{h,k}^2=(\zeta_k r_{h,k})^2\to 0$ as $k\to\infty$, so

$$\mathrm{E}\left[\|Q_k^{-1}\sqrt{n}\mathbb{E}_n\left[p^k(W_i)\{Y_i-h_k(W_i)\}\right]\|^2\right]\lesssim k.$$

M now implies

$$\|Q_k^{-1}\sqrt{n}\mathbb{E}_n\left\{p^{k_n}(W_i)\left[Y_i-h_{k_n}(W_i)\right]\right\}\|\lesssim_{\mathrm{P}}\sqrt{k_n}.$$

Using CS we therefore arrive at

$$\begin{aligned}&\|\mathrm{IV}_{b,n}\|_{\mathcal{T}}\\&=\sup_{t\in\mathcal{T}}\left|\psi_{k_n}(t)^\top\widehat{Q}_{k_n}^{-1}(Q_{k_n}-\widehat{Q}_{k_n})Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}(W_i)\{Y_i-h_{k_n}(W_i)\}\right]\right|\\&\leqslant\left\|Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}(W_i)\{Y_i-h_{k_n}(W_i)\}\right]\right\|\sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top\widehat{Q}_{k_n}^{-1}(Q_{k_n}-\widehat{Q}_{k_n})\|\\&\leqslant\left\|Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}(W_i)\{Y_i-h_{k_n}(W_i)\}\right]\right\|\|\widehat{Q}_{k_n}-Q_{k_n}\|_{\mathrm{op}}\sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top\widehat{Q}_{k_n}^{-1}\|\\&\lesssim_{\mathrm{P}}\sqrt{k_n}\sqrt{\zeta_{k_n}^2\ln(k_n)/n}.\end{aligned}$$

Step 4c: $\|\mathbf{IV}_{c,n}\|_{\mathcal{T}}$ In this section I show that

$$\|\mathrm{IV}_{c,n}\|_{\mathcal{T}}\lesssim_{\mathrm{P}} R_{\delta,k_n}\sqrt{\ln(k_n/R_{\delta,k_n})}+\zeta_{k_n}r_{h,k_n}.$$

Letting $U_i\coloneqq Y_i-h^*(W_i)$, we may decompose $\mathrm{IV}_{c,n}(t)$ as

$$\begin{aligned}&\mathrm{IV}_{c,n}(t)\\&=\sqrt{n}\mathbb{E}_n\left[U_i\{\delta_{k_n}(t,W_i)-\delta(t,W_i)\}\right]-\sqrt{n}\mathbb{E}_n\left[\delta_{k_n}(t,W_i)\{h_{k_n}(W_i)-h^*(W_i)\}\right]\\&=:\mathrm{IV}_{d,n}(t)+\mathrm{IV}_{e,n}(t).\end{aligned}$$

By T it therefore suffices to show that

$$\|\mathrm{IV}_{d,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} R_{\delta,k_n}\sqrt{\ln\left(k_n/R_{\delta,k_n}\right)} \quad \text{and} \quad \|\mathrm{IV}_{e,n}\|_{T} \lesssim_{\mathrm{P}} \zeta_{k_n} r_{h,k_n}.$$

For the purpose of bounding $\|\mathrm{IV}_{d,n}\|_{\mathcal{T}}$, consider the function class $\mathcal{F}_k \coloneqq \mathcal{F}_k(\mathcal{T}) \coloneqq \{f : z \mapsto \{y - h^*(w)\}\{\delta_k(t,w) - \delta^*(t,w)\}; t \in \mathcal{T}\}$. Note that $\mathrm{E}[f(Z)] = 0$ for any $f \in \mathcal{F}_k$, so we may view the stochastic process $\{\mathrm{IV}_{d,n}(t); t \in \mathcal{T}\}$ as an empirical process $\{\mathbb{G}_n(f) \,|\, f \in \mathcal{F}_k\}$. For any $t_1, t_2 \in \mathcal{T}$, by J we have

$$\begin{aligned}|\delta(t,w) - \delta(t,w)| &= |\mathrm{E}[\{\omega(t_1,X) - \omega(t_2,X)\}(\partial/\partial h)\rho(Z,h^*(W))]| \\ &\lesssim \mathrm{E}[|(\partial/\partial h)\rho(Z,h^*(W))|\,|W=w]\,\|t_1 - t_2\|.\end{aligned}$$

Consequently, using Assumption 3 and the fact that conditional expectations are $L^2(P)$ contractions,

$$\begin{aligned}\mathrm{E}[\{\delta(t_1,W) - \delta(t_2,W)\}^2] &\lesssim \mathrm{E}[\{\mathrm{E}[|(\partial/\partial h)\rho(Z,h^*(W))|\,|W=w]\}^2]\,\|t_1 - t_2\|^2 \\ &\leqslant \mathrm{E}\Big[(\partial/\partial h)\rho(Z,h^*(W))^2\Big]\,\|t_1 - t_2\|^2 \lesssim \|t_1 - t_2\|^2.\end{aligned}$$

Given that mean-square projections are also $L^2(P)$ contractions,

$$\begin{aligned}&\|Q_k^{-1/2}\mathrm{E}\left[p^k(W)\{\delta(t_1,W) - \delta(t_2,W)\}\right]\|^2 \\ &= \mathrm{E}\Big[\left(p^k(W)^\top Q_k^{-1}\mathrm{E}\left[p^k(W)\{\delta(t_1,W) - \delta(t_2,W)\}\right]\right)^2\Big] \\ &\leqslant \mathrm{E}[\{\delta(t_1,W) - \delta(t_2,W)\}^2]\end{aligned}$$

so by CS and the previous two displays,

$$\begin{aligned}|\delta_k(t_1,w) - \delta_k(t_2,w)| &= |p^k(w)^\top Q_k^{-1}\mathrm{E}\left[p^k(W)\{\delta(t_1,W) - \delta(t_2,W)\}\right]| \\ &\leqslant \|p^k(w)^\top Q_k^{-1/2}\|\|Q_k^{-1/2}\mathrm{E}\left[p^k(W)\{\delta(t_1,W) - \delta(t_2,W)\}\right]\| \\ &\lesssim \|p^k(w)^\top Q_k^{-1/2}\|\,\|t_1 - t_2\|. \qquad \text{(B.10)}\end{aligned}$$

Thus, for any $f_1 \coloneqq f(\cdot, t_1), f_2 \coloneqq f(\cdot, t_2) \in \mathcal{F}_k$, by T,

$$
\begin{aligned}
&|f_1(z) - f_2(z)| \\
&\leqslant |y - h^*(w)| \left(|\delta_k(t_1, w) - \delta_k(t_2, w)| + |\delta(t_1, w) - \delta(t_2, w)|\right) \\
&\leqslant C\,|y - h^*(w)| \left\{ \|p^k(w)^\top Q_k^{-1/2}\| + \mathrm{E}\left[|(\partial/\partial h)\,\rho(Z, h^*(W))|\,|W = w\right] \right\} \|t_1 - t_2\| \\
&=: F_{1k}(z)\,\|t_1 - t_2\|.
\end{aligned}
$$

Moreover, for any $f \coloneqq f(\cdot, t) \in \mathcal{F}_k$,

$$
|f(z)| = |y - h^*(w)|\,|\delta_k(t, w) - \delta(t, w)| \leqslant |y - h^*(w)|\,\|\delta_k(\cdot, w) - \delta(\cdot, w)\|_{\mathcal{T}} =: F_{2k}(z).
$$

Using Assumptions 3 and 4, the inequality $(a+b)^2 \leqslant 2a^2 + 2b^2$, and the fact that conditional expectations are $L^2(P)$ contractions, we see that

$$
\begin{aligned}
\mathrm{E}[F_{1k}(Z)^2] &\lesssim \mathrm{E}\left[U^2 \left\{ \|p^k(W)^\top Q_k^{-1/2}\| + \mathrm{E}\left[|(\partial/\partial h)\,\rho(Z, h^*(W))|\,|W\right] \right\}^2 \right] \\
&\lesssim \mathrm{E}[\|p^k(W)^\top Q_k^{-1/2}\|^2] + \mathrm{E}\left(\left\{\mathrm{E}\left[|(\partial/\partial h)\,\rho(Z, h^*(W))|\,|W\right]\right\}^2\right) \\
&\leqslant k + \mathrm{E}\left[(\partial/\partial h)\,\rho(Z, h^*(W))^2\right] \lesssim k \quad \text{as } k \to \infty.
\end{aligned}
$$

Given Assumptions 4 and 7, we get

$$
\mathrm{E}[F_{2k}(Z)^2] = \mathrm{E}\left[U^2 \|\delta_k(\cdot, W) - \delta(\cdot, W)\|_{\mathcal{T}}^2\right] \lesssim \mathrm{E}\left[\|\delta_k(\cdot, W) - \delta(\cdot, W)\|_{\mathcal{T}}^2\right] = R_{\delta,k}^2 \to 0
$$

as $k \to \infty$. Thus, defining $F_k \coloneqq F_{1,k} + F_{2,k}$ we must have

$$
\mathrm{E}[F_k(Z)^2] \lesssim k + R_{\delta,k}^2 \lesssim k \quad \text{as } k \to \infty,
$$

and it follows that F_k is a square-integrable envelope for $\mathcal{F}_k$ satisfying

$$
|f_1(z) - f_2(z)| \leqslant F_k(z)\,\|t_1 - t_2\| \quad \text{and} \quad \|F_k\|_{P,2} \lesssim k^{1/2} \quad \text{as } k \to \infty.
$$

Using $\mathcal{T}$ compact and the previous display, van der Vaart and Wellner (1996, Theorem 2.7.11) implies that

$$
N_{[\,]}(\varepsilon \|F_k\|_{P,2}, \mathcal{F}_k, L^2(P)) \leqslant (C/\varepsilon)^d, \quad \varepsilon \in (0, 1],
$$

and thus

$$J_{[]}\left(\delta, \mathcal{F}_k, L^2\left(P\right)\right) \leqslant \int_0^\delta \sqrt{1 + d \ln\left(C/\varepsilon\right)} \mathrm{d}\varepsilon, \quad \delta \in (0, 1].$$

where the right-hand side does not depend on k. In particular, $J_{[]}\left(1, \mathcal{F}_{k_n}, L^2\left(P\right)\right) \lesssim 1$

Defining

$$\sigma_n^2 \coloneqq \sup_{f \in \mathcal{F}_{k_n}} \mathbb{E}_n\left[f\left(Z_i\right)^2\right]$$

we see that

$$\sigma_n^2 = \sup_{t \in \mathcal{T}} \mathbb{E}_n\left[U_i^2\left\{\delta_{k_n}\left(t, W_i\right) - \delta\left(t, W_i\right)\right\}^2\right] \leqslant \mathbb{E}_n\left[U_i^2 \|\delta_{k_n}\left(\cdot, W_i\right) - \delta\left(\cdot, W_i\right)\|_{\mathcal{T}}^2\right]$$

such that

$$\mathrm{E}\left[\sigma_n^2\right] \leqslant \mathrm{E}\left[U^2 \|\delta_{k_n}\left(\cdot, W\right) - \delta\left(\cdot, W\right)\|_{\mathcal{T}}^2\right] \lesssim \mathrm{E}\left[\|\delta_{k_n}\left(\cdot, W\right) - \delta\left(\cdot, W\right)\|_{\mathcal{T}}^2\right] = R_{\delta, k_n}^2.$$

There are two cases: (1) $R_{\delta,k_n}/\|F_{k_n}\|_{P,2} \to 0$ and (2) $R_{\delta,k_n}/\|F_{k_n}\|_{P,2} \nrightarrow 0$.

Case 1: $R_{\delta,k_n}/\|F_{k_n}\|_{P,2} \to 0$. Given that $\sqrt{\mathrm{E}\left[\sigma_n^2\right]} \leqslant C_1 R_{\delta,k_n}$, by the change of variables $\varepsilon' \coloneqq \varepsilon/C_1$ we have

$$\begin{aligned} J_{[]}\left(\sqrt{\mathrm{E}\left[\sigma_n^2\right]}/\|F_{k_n}\|_{P,2}, \mathcal{F}_{k_n}, L^2\left(P\right)\right) &\leqslant J_{[]}\left(C_1 R_{\delta,k_n}/\|F_{k_n}\|_{P,2}, \mathcal{F}_{k_n}, L^2\left(P\right)\right) \\ &= C_1 \int_0^{R_{\delta,k_n}/\|F_{k_n}\|_{P,2}} \sqrt{1 + d_t \ln\left(C_3/\varepsilon'\right)} \mathrm{d}\varepsilon' \\ &\eqqcolon C_1 \overline{J}_{[]}\left(\Delta_{k_n}/\|F_{k_n}\|_{P,2}\right) \end{aligned} \tag{B.11}$$

van der Vaart and Wellner (2011, p. 196) establishes the maximal inequality

$$\mathrm{E}\left[\|\mathbb{G}_n\|_{\mathcal{F}_{k_n}}\right] \lesssim J_{[]}\left(\sqrt{\mathrm{E}\left[\sigma_n^2\right]}/\|F_{k_n}\|_{P,2}, \mathcal{F}_{k_n}, L^2\left(P\right)\right) \|F_{k_n}\|_{P,2}.$$

The previous two displays show that

$$\mathrm{E}\left[\|\mathbb{G}_n\|_{\mathcal{F}_{k_n}}\right]) \lesssim \overline{J}_{[]}\left(\Delta_{k_n}/\|F_{k_n}\|_{P,2}\right) \|F_{k_n}\|_{P,2}$$

and from van der Vaart and Wellner (1996, p. 239) we know that an integral of the

form $\int_0^\delta [1+\ln(1/u)]^{1/2}\mathrm{d}u$—as in (B.11)—satisfies $\int_0^\delta [1+\ln(1/u)]^{1/2}\mathrm{d}u \lesssim \delta\sqrt{\ln(1/\delta)}$ as $\delta \downarrow 0$. Since $R_{\delta,k_n}/\|F_{\delta,k_n}\|_{P,2} \to 0$ holds by hypothesis, the previous display combined with $\|F_{k_n}\|_{P,2} \lesssim \sqrt{k_n}$ and M yields

$$\begin{aligned}\|\mathbb{G}_n\|_{\mathcal{F}_{k_n}} &\lesssim_{\mathrm{P}} (R_{\delta,k_n}/\|F_{k_n}\|_{\mathrm{P},2})\sqrt{\ln(\|F_{k_n}\|_{P,2}/R_{\delta,k_n})}\|F_{k_n}\|_{P,2} \\ &= R_{\delta,k_n}\sqrt{\ln(\|F_{k_n}\|_{P,2}/R_{\delta,k_n})} \lesssim R_{\delta,k_n}\sqrt{\ln(k_n/R_{\delta,k_n})}.\end{aligned}$$

Case 2. $R_{\delta,k_n}/\|F_{k_n}\|_{P,2} \nrightarrow 0$. Given that $R_{\delta,k_n} \to 0$ (Assumption 7), we must have $\|F_{k_n}\|_{P,2} \lesssim R_{\delta,k}$. van der Vaart and Wellner (1996, Theorem 2.14.2) and $J_{[\,]}(1,\mathcal{F}_{k_n},L^2(P)) \lesssim 1$ yield

$$\mathrm{E}\left[\|\mathbb{G}_n\|_{\mathcal{F}_{k_n}}\right] \lesssim J_{[\,]}\left(1,\mathcal{F}_{k_n},L^2(P)\right)\|F_{k_n}\|_{P,2} \lesssim \|F_{k_n}\|_{P,2} \lesssim R_{\delta,k_n} \lesssim R_{\delta,k_n}\sqrt{\ln\left(\frac{k_n}{R_{\delta,k_n}}\right)}.$$

M now yields the same rate as in Case 1. In either case, we observe that $\|\mathrm{IV}_{d,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} R_{\delta,k_n}\sqrt{\ln(k_n/R_{\delta,k_n})}$.

For the purpose of bounding $\|\mathrm{IV}_{e,n}\|_{\mathcal{T}}$, consider the function class $\mathcal{F}_k \coloneqq \{f : z \mapsto \delta_k(t,w)\{h_k(w)-h^*(w)\}; t \in \mathcal{T}\}$. Note that, by orthogonality of mean–square projections we have $\mathrm{E}[f(Z)] = 0$ for any $f \in \mathcal{F}_k$, so we may view the stochastic process $\{\mathrm{IV}_{e,n}(t); t \in \mathcal{T}\}$ as an empirical process $\{\mathbb{G}_n(f); f \in \mathcal{F}_{k_n}\}$. For any $t_1, t_2 \in \mathcal{T}$, using the bound in (B.10) we have that $f_1 \coloneqq f(\cdot;t_1), f_2 \coloneqq f(\cdot;t_2) \in \mathcal{F}_k$, satisfy

$$\begin{aligned}|f_1(z)-f_2(z)| &= |\delta_k(t_1,w)-\delta_k(t_2,w)|\,|h_k(w)-h^*(w)| \\ &\lesssim \|p^k(w)^\top Q_k^{-1/2}\|\,|h_k(w)-h^*(w)|\,\|t_1-t_2\| \\ &\lesssim \zeta_k\,|h_k(w)-h^*(w)|\,\|t_1-t_2\|.\end{aligned}$$

The previous display implies

$$|f_1(z)-f_2(z)| \leqslant F_{1,k}(z)\,\|t_1-t_2\|,$$

for $F_{1k}(z) \coloneqq C_1\zeta_k|h_k(w)-h^*(w)|$ and some $C_1 \in (0,\infty)$. Since conditional expec-

tations are $L^2(P)$ contractions, by Assumptions 2 and 3,

$$\begin{aligned}\mathrm{E}[\delta(t,W)^2] &= \mathrm{E}\Big(\Big\{\mathrm{E}\Big[\omega(t,X)\Big|\frac{\partial}{\partial h}\rho(Z,h^*(W))\Big|W\Big]\Big\}^2\Big)\\ &\leqslant \mathrm{E}\left[\omega(t,X)^2\frac{\partial}{\partial h}\rho(Z,h^*(W))^2\right] \lesssim \mathrm{E}\Big[\frac{\partial}{\partial h}\rho(Z,h^*(W))^2\Big] < \infty,\end{aligned}$$

thus implying $\sup_{t\in\mathcal{T}}\mathrm{E}[\delta(t,W)^2] < \infty$. By CS and using that mean–square projections are $L^2(P)$ contractions as well, we get

$$\begin{aligned}|\delta_k(t,w)| &= |p^k(w)^\top Q_k^{-1}\mathrm{E}[p^k(W)\delta(t,W)]|\\ &\leqslant \|p^k(w)^\top Q_k^{-1/2}\|\|Q_k^{-1/2}\mathrm{E}[p^k(W)\delta(t,W)]\|\\ &\lesssim \|p^k(w)\|\mathrm{E}[\delta(t,W)^2] \lesssim \zeta_k,\end{aligned}$$

which implies that for any $f \coloneqq f(\cdot;t) \in \mathcal{F}_k$,

$$|f(z)| = |\delta_k(t,w)|\,|h_k(w) - h^*(w)| \lesssim \zeta_k\,|h_k(w) - h^*(w)|.$$

The previous diplay shows that $|f(z)| \leqslant F_{2k}(z)$ for $F_{2k}(z) \coloneqq C_2\zeta_k\,|h_k(w) - h^*(w)|$ and some $C_2 \in (0,\infty)$. Let $C_3 \coloneqq C_1 \vee C_2$, and define $F_k(z) \coloneqq C_3\zeta_k\,|h_k(w) - h^*(w)|$. Then by Assumption 7,

$$\|F_k\|_{P,2} = C_3\zeta_k\|h_k - h^*\|_{P,2} = C_3\zeta_k r_{h,k} \to 0 \text{ as } k\to\infty,$$

In particular, $\|F_k\|_{P,2} \lesssim 1$. Now, F_k is a square-integrable envelope for $\mathcal{F}_k$ satisfying

$$|f_1(z) - f_2(z)| \leqslant F_k(z)\,\|t_1 - t_2\|.$$

Using $\mathcal{T}$ compact and the previous display, by van der Vaart and Wellner (1996, Theorem 2.7.11) we see that

$$N_{[\,]}(\varepsilon\|F_k\|_{P,2}, \mathcal{F}_k, L^2(\mathrm{P})) \leqslant (C/\varepsilon)^{d_t}, \quad \varepsilon\in(0,1],$$

and thus

$$J_{[\,]}\left(\delta, \mathcal{F}_k, L^2(P)\right) \leqslant \int_0^\delta \sqrt{1 + d\ln(C/\varepsilon)}\mathrm{d}\varepsilon, \quad \delta\in(0,1],$$

where the right-hand side does not depend on k. In particular, $J_{[\,]}\left(1, \mathcal{F}_k, L^2\left(P\right)\right) \lesssim 1$. Using van der Vaart and Wellner (1996, Theorem 2.14.2) $J_{[\,]}\left(1, \mathcal{F}_{k_n}, L^2\left(P\right)\right) \lesssim 1$, we arrive at

$$\mathrm{E}\left[\|\mathbb{G}_n\|_{\mathcal{F}_{k_n}}\right] \lesssim J_{[\,]}\left(1, \mathcal{F}_{k_n}, L^2\left(P\right)\right) \|F_{k_n}\|_{P,2} \lesssim \|F_{k_n}\|_{P,2} \lesssim \zeta_{k_n} r_{h,k_n},$$

so $\|\mathrm{IV}_{e,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} \zeta_{k_n} r_{h,k_n}$ by M. □

B.2 Supporting Lemmas

For now, let Q and $\widehat{Q}$ be symmetric but otherwise arbitrary random matrices of possibly growing dimension. Also, denote the smallest and largest eigenvalue of a matrix A by $\lambda_{\min}\left(A\right)$ and $\lambda_{\max}\left(A\right)$, respectively.

Lemma B1. *If* $\lambda_{\min}\left(Q\right) \geqslant c$ wp $\to 1$ *for some constant* $c \in (0, \infty)$ *and* $\|\widehat{Q} - Q\|_{\mathrm{op}} \to_{\mathrm{P}} 0$*, then* $\widehat{Q}$ *is invertible* wp $\to 1$ *and* $\lambda_{\min}(\widehat{Q})^{-1} \lesssim_{\mathrm{P}} 1$.[26]

Proof. Given that the eigenvalues of a symmetric (hence square) matrix A are bounded in absolute value by the operator norm, for conformable vectors v,

$$\lambda_{\min}(\widehat{Q}) = \min_{\|v\|=1} \left\{v^\top \widehat{Q} v\right\} \geqslant \lambda_{\min}\left(A\right) - \lambda_{\max}(\widehat{Q} - Q) \geqslant \lambda_{\min}\left(A\right) - \|\widehat{Q} - Q\|_{\mathrm{op}}.$$

It follows that

$$\begin{aligned} \mathrm{P}\left(\lambda_{\min}(\widehat{Q}) < c/2\right) &\leqslant \mathrm{P}\left(\lambda_{\min}\left(Q\right) - \|\widehat{Q} - Q\|_{\mathrm{op}} < c/2\right) \\ &\leqslant \mathrm{P}\left(\|\widehat{Q} - Q\|_{\mathrm{op}} \geqslant c/2\right) + \mathrm{P}\left(\lambda_{\min}\left(Q\right) < c\right) \to 0, \end{aligned}$$

so $\mathrm{P}(\lambda_{\min}(\widehat{Q}) \geqslant c/2) \to 1$. Hence, $\widehat{Q}$ is invertible wp $\to 1$. Given that $\lambda_{\min}(\widehat{Q} - Q) \geqslant c/2$ wp $\to 1$, for any $C > 2/c$, we have

$$\overline{\lim_{n\to\infty}}\, \mathrm{P}\left(\lambda_{\min}(\widehat{Q})^{-1} > C\right) \leqslant \overline{\lim_{n\to\infty}}\, \mathrm{P}\left(\lambda_{\min}(\widehat{Q})^{-1} < c/2\right) = 0.$$

In particular, $\lim_{C\to\infty} \overline{\lim}_{n\to\infty} \mathrm{P}(\lambda_{\min}(\widehat{Q})^{-1} > C) = 0$. □

[26] This is Newey (1995, Lemma A.4) except that I state convergence in terms of the (weaker) operator matrix norm instead of the (stronger) Frobenius norm.

For now, let $Y, H \in \mathbf{R}^n$, $P \in \mathbf{R}^{n\times k}$ be arbitrary and of possibly growing dimensions n and k and abbreviate $U \coloneqq Y - H$, $\widehat{\pi} \coloneqq \left(P^\top P\right)^- P^\top Y$ and $\widehat{H} \coloneqq P\widehat{\pi}$.

Lemma B2. *For any* $\pi \in \mathbf{R}^k$,

$$\begin{aligned}\|\widehat{H} - H\|^2 &\leqslant U^\top P\left(P^\top P\right)^- P^\top U + \|P\pi - H\|^2,\\ \|\widehat{H} - P\pi\|^2 &\leqslant 2U^\top P\left(P^\top P\right)^- P^\top U + 2\|P\pi - H\|^2.\end{aligned}$$

Proof. Generalized inversion preserves symmetry, so $\mathcal{P}_P \coloneqq P\left(P^\top P\right)^- P^\top$ and $\mathcal{M}_P \coloneqq I - \mathcal{P}_P$ are symmetric idempotent. Given that also $\mathcal{P}_A P = P$ [see, e.g., Rao (1973, 1b.5(vi)(a))], for any fixed $\pi \in \mathbf{R}^k$, we must have

$$\begin{aligned}\|\widehat{H} - H\|^2 = \|\mathcal{P}_P Y - H\|^2 &= \|\mathcal{P}_P U - \mathcal{M}_P H\|^2 = U^\top \mathcal{P}_P U + H^\top \mathcal{M}_P H\\ &= U^\top \mathcal{P}_P U + (H - A\pi)^\top \mathcal{M}_P (H - A\pi) \leqslant U^\top \mathcal{P}_P U + \|P\pi - H\|^2,\end{aligned}$$

where the inequality follows from an idempotent matrix having only zero or one eigenvalues. Similarly, abbreviating $H_\pi \coloneqq P\pi$,

$$\begin{aligned}\|\widehat{H} - P\pi\|^2 = \|\mathcal{P}_P Y - H_\pi\|^2 &= \|\mathcal{P}_P(U + H - H_\pi)\|^2\\ &= (U + H - H_\pi)^\top \mathcal{P}_P (U + H - H_\pi)\\ &\leqslant 2U^\top \mathcal{P}_P U + 2(H - H_\pi)^\top \mathcal{P}_P (H - H_\pi)\\ &\leqslant 2U^\top \mathcal{P}_P U + 2\|H_\pi - H\|^2,\end{aligned}$$

where the first inequality follows from $(v+w)^\top A(v+w) \leqslant 2v^\top Av + 2w^\top Aw$ for A p.s.d. □

Next, interpret $\{(Y_i, W_i)\}_1^n$ as i.i.d. $\mathbf{R}^{1+d}$-valued random variables with $d \in \mathbf{N}$ (fixed), $\mathrm{E}[Y^2] < \infty$, $\mathcal{W} \coloneqq \mathrm{supp}(W)$, and let $p^k : \mathbf{R}^d \to \mathbf{R}^k$ be a nonrandom vector function of possibly growing length satisfying $\zeta_k \coloneqq \sup_{w\in\mathcal{W}} \left\|p^k(w)\right\| < \infty$ for all $k \in \mathbf{N}$. Also, define $h(w) \coloneqq \mathrm{E}[Y|W = w]$, $\sigma^2(w) \coloneqq \mathrm{var}(Y|W = w), w \in \mathcal{W}$, and $U_i \coloneqq Y_i - h(W_i)$ and let $\mathbf{U}$ and P be the $n \times 1$ vector and $n \times k$ matrix of U_i's and $p^k(W_i)^\top$'s, respectively.

Lemma B3. $\mathrm{E}[\mathbf{U}^\top P\left(P^\top P\right)^- P^\top \mathbf{U}] \leqslant k\|\sigma^2\|_{\mathcal{W}}$.

Proof. By the i.i.d. assumption, the positive semidefinite (p.s.d.) matrix

$$\mathrm{E}\left[\mathbf{U}\mathbf{U}^{\top}\middle|\left\{W_i\right\}_1^n\right]=\operatorname{diag}\left\{\sigma^2\left(W_i\right)\right\}_1^n.$$

Given that $\mathcal{P}_P \coloneqq P\left(P^{\top}P\right)^{-}P^{\top}$ is also p.s.d., using $\operatorname{tr}(AB)\leqslant\lambda_{\max}(A)\operatorname{tr}(B)$ for A,B p.s.d., we get

$$\begin{aligned}\mathrm{E}\left[\mathbf{U}^{\top}P\left(P^{\top}P\right)^{-}P^{\top}\mathbf{U}\middle|\left\{W_i\right\}_1^n\right]&=\operatorname{tr}\left(\mathrm{E}\left[\mathbf{U}\mathbf{U}^{\top}\middle|\left\{W_i\right\}_1^n\right]\mathcal{P}_P\right)\\&\leqslant\max_{1\leqslant i\leqslant n}\sigma^2\left(W_i\right)\operatorname{tr}\left(\mathcal{P}_P\right)\leqslant\|\sigma^2\|_{\mathcal{W}}\operatorname{tr}\left(\mathcal{P}_P\right).\end{aligned}$$

The matrix $\left(P^{\top}P\right)^{-}P^{\top}P$ is idempotent so its trace $\operatorname{tr}\left(\mathcal{P}_P\right)=\operatorname{tr}((P^{\top}P)^{-}P^{\top}P)$ equals $\operatorname{rank}((P^{\top}P)^{-}P^{\top}P)=\operatorname{rank}(P^{\top}P)$ [see Rao (1973, 1b(ii)(a))]. Applying also the bound $\operatorname{rank}(P^{\top}P)=\operatorname{rank}(P)\leqslant n\wedge k\leqslant k$, the first claim now follows from the previous display by taking the expectation over the W_i's. □

Lemma B4. *If the eigenvalues of* $Q_k\coloneqq\mathrm{E}[p^k(W)\,p^k(W)^{\top}]$ *are bounded from above uniformly in* k*, then*

$$\mathrm{E}\left[\|P^{\top}P/n-Q_{k_n}\|_{op}\right]\lesssim\frac{\zeta_{k_n}^2\ln k_n}{n}+\sqrt{\frac{\zeta_{k_n}^2\ln k_n}{n}}.$$

Proof. The matrix $\widehat{Q}_k=\mathbb{E}_n[p^k(W_i)p^k(W_i)]$ is the average of the n independent p.s.d. $k\times k$-matrix valued random variables $p^k\left(W_i\right)p^k\left(W_i\right)^{\top}$ with the matrix Q_k as their common mean. Given that

$$\begin{aligned}\|p^k\left(W_i\right)p^k\left(W_i\right)^{\top}\|_{\mathrm{op}}&\leqslant\|p^k\left(W_i\right)p^k\left(W_i\right)^{\top}\|_F\\&=[\operatorname{tr}(p^k\left(W_i\right)p^k\left(W_i\right)^{\top}p^k\left(W_i\right)p^k\left(W_i\right)^{\top})]^{1/2}\\&=\|p^k\left(W_i\right)\|^2\leqslant\zeta_k^2,\end{aligned}$$

these n random matrices are bounded in operator norm by ζ_k^2. By hypothesis, $\|Q_k\|_{\mathrm{op}}=[\lambda_{\max}\left(Q_k^{\top}Q_k\right)]^{1/2}=\lambda_{\max}\left(Q_k\right)\lesssim 1$.The claim now follows from Belloni, Chernozhukov, Chetverikov, and Kato (2015, Lemma 6.2), which builds on a fundamental result obtained by Rudelson (1999). □

Lemma B5. *Let* σ^2 *be bounded on* $\mathcal{W}$*, the eigenvalues of* $Q_k\coloneqq\mathrm{E}[p^k(W)\,p^k(W)^{\top}]$ *bounded from above and below uniformly in* k*, let* $\widetilde{\pi}\in\mathbf{R}^k$ *satisfy* $\left\|p^k\widetilde{\pi}-h\right\|_{\mathcal{W}}\lesssim k^{-\alpha}$

for some $\alpha \in \mathbf{R}_{++}$, *and define* $\widehat{h} \coloneqq p^{k\top}\widehat{\pi}$ *and* $\widetilde{h} \coloneqq p^{k\top}\widetilde{\pi}$. *Then, provided* $k_n/n \to 0$ *and* $\zeta_{k_n}^2 \ln(k_n)/n \to 0$, *we have*

1. $\|\widehat{h} - h\|_{n,2} \lesssim_{\mathrm{P}} \sqrt{k_n/n} + k_n^{-\alpha}$
2. $\|\widehat{h} - \widetilde{h}\|_{n,2} \lesssim_{\mathrm{P}} \sqrt{k_n/n} + k_n^{-\alpha}$
3. $\|\widehat{\pi} - \widetilde{\pi}\| \lesssim_{\mathrm{P}} \sqrt{k_n/n} + k_n^{-\alpha}$
4. $\|\widehat{h} - h\|_{\mathcal{W}} \lesssim_{\mathrm{P}} \zeta_{k_n}(\sqrt{k_n/n} + k_n^{-\alpha})$

Proof. By Lemma B2,

$$\|\widehat{h} - h\|_{n,2}^2 \leqslant U^\top P \left(P^\top P\right)^- P^\top U/n + \|\widetilde{h} - h\|_{n,2}^2,$$
$$\|\widehat{h} - \widetilde{h}\|_{n,2}^2 \leqslant 2U^\top P \left(P^\top P\right)^- P^\top U/n + 2\|\widetilde{h} - h\|_{n,2}^2.$$

By hypothesis $\|\widetilde{h} - h\|_{n,2} \leqslant \|\widetilde{h} - h\|_{\mathcal{W}} \lesssim k^{-\alpha}$. Moreover, via M, Lemma B3 and $\|\sigma^2\|_{\mathcal{W}} < \infty$ imply $U^\top P \left(P^\top P\right)^- P^\top U \lesssim_{\mathrm{P}} k_n$. The first two claims now follow from the previous display.

Via M, given that $\lambda_{\max}(Q_k) \lesssim 1$, Lemma B2 and $\zeta_{k_n}^2 \ln(k_n)/n \to 0$ imply $\|P^\top P/n - Q_{k_n}\|_{\mathrm{op}} \to_{\mathrm{P}} 0$. Given that also $\lambda_{\min}(Q_k)^{-1} \lesssim 1$, Lemma B2 implies that then $P^\top P/n$ is invertible wp $\to 1$ and $\lambda_{\min}(P^\top P/n)^{-1} \lesssim_{\mathrm{P}} 1$. Hence

$$\begin{aligned}\|\widehat{\pi} - \widetilde{\pi}\|^2 &\leqslant \lambda_{\min}(P^\top P/n)^{-1}\|P(\widehat{\pi} - \widetilde{\pi})\|^2/n \\ &= \lambda_{\min}(P^\top P/n)^{-1}\|\widehat{h} - \widetilde{h}\|_{\mathbb{P}_n,2}^2 \\ &\lesssim_{\mathrm{P}} \|\widehat{h} - \widetilde{h}\|_{\mathbb{P}_n,2}^2.\end{aligned}$$

The third claim now follows from the second. Given that $\|\widehat{h} - \widetilde{h}\|_{\mathcal{W}} = \sup_{w\in\mathcal{W}}|p^k(w)^\top(\widehat{\pi} - \widetilde{\pi})| \leqslant \zeta_k\|\widehat{\pi} - \widetilde{\pi}\|$ and $\|\widetilde{h} - h\|_{\mathcal{W}} \lesssim k^{-\alpha}$, by T and the third claim,

$$\|\widehat{h} - h\|_{\mathcal{W}} \lesssim_{\mathrm{P}} \zeta_{k_n}\|\widehat{\pi} - \widetilde{\pi}\| + k_n^{-\alpha} \lesssim_{\mathrm{P}} \zeta_{k_n}(\sqrt{k_n/n} + k_n^{-\alpha}).$$

□

Lemma B6. *Let* X_n *and* Y_n *be sequences of stochastic processes defined on a common probability space* $(\Omega, \mathcal{F}, P)$ *and taking values in a separable metric space* $(\mathbb{D}, d)$, *and let* $\mathcal{F}_n$ *be a sequence of sub-*σ*-algebras. If* $X_n \rightsquigarrow_{\mathrm{P},\mathcal{F}} X$ *in* $\mathbb{D}$ *and* $d(X_n, Y_n) \to_{\mathrm{P}} 0$, *then* $Y_n \rightsquigarrow_{\mathrm{P},\mathcal{F}} X$ *in* $\mathbb{D}$.

Proof. By T,

$$\begin{aligned}
&\sup_{h\in \mathrm{BL}_1(\mathbb{D})} \left|\mathrm{E}\left[h\left(Y_n\right)\middle|\mathcal{F}_n\right] - \mathrm{E}\left[h\left(X\right)\right]\right| \\
&\leqslant \sup_{h\in \mathrm{BL}_1(\mathbb{D})} \left|\mathrm{E}\left[h\left(Y_n\right) - h\left(X_n\right)\middle|\mathcal{F}_n\right]\right| + \sup_{h\in \mathrm{BL}_1(\mathbb{D})} \left|\mathrm{E}\left[h\left(X_n\right)\middle|\mathcal{F}_n\right] - \mathrm{E}\left[h\left(X\right)\right]\right| \\
&\leqslant d\left(X_n, Y_n\right) \wedge 2 + o_{\mathrm{P}}\left(1\right) = o_{\mathrm{P}}\left(1\right).
\end{aligned}$$

□

C Online Supplement (Not Intended For Publication)

This supplement contains the details of the proof of the bootstrap equivalence claimed in Lemma 3 (which was omitted from the online appendix intended for publication due to space constraints and its similarity with the proof of Lemma A2; see Section B.1) and additional supporting lemmas.

C.1 Omitted Proofs for Section 3.2

Define the stochastic processes $\widehat{G}^u$ and G_n^{*u} by

$$\widehat{G}^u(t) \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \widehat{g}(t, Z_i), \quad G_n^{*u}(t) \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i g(t, Z_i).$$

which are the "*u*ncentered" versions of $\widehat{G}$ and G_n^*, respectively, i.e., the displayed processes are not centered at the sample mean. The following lemma shows that the uncentered processes are asymptotically equivalent.

Lemma C1. *If Assumptions 1–8 hold, then* $\max_{1 \leqslant \ell \leqslant L} \|\widehat{G}_\ell^u - G_{\ell n}^{*u}\|_{\mathcal{X}_\ell} \to_{\mathrm{P}} 0$.

Proof of Lemma C1. The proof proceeds in a number of steps parallelling the proof of Lemma A2. It suffices to establish the claimed convergence for given ℓ. I therefore drop the ℓ subscripts throughout and refer to the (ℓth) index set $\mathcal{X}_\ell$ as $\mathcal{T}$ itself.

Step 0 (Main)

For fixed $t \in \mathcal{T}$ a decomposition yields

$$\begin{aligned}
\widehat{G}^u(t) - G_n^{*u}(t) &= \sqrt{n}\mathbb{E}_n\left[\xi_i\{\widehat{g}(t, Z_i) - g(t, Z_i)\}\right] \\
&= \sqrt{n}\mathbb{E}_n\left[\xi_i\omega(t, X_i)\left\{\rho(Z_i, \widehat{\beta}, \widehat{h}(W_i)) - \rho(Z_i, \beta_0, h^*(W_i))\right\}\right] \\
&\quad - \left[\widehat{b}(t) - b(t)\right]^\top \sqrt{n}\mathbb{E}_n\left[\xi_i s(Z_i)\right] \\
&\quad - \widehat{b}(t)^\top \sqrt{n}\mathbb{E}_n\left[\xi_i\{\widehat{s}(Z_i) - s(Z_i)\}\right] \\
&\quad + \sqrt{n}\mathbb{E}_n\left[\xi_i(\widehat{\delta}(t, W_i)^\top\{Y_i - \widehat{h}(W_i)\} - \delta(t, W_i)^\top U_i)\right], \\
&=: \mathrm{I}_n(t) + \mathrm{II}_n(t) + \mathrm{III}_n(t) + \mathrm{IV}_n(t). \qquad \text{(C.1)}
\end{aligned}$$

where $U_i = Y_i - h^*(W_i)$. The following steps show that the four remainder terms $\to_{\mathrm{P}} 0$ uniformly over $\mathcal{T}$. The claim therefore follows from T.

Step 1: $\|\mathrm{I}_n\|_{\mathcal{T}} \to_{\mathrm{P}} 0$

Assumption 1 and M implies that $\|\widehat{\beta} - \beta_0\| \lesssim_{\mathrm{P}} n^{-1/2} \to 0$. Let $\mathcal{N}_0$ be any open neighborhood of β_0 (Assumption 3). Then $\widehat{\beta} \in \mathcal{N}_0$ wp $\to 1$. To simplify notation and ensure that objects are globally well defined, in what follows I will—without loss of generality—assume that $\widehat{\beta} \in \mathcal{N}_0$ with probability one for all n. A MVE of $\beta \mapsto \rho(Z_i, \beta, \widehat{h}(W_i))$ at $\widehat{\beta}$ around β_0 and CS show that

$$\begin{aligned}
\|\mathrm{I}_n\|_{\mathcal{T}} &\leqslant \sup_{t\in\mathcal{T}}\left|\sqrt{n}\mathbb{E}_n\left[\xi_i\omega(t, X_i)\left\{\rho(Z_i, \beta_0, \widehat{h}(W_i)) - \rho(Z_i, \beta_0, h^*(W_i))\right\}\right]\right| \\
&\quad + \sqrt{n}\|\widehat{\beta} - \beta_0\| \sup_{t\in\mathcal{T}}\left\|\mathbb{E}_n\left[\xi_i\omega(t, X_i)(\partial/\partial\beta)(Z_i, \overline{\beta}, \widehat{h}(W_i))\right]\right\| \\
&=: \|\mathrm{I}_{a,n}\|_{\mathcal{T}} + \sqrt{n}\|\widehat{\beta} - \beta_0\|\|\mathrm{I}_{b,n}\|_{\mathcal{T}}, \qquad \text{(C.2)}
\end{aligned}$$

where $\overline{\beta}$ satisfies $\|\overline{\beta} - \beta_0\| \leqslant \|\widehat{\beta} - \beta_0\|$ such that $\overline{\beta} \in \mathcal{N}_0$ for n sufficiently large. Since $\sqrt{n}\|\widehat{\beta} - \beta_0\| \lesssim_{\mathrm{P}} 1$ it suffices to show that $\|\mathrm{I}_{a,n}\|_{\mathcal{T}}$ and $\|\mathrm{I}_{b,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$.

Step 1a: $\|\mathrm{I}_{a,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$**.** Abbreviate $(z, v) \mapsto \rho(z, \beta_0, v)$ by ρ. By a MVE of $s \mapsto \rho(Z_i, s)$ at $\widehat{h}(W_i)$ around $h^*(W_i)$ and T we may be bound $\|\mathrm{I}_{a,n}\|_{\mathcal{T}}$ defined in

(C.2) by

$$\begin{aligned}
&\sup_{t\in\mathcal{T}}\left|\sqrt{n}\mathbb{E}_n\left[\xi_i\omega\left(t,X_i\right)\{\frac{\partial}{\partial h^\top}\rho(Z_i,\overline{h}\left(W_i\right))-\frac{\partial}{\partial h^\top}\rho(Z_i,h^*\left(W_i\right))\}\{\widehat{h}\left(W_i\right)-h^*\left(W_i\right)\}\right]\right| \\
&\quad+\sup_{t\in\mathcal{T}}\left|\sqrt{n}\mathbb{E}_n\left[\xi_i\omega\left(t,X_i\right)\frac{\partial}{\partial h^\top}\rho(Z_i,h^*\left(W_i\right))\{\widehat{h}\left(W_i\right)-h^*\left(W_i\right)\}\right]\right| \\
&=:\|\mathrm{I}_{a,1,n}\|_{\mathcal{T}}+\|\mathrm{I}_{a,2,n}\|_{\mathcal{T}}. \qquad \text{(C.3)}
\end{aligned}$$

Step 1a(1): $\|\mathrm{I}_{a,1,n}\|_{\mathcal{T}}\to_{\mathrm{P}}0$. By T and Assumptions 2 and 8

$$\begin{aligned}
\|\mathrm{I}_{a,1,n}\|_{\mathcal{T}}&\lesssim\sqrt{n}\mathbb{E}_n\left[|\xi_i|R'\left(Z_i\right)\|\widehat{h}\left(W_i\right)-h^*\left(W_i\right)\|^2\right] \\
&\leqslant d\mathbb{E}_n\left[|\xi_i|\,R'\left(Z_i\right)\right]\sqrt{n}\max_{1\leqslant m\leqslant d}\|\widehat{h}_m-h_m^*\|_{\mathcal{W}}^2 \\
&\lesssim_{\mathrm{P}}\mathrm{E}[R'(Z)]\sqrt{n}\max_{1\leqslant m\leqslant d}\|\widehat{h}_m-h_m^*\|_{\mathcal{W}}^2\to_{\mathrm{P}}0,
\end{aligned}$$

with R' from and the $\to_{\mathrm{P}}0$ guaranteed by Assumption 8.

Step 1a(2): $\|\mathrm{I}_{a,2,n}\|_{\mathcal{T}}\to_{\mathrm{P}}0$. Given that

$$\mathrm{I}_{a,2,n}\left(t\right)=\sum_{m=1}^{d}\sqrt{n}\mathbb{E}_n\left[\xi_i\omega\left(t,X_i\right)\frac{\partial}{\partial h_m}\rho(Z_i,h^*\left(W_i\right))\{\widehat{h}_m\left(W_i\right)-h_m^*\left(W_i\right)\}\right] \qquad \text{(C.4)}$$

by T, it suffices to bound each summand uniformly over $\mathcal{T}$ in probability. I therefore omit the m subscript for the remainder of this paragraph and interpret $(\partial/\partial h)\,\rho$ as a scalar. (I still evaluate it at values for the vector h^*, though.) Let $\widetilde{h}\coloneqq p^{k\top}\widetilde{\pi}$ for $\widetilde{\pi}$ provided by Assumption 6. Then we may bound such a summand uniformly over $\mathcal{T}$ by

$$\begin{aligned}
&\sup_{t\in\mathcal{T}}\left|\sqrt{n}\mathbb{E}_n\left[\xi_i\omega\left(t,X_i\right)(\partial/\partial h)\,\rho(Z_i,h^*\left(W_i\right))\{\widehat{h}\left(W_i\right)-\widetilde{h}\left(W_i\right)\}\right]\right| \\
&\quad+\sup_{t\in\mathcal{T}}\left|\sqrt{n}\mathbb{E}_n\left[\xi_i\omega\left(t,X_i\right)(\partial/\partial h)\,\rho(Z_i,h^*\left(W_i\right))\{\widetilde{h}\left(W_i\right)-h^*\left(W_i\right)\}\right]\right| \\
&=:\|\mathrm{I}_{a,2,1,n}\|_{\mathcal{T}}+\|\mathrm{I}_{a,2,2,n}\|_{\mathcal{T}}.
\end{aligned}$$

I consider $\|\mathrm{I}_{a,2,1,n}\|_{\mathcal{T}}$ and $\|\mathrm{I}_{a,2,2,n}\|_{\mathcal{T}}$ in turn. By CS $\|\mathrm{I}_{a,2,1,n}\|_{\mathcal{T}}$ is bounded by

$$\begin{aligned}\|\mathrm{I}_{a,2,1,n}\|_{T} &\leqslant \|\widehat{\pi}-\widetilde{\pi}\| \sup_{t\in\mathcal{T}} \left\|\sqrt{n}\mathbb{E}_n\left[\xi_i\omega\left(t,X_i\right)\left(\partial/\partial h\right)\rho(Z_i,h^*\left(W_i\right))p^{k_n}\left(W_i\right)\right]\right\| \\ &\leqslant \|\widehat{\pi}-\widetilde{\pi}\|\Big(\sum_{j=1}^{k_n}\sup_{t\in\mathcal{T}}\left\{\sqrt{n}\mathbb{E}_n\left[\xi_i\omega\left(t,X_i\right)\left(\partial/\partial h\right)\rho(Z_i,h^*\left(W_i\right))p_j\left(W_i\right)\right]\right\}^2\Big)^{1/2}.\end{aligned}$$

Fix k and let

$$\mathcal{F}_j' \coloneqq \left\{f:(v,z)\mapsto v\omega\left(t,x\right)\left(\partial/\partial h\right)\rho\left(z,h^*\left(w\right)\right)p_j\left(w\right);t\in\mathcal{T}\right\}.$$

Note $\mathrm{E}[f(\xi,Z)]=0$ for every $f\in\mathcal{F}_j'$, so $\left\{\sqrt{n}\mathbb{E}_n\left[f\left(\xi_i,Z_i\right)\right];f\in\mathcal{F}_j'\right\}$ is an empirical process. For $f\coloneqq f\left(\cdot,t\right),f_1\coloneqq f\left(\cdot;t_1\right),f_2\coloneqq f\left(\cdot;t_2\right)\in\mathcal{F}_j'$ arbitrary, by Assumption 2 we have

$$\begin{aligned}\left|f\left(v,z\right)\right| &\leqslant C_1\left|v\right|\left|\left(\partial/\partial h\right)\rho\left(z,h^*\left(w\right)\right)\right|\left\|p_j\right\|_{\mathcal{W}} \\ \left|f_1\left(v,z\right)-f_2\left(v,z\right)\right| &\leqslant C_2\left|v\right|\left|\left(\partial/\partial h\right)\rho\left(z,h^*\left(w\right)\right)\right|\left\|p_j\right\|_{\mathcal{W}}\|t_1-t_2\|.\end{aligned}$$

It follows from Assumption 3 and the previous display that

$$F_j'\left(v,z\right)\coloneqq\left(C_1\vee C_2\right)\left|v\right|\left|\left(\partial/\partial h\right)\rho\left(z,h^*\left(w\right)\right)\right|\left\|p_j\right\|_{\mathcal{W}}$$

is an envelope for $\mathcal{F}_j'$ satisfying $\mathrm{E}[F_j'(\xi,Z)^2]\propto\|p_j\|_{\mathcal{W}}^2$, which is finite for every j by Assumption 7. Moreover, by compactness of $\mathcal{T}$ (Assumption 2) and the previous display,

$$N_{[\,]}(\varepsilon(\mathrm{E}[F_j'(\xi,Z)^2])^{1/2},\mathcal{F}_j',L^2\left(\xi,Z\right))\leqslant N\left(\varepsilon,\mathcal{T},\|\cdot\|\right)\lesssim\varepsilon^{-d_x},\quad\varepsilon\in(0,1].$$

It follows that the bracketing entropy integral $J_{[\,]}(1,\mathcal{F}_j',L^2\left(\xi,Z\right))$ is bounded by a constant independent of j, so by van der Vaart and Wellner (1996, Theorem 2.14.2)

$$\mathrm{E}[\|\mathbb{G}_n\|_{\mathcal{F}_j'}]\lesssim J_{[\,]}(1,\mathcal{F}_j',L^2\left(\xi,Z\right))\mathrm{E}[F_j'(\xi,Z)^2])^{1/2}\lesssim(\mathrm{E}[F_j'(\xi,Z)^2])^{1/2}\propto\|p_j\|_{\mathcal{W}}.$$

van der Vaart and Wellner (1996, Theorem 2.14.5) and the previous display show that

$$(\mathrm{E}[\|\mathbb{G}_n\|_{\mathcal{F}_j'}^2])^{1/2}\lesssim\mathrm{E}[\|\mathbb{G}_n\|_{\mathcal{F}_j'}]+(\mathrm{E}[F_j'(\xi,Z)^2])^{1/2}\lesssim\|p_j\|_{\mathcal{W}},$$

Allowing $k = k_n$, the previous display, in turn, implies

$$\mathrm{E}\Big[\sum_{j=1}^{k_n}\|\mathbb{G}_n\|_{\mathcal{F}_j'}^2\Big] = \sum_{j=1}^{k_n}\mathrm{E}[\|\mathbb{G}_n\|_{\mathcal{F}_j'}^2] \lesssim \sum_{j=1}^{k_n}\|p_j\|_{\mathcal{W}}^2,$$

so by M we get

$$\sum_{j=1}^{k_n}\|\mathbb{G}_n\|_{\mathcal{F}_j'}^2 \lesssim_{\mathrm{P}} \sum_{j=1}^{k_n}\|p_j\|_{\mathcal{W}}^2.$$

From Lemma B5, M and Assumption 7 it now follows that

$$\|\mathrm{I}_{a,2,1,n}\|_{\mathcal{T}} \leqslant \|\widehat{\pi} - \widetilde{\pi}\|\Big(\sum_{j=1}^{k_n}\|\mathbb{G}_n\|_{\mathcal{F}_j'}^2\Big)^{1/2} \lesssim_{\mathrm{P}} (\sqrt{k_n/n} + k_n^{-\alpha})\Big(\sum_{j=1}^{k_n}\|p_j\|_{\mathcal{W}}^2\Big)^{1/2} \to 0.$$

Similarly, fix k and let

$$\mathcal{F}_k' \coloneqq \{f : (v,z) \mapsto v\omega\,(t,x)\,(\partial/\partial h)\,\rho\,(z, h^*\,(w))\,\{\widetilde{h}\,(w) - h^*\,(w)\}; t \in \mathcal{T}\}.$$

Note $\mathrm{E}[f(\xi, Z)] = 0$ for every $f \in \mathcal{F}_k'$, so $\{\sqrt{n}\mathbb{E}_n\,[f\,(\xi_i, Z_i)]\,; f \in \mathcal{F}_k'\}$ is an empirical process. For $f \coloneqq f\,(\cdot; t)\,, f_1 \coloneqq f\,(\cdot; t_1)\,, f_2 \coloneqq f\,(\cdot; t_2) \in \mathcal{F}_{jk}'$ arbitrary, by Assumption 2 we have

$$\begin{aligned} |f\,(v,z)| &\leqslant C_1|v||(\partial/\partial h)\,\rho\,(z, h^*\,(w))|\|\widetilde{h} - h^*\|_{\mathcal{W}}, \\ |f_1\,(v,z) - f_2\,(v,z)| &\leqslant C_2|v||(\partial/\partial h)\,\rho\,(z, h^*\,(w))|\|\widetilde{h} - h^*\|_{\mathcal{W}}\|t_1 - t_2\|. \end{aligned}$$

By Assumption 8, CS and the previous display we see that

$$F_k'\,(v,z) \coloneqq (C_1 \vee C_2)\,|v|\,|(\partial/\partial h)\,\rho\,(z, h^*\,(w))|\,\|\widetilde{h} - h^*\|_{\mathcal{W}}$$

is an envelope for $\mathcal{F}_k'$ satisfying $\mathrm{E}[F_k'(\xi, Z)^2] \propto \|\widetilde{h} - h^*\|_{\mathcal{W}}^2$, which by Assumption 6 is finite for every k. Moreover, by compactness of $\mathcal{T}$ (Assumption 2) and the previous display,

$$N_{[\,]}(\varepsilon(\mathrm{E}[F_k'(\xi, Z)^2])^{1/2}, \mathcal{F}_k', L^2\,(\xi, Z)) \leqslant N\,(\varepsilon, \mathcal{T}, \|\cdot\|) \lesssim \varepsilon^{-d_x}, \quad \varepsilon \in (0, 1].$$

which implies that the bracketing entropy integral $J_{[]}(1, \mathcal{F}'_k, L^2(\xi, Z))$ is bounded by a constant independent of k. Using van der Vaart and Wellner (1996, Theorem 2.14.2) and Assumption 6, we therefore get

$$\begin{aligned}\mathrm{E}[\|\mathbb{G}_n\|_{\mathcal{F}'_k}] &\lesssim J_{[]}(1, \mathcal{F}'_k, L^2(\xi, Z))\mathrm{E}[F'_k(\xi, Z)^2])^{1/2} \\ &\lesssim \mathrm{E}[F'_k(\xi, Z)^2])^{1/2} \propto \|\widetilde{h} - h^*\|_{\mathcal{W}} \lesssim k^{-\alpha}.\end{aligned}$$

By M it follows that $\|\mathrm{I}_{a,2,2,n}\|_{\mathcal{T}} = \|\mathbb{G}_n\|_{\mathcal{F}'_{k_n}} \lesssim_{\mathrm{P}} k_n^{-\alpha} \to 0$, which completes the proof that each summand in (C.4) $\to_{\mathrm{P}} 0$, which via (C.3), in turn, shows that $\|\mathrm{I}_{a,n}\|_{\mathcal{T}}$ defined in (C.2) $\to_{\mathrm{P}} 0$.

Step 1b: $\|\mathrm{I}_{b,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$**.** By T we may bound $\|\mathrm{I}_{b,n}\|_{\mathcal{T}}$ defined in (C.2) by

$$\begin{aligned}&\sup_{t\in\mathcal{T}} \left\|\mathbb{E}_n\left[\xi_i \omega(t, X_i)(\partial/\partial\beta)\rho(Z_i, \overline{\beta}, h^*(W_i))\right]\right\| \\ &\quad + \sup_{t\in\mathcal{T}} \left\|\mathbb{E}_n\left[\xi_i \omega(t, X_i)\{(\partial/\partial\beta)\rho(Z_i, \overline{\beta}, \widehat{h}(W_i)) - (\partial/\partial\beta)\rho(Z_i, \overline{\beta}, h^*(W_i))\}\right]\right\| \\ &=: \|\mathrm{I}_{b,1,n}\|_{\mathcal{T}} + \|\mathrm{I}_{b,2,n}\|_{\mathcal{T}}. \qquad \text{(C.5)}\end{aligned}$$

The second term $\|\mathrm{I}_{b,2,n}\|_{\mathcal{T}}$ satisfies

$$\begin{aligned}\|\mathrm{I}_{b,2,n}\|_{\mathcal{T}} &\lesssim \mathbb{E}_n\left[|\xi_i| a(Z_i) \|\widehat{h}(W_i) - h^*(W_i)\|^c\right] \\ &\leqslant \mathbb{E}_n[|\xi_i| a(Z_i)] \max_{1\leqslant m'\leqslant d} \|\widehat{h}_{m'} - h^*_{m'}\|^c_{\mathcal{W}} \lesssim_{\mathrm{P}} \max_{1\leqslant m'\leqslant d} \|\widehat{h}_{m'} - h^*_{m'}\|^c_{\mathcal{W}} \xrightarrow{\mathrm{P}} 0,\end{aligned}$$

where the $\lesssim$ follows from Assumptions 2 and 3, the $\lesssim_{\mathrm{P}}$ from the ξ_i's being i.i.d., zero mean, unit variance (hence having finite first moment) and independent of the data, and the $\to_{\mathrm{P}} 0$ stems from Lemma B5 and Assumption 7.

To show that $\|\mathrm{I}_{b,1,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$, observe that the $\{(\xi_i, Z_i)\}_1^n$ are i.i.d., the map $(t, \beta) \mapsto \xi\omega(t, X)(\partial/\partial\beta)\rho(Z, \beta, h^*(W))$ is continuous on $\mathcal{T} \times \mathcal{N}_0$ (Assumptions 2 and 3) and therefore continuous on the product $\mathcal{T} \times \overline{B}$, where $\overline{B} \subset \mathcal{N}_0$ is any closed ball with center β_0 and sufficiently small radius (Assumption 1). Moreover, $\mathcal{T} \times \overline{B}$ is compact (Assumption 2), and $\sup_{\mathcal{T}\times\overline{B}} \|\xi\omega(t, X)(\partial/\partial\beta)\rho(Z, \beta, h^*(W))\| \lesssim |\xi| \sup_{\overline{B}} \|(\partial/\partial\beta)\rho(Z, \beta, h^*(W))\|$, where by independence, CS, and Assumption 3,

$$\mathrm{E}\left[|\xi| \sup_{\beta\in\overline{B}} \|(\partial/\partial\beta)\rho(Z, \beta, h^*(W))\|\right] \leqslant \mathrm{E}\left[\sup_{\beta\in\overline{B}} \|(\partial/\partial\beta)\rho(Z, \beta, h^*(W))\|\right] < \infty,$$

Given that the ξ_i's are centered and independent of the data, Newey and McFadden (1994, Lemma 2.4) shows that

$$\sup_{\mathcal{T}\times\overline{B}} \|\mathbb{E}_n\left[\xi_i \omega\left(t, X_i\right)\left(\partial/\partial\beta\right)\rho\left(Z_i, \beta, h^*\left(W_i\right)\right)\right]\| \xrightarrow{\mathrm{P}} 0.$$

That $\|\mathrm{I}_{b,1,n}\|_{\mathcal{T}}$ defined in (C.5) $\to_{\mathrm{P}} 0$ now follows from $\overline{\beta} \in \overline{B}$ wp $\to 1$ and the previous display. Via (C.2) and (C.5), this $\to_{\mathrm{P}} 0$ in turn implies $\|\mathrm{I}_n\|_{\mathcal{T}} \to_{\mathrm{P}} 0$.

Step 2: $\|\mathrm{II}_n\|_{\mathcal{T}} \to_{\mathrm{P}} 0$**.**

By CS, II_n defined in (C.1) satisfies

$$\|\mathrm{II}_n\|_{\mathcal{T}} \leqslant \left\|\sqrt{n}\mathbb{E}_n\left[\xi_i s\left(Z_i\right)\right]\right\| \sup_{t\in\mathcal{T}} \|\widehat{b}\left(t\right) - b\left(t\right)\|,$$

To show $\|\mathrm{II}_n\|_{\mathcal{T}} \to_{\mathrm{P}} 0$, it therefore suffices to show $\|\sqrt{n}\mathbb{E}_n\left[\xi_i s\left(Z_i\right)\right]\| \lesssim_{\mathrm{P}} 1$ and $\sup_{t\in\mathcal{T}} \|\widehat{b}\left(t\right) - b\left(t\right)\| \to_{\mathrm{P}} 0$.

Step 2a: $\|\sqrt{n}\mathbb{E}_n\left[\xi_i s\left(Z_i\right)\right]\| \lesssim_{\mathrm{P}} 1$**.** Given that the ξ_i's are i.i.d., zero-mean, unit variance and independent of the data we have

$$\mathrm{E}\left[\left\|\sqrt{n}\mathbb{E}_n\left[\xi_i s\left(Z_i\right)\right]\right\|^2 \middle| \left\{Z_i\right\}_1^n\right] = \mathbb{E}_n\left[\left\|s\left(Z_i\right)\right\|^2\right].$$

The desired $\|\sqrt{n}\mathbb{E}_n\left[\xi_i s\left(Z_i\right)\right]\| \lesssim_{\mathrm{P}} 1$ now follows from iterated expectations, square integrability of $s\left(Z\right)$ (Assumption 1) and M.

Step 2b: Behavior of $\widehat{b}$**.** In this step I show that

$$\text{(a) } \sup_{t\in\mathcal{T}} \|\widehat{b}\left(t\right) - b\left(t\right)\| \xrightarrow{\mathrm{P}} 0 \quad \text{and} \quad \text{(b) } \sup_{t\in\mathcal{T}} \|\widehat{b}\left(t\right)\| \lesssim_{\mathrm{P}} 1,$$

with b and $\widehat{b}$ defined in (3.8) and (3.21), respectively. To show (a), note that the argument used in Step 1 of the proof of Lemma A2 shows that

$$\begin{aligned}&(t,\beta) \mapsto \mathrm{E}\left[\omega\left(t, X\right)\left(\partial/\partial\beta\right)\rho\left(Z, \beta, h^*\left(W\right)\right)\right] \text{ is uniformly continuous on } \mathcal{T}\times\overline{B},\\ &\text{and } \sup_{\mathcal{T}\times\overline{B}} \|\left(\mathbb{E}_n - \mathrm{E}\right)\omega\left(t, X_i\right)\left(\partial/\partial\beta\right)\rho\left(Z_i, \beta, h^*\left(W_i\right)\right)\| \xrightarrow{\mathrm{P}} 0,\end{aligned}$$

where $\overline{B} \subset \mathcal{N}_0$ is any closed set containing β_0 in its interior (Assumption 1). By T we have

$$\begin{aligned}\sup_{t\in\mathcal{T}}\|\widehat{b}(t)-b(t)\| \leqslant & \sup_{t\in\mathcal{T}}\left\|\mathbb{E}_n\left[\omega(t,X_i)\{\frac{\partial}{\partial\beta}\rho(Z_i,\widehat{\beta},\widehat{h}(W_i))-\frac{\partial}{\partial\beta}\rho(Z_i,\widehat{\beta},h^*(W_i))\}\right]\right\| \\ & + \sup_{t\in\mathcal{T}}\left\|(\mathbb{E}_n-\mathrm{E}_Z)\left[\omega(t,X_i)\frac{\partial}{\partial\beta}\rho(Z_i,\widehat{\beta},h^*(W_i))\right]\right\| \\ & + \sup_{t\in\mathcal{T}}\left\|\mathrm{E}_Z\left[\omega(t,X)\frac{\partial}{\partial\beta}\rho(Z,\widehat{\beta},h^*(W))\right]-b(t)\right\|.\end{aligned}$$

Given that $\widehat{\beta} \in \overline{B}$ wp $\to 1$, the second and third term on the right $\to_{\mathrm{P}} 0$ due to uniform convergence and uniform continuity, respectively. By T and Assumptions 2 and 3, the first term is bounded by a constant multiple of

$$\begin{aligned}\mathbb{E}_n[a(Z_i)\|\widehat{h}(Z_i)-h^*(Z_i)\|^c] &\leqslant d^{c/2}\mathbb{E}_n[a(Z_i)]\max_{1\leqslant m\leqslant d}\|\widehat{h}_m-h_m^*\|_{\mathcal{W}}^c \\ &\lesssim_{\mathrm{P}} \max_{1\leqslant m\leqslant d}\|\widehat{h}_m-h_m^*\|_{\mathcal{W}}^c \xrightarrow{\mathrm{P}} 0,\end{aligned}$$

where the $\lesssim_{\mathrm{P}}$ follows from M and the $\to_{\mathrm{P}} 0$ from Lemma B5. The previous display finishes the proof of (a) and therefore the proof of Step 2 ($\|\mathrm{II}_n\|_{\mathcal{T}} \to_{\mathrm{P}} 0$).

To show (b), note that the argument in used in Step 1 of the proof of Lemma A2 also shows that $\sup_{t\in\mathcal{T}}\|b(t)\| \lesssim 1$. Two applications of T yield

$$\left|\sup_{t\in\mathcal{T}}\|\widehat{b}(t)\|-\sup_{t\in\mathcal{T}}\|b(t)\|\right| \leqslant \sup_{t\in\mathcal{T}}|\|\widehat{b}(t)\|-\|b(t)\|| \leqslant \sup_{t\in\mathcal{T}}\|\widehat{b}(t)-b(t)\| \xrightarrow{\mathrm{P}} 0,$$

which combined with $\sup_{t\in\mathcal{T}}\|b(t)\| \lesssim 1$ implies $\sup_{t\in\mathcal{T}}\|\widehat{b}(t)\| \lesssim_{\mathrm{P}} 1$.

Step 3: $\|\mathrm{III}_n\|_{\mathcal{T}} \to_{\mathrm{P}} 0$

By CS,

$$\|\mathrm{III}_n\|_{\mathcal{T}} \leqslant \left\|\sqrt{n}\mathbb{E}_n[\xi_i\{\widehat{s}(Z_i)-s(Z_i)\}]\right\| \sup_{t\in\mathcal{T}}\|\widehat{b}(t)\|. \tag{C.6}$$

By Step 2b, $\sup_{t\in\mathcal{T}}\|\widehat{b}(t)\| \lesssim_{\mathrm{P}} 1$, so to show $\|\mathrm{III}_n\|_{\mathcal{T}} \to_{\mathrm{P}} 0$, it suffices to show that

$$\|\sqrt{n}\mathbb{E}_n[\xi_i\{\widehat{s}(Z_i)-s(Z_i)\}]\| \xrightarrow{\mathrm{P}} 0.$$

To this end, note that by the ξ_i's being i.i.d., zero-mean, unit variance and independent of the data, and $\widehat{s}$ being $\{Z_i\}_1^n$-measurable (Assumption 8), we have

$$\mathrm{E}\left[\left\|\sqrt{n}\mathbb{E}_n\left[\xi_i\{\widehat{s}\left(Z_i\right)-s\left(Z_i\right)\}\right]\right\|\middle|\left\{Z_i\right\}_1^n\right]=\mathbb{E}_n\left[\left\|\widehat{s}\left(Z_i\right)-s\left(Z_i\right)\right\|^2\right]=\left\|\widehat{s}-s\right\|_{n,2}^2.$$

By Assumption 8, the right-hand side $\to_{\mathrm{P}} 0$, so $\left\|\sqrt{n}\mathbb{E}_n\left[\xi_i\{\widehat{s}\left(Z_i\right)-s\left(Z_i\right)\}\right]\right\| \xrightarrow{\mathrm{P}} 0$ follows from Lemma C3. Via (C.6), this $\to_{\mathrm{P}} 0$ finishes the proof of the claim that $\|\mathrm{III}_n\|_{\mathcal{T}}$ as defined in (C.1) $\to_{\mathrm{P}} 0$.

Step 4: $\|\mathrm{IV}_n\|_{\mathcal{T}} \to_{\mathrm{P}} 0$

Given that IV_n defined in (C.1) may be written as the sum

$$\mathrm{IV}_n\left(t\right)=\sum_{m=1}^{d}\sqrt{n}\mathbb{E}_n\left[\xi_i\{\widehat{\delta}_m\left(t,W_i\right)\left\{Y_{mi}-\widehat{h}_m\left(W_i\right)\right\}-\delta_m\left(t,W_i\right)U_{mi}\}\right],$$

it suffices to bound each summand uniformly over $\mathcal{T}$ in probability. I therefore omit the m subscript for the remainder of Step 4 and interpret $\left(\partial/\partial h\right)\rho$ as a scalar derivative. By T, the (mth) summand satisfies the uniform bound

$$\begin{aligned}
&\sup_{t\in\mathcal{T}}\left|\sqrt{n}\mathbb{E}_n\big(\xi_i\{\widehat{\delta}\left(t,W_i\right)\left[Y_i-\widehat{h}\left(W_i\right)\right]-\delta\left(t,W_i\right)U_i\}\big)\right| \\
&\leqslant \sup_{t\in\mathcal{T}}\left|\sqrt{n}\mathbb{E}_n\left[\xi_i U_i\{\widehat{\delta}\left(t,W_i\right)-\delta\left(t,W_i\right)\}\right]\right| \\
&\quad+\sup_{t\in\mathcal{T}}\left|\sqrt{n}\mathbb{E}_n\left[\xi_i\widehat{\delta}\left(t,W_i\right)\left\{\widehat{h}\left(W_i\right)-h^*\left(W_i\right)\right\}\right]\right| =: \|\mathrm{IV}_{a,n}\|_{\mathcal{T}}+\|\mathrm{IV}_{b,n}\|_{\mathcal{T}}. \qquad \text{(C.7)}
\end{aligned}$$

I consider each term on the right-hand side in turn.

Step 4a: $\|\mathrm{IV}_{a,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$**.** Recalling the definitions of δ_k and ψ_k in (B.5) and (B.9), respectively, we may write $\delta_k(t,w) = p^k(w)^\top Q_k^{-1}\psi_k(t)$, such that by T,

$$\begin{aligned}
\|\mathrm{IV}_{a,n}\|_{\mathcal{T}} &= \sup_{t\in\mathcal{T}} \left| \sqrt{n}\mathbb{E}_n \left[\xi_i U_i p^{k_n}(W_i)^\top \widehat{Q}_{k_n}^{-1} \widehat{\psi}_{k_n}(t) \right] \right| \\
&\leqslant \sup_{t\in\mathcal{T}} \left| \widehat{\psi}_{k_n}(t)^\top (\widehat{Q}_{k_n}^{-1} - Q_{k_n}^{-1}) \sqrt{n}\mathbb{E}_n \left[p^{k_n}(W_i)\, \xi_i U_i \right] \right| \\
&\quad + \sup_{t\in\mathcal{T}} \left| \{\widehat{\psi}_{k_n}(t) - \psi_{k_n}(t)\}^\top Q_{k_n}^{-1} \sqrt{n}\mathbb{E}_n \left[p^{k_n}(W_i)\, \xi_i U_i \right] \right| \\
&\quad + \sup_{t\in\mathcal{T}} \left| \sqrt{n}\mathbb{E}_n \left[\xi_i U_i \{\delta_{k_n}(t,W_i) - \delta(t,W_i)\} \right] \right| \\
&=: \|\mathrm{IV}_{a,1,n}\|_{\mathcal{T}} + \|\mathrm{IV}_{a,2,n}\|_{\mathcal{T}} + \|\mathrm{IV}_{a,3,n}\|_{\mathcal{T}}
\end{aligned} \tag{C.8}$$

where I employ the convenient shorthand

$$\widehat{\psi}_k(t) = \mathbb{E}_n[p^k(W_i)\,\omega(t,X_i)\,(\partial/\partial\beta)\,\rho(Z_i,\widehat{\beta},\widehat{h}(W_i))]$$

as defined in the (auxilliary) Step 4c below.

Step 4a(1): $\|\mathrm{IV}_{a,1,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$**.** By the ξ_i's being i.i.d., zero-mean, unit variance and independent of the data,

$$\begin{aligned}
\mathrm{E}\left[\left\| Q_{k_n}^{-1/2} \sqrt{n}\mathbb{E}_n \left[p^{k_n}(W_i)\, \xi_i U_i \right] \right\|^2 \right] &= \mathrm{E}[\xi^2 U^2 p^{k_n}(W)^\top Q_{k_n}^{-1} p^{k_n}(W)] \\
&= \mathrm{E}[U^2 p^{k_n}(W)^\top Q_{k_n}^{-1} p^{k_n}(W)] \\
&\lesssim \mathrm{E}[p^{k_n}(W)^\top Q_{k_n}^{-1} p^{k_n}(W)] = k_n,
\end{aligned}$$

so by M we have

$$\left\| Q_{k_n}^{-1/2} \sqrt{n}\mathbb{E}_n \left[p^{k_n}(W_i)\, \xi_i U_i \right] \right\| \lesssim_{\mathrm{P}} \sqrt{k_n}. \tag{C.9}$$

Step 4c shows that $\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1}\| \lesssim_{\mathrm{P}} 1$, so by CS, Assumption 5, Lemma B4, and the previous display,

$$\begin{aligned}
\|\mathrm{IV}_{a,1,n}\|_{\mathcal{T}} &= \sup_{t\in\mathcal{T}}\left|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1}(Q_{k_n}-\widehat{Q}_{k_n})Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}(W_i)\,\xi_i U_i\right]\right| \\
&\leqslant \left\|Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}(W_i)\,\xi_i U_i\right]\right\| \sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1}(Q_{k_n}-\widehat{Q}_{k_n})\| \\
&\leqslant \left\|Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}(W_i)\,\xi_i U_i\right]\right\| \|\widehat{Q}_{k_n}-Q_{k_n}\|_{\mathrm{op}} \sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1}\| \\
&\lesssim \left\|Q_{k_n}^{-1/2}\sqrt{n}\mathbb{E}_n\left[p^{k_n}(W_i)\,\xi_i U_i\right]\right\| \|\widehat{Q}_{k_n}-Q_{k_n}\|_{\mathrm{op}} \sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1}\| \\
&\lesssim_{\mathrm{P}} \sqrt{k_n}[\zeta_{k_n}^2 \ln(k_n)/n]^{1/2} = [\zeta_{k_n}^2 k_n \ln(k_n)/n]^{1/2} \to 0.
\end{aligned}$$

Step 4a(2): $\|\mathrm{IV}_{a,2,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$. By CS, Assumption 5, (C.9), Step 4c and Assumption 8,

$$\begin{aligned}
\|\mathrm{IV}_{a,2,n}\|_{\mathcal{T}} &\leqslant \left\|Q_{k_n}^{-1}\sqrt{n}\mathbb{E}_n\left[p^{k_n}(W_i)\,\xi_i U_i\right]\right\| \sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)-\psi_{k_n}(t)\| \\
&\lesssim \left\|Q_{k_n}^{-1/2}\sqrt{n}\mathbb{E}_n\left[p^{k_n}(W_i)\,\xi_i U_i\right]\right\| \sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)-\psi_{k_n}(t)\| \\
&\lesssim_{\mathrm{P}} \sqrt{k_n}\Big[\zeta_{k_n}(\sqrt{k_n/n}+k_n^{-\alpha}) + \Big(\sum_{j=1}^{k_n}\|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2}/\sqrt{n}\Big] \\
&= \zeta_{k_n}\sqrt{k_n}(\sqrt{k_n/n}+k_n^{-\alpha}) + \Big(\sum_{j=1}^{k_n}\|p_{jk_n}\|_{\mathcal{W}}^2\Big)^{1/2}\sqrt{k_n/n} \to 0.
\end{aligned}$$

Step 4a(3): $\|\mathrm{IV}_{a,3,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$. Fix k and let

$$\mathcal{F}_k' := \{(v,z)\mapsto v\{y-h^*(w)\}\{\delta_k(t,w)-\delta(t,w)\}; t\in\mathcal{T}\}.$$

Given that each $\mathrm{E}[f(\xi,Z)]=0$ for each $f\in\mathcal{F}_k'$, the stochastic process IV_n may be viewed as an empirical process $\mathbb{G}_n$ indexed by the changing classes $\mathcal{F}_{k_n}'$. For $f=f_t, f_1=f_{t_1}, f_2=f_{t_2}\in\mathcal{F}_{k_n}'$ arbitrary, by arguments parallel to those used in Step 4c in the proof of Lemma A2, there exists a function $z\mapsto F_k(z)$ such that

$$\begin{aligned}
|f(v,z)| &\leqslant |v|\,F_k(z), \\
|f_1(v,z)-f_2(v,z)| &\leqslant |v|\,F_k(z)\,\|t_1-t_2\|,
\end{aligned}$$

and $\|F_k\|_{P,2} \lesssim \sqrt{k}$. The ξ_i's being zero mean, unit variance and independent of the data implies that $F'_k : (v, z) \mapsto |v| F_k(z)$ is an envelope for $\mathcal{F}'_k$ with $(\mathrm{E}[F'_k(\xi, Z)^2])^{1/2} = \|F_k\|_{P,2} \lesssim \sqrt{k}$ as $k \to \infty$, satisfying

$$|f_1(s, z) - f_2(s, z)| \leqslant F'_k(s, z) \|t_1 - t_2\|.$$

Using $\mathcal{T}$ compact and the previous display, by van der Vaart and Wellner (1996, Theorem 2.7.11) we see that

$$N_{[\,]}(\varepsilon(\mathrm{E}[F'_k(\xi, Z)^2])^{1/2}, \mathcal{F}'_k, L^2(\xi, Z)) \leqslant (C/\varepsilon)^{d_x}, \quad \varepsilon \in (0, 1].$$

and thus

$$J_{[\,]}\left(\delta, \mathcal{F}'_k, L^2(\xi, Z)\right) \leqslant \int_0^\delta \sqrt{1 + d \ln(C/\varepsilon)} \mathrm{d}\varepsilon, \quad \delta \in (0, 1].$$

where the right-hand side does not depend on k. In particular, $J_{[\,]}\left(1, \mathcal{F}'_{k_n}, L^2(\xi, Z)\right) \lesssim 1$. Defining

$$\sigma_n^2 \coloneqq \sup_{f \in \mathcal{F}'_{k_n}} \mathbb{E}_n[f(\xi_i, Z_i)^2]$$

we see that

$$\sigma_n^2 = \sup_{t \in \mathcal{T}} \mathbb{E}_n\left[\xi_i^2 U_i^2 \left\{\delta_{k_n}(t, W_i) - \delta(t, W_i)\right\}^2\right] \leqslant \mathbb{E}_n\left[\xi_i^2 U_i^2 \|\delta_{k_n}(\cdot, W_i) - \delta(\cdot, W_i)\|_{\mathcal{T}}^2\right],$$

thus implying

$$\mathrm{E}\left[\sigma_n^2\right] \leqslant \mathrm{E}\left[\xi^2 U^2 \|\delta_{k_n}(\cdot, W) - \delta(\cdot, W)\|_{\mathcal{T}}^2\right] \leqslant C\mathrm{E}\left[\|\delta_{k_n}(\cdot, W) - \delta(\cdot, W)\|_{\mathcal{T}}^2\right] = C R_{\delta, k_n}^2,$$

where the $\lesssim$ follows from the ξ_i's being zero mean, unit variance, and independent of the data and Assumption 4, and the last equality follow from the definitions of δ_k and $R_{\delta,k}$.

It suffices to consider the two cases (1) $R_{\delta,k_n}/\|F_{k_n}\|_{P,2} \to 0$ and (2) $R_{\delta,k_n}/\|F_{k_n}\|_{P,2} \nrightarrow 0$ in turn. *Case 1*: $R_{\delta,k_n}/\|F_{k_n}\|_{P,2} \to 0$. Given that $\sqrt{\mathrm{E}\left[\sigma_n^2\right]} \leqslant C R_{\delta,k_n}$, by the change

of variables $\varepsilon' \coloneqq \varepsilon/C$ we have

$$\begin{aligned} J_{[\,]}\left(\sqrt{\mathrm{E}\left[\sigma_n^2\right]}/\|F_{k_n}\|_{P,2}, \mathcal{F}'_k, L^2\left(\xi, Z\right)\right) &\leqslant J_{[\,]}\left(CR_{\delta,k_n}/\|F_{k_n}\|_{P,2}, \mathcal{F}'_k, L^2\left(\xi, Z\right)\right) \\ &= C\int_0^{R_{\delta,k_n}/\|F_{k_n}\|_{P,2}} \sqrt{1 + d_t \ln\left(C'/\varepsilon'\right)}\mathrm{d}\varepsilon' \\ &=: C\overline{J}_{[\,]}\left(R_{\delta,k_n}/\|F_{k_n}\|_{P,2}\right) \end{aligned} \tag{C.10}$$

By van der Vaart and Wellner (2011, p. 196) we have the maximal inequality

$$\begin{aligned} \mathrm{E}[\|\mathbb{G}_n\|_{\mathcal{F}'_{k_n}}] &\lesssim J_{[\,]}\left(\sqrt{\mathrm{E}\left[\sigma_n^2\right]}/\|F_{k_n}\|_{P,2}, \mathcal{F}'_{k_n}, L^2\left(\xi, Z\right)\right)\|F_{k_n}\|_{P,2} \\ &\lesssim \overline{J}_{[\,]}\left(R_{\delta,k_n}/\|F_{k_n}\|_{P,2}\right)\|F_{k_n}\|_{P,2}, \end{aligned}$$

and from van der Vaart and Wellner (1996, p. 239) we know that an entropy integral (bound) of the form (C.10) satisfies $\overline{J}_{[\,]}(\delta) \lesssim \delta\sqrt{\ln(1/\delta)}$ as $\delta \downarrow 0$. Since $R_{\delta,k_n}/\|F_{k_n}\|_{P,2} \to 0$ holds by hypothesis, the previous display combined with $\|F_{k_n}\|_{P,2} \lesssim \sqrt{k_n}$ yields

$$\begin{aligned} \mathrm{E}[\|\mathbb{G}_n\|_{\mathcal{F}'_{k_n}}] &\lesssim \left(R_{\delta,k_n}/\|F_{k_n}\|_{P,2}\right)\sqrt{\ln\left(\frac{\|F_{k_n}\|_{P,2}}{R_{\delta,k_n}}\right)}\|F_{k_n}\|_{P,2} = R_{\delta,k_n}\sqrt{\ln\left(\frac{\|F_{k_n}\|_{P,2}}{R_{\delta,k_n}}\right)} \\ &\lesssim R_{\delta,k}\sqrt{\ln\left(k_n/R_{\delta,k_n}\right)}. \end{aligned}$$

Case 2. Suppose that $R_{\delta,k_n}/\|F_{k_n}\|_{P,2} \nrightarrow 0$. Given that $R_{\delta,k_n} \to 0$ (Assumption 7), we must have $\|F_{k_n}\|_{P,2} \lesssim R_k$. van der Vaart and Wellner (1996, Theorem 2.14.2) and $J_{[\,]}\left(1, \mathcal{F}'_{k_n}, L^2\left(\xi, Z\right)\right) \lesssim 1$ yield

$$\begin{aligned} \mathrm{E}[\|\mathbb{G}_n\|_{\mathcal{F}'_{k_n}}] &\lesssim J_{[\,]}\left(1, \mathcal{F}_{k'_n}, L^2\left(\xi, Z\right)\right)\|F_{k_n}\|_{P,2} \\ &\lesssim \|F_{k_n}\|_{P,2} \lesssim R_{\delta,k_n} \lesssim R_{\delta,k_n}\sqrt{\ln\left(k_n/R_{\delta,k_n}\right)} \end{aligned}$$

as in Case 1. The claim $\|\mathrm{IV}_{a,3,n}\|_{\mathcal{T}}$ as defined in (C.8) $\to_{\mathrm{P}} 0$ now follows from M and $R_{\delta,k_n}\sqrt{\ln\left(k_n/R_{\delta,k_n}\right)} \to 0$ (Assumption 7). Via (C.8), this $\to_{\mathrm{P}} 0$ in turn shows that $\|\mathrm{IV}_{a,n}\|_{\mathcal{T}}$ as defined in (C.7) $\to_{\mathrm{P}} 0$.

Step 4b: $\|\mathrm{IV}_{b,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$. Step 4c shows that $\sup_{\mathcal{T}} \|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1}\| \lesssim_{\mathrm{P}} 1$, so by CS it follows that

$$\begin{aligned}\|\mathrm{IV}_{b,n}\|_{\mathcal{T}} &= \sup_{t\in\mathcal{T}} \left| \widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1} \sqrt{n}\mathbb{E}_n \left[p^{k_n}(W_i)\, \xi_i \{\widehat{h}(W_i) - h^*(W_i)\} \right] \right| \\ &\leqslant \left\| \sqrt{n}\mathbb{E}_n \left[p^{k_n}(W_i)\, \xi_i \{\widehat{h}(W_i) - h^*(W_i)\} \right] \right\| \sup_{t\in\mathcal{T}} \|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1}\| \\ &\lesssim_{\mathrm{P}} \left\| \sqrt{n}\mathbb{E}_n \left[p^{k_n}(W_i)\, \xi_i \{\widehat{h}(W_i) - h^*(W_i)\} \right] \right\|.\end{aligned}$$

To show that the right-hand side $\to_{\mathrm{P}} 0$, note that by the ξ_i's being i.i.d., zero-mean, unit variance and independent of $\{Z_i\}_1^n$, and $\widehat{h}$ being $\{Z_i\}_1^n$-measurable,

$$\begin{aligned}&\mathrm{E}\left[\left\| \sqrt{n}\mathbb{E}_n \left[p^{k_n}(W_i)\, \xi_i [\widehat{h}(W_i) - h^*(W_i)] \right] \right\|^2 \middle| \{Z_i\}_1^n \right] \\ &= \mathbb{E}_n \left\{ \|p^{k_n}(W_i)\|^2 [\widehat{h}(W_i) - h^*(W_i)]^2 \right\} \leqslant \Big(\sum_{j=1}^{k_n} \|p_j\|_{\mathcal{W}}^2 \Big) \|\widehat{h} - h^*\|_{\mathbb{P}_n,2}^2 \\ &\lesssim_{\mathrm{P}} \left[\Big(\sum_{j=1}^{k_n} \|p_{jk_n}\|_{\mathcal{W}}^2 \Big)^{1/2} (\sqrt{k_n/n} + k_n^{-\alpha}) \right]^2 \to 0,\end{aligned}$$

where the $\lesssim_{\mathrm{P}}$ follows from Lemma B5 and the $\to 0$ from Assumption 7. It follows by conditional CS that $\mathrm{E}[\|\sqrt{n}\mathbb{E}_n\{p^{k_n}(W_i)\, \xi_i[\widehat{h}(W_i) - h_*(W_i)]\}\| | \{Z_i\}_1^n] \to_{\mathrm{P}} 0$, so $\|\sqrt{n}\mathbb{E}_n[p^{k_n}(W_i)\, \xi_i\{\widehat{h}(W_i) - h_*(W_i)\}]\| \to_{\mathrm{P}} 0$ by Lemma C3. This $\to_{\mathrm{P}} 0$ finishes the proof of the claim that $\|\mathrm{IV}_{b,n}\|_{\mathcal{T}}$ as defined in (C.7) $\to_{\mathrm{P}} 0$, which in turn shows that $\|\mathrm{IV}_n\|_{\mathcal{T}}$ as defined in (C.1) $\to_{\mathrm{P}} 0$.

Step 4c (auxilliary): Behavior of $\widehat{\psi}_{k_n}$ and $\widehat{Q}_{k_n}^-$. Motivated by the LOIE, I estimate $\psi_k(\cdot) = \mathrm{E}[p^k(W)\, \delta(\cdot, W)]$ as defined in (B.9) by

$$\widehat{\psi}_k(\cdot) := \mathbb{E}_n[p^k(W_i)\, \omega(\cdot, X_i)\, (\partial/\partial h)\, \rho(Z_i, \widehat{\beta}, \widehat{h}(W_i))]. \tag{C.11}$$

Note that this definition allows us to write $\widehat{\delta}$ defined in (3.22) as

$$(t, w) \mapsto \widehat{\delta}(t, w) = p^{k_n}(w)^\top \widehat{Q}_{k_n}^- \widehat{\psi}_{k_n}(t).$$

This section shows that

$$\begin{aligned}
&\text{(a) } \sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)-\psi_{k_n}(t)\| \\
&\quad\lesssim_{\mathrm{P}} \zeta_{k_n}\max_{1\leqslant m'\leqslant d}\left(\sqrt{\frac{k_{m',n}}{n}}+k_{m',n}^{-\alpha_{m'}}\right)+\frac{1}{\sqrt{n}}\Big(\sum_{j=1}^{k_n}\|p_j\|_{\mathcal{W}}^2\Big)^{1/2}\to 0, \\
&\text{(b) } \sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^{\top}\widehat{Q}_{k_n}^{-}-\psi_{k_n}(t)^{\top}Q_{k_n}^{-1}\|\xrightarrow{\mathrm{P}}0, \\
&\text{(c) } \sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^{\top}\widehat{Q}_{k_n}^{-}\|\lesssim_{\mathrm{P}}1.
\end{aligned}$$

To show (a), recall $\Delta(t,z,h)$ from (B.7) and define

$$\Delta_i^k(t)\coloneqq(\Delta(t,Z_i,p_1),\ldots,\Delta(t,Z_i,p_k))^{\top}.$$

Then by T we have

$$\begin{aligned}
&\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)-\psi_{k_n}(t)\| \\
&\leqslant \sup_{t\in\mathcal{T}}\|\mathbb{E}_n[\omega(t,X_i)\{(\partial/\partial h)\,\rho(Z_i,\widehat{\beta},\widehat{h}(W_i))-(\partial/\partial h)\,\rho(Z_i,\beta_0,h^*(W_i))\}p^{k_n}(W_i)]\| \\
&\quad+\sup_{t\in\mathcal{T}}\|(\mathbb{E}_n-\mathrm{E})\,\Delta_i^{k_n}(t)\|.
\end{aligned}$$

By Assumptions 1, 2 and 8 and T followed by CS

$$\begin{aligned}
&\sup_{t\in\mathcal{T}}\|\mathbb{E}_n[\omega(t,X_i)\{(\partial/\partial h)\,\rho(Z_i,\widehat{\beta},\widehat{h}(W_i))-(\partial/\partial h)\,\rho(Z_i,\beta_0,h^*(W_i))\}p^{k_n}(W_i)]\| \\
&\lesssim \mathbb{E}_n[\|p^{k_n}(W_i)\|R'(Z_i)\{\|\widehat{\beta}-\beta_0\|+\|\widehat{h}(W_i)-h^*(W_i)\|\}] \\
&\leqslant \zeta_{k_n}\left\{\mathbb{E}_n\left[R'(Z_i)^2\right]\right\}^{1/2}\left(\|\widehat{\beta}-\beta_0\|+\max_{1\leqslant m'\leqslant d}\|\widehat{h}_{m'}-h^*_{m'}\|_{n,2}\right) \\
&\lesssim_{\mathrm{P}} \zeta_{k_n}\left\{\mathbb{E}_n\left[R'(Z_i)^2\right]\right\}^{1/2}(n^{-1/2}+\max_{1\leqslant m'\leqslant d}(\sqrt{k_{m',n}/n}+k_{m',n}^{-\alpha_{m'}})) \\
&\lesssim \left\{\mathbb{E}_n\left[R'(Z_i)^2\right]\right\}^{1/2}\zeta_{k_n}\max_{1\leqslant m'\leqslant d}(\sqrt{k_{m',n}/n}+k_{m',n}^{-\alpha_{m'}})\to 0,
\end{aligned}$$

where the $\lesssim_{\mathrm{P}}$ follows from Lemma B5 and the $\to 0$ from Assumption 8.

Moreover, the argument used in Step 3a of the proof of Lemma A2 shows that

$$\sup_{t\in\mathcal{T}}\|\mathbb{E}_n\{\Delta_i^{k_n}(t)\}\| \lesssim_{\mathrm{P}} \Big(\sum_{j=1}^{k_n}\|p_j\|_{\mathcal{W}}^2\Big)^{1/2}/\sqrt{n}.$$

Lemmas B1 and B4 and Assumptions 5 and 7 show that $\widehat{Q}_{k_n}$ is invertible wp $\to 1$ and $\lambda_{\min}(\widehat{Q}_{k_n})^{-1} \lesssim_{\mathrm{P}} 1$. To ease notation I will (without loss of generality) assume that $\widehat{Q}_{k_n}^{-1}$ exists with probability one for all n, such that $\widehat{Q}_{k_n}^{-} = \widehat{Q}_{k_n}^{-1}$. The argument used in Step 4 of the proof of Lemma A2 shows that $\sup_{\mathcal{T}}\|\psi_{k_n}(t)^\top Q_{k_n}^{-1}\| \lesssim 1$, so by (a) and T,

$$\begin{aligned}
&\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^\top\widehat{Q}_{k_n}^{-1} - \psi_{k_n}(t)^\top Q_{k_n}^{-1}\| \\
&\leqslant \sup_{t\in\mathcal{T}}\|[\widehat{\psi}_{k_n}(t) - \psi_{k_n}(t)]^\top\widehat{Q}_{k_n}^{-1}\| + \sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top(\widehat{Q}_{k_n}^{-1} - Q_{k_n}^{-1})\| \\
&\leqslant \|\widehat{Q}_{k_n}^{-1}\|_{\mathrm{op}}\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t) - \psi_{k_n}(t)\| + \sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top Q_{k_n}^{-1}(\widehat{Q}_{k_n} - Q_{k_n})\widehat{Q}_{k_n}^{-1}\| \\
&\leqslant \|\widehat{Q}_{k_n}^{-1}\|_{\mathrm{op}}\Big(\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t) - \psi_{k_n}(t)\| + \|\widehat{Q}_{k_n} - Q_{k_n}\|_{\mathrm{op}}\sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top Q_{k_n}^{-1}\|\Big) \xrightarrow{\mathrm{P}} 0,
\end{aligned}$$

which shows (b). Part (c) follows from (b) and $\sup_{t\in\mathcal{T}}\|\psi_{k_n}(t)^\top Q_{k_n}^{-1}\| \lesssim 1$. This concludes the proof of the claim that $\|\mathrm{IV}_n\|_{\mathcal{T}}$ as defined in (C.1) $\to_{\mathrm{P}} 0$ and hence the proof of Lemma C1. □

Lemma C2. *If Assumptions 1–8 hold, then*

$$\max_{1\leqslant\ell\leqslant L}\|\mathbb{E}_n[\widehat{g}_\ell(\cdot, Z_i) - g_\ell(\cdot, Z_i)]\|_{\mathcal{X}_\ell} \xrightarrow{\mathrm{P}} 0.$$

Proof of Lemma C2. The proof proceeds in a number of steps. Since the lemma is stated for a given ℓ, for notational convenience I drop the ℓ subscripts throughout and refer to the (ℓth) index set $\mathcal{X}_\ell$ as $\mathcal{T}$ itself.

Step 0 (Main)

For fixed $t \in \mathcal{T}$ we may write

$$\begin{aligned}\mathbb{E}_n[\widehat{g}(t, Z_i) - g(t, Z_i)] &= \mathbb{E}_n\left[\omega(t, X_i)\left\{\rho(Z_i, \widehat{\beta}, \widehat{h}(W_i)) - \rho(Z_i, \beta_0, h^*(W_i))\right\}\right] \\ &\quad - [\widehat{b}(t) - b(t)^\top]^\top \mathbb{E}_n[s(Z_i)] - \widehat{b}(t)^\top \mathbb{E}_n[\widehat{s}(Z_i) - s(Z_i)] \\ &\quad + \mathbb{E}_n[\widehat{\delta}(t, W_i)^\top \{Y_i - \widehat{h}(W_i)\} - \delta(t, W_i)^\top U_i] \\ &=: \mathrm{I}_n(t) + \mathrm{II}_n(t) + \mathrm{III}_n(t) + \mathrm{IV}_n(t).\end{aligned}$$

The following steps show that the four remainder terms $\to_{\mathrm{P}} 0$ uniformly over $\mathcal{T}$. The claim therefore follows from T.

Step 1: $\|\mathrm{I}_n\|_{\mathcal{T}} \to_{\mathrm{P}} 0$

Assumption 1 implies that $\|\widehat{\beta} - \beta_0\| \lesssim_{\mathrm{P}} n^{-1/2} \to 0$. Letting $\mathcal{N}_0$ be any open neighborhood of β_0 (which exists by Assumption 3), we have $\widehat{\beta} \in \mathcal{N}_0$ wp $\to 1$. To simplify notation and ensure that objects are globally well defined, in what follows I will—without loss of generality—assume that $\widehat{\beta} \in \mathcal{N}_0$ with probability equal to one for all n. An MVE of $\beta \mapsto \rho(Z_i, \beta, \widehat{h}(W_i))$ at $\widehat{\beta}$ around β_0 followed by CS show that

$$\begin{aligned}\|\mathrm{I}_n\|_{\mathcal{T}} &\leqslant \sup_{t\in\mathcal{T}} |\mathbb{E}_n[\omega(t, X_i)\{\rho(Z_i, \beta_0, \widehat{h}(W_i)) - \rho(Z_i, \beta_0, h^*(W_i))\}]| \\ &\quad + \|\widehat{\beta} - \beta_0\| \sup_{t\in\mathcal{T}} \|\mathbb{E}_n[\omega(t, X_i)(\partial/\partial\beta)\rho(Z_i, \overline{\beta}, \widehat{h}(W_i))]\| \\ &=: \|\mathrm{I}_{a,n}\|_{\mathcal{T}} + \|\widehat{\beta} - \beta_0\| \|\mathrm{I}_{b,n}\|_{\mathcal{T}},\end{aligned}$$

where $\overline{\beta}$ satisfies $\|\overline{\beta} - \beta_0\| \leqslant \|\widehat{\beta} - \beta_0\|$ such that $\overline{\beta} \in \mathcal{N}_0$ for n sufficiently large. Since $\|\widehat{\beta} - \beta_0\| \to_{\mathrm{P}} 0$ it suffices to show that $\|\mathrm{I}_{a,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$ and $\|\mathrm{I}_{b,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} 1$. Step 1 in the proof of Lemma A2 shows that

$$\begin{aligned}&\sup_{t\in\mathcal{T}} \|\mathrm{I}_{b,n}(t) - \mathrm{E}_Z[\omega(t, X)(\partial/\partial\beta)\rho(Z, \beta_0, h^*(W_i))]\| \xrightarrow{\mathrm{P}} 0, \\ &\quad \text{and } \sup_{t\in\mathcal{T}} \|\mathrm{E}_Z[\omega(t, X)(\partial/\partial\beta)\rho(Z, \beta_0, h^*(W_i))]\| < \infty.\end{aligned}$$

which combine to yield $\|\mathrm{I}_{b,n}\|_{\mathcal{T}} \lesssim_{\mathrm{P}} 1$.

Step 1a: $\|\mathrm{I}_{a,n}\|_{\mathcal{T}} \to_{\mathrm{P}} 0$**.** Abbreviate $(z,v) \mapsto \rho(z,\beta_0,v)$ by ρ. By an MVE of $v \mapsto \rho(Z_i,v)$ at $\widehat{h}(W_i)$ around $h^*(W_i)$ and T we may be bound $\|\mathrm{I}_{a,n}\|_{\mathcal{T}}$ by

$$\begin{aligned}
&\sup_{t\in\mathcal{T}}|\mathbb{E}_n[\omega(t,X_i)\{\left(\partial/\partial h^\top\right)\rho(Z_i,\overline{h}(W_i)) - \left(\partial/\partial h^\top\right)\rho(Z_i,h^*(W_i))\}\{\widehat{h}(W_i) - h^*(W_i)\}]| \\
&\quad + \sup_{t\in\mathcal{T}}|\mathbb{E}_n[\omega(t,X_i)\left(\partial/\partial h^\top\right)\rho(Z_i,h^*(W_i))\{\widehat{h}(W_i) - h^*(W_i)\}]| \\
&=: \|\mathrm{I}_{a,1,n}\|_{\mathcal{T}} + \|\mathrm{I}_{a,2,n}\|_{\mathcal{T}},
\end{aligned}$$

where $\|\overline{h}(W_i) - h^*(W_i)\| \leqslant \|\widehat{h}(W_i) - h^*(W_i)\|$. By T, CS and Assumptions 2 and 8

$$\|\mathrm{I}_{a,1,n}\|_{\mathcal{T}} \lesssim \mathbb{E}_n[R'(Z_i)\|\widehat{h}(W_i) - h^*(W_i)\|^2] \lesssim \mathbb{E}_n[R'(Z_i)] \max_{1\leqslant m\leqslant d}\|\widehat{h}_m - h^*_m\|^2_{\mathcal{W}} \xrightarrow{\mathrm{P}} 0.$$

Similarly, by T, CS and Assumptions 2 and 3,

$$\begin{aligned}
\|\mathrm{I}_{a,2,n}\|_{\mathcal{T}} &\lesssim \mathbb{E}_n[|\left(\partial/\partial h^\top\right)\rho(Z_i,h^*(W_i))\{\widehat{h}(W_i) - h^*(W_i)\}|] \\
&\lesssim \mathbb{E}_n[\|(\partial/\partial h)\rho(Z_i,h^*(W_i))\|] \lesssim_{\mathrm{P}} \max_{1\leqslant m\leqslant d}\|\widehat{h}_m - h^*_m\|_{\mathcal{W}} \xrightarrow{\mathrm{P}} 0,
\end{aligned}$$

where the $\lesssim_{\mathrm{P}}$ follows from $\|(\partial/\partial h)\rho(Z,h^*(W))\|$ being square-integrable and the $\to_{\mathrm{P}} 0$ from Lemma B5.

Step 2: $\|\mathrm{II}_n\|_{\mathcal{T}} \to_{\mathrm{P}} 0$

Step 2b in the proof of Lemma C1 shows that $\sup_{t\in\mathcal{T}}\|\widehat{b}(t) - b(t)\| \to_{\mathrm{P}} 0$, so by CS, Assumption 1, and M

$$\|\mathrm{II}_n\|_{\mathcal{T}} \leqslant \|\mathbb{E}_n[s(Z_i)]\| \sup_{t\in\mathcal{T}}\|\widehat{b}(t) - b(t)\| \xrightarrow{\mathrm{P}} 0.$$

Step 3: $\|\mathrm{III}_n\|_{\mathcal{T}} \to_{\mathrm{P}} 0$

Step 2b in the proof of Lemma C1 also shows $\sup_{t\in\mathcal{T}}\|\widehat{b}(t)\| \lesssim_{\mathrm{P}} 1$, so by CS and Assumption 8,

$$\|\mathrm{III}_n\|_{\mathcal{T}} \leqslant \|\mathbb{E}_n[\widehat{s}(Z_i) - s(Z_i)]\| \sup_{t\in\mathcal{T}}\|\widehat{b}(t)\| \leqslant \|\widehat{s} - s\|_{n,2} \sup_{t\in\mathcal{T}}\|\widehat{b}(t)\| \xrightarrow{\mathrm{P}} 0.$$

Step 4: $\|\mathrm{IV}_n\|_{\mathcal{T}} \to_{\mathrm{P}} 0$

Given that

$$\begin{aligned}
&\mathbb{E}_n[\widehat{\delta}\left(t, W_i\right)^{\top}\left\{Y_i - \widehat{h}\left(W_i\right)\right\} - \delta\left(t, W_i\right)^{\top} U_i] \\
&= \sum_{m=1}^{d} \mathbb{E}_n[\widehat{\delta}_m\left(t, W_i\right)\left\{Y_{mi} - \widehat{h}_m\left(W_i\right)\right\} - \delta_m\left(t, W_i\right) U_{mi}],
\end{aligned}$$

by T, it suffices to bound each right-hand side summand uniformly over $\mathcal{T}$ in probability. I therefore drop the m subscript for the remainder of this step. Now, for fixed $t \in \mathcal{T}$, adding and subtracting $p^{k_n}\left(W_i\right)^{\top} Q_{k_n}^{-1}\widehat{\psi}_{k_n}\left(t\right) U_i$ [with $\widehat{\psi}_k$ defined in (C.11)], recalling that $\delta_k\left(t, w\right) = p^k\left(w\right)^{\top} Q_k^{-1}\psi_k\left(t\right)$ we may decompose (the mth summand) as follows:

$$\begin{aligned}
&\mathbb{E}_n[U_i\{\widehat{\delta}\left(t, W_i\right) - \delta\left(t, W_i\right)\}] - \mathbb{E}_n[\widehat{\delta}\left(t, W_i\right)\{\widehat{h}\left(W_i\right) - h^*\left(W_i\right)\}] \\
&= \widehat{\psi}_{k_n}\left(t\right)^{\top}\left(\widehat{Q}_{k_n}^{-1} - Q_{k_n}^{-1}\right)\mathbb{E}_n[p^{k_n}\left(W_i\right) U_i] + [\widehat{\psi}_{k_n}\left(t\right) - \psi_{k_n}\left(t\right)]^{\top} Q_{k_n}^{-1}\mathbb{E}_n[p^{k_n}\left(W_i\right) U_i] \\
&\quad + \mathbb{E}_n[U_i\{\delta_{k_n}\left(t, W_i\right) - \delta\left(t, W_i\right)\}] - \widehat{\psi}_{k_n}\left(t\right)^{\top}\widehat{Q}_{k_n}^{-1}\mathbb{E}_n[p^{k_n}\left(W_i\right)\{\widehat{h}\left(W_i\right) - h^*\left(W_i\right)\}] \\
&=: \mathrm{IV}_{a,n}\left(t\right) + \mathrm{IV}_{b,n}\left(t\right) + \mathrm{IV}_{c,n}\left(t\right) + \mathrm{IV}_{d,n}\left(t\right).
\end{aligned}$$

The desired $\|\mathrm{IV}_n\|_{\mathcal{T}} \to_{\mathrm{P}} 0$ will follow by T if we can show that the four remainder terms $\to_{\mathrm{P}} 0$. To this end, note first that by Assumption 5,

$$\begin{aligned}
\mathrm{E}\left[\|Q_k^{-1}\mathbb{E}_n[p^k\left(W_i\right) U_i]\|^2\right] &\lesssim \mathrm{E}\left[\|Q_k^{-1/2}\mathbb{E}_n[p^k\left(W_i\right) U_i]\|^2\right] = \mathrm{E}[U^2 p^k\left(W\right)^{\top} Q_k^{-1} p^k\left(W\right)]/n \\
&\lesssim \mathrm{E}[p^k\left(W\right)^{\top} Q_k^{-1} p^k\left(W\right)]/n = k/n,
\end{aligned}$$

so by M we have

$$\|Q_{k_n}^{-1}\mathbb{E}_n[p^{k_n}\left(W_i\right) U_i]\| \lesssim_{\mathrm{P}} \sqrt{k_n/n} \to 0.$$

Step 4c in the proof of Lemma C1 shows that $\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}\left(t\right)^{\top}\widehat{Q}_{k_n}^{-1}\| \lesssim_{\mathrm{P}} 1$. Moreover, Lemma B4 show that $\|\widehat{Q}_{k_n} - Q_{k_n}\|_{\mathrm{op}} \lesssim_{\mathrm{P}} [\zeta_{k_n}^2 \ln(k_n)/n]^{1/2} \to 0$, so by the previous

display and CS,

$$\begin{aligned}\|\mathrm{IV}_{a,n}\|_{\mathcal{T}} &= \|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1}(Q_{k_n}-\widehat{Q}_{k_n})Q_{k_n}^{-1}\mathbb{E}_n[p^{k_n}(W_i)U_i]\|_{\mathcal{T}} \\ &\leqslant \|Q_{k_n}^{-1}\mathbb{E}_n[p^{k_n}(W_i)U_i]\|\|\widehat{Q}_{k_n}-Q_{k_n}\|_{\mathrm{op}}\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1}\| \\ &\lesssim_{\mathrm{P}} (k_n/n)^{1/2}\left\{\zeta_{k_n}^2\ln(k_n)/n\right\}^{1/2}\to 0.\end{aligned}$$

Step 4c in the proof of Lemma C1 also shows that $\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)-\psi_{k_n}(t)\|\to_{\mathrm{P}} 0$, so by CS,

$$\|\mathrm{IV}_{b,n}\|_{\mathcal{T}} \leqslant \|Q_{k_n}^{-1}\mathbb{E}_n[p^{k_n}(W_i)U_i]\|\sup_{\mathcal{T}}\|\widehat{\psi}_{k_n}(t)-\psi_{k_n}(t)\| \xrightarrow{\mathrm{P}} 0.$$

Step 4c in the proof of Lemma A2 shows that

$$\|\mathrm{IV}_{c,n}\|_{t\in\mathcal{T}} = \sup_{t\in\mathcal{T}}|\mathbb{E}_n[U_i\{\delta_{k_n}(t,W_i)-\delta(t,W_i)\}]| \lesssim_{\mathrm{P}} R_{\delta,k_n}\sqrt{\ln(k_n/R_{\delta,k_n})}\to 0.$$

Lastly, by CS, Lemma B5 and $\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1}\|\lesssim_{\mathrm{P}} 1$ we get

$$\begin{aligned}\|\mathrm{IV}_{d,n}\|_{\mathcal{T}} &\leqslant \|\mathbb{E}_n[p^{k_n}(W_i)\{\widehat{h}(W_i)-h^*(W_i)\}]\|\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1}\| \\ &\leqslant \|\mathbb{E}_n[p^{k_n}(W_i)\{\widehat{h}(W_i)-h^*(W_i)\}]\|\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1}\| \\ &\lesssim \Big(\sum_{j=1}^{k_n}\|p_j\|_{\mathcal{W}}^2\Big)^{1/2}\max_{1\leqslant m'\leqslant d}\|\widehat{h}_{m'}-h^*_{m'}\|_{n,2}\sup_{t\in\mathcal{T}}\|\widehat{\psi}_{k_n}(t)^\top \widehat{Q}_{k_n}^{-1}\| \\ &\lesssim_{\mathrm{P}} \Big(\sum_{j=1}^{k_n}\|p_j\|_{\mathcal{W}}^2\Big)^{1/2}\max_{1\leqslant m'\leqslant d}\left(\sqrt{k_{m',n}/n}+k_{m',n}^{-\alpha_{m'}}\right)\to 0.\end{aligned}$$

This finishes the proof of $\|\mathbb{E}_n[\widehat{g}(\cdot,Z_i)-g(\cdot,Z_i)]\|_{\mathcal{T}}\to_{\mathrm{P}} 0$. □

Proof of Lemma 3. Since the lemma is stated for a given ℓ, I drop the ℓ subscripts throughout and refer to the (ℓth) index set $\mathcal{X}_\ell$ as $\mathcal{T}$ itself. Then by T,

$$\begin{aligned}\|\widehat{G}-G_n^*\|_{\mathcal{T}} &\equiv \left\|\sqrt{n}\mathbb{E}_n\left[\left(\xi_i-\overline{\xi}\right)\widehat{g}(\cdot,Z_i)\right]-\sqrt{n}\mathbb{E}_n\left[\left(\xi_i-\overline{\xi}\right)g(\cdot,Z_i)\right]\right\|_{\mathcal{T}} \\ &= \left\|\sqrt{n}\mathbb{E}_n[\xi_i\widehat{g}(\cdot,Z_i)]-\sqrt{n}\mathbb{E}_n[\xi_i g(\cdot,Z_i)]-\sqrt{n}\overline{\xi}\{\mathbb{E}_n[\widehat{g}(\cdot,Z_i)-g(\cdot,Z_i)]\}\right\|_{\mathcal{T}} \\ &= \left\|\widehat{G}^u-G_n^{*u}-\sqrt{n}\overline{\xi}\{\mathbb{E}_n[\widehat{g}(\cdot,Z_i)-g(\cdot,Z_i)]\}\right\|_{\mathcal{T}} \\ &\leqslant \|\widehat{G}^u-G_n^{*u}\|_{\mathcal{T}}+|\sqrt{n}\overline{\xi}|\,\|\mathbb{E}_n[\widehat{g}(\cdot,Z_i)]-\mathbb{E}_n[g(\cdot,Z_i)]\|_{\mathcal{T}}.\end{aligned}$$

The first term on the right $\to_{\mathrm{P}} 0$ by Lemma C1. Given that $\sqrt{n}\overline{\xi} \sim \mathrm{N}(0,1)$, certainly $|\sqrt{n}\overline{\xi}| \lesssim_{\mathrm{P}} 1$. The second term therefore $\to_{\mathrm{P}} 0$ by Lemma C2. □

C.2 Additional Supporting Lemmas

Lemma C3. *If X_n is a sequence of nonnegative random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathrm{P})$, $\mathcal{F}_n$ is a sequence of sub-σ-algebras, and $\mathrm{E}[X_n|\mathcal{F}_n] \to_{\mathrm{P}} 0$, then $X_n \to_{\mathrm{P}} 0$.*

Proof. Fix $n \in \mathbf{N}$, let $Y_n \coloneqq \mathrm{E}[X_n|\mathcal{F}_n]$ and let $A_n \coloneqq \{Y_n = 0\}$. Then $X_n = 0$ almost everywhere on A_n. Indeed, if X_n is not zero almost everywhere on A_n, then there exists $C \in (0,\infty)$ such that $B_{n,C} \coloneqq \{\omega \in A_n | X_n(\omega) > 1/C\}$ satisfies $\mathrm{P}(B_{n,C}) > 0$. By definition of (a version of) the conditional expectation of X_n given $\mathcal{F}_n$, we must have $\int_A X_n \mathrm{d}P = \int_A Y_n \mathrm{d}P$ for every $A \in \mathcal{F}_n$ and, in particular, for A_n. Since $Y_n = 0$ on A_n and $B_{n,C} \subset A_n$, it follows that

$$0 = \int_{A_n} Y_n \mathrm{d}P = \int_{A_n} X_n \mathrm{d}P \geqslant \int_{B_{n,C}} X_n \mathrm{d}P \geqslant \mathrm{P}(B_{n,C})/C,$$

which contradicts $\mathrm{P}(B_{n,C}) > 0$. Since $n \in \mathbf{N}$ was arbitrary, we have shown that $X_n = 0$ on A_n for each $n \in \mathbf{N}$. Now, fix $\varepsilon, \delta > 0$. Then $\mathrm{P}(X_n > \varepsilon \cap Y_n = 0) = 0$ by the previous claim, and it follows that

$$\begin{aligned}\mathrm{P}(X_n > \varepsilon) &= \mathrm{P}(X_n > \varepsilon \cap Y_n = 0) + \mathrm{P}(X_n > \varepsilon \cap 0 < Y_n \leqslant \delta\varepsilon) + \mathrm{P}(X_n > \varepsilon \cap Y_n > \delta\varepsilon) \\ &\leqslant \mathrm{P}\left(X_n > \delta^{-1} Y_n > 0\right) + \mathrm{P}(Y_n > \delta\varepsilon).\end{aligned}$$

Given that Y_n is $\mathcal{F}_n$ measurable, by conditional M we have

$$\begin{aligned}\mathrm{P}\left(X_n > \delta^{-1} Y_n > 0\right) &= \mathrm{E}\left[\mathbf{1}_{Y_n > 0} \mathrm{P}\left(X_n > \delta^{-1} Y_n \middle| \mathcal{F}_n\right)\right] \leqslant \mathrm{E}\left[\mathbf{1}_{Y_n > 0} \delta \mathrm{E}[X_n|\mathcal{F}_n]/Y_n\right] \\ &= \delta \mathrm{P}(Y_n > 0) \leqslant \delta.\end{aligned}$$

By $Y_n \to_{\mathrm{P}} 0$ and the previous two displays we see that for any $\varepsilon, \delta > 0$, $\overline{\lim}\, \mathrm{P}(X_n > \varepsilon) \leqslant \delta$, so the claim follows from letting $\delta \to 0$. □