NBER WORKING PAPER SERIES

MINIMAX RISK AND UNIFORM CONVERGENCE RATES FOR NONPARAMETRIC DYADIC REGRESSION

Bryan S. Graham
Fengshi Niu
James L. Powell

Working Paper 28548
http://www.nber.org/papers/w28548

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2021

We thank seminar audiences at University of California - Berkeley for helpful feedback. We also thank Matias Cattaneo, Michael Jansson and Harold Chiang for useful comments and discussion. All the usual disclaimers apply. Financial support from the National Science Foundation (SES #1357499, SES #1851647) is gratefully acknowledged. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Minimax Risk and Uniform Convergence Rates for Nonparametric Dyadic Regression
Bryan S. Graham, Fengshi Niu, and James L. Powell
NBER Working Paper No. 28548
March 2021
JEL No. C14

ABSTRACT

Let $i = 1, \ldots, N$ index a simple random sample of units drawn from some large population. For each unit we observe the vector of regressors X_i and, for each of the $N(N-1)$ ordered pairs of units, an outcome Y_{ij}. The outcomes Y_{ij} and Y_{kl} are independent if their indices are disjoint, but dependent otherwise (i.e., "dyadically dependent"). Let $W_{ij} = \left(X_i' X_j'\right)'$; using the sampled data we seek to construct a nonparametric estimate of the mean regression function $g(W_{ij}) \stackrel{def}{=} E[Y_{ij} \mid X_i, X_j]$.

We present two sets of results. First, we calculate lower bounds on the minimax risk for estimating the regression function at (i) a point and (ii) under the infinity norm. Second, we calculate (i) pointwise and (ii) uniform convergence rates for the dyadic analog of the familiar Nadaraya-Watson (NW) kernel regression estimator. We show that the NW kernel regression estimator achieves the optimal rates suggested by our risk bounds when an appropriate bandwidth sequence is chosen. This optimal rate differs from the one available under iid data: the effective sample size is smaller and $d_W = \dim(W_{ij})$ influences the rate differently.

Bryan S. Graham
University of California - Berkeley
530 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
bgraham@econ.berkeley.edu

James L. Powell
University of California - Berkeley
Department of Economics
508-1 Evans Hall #3880
Berkeley, CA 94720-3880
powell@econ.berkeley.edu

Fengshi Niu
University of California - Berkeley
530 Evans Hall #3880
Berkeley, CA 94720-3880
fniu@berkeley.edu

1 Introduction

Let $i = 1, \ldots, N$ index a simple random sample of units drawn from some large population. For each unit we observe the vector of regressors X_i and, for each of the $N(N-1)$ ordered pairs of units, or *directed dyads*, we observe the "dyadic" outcome Y_{ij} (e.g., total exports from country i to country j). The outcomes Y_{ij} and Y_{kl} are independent if their indices are disjoint, but dependent otherwise (e.g., exports from Japan to Korea may covary with those from Japan to Vietnam).

Let $W_{ij} = \left(X_i', X_j'\right)'$; using the sampled data we seek to construct a nonparametric estimate of the mean regression function

$$g(W_{ij}) \stackrel{def}{\equiv} \mathbb{E}\left[Y_{ij} \middle| X_i, X_j\right]. \tag{1}$$

We present two sets of results. First, we calculate lower bounds on the minimax risk for estimating the regression function at (i) a point and (ii) under the infinity norm. Second, we calculate (i) pointwise and (ii) uniform convergence rates for the dyadic analog of the familiar Nadaraya-Watson (NW) kernel regression estimator. We show that the NW kernel regression estimator achieves the optimal rates suggested by our risk bounds when an appropriate bandwidth sequence is chosen.

Analogous results are widely available in the i.i.d. setting. For nonparametric regression risk bounds see, for example, Stone (1980, 1982) and Ibragimov and Has' Minskii (1982, 1984). Tsybakov (2008) provides a masterful synthesis of these results, from which we draw in formulating our own proofs.

Uniform convergence of kernel averages with i.i.d. data, as well as stationary strong mixing data, have been studied by, for example, Newey (1994) and Hansen (2008) respectively. The latter paper includes additional references to the extensive literature in this area. Our uniform convergence proofs build upon those of Hansen (2008). Nonparametric density estimation with dyadic data was first considered by Graham et al. (2019); Chiang et al. (2019) present uniform convergence results for dyadic density estimators.[1]

Our results provide insight in the structure of dyadic nonparametric estimation problems. Our minimax risk bounds suggest that, N, the number of units, *not* $n \stackrel{def}{\equiv} N \times (N-1)$, the number of dyadic outcomes, is the relevant "sample size" for dyadic estimation problems. This is consistent with the long standing intuition among empirical researchers that dyadic dependence makes inference less precise (see Aronow et al. (2017) and the references cited therein), as well as with a small, but growing, number of more formal rates-of-convergence

[1] It is possible that the methods of inference presented in Chiang et al. (2019) could be adapted to our setting.

results (cf., Graham, 2020a).

More surprisingly, we find that the relevant dimension of our estimation problem is just $d_X = \dim(X_i)$, not $d_W = 2d_X$. We provide two intuitions for this fact. The first, described below, stems from the thought experiment underlying our minimax risk bound calculations. The second, arises from the fact that the Hájek projection of the NW estimator has a "partial-mean-like" structure. As is well known, averaging over the marginal distribution of some regressors, while holding the remaining ones fixed, improves rates-of-convergence (e.g., Newey, 1994; Linton and Nielsen, 1995).

Graham (2020a) surveys empirical studies in economics utilizing dyadic data. Interest in, as well as the availability of, such data are growing in economics, other academic fields, and in enterprise settings. This paper provides an initial set of results for nonparametric regression with dyadic data. These results are, of course, of direct interest. They should, as has been true with their i.i.d. predecessors, also be useful for proving consistency of two-step semiparametric M-estimators under dyadic dependence (see Chiang et al. (2019) for some results on double machine learning with dyadic data).

2 Lower Bounds on the Minimax Risk

Let $i = 1, \ldots, N$ index a simple random sample of units drawn from some large population. The econometrician observes the vector of regressors, X_i, for each sampled unit as well as the scalar outcome, Y_{ij}, for each directed pair of sampled units (i.e., each directed dyad). Let $\mathbf{Z}_N = (X_1, \ldots, X_N, Y_{ij}, 1 \leq i \neq j \leq N)$ be the observable data when N units are sampled. The regression function of interest is (1) above. The goal is to construct a nonparametric estimate of $g : \mathbb{R}^{d_W} \to \mathbb{R}$ where $d_W = 2d_x$.

We assume that Y_{ij} is generated according to the following conditionally independent dyad (CID) model (cf., Graham, 2020a, Section 3.3).

$$Y_{ij} = h(X_i, X_j, U_i, U_j, V_{ij}). \tag{2}$$

Random sampling ensures that (X_i, U_i) is independent and identically distributed for $i = 1, \ldots, N$. We further assume that $\{(V_{ij}, V_{ji})\}_{1 \leq i < j \leq N}$ are i.i.d. and indepenent of $\mathbf{X} = (X_1, \ldots, X_N)'$ and $\mathbf{U} = (U_1, \ldots, U_N)$. Here h is an unknown function, often called the *graphon*. This set-up, which can also be derived as an implication of more primitive exchangeability assumptions, has the following implications (see Graham (2020a,b) for additional discussion):

1. The Y_{ij} are relatively exchangeable given the W_{ij}. Namely, the conditional distribution

of $\mathbf{Y}$ is invariant across permutations of the indices $\sigma : \mathbb{N} \to \mathbb{N}$ satisfying the restriction $[W_{\sigma(i)\sigma(j)}] \stackrel{d}{=} [W_{ij}]$:

$$[Y_{ij}] \stackrel{d}{=} [Y_{\sigma(i)\sigma(j)}].$$

2. Y_{ij} and Y_{kl} are independent if their indices are disjoint.

3. Y_{ij} and Y_{kl} are dependent (unconditionally or conditionally given $X_1, \ldots, X_N$) if they share at least one index in common.

The statistical problem is to estimate the regression function g when the only prior restriction on it is that it belongs to the Hölder class of functions.

Definition 2.1. (Hölder Class) Given a vector $s = (s_1, \ldots, s_d)$, define $|s| = s_1 + \cdots + s_d$ and

$$D^s = \frac{\partial^{s_1 + \cdots + s_d}}{\partial^{s_1} w_1 \cdots \partial^{s_d} w_d}.$$

Let β and L be two positive numbers. The *Hölder class* $\Sigma(\beta, L)$ on $\mathbb{R}^d$ is defined as the set of $l = \lfloor \beta \rfloor$ times differentiable functions $g : \mathbb{R}^d \to \mathbb{R}$ whose partial derivative $D^s g$ satisfies

$$|D^s g(w) - D^s g(w')| \leq L||w - w'||_\infty^{\beta - l}, \quad \forall w, w' \in \mathbb{R}^d$$

for all s such that $|s| = \lfloor \beta \rfloor$. $\lfloor \beta \rfloor$ denotes the greatest integer strictly less than the real number β.

An estimator $\hat{g}_N$ is a function $w \mapsto \hat{g}_N(w) = \hat{g}_N(w, \mathbf{Z}_N)$ measurable with respect to $\mathbf{Z}$. Our first result establishes a lower bound on the minimax risk for estimating the regression function at a single point and under the infinity norm. We state this result under a Gaussian error assumption, which simplifies the proof.

Theorem 2.1. (Minimax Risk Lower Bound) *Suppose that $\beta > 0$ and $L > 0$; X_i is continuously distributed on $\mathbb{R}^{d_X}$ with density f and $\sup_x f(x) \leq B_3 < \infty$; and Y_{ij} is generated according to the following nonparametric regression model:*

$$Y_{ij} = g(W_{ij}) + e_{ij}, \quad i \neq j,$$

with $e_{ij} = U_i + U_j + V_{ij}$, $U_i \overset{\text{iid}}{\sim} N(0,1)$, and $V_{ij} \overset{\text{iid}}{\sim} N(0,1)$, then

(i) For all $w \in \mathbb{R}^{d_W}$,

$$\liminf_{N \to \infty} \inf_{\hat{g}_N} \sup_{g \in \Sigma(\beta, L)} \mathbb{E}_g \left[N^{\frac{2\beta}{2\beta + d_X}} \left(\hat{g}_N(w) - g(w) \right)^2 \right] \geq c_1,$$

where $c_1 > 0$ depends only on β and L.

(ii)

$$\liminf_{N\to\infty} \inf_{\hat{g}_N} \sup_{g\in\Sigma(\beta,L)} \mathbb{E}_g \left[\left(\frac{N}{\ln N} \right)^{\frac{2\beta}{2\beta+d_X}} ||\hat{g}_N - g||_\infty^2 \right] \geq c_2,$$

where $c_2 > 0$ also depends only on β and L.

Our proof follows the general recipe outlined in Chapter 2 of Tsybakov (2008). The lower bound at a point is based on Le Cam's method of two hypotheses. The lower bound under the infinity norm is based on Fano's method of multiple hypotheses.

The key, and novel, step in our proof involves constructing hypotheses close enough to one other in terms of Kullback-Leibler (KL) divergence while being at the same time different enough in terms of the target regression function.

An essential feature of our construction is additive separability of the regression functions. In the hypotheses we consider, $Y_{ij} = k(X_i) + k(X_j) + U_i + U_j + V_{ij}$. Next suppose we also observe $T_i \stackrel{def}{\equiv} k(X_i) + U_i$. Observe that $(X_i, T_i, i = 1, \ldots, N)$ is sufficient with respect to $(X_i, T_i, i = 1, \ldots, N, Y_{kl}, 1 \leq k \neq l \leq N)$ for the parameter k.

It is well-known that the optimal rates of convergence for estimating k using iid data $(X_i, T_i, i = 1, \ldots, N)$ are $N^{-\frac{\beta}{2\beta+d_X}}$ pointwise and $\left(\frac{N}{\ln N}\right)^{-\frac{\beta}{2\beta+d_X}}$ for the infinity norm. We expect the rates for estimating g to be no faster than these. The proof of Theorem 2.1 makes this intuition rigorous.

Relative to its iid counterpart, there are two distinctive features of Theorem 2.1. First, the relevant sample size is not the number of observed dyadic outcomes $n = N \times (N-1)$, but instead the number of sampled units, N. Dependence across outcomes sharing indices in common is strong enough to slow down the feasible rate of convergence. Second, although the regression function has $d_W = 2d_X$ arguments, the relevant dimension reflected in the rate of convergence result is just d_X (i.e., just half of what might naively be expected).

The form of our constructed hypotheses provides one intuition for this second finding: clearly the relevant dimension of the problem of estimating $k(x)$ is just d_X. Relatedly this finding is consistent with those of Linton and Nielsen (1995) in their analysis of additively separable, but otherwise nonparametric, regression functions (see also Newey (1994)).

The pairwise structure of dyadic data results in apparent data abundance (sample N agents, but observe $O(N^2)$ outcomes!). This abundance is both illusory, in the sense that the effective sample size is indeed just N, and real, in the sense that availability of the pairwise outcome data allows for an effective reduction in the dimensionality of the problem via partial mean like average (as in Newey (1994) and Linton and Nielsen (1995) in a different context).

3 Kernel Estimator of Dyadic Regression

In this section we study the properties of a specific nonparametric regression estimator. Namely, the dyadic analog of the well-known Nadaraya-Watson (NW) kernel regression estimator. While our results are specific to this estimator, they could, for example, be extended to apply to local linear regression (e.g., Hansen, 2008).

The dyadic NW kernel regression estimator is

$$\hat{g}_N(w) := \frac{\sum_{1\le i\neq j\le N} K_{ij,N}(w)Y_{ij}}{\sum_{1\le i\neq j\le N} K_{ij,N}(w)}, \tag{3}$$

where

$$K_{ij,N}(w) := \frac{1}{h_N^{d_W}} K\left(\frac{W_{ij}-w}{h_N}\right),$$

K is a fixed multivariate kernel function, and h_N is a vanishing bandwidth sequence.

We first develop a sequence of results useful for bounding the variance of kernel objects of the form

$$\hat{\Psi}_N(w) := \frac{1}{N(N-1)} \sum_{1\le i\neq j\le N} Y_{ij}K_{ij,N}(w) \tag{4}$$

and then apply these results to the NW regression estimator. We then bound the NW estimator's bias and combine the two sets of results to formulate a risk bound.

3.1 Variance Bound and Uniform Convergence

Here we are interested in bounding the deviation of $\hat{\Psi}_N(w)$ from its mean. We begin with a presentation of our maintained assumptions.

Assumption 3.1 (Model)**.** The data generating process is as described in Section 2 with

(i) X_i continuously distributed with marginal density $f(x)$ s.t. $\sup_{x\in\mathbb{R}^{d_X}} f(x) \le B_3 < \infty$;

(ii) $\sup_{x_1,x_2\in\mathbb{R}^{d_X}} \mathbb{E}\left[|Y_{12}|^2 \middle| (X_1,X_2)=(x_1,x_2)\right]\cdot f(x_1)f(x_2) \le B_4 < \infty$,
$\sup_{x_1,x_2,x_3\in\mathbb{R}^{d_X}} \mathbb{E}\left[|Y_{12}Y_{13}| \middle| (X_1,X_2,X_3)=(x_1,x_2,x_3)\right]\cdot f(x_1)f(x_2)f(x_3) \le B_5 < \infty$.

Condition (i) is a standard condition in the context of kernel estimation, while (ii) ensures that various second moments appearing in our variance calculations are finite.

Assumption 3.2 (Kernel, Part A)**.** $\sup_{w\in\mathbb{R}^{d_W}} |K(w)| \le K_{\max} < \infty$, $\int_{w\in\mathbb{R}^{d_W}} |K(w)|\mathrm{d}w \le B_1 < \infty$, and $\sup_{x\in\mathbb{R}^{d_X}} \int |K(x,x')|\mathrm{d}x' \le B_2 < \infty$.

Assumption 3.2 is satisfied by many widely-used multivariate kernel functions. Our first result holds under Assumptions 3.1 and 3.2.

Theorem 3.1 (Variance Bound)**.** *Under Assumptions 3.1 and 3.2, and the bandwidth condition $Nh_N^{d_X} \to \infty$ as $N \to \infty$, there exists a constant $M_0 < \infty$ such that for N sufficiently large*

$$\text{Var}\left(\hat{\Psi}_N(w)\right) \leq \frac{M_0}{Nh_N^{d_X}}$$

for all $w \in \mathbb{R}^{d_W}$.

A proof is available in the appendix. Mirroring our risk bound results, two features of Theorem 3.1 merit comment. First, N not $n = N \times (N-1)$ appears in the denominator. This is due to the effects of dependence across dyads sharing units in common. Second, the relevant dimension of the problem is d_X, not $d_W = 2d_X$, this reflects the U-statistic like structure of kernel weighted averages and the partial mean like averaging this structure induces.

To establish uniform convergence, we need additional moment conditions on Y_{ij} as well as some smoothness conditions on the kernel K. As in Hansen (2008), we require the kernel to either have bounded support and be Lipschitz or have bounded derivatives and an integrable tail. See Hansen (2008) for additional discussion about these conditions. As with Assumption 3.2 above, most commonly used kernels satisfy these conditions.

Assumption 3.3 (Regularity Condition)**.** (i) For some $s > 2$, $\mathbb{E}|Y_{12}|^s < \infty$ and $\sup_{x_1,x_2 \in \mathbb{R}^{d_X}} \mathbb{E}\left[|Y_{12}|^s \middle| (X_1, X_2) = (x_1, x_2)\right] \cdot f(x_1, x_2) \leq B_{4,s} < \infty$;

(ii) For some $\Lambda_1 < \infty$ and $L < \infty$, either (a) or (b) holds

(a) $K(w) = 0$ for $||w|| > L$, and $|K(w) - K(w')| \leq \Lambda_1 ||w - w'||$ for all $w, w' \in \mathbb{R}^{2d}$

(b) $K(w)$ is differentiable, $\left|\left|\frac{\partial}{\partial w} K(w)\right|\right| \leq \Lambda_1$, where $\left|\left|\frac{\partial}{\partial w} K(w)\right|\right| = \left|\left|\left(\frac{\partial}{\partial w_1} K(w), \ldots \frac{\partial}{\partial w_{2d}} K(w)\right)\right|\right|_\infty$, and for some $\nu > 1$, $\left|\left|\frac{\partial}{\partial w} K(w)\right|\right| \leq \Lambda_1 ||w||^{-\nu}$ for $||w|| > L$.

Part (ii) coincides with Assumption 3 in Hansen (2008). This assumption implies that for all $||w_1 - w_2|| \leq \delta \leq L$,

$$|K(w_2) - K(w_1)| \leq \delta K^*(w_1),$$

where $K^*(u)$ satisfies Assumption 3.1. If case (a) holds, then $K^*(u) = 2d\Lambda_1 \mathbb{1}(||u|| \leq 2L)$. If case (b) holds, then, $K^*(u) = 2d[\Lambda_1 \mathbb{1}(||u|| \leq 2L) + (||u|| - L)^{-\nu} \mathbb{1}(||u|| > 2L)]$. In both cases K^* is bounded and integrable and therefore satisfies Assumption 3.1.

Define

$$a_N := \left(\frac{\ln N}{N h_N^{d_X}}\right)^{1/2}.$$

Theorem 3.2 (Weak Uniform Convergence)**.** *Under Assumptions 3.1, 3.2, 3.3, and the bandwidth conditions* $\max\left\{\min\left\{(a_N h_N^{2d_X})^{-\frac{1}{s-1}}, [N^2(\ln(\ln N))^2 \ln N]^{\frac{1}{s}}\right\}, a_N^{-\frac{1}{s-1}}\right\} \ll \min\left\{a_N^{-1}, \frac{N}{\ln N} h_N^{\frac{3}{2}d_X}\right\}$ *and* $\frac{N}{\ln N} h_N^{d_X} \to \infty$*, we have for any* $q > 0$*,* $c_N = N^q$*,*

$$\sup_{||w|| \le c_N} \left|\hat{\Psi}_N(w) - \mathbb{E}\hat{\Psi}_N(w)\right| = O_P(a_N).$$

This theorem establishes uniform convergence of $\hat{\Psi}_N(w)$ to its mean in probability over an expanding set with radius growing at a polynomial rate.

In the proof, we decompose $\hat{\Psi}_N(w)$ into two parts

$$\hat{\Psi}_N(w) = \tilde{\Psi}_N(w) + R_N(w),$$

in which $\tilde{\Psi}_N(w) = \frac{1}{N(N-1)} \sum_{1 \le i \ne j \le N} Y_{ij} \cdot \mathbb{1}\left(|Y_{ij}| < \tau_N\right) K_{ij,N}$ is a truncated version of $\hat{\Psi}_N(w)$ with a carefully chosen threshold parameter τ_N and $R_N(w)$ is a residual. The boundedness induced by this truncation is technically convenient as it facilitates the application of various concentration inequalities. To establish concentration of $\tilde{\Psi}_N$, we apply Bernstein inequality to its Hájek Projection (i.e., to the first-order terms in the Hoeffding decomposition) and apply Arcones and Gine (1993)'s concentration inequalities for degenerate U-statistics to the second-order terms in the Hoeffding decompositon. Both these bounds requires the truncation threshold to be small enough. To bound the magnitude of the residual R_N, we can either apply a triangular inequality to bound the sup of its first moment or use the Borel-Cantelli Lemma to bound its probability of being nonzero. Both these bounds requires the truncation threshold to be large.

A proper truncation threshold satisfying both requirements exists only if the bandwidth sequence satisfies the condition

$$\max\left\{\min\left\{(a_N h_N^{2d_X})^{-\frac{1}{s-1}}, [N^2(\ln(\ln N))^2 \ln N]^{\frac{1}{s}}\right\}, a_N^{-\frac{1}{s-1}}\right\} \ll \min\left\{a_N^{-1}, \frac{N}{\ln N} h_N^{\frac{3}{2}d_X}\right\}.$$

The complicated form of this condition is technical in nature. When all (conditional) moments of Y_{12} are bounded, such that $s = \infty$ (of Assumption 3.3 above), this condition simplifies to $\frac{N}{\ln N} h_N^{\frac{3}{2}d_X} \gg 1$.

In order to state the weak uniform convergence result for the kernel regression estimator $\hat{g}_N$, we need additional smoothness assumptions on the kernel. As in other applications of

kernel estimation, these assumptions are employed for bias reduction purpose.

Assumption 3.4 (Kernel, Part B)**.**

$$\int_{\mathbb{R}^{d_W}} w_1^{l_1} \cdots w_{d_W}^{l_{d_W}} K(w)dw = \begin{cases} 1, & \text{if } l_1 = \cdots = l_{d_W} = 0 \\ 0, & \text{if } (l_1, \ldots, l_{d_W})' \in \mathbb{Z}_+^{d_W} \text{ and } l_1 + \cdots + l_{d_X} < \beta \end{cases}$$

We can now give a uniform convergence result for the NW regression estimator under dyadic dependence over a sequence of expanding sets.

Theorem 3.3. *Suppose* $f_W, g \in \Sigma(\beta, L)$ *and* $\delta_N = \inf_{||w|| \leq C_N} f_W(w) > 0$, $\delta_N^{-1} a_N^* \to 0$ *where* $a_N^* := \left(\frac{\ln N}{N h_N^{d_X}}\right)^{1/2} + h_N^\beta$. *Under the Assumptions of Theorem, 3.2 and Assumption 3.4*

$$\sup_{||w|| \leq C_N} |\hat{g}_N(w) - g(w)| = O_p(\delta_N^{-1} a_N^*).$$

The optimal convergence rate is

$$\sup_{||w|| \leq C_N} |\hat{g}_N(w) - g(w)| = O_p\left(\delta_N^{-1} \left(\frac{\ln N}{N}\right)^{\frac{\beta}{2\beta + d_X}}\right).$$

As in the iid case, the KW estimator achieves the optimal rate suggested by Theorem 2.1 for a compact set with $C_N = C$. If we look at a sequence of expanding sets approaching the entire space $\mathbb{R}^{d_W}$, then there is an additional penalty term δ_N due to the presence of the denominator $f_W(w)$.

References

Arcones, M. A. and Gine, E. (1993). Limit theorems for u-processes. *The Annals of Probability*, 21(3):1494–1542.

Aronow, P. M., Samii, C., and Assenova, V. A. (2017). Cluster–robust variance estimation for dyadic data. *Political Analysis*, 23(4):564–577.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Chiang, H. D., Kato, K., Ma, Y., and Sasaki, Y. (2019). Multiway cluster robust double/debiased machine learning. *arXiv preprint arXiv:1909.03489*.

Graham, B. S. (2020a). Network data. In *Handbook of Econometrics volume 7*. North-Holland, Amsterdam.

Graham, B. S. (2020b). Sparse network asymptotics for logistic regression. *arXiv preprint arXiv:2010.04703*.

Graham, B. S., Niu, F., and Powell, J. L. (2019). Kernel density estimation for undirected dyadic data. *arXiv preprint arXiv:1907.13630*.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(3):726–748.

Ibragimov, I. A. and Has' Minskii, R. Z. (1982). Bounds for the risks of nonparametric regression estimates. *Theory of Probability and Its Applications*, 16:84–99.

Ibragimov, I. A. and Has' Minskii, R. Z. (1984). Asymptotic bounds on the quality of the nonparametric regression estimation in l_o. *Journal of Soviet Mathematics*, 25:540–550.

Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82:93–100.

Newey, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, pages 233–253.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer.

Appendix

All notation is as established in the main text unless noted otherwise. Equation numbering continues in sequence with that of the main text.

Proof of Theorem 2.1

Our method of proof follows the general approach outlined in Chapter 2 of Tsybakov (2008). To prove part (i) we use Le Cam's two-point method to find a lower risk bound for estimation of the regression function at a point. To prove statement (ii), which involves the infinity-norm metric, we use Fano's method.

Proof of statement (i)

Our proof of statement (i) essentially involves checking the conditions, as specially formulated for our dyadic regression problem, of Theorem 2.3 of (Tsybakov, 2008).

For $k = 0, 1$, let P_{kN} be a probability measure for the observed data $\{(X_i', Y_{ij})\}_{1 \le i \ne j \le N}$ with regression function g_{kN}. The general reduction scheme outlined in Section 2.2 of Tsybakov (2008), as well as his Theorems 2.1 and 2.2, imply that our Theorem 2.1 will hold if we can construct two sequences of hypotheses g_{0N}, g_{1N} such that

(a) the regression functions g_{0N}, g_{1N} are in the Hölder class $\Sigma(\beta, L)$;

(b) $d(\theta_1, \theta_0) = |g_{1N}(w) - g_{0N}(w)| \ge 2A\psi_N$ with $\psi_N = N^{-\frac{\beta}{2\beta + d_X}}$ and $\theta_0 = g_{0N}(w)$ and $\theta_1 = g_{1N}(w)$ for some fixed $w \in \mathbb{X} \times \mathbb{X}$;

(c) the Kullback-Leibler divergence of P_{0N} from P_{1N} is bounded: $\mathrm{KL}(P_{0N}, P_{1N}) \le \alpha < \infty$.

The "trick" of the proof is choosing these two sequences of hypotheses appropriately. Letting $w = (x_{10}, x_{20})$ we choose the sequences:

$$g_{0N}(x_1, x_2) \equiv 0$$
$$g_{1N}(x_1, x_2) = \frac{L h_N^\beta}{2} \left[K\left(\frac{x_1 - x_{10}}{h_N}\right) + K\left(\frac{x_1 - x_{20}}{h_N}\right) + K\left(\frac{x_2 - x_{10}}{h_N}\right) + K\left(\frac{x_2 - x_{20}}{h_N}\right) \right]$$

where $h_N = c_0 N^{-\frac{1}{2\beta + d_X}}$ and the function $K : \mathbb{R}^{d_W} \to [0, \infty)$ satisfies

$$K \in \Sigma(\beta, 1/2) \cap C^\infty(\mathbb{R}^{d_X}) \text{ and } K(x) > 0 \iff ||x||_\infty \in (-1/2, 1/2). \tag{5}$$

There exist functions K satisfying this condition. For example, for a sufficiently small $a > 0$, we can take

$$K(x) = \Pi_{i=1}^{d_X} \lambda(x_i), \quad \text{where } \lambda(u) = a\eta(2u) \text{ and } \eta(u) = \exp\left(-\frac{1}{1 - u^2}\right) \mathbb{1}(|u| \le 1).$$

See also Equation (2.34) in Tsybakov (2008).

We verify conditions (a), (b) and (c) in sequence.

Verification of (a) $g_{0N}, g_{1N} \in \Sigma(\beta, L)$

For $s = (\underbrace{s_1, \ldots, s_{d_X}}_{\mathcal{S}_1}, \underbrace{s_{d_{X+1}}, \ldots, s_{2d_X}}_{\mathcal{S}_2})$ with $|s| = \lfloor \beta \rfloor$, $w = (x_1, x_2)$ and $w' = (x_1', x_2')$, the s^{th} order derivative of g_{1N} is

$$
\begin{aligned}
D^s g_{1N}(w) &= L h_N^{\beta} \left[D^s K\left(\frac{x_1 - x_{10}}{h_N}\right) + D^s K\left(\frac{x_1 - x_{20}}{h_N}\right) + D^s K\left(\frac{x_2 - x_{10}}{h_N}\right) + D^s K\left(\frac{x_2 - x_{20}}{h_N}\right) \right] \\
&= \begin{cases} 0 & \text{if } |\mathcal{S}_1| \notin \{0, |s|\} \\ \frac{L h_N^{\beta - \lfloor \beta \rfloor}}{2} \left[D^{\mathcal{S}_1} K\left(\frac{x_1 - x_{10}}{h_N}\right) + D^{\mathcal{S}_1} K\left(\frac{x_1 - x_{20}}{h_N}\right) \right] & \text{if } |\mathcal{S}_1| = |s| \\ \frac{L h_N^{\beta - \lfloor \beta \rfloor}}{2} \left[D^{\mathcal{S}_2} K\left(\frac{x_2 - x_{10}}{h_N}\right) + D^{\mathcal{S}_2} K\left(\frac{x_2 - x_{20}}{h_N}\right) \right] & \text{if } |\mathcal{S}_1| = 0 \end{cases}.
\end{aligned}
$$

Therefore, if $|\mathcal{S}_1| \notin \{0, |s|\}$, then $|D^s g_{1N}(w) - D^s g_{1N}(w')| = 0$; if $|\mathcal{S}_1| = |s|$, then

$$
\begin{aligned}
&|D^s g_{1N}(w) - D^s g_{1N}(w')| \\
&= \frac{L h_N^{\beta - \lfloor \beta \rfloor}}{2} \left[\left| D^{\mathcal{S}_1} K\left(\frac{x_1 - x_{10}}{h_N}\right) - D^{\mathcal{S}_1} K\left(\frac{x_1' - x_{10}}{h_N}\right) \right| + \left| D^{\mathcal{S}_1} K\left(\frac{x_1 - x_{20}}{h_N}\right) - D^{\mathcal{S}_1} K\left(\frac{x_1' - x_{20}}{h_N}\right) \right| \right] \\
&\leq L ||x_1 - x_1'||_\infty^{\beta - \lfloor \beta \rfloor} \\
&\leq L ||w - w'||_\infty^{\beta - \lfloor \beta \rfloor};
\end{aligned}
$$

and, finally, if $|\mathcal{S}_1| = 0$, then

$$
\begin{aligned}
&|D^s g_{1N}(w) - D^s g_{1N}(w')| \\
&= \frac{L h_N^{\beta - \lfloor \beta \rfloor}}{2} \left[\left| D^{\mathcal{S}_2} K\left(\frac{x_2 - x_{10}}{h_N}\right) - D^{\mathcal{S}_2} K\left(\frac{x_2' - x_{10}}{h_N}\right) \right| + \left| D^{\mathcal{S}_2} K\left(\frac{x_2 - x_{20}}{h_N}\right) - D^{\mathcal{S}_2} K\left(\frac{x_2' - x_{20}}{h_N}\right) \right| \right] \\
&\leq L ||x_2 - x_2'||_\infty^{\beta - \lfloor \beta \rfloor} \\
&\leq L ||w - w'||_\infty^{\beta - \lfloor \beta \rfloor}.
\end{aligned}
$$

Hence $g_{1N} \in \Sigma(\beta, L)$. We also have that $g_{0N} \in \Sigma(\beta, L)$ by inspection.

Verification of (b): $d\left(\theta\left(P_{0N}\right), \theta\left(P_{1N}\right)\right) = |g_{1N}(w) - g_{0N}(w)| \geq 2A\psi_N$ **with** $\psi_N = N^{-\frac{\beta}{2\beta + d}}$

Here we check that our hypotheses are $2s$-separated. We have that

$$
\begin{aligned}
|g_{1N}(w) - g_{0N}(w)| &= \frac{L h_N^{\beta}}{2} \left[2K(0) + K\left(\frac{x_{10} - x_{20}}{h_N}\right) + K\left(\frac{x_{20} - x_{10}}{h_N}\right) \right] \geq 2 L h_N^{\beta} K(0) \\
&= L K(0) c_0^{\beta} \psi_N,
\end{aligned}
$$

and hence condition (b) holds with $A = \frac{LK(0)c_0^\beta}{2}$.

Verification of (c): $\mathrm{KL}(P_{0N}, P_{1N}) \leq \alpha < \infty$

This condition allows for the application of part (iii) of Theorem 2.2 in Tsybakov (2008). We begin by establishing some helpful notation. Let $\mathbf{Y} = [Y_{ij}]_{1\leq i,j\leq N}$ be the $N \times N$ adjacency matrix; $\mathbf{G}_k = [g_{kN}(W_{ij})]_{1\leq i,j\leq N}$ for $k = 0, 1$ the associated matrices of regression functions for the two sequences of hypotheses; and $\mathbf{V} = [V_{ij}]_{1\leq i,j\leq N}$ the corresponding matrix of dyadic-specific disturbances. Note the diagonals of each of these matrices consist of "structural" zeros. Further let $\mathbf{U} = [U_i]_{1\leq i\leq N}$ be the $N \times 1$ vector of agent-specific disturbances. Finally let $\mathbf{K}$ be the $N \times 1$ vector with i^{th} element $\frac{Lh_N^\beta}{2}\left[K\left(\frac{X_i - x_{10}}{h_N}\right) + K\left(\frac{X_i - x_{20}}{h_N}\right)\right]$.

Let ι_J denote a $J \times 1$ vector of ones, $\underset{\sim}{0}_{K,J}$ a $K \times J$ matrix of zeros, and I_J the $J \times J$ identity matrix. We also define the following selection matrices:

$$\mathcal{T}_1 = \begin{pmatrix} \iota_{N-1} & 0 & 0 & \cdots & 0 & 0 \\ \underset{\sim}{0} & \iota_{N-2} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}_{\binom{N}{2}\times N}, \quad \mathcal{T}_2 = \begin{pmatrix} \underset{\sim}{0}_{N-1,1} & I_{N-1} \\ \underset{\sim}{0}_{N-2,2} & I_{N-2} \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}_{\binom{N}{2}\times N},$$

from which we form $\mathcal{T} = \mathcal{T}_1 + \mathcal{T}_2$ and, finally, $\mathbf{T} = \iota_2 \otimes \mathcal{T}$. Next let $\mathbf{y} = (\mathrm{vech}(\mathbf{Y}')', \mathrm{vech}(\mathbf{Y})')'$ be the $N(N-1) \times 1$ vectorization of the dyadic outcomes. Similarly let $\mathbf{g}_k$ for $k = 0, 1$ and $\mathbf{v}$ be the corresponding vectorizations of, respectively, $\mathbf{G}_k$ and $\mathbf{V}$.

Using this notation we can write the $N(N-1) \times 1$ vector of composite regression errors $e_{ij} = U_i + U_j + V_{ij}$ as $\mathbf{e} = \mathbf{TU} + \mathbf{v}$ and its variance covariance matrix as

$$\Omega = \mathrm{Var}(\mathbf{e}) = \mathbf{I}_{N(N-1)\times N(N-1)} + \mathbf{TT}^T.$$

Under P_{0N} we have that

$$\mathbf{g}_0 = 0, \ \mathbf{y} = \mathbf{e}, \ \mathbf{y}|\mathbf{X} \sim \mathrm{N}(0, \Omega).$$

While under P_{1N} we instead have that

$$\mathbf{g}_1 = \mathbf{TK}, \ \mathbf{y} = \mathbf{TK} + \mathbf{e}, \ \mathbf{y}|\mathbf{X} \sim \mathrm{N}(\mathbf{TK}, \Omega).$$

Let $K_{\max} = \max_u K(u)$ and recall that $h_N = c_0 N^{-\frac{1}{2\beta + d_X}}$. We can now evaluate the KL

divergence as follows:

$$\begin{aligned}
\mathrm{KL}\left(P_{0N}, P_{1N}\right) &= \int \log \frac{\mathrm{d}P_{0N}}{\mathrm{d}P_{1N}} \mathrm{d}P_{0N} \\
&= \int \log \frac{p_{0N}(\mathbf{y}|\mathbf{X})}{p_{1N}(\mathbf{y}|\mathbf{X})} \mathrm{d}P_{0N} \\
&= -\frac{1}{2} \int \mathbf{y}^\top \Omega^{-1} \mathbf{y} - (\mathbf{y} - \mathbf{g}_1)^\top \Omega^{-1} (\mathbf{y} - \mathbf{g}_1) \mathrm{d}P_{0N} \\
&= \frac{1}{2} \int \mathbf{g}_1^\top \Omega^{-1} \mathbf{g}_1 \mathrm{d}P_{0N} \\
&= \frac{1}{2} \mathbb{E}_{P_{0N}} \left[\mathbf{K}^\top \mathbf{T}^\top (\mathbf{I} + \mathbf{T}\mathbf{T}^\top)^{-1} \mathbf{T}\mathbf{K} \right] \\
&\leq \frac{1}{2} \mathbb{E}_{P_{0N}} \left[\mathbf{K}^\top \mathbf{K} \right] \\
&\leq \frac{1}{2} L^2 K_{\max}^2 B_3 h_N^{2\beta + d_X} N \\
&= \frac{1}{2} L^2 K_{\max}^2 B_3 c_0^{2\beta + d_X},
\end{aligned} \tag{6}$$

for N large enough such that $N h_N^{d_X} \geq 1$ and $L K_{\max} h_N^{2\beta}$ bounded above.

In the derivation above, the third equality follows from the form of the multivariate normal density. The weak inequality in line six holds because

$$\begin{aligned}
\mathbf{K}^\top \mathbf{K} - \mathbf{K}^\top \mathbf{T}^\top (\mathbf{I} + \mathbf{T}\mathbf{T}^\top)^{-1} \mathbf{T}\mathbf{K} &= \mathbf{K}^\top \left[\mathbf{I}_N - \mathbf{T}^\top (\mathbf{I} + \mathbf{T}\mathbf{T}^\top)^{-1} \mathbf{T} \right] \mathbf{K} \\
&= \mathbf{K}^\top \left[\mathbf{I}_N + \mathbf{T}^\top \mathbf{T} \right]^{-1} \mathbf{K} \\
&\geq 0.
\end{aligned}$$

Finally, the weak inequality in line seven holds because, using condition (5) above,

$$\begin{aligned}
&\mathbb{E}\left[\left(K\left(\frac{X_i - x_{10}}{h_N} \right) + K\left(\frac{X_i - x_{20}}{h_N} \right) \right)^2 \right] \\
&\leq 2\mathbb{E}\left[\left(K\left(\frac{X_i - x_{10}}{h_N} \right) \right)^2 + \left(K\left(\frac{X_i - x_{20}}{h_N} \right) \right)^2 \right] \\
&= 2 \int \left(K\left(\frac{x - x_{10}}{h_N} \right) \right)^2 + \left(K\left(\frac{x - x_{20}}{h_N} \right) \right)^2 \mathrm{d}F(x) \\
&\leq 2K_{\max}^2 \int \mathbb{1}\left(\left| \frac{x - x_{10}}{h_N} \right| \leq \frac{1}{2} \right) + \mathbb{1}\left(\left| \frac{x - x_{20}}{h_N} \right| \leq \frac{1}{2} \right) \mathrm{d}F(x) \\
&= 2K_{\max}^2 h_N^{d_X} \left[\int \mathbb{1}\left(|u| \leq \frac{1}{2} \right) \left[f(x_{10} + h_N u) + f(x_{20} + h_N u) \right] \mathrm{d}u \right] \\
&\leq 4 h_N^{d_X} B_3 K_{\max}^2,
\end{aligned}$$

and where it is also helpful to remind oneself of the definition of $\mathbf{K}$ given earlier.

If we take $c_0 = \left(\frac{2\alpha}{L^2 K_{max}^2 B_3}\right)^{\frac{1}{2\beta+d_X}}$, then we obtain $\mathrm{KL}\left(P_{0N}, P_{1N}\right) \leq \alpha$. This result, and condition (b) above, gives – invoking equations (2.7) and (2.9) on p. 29 of Tsybakov (2008) as well as part (iii) of his Theorem 2.2:

$$\inf_{\hat{g}_N} \sup_{g \in \Sigma(\beta, L)} \mathbb{E}_g \left[1\left(|g_{1N}(w) - g_{0N}(w)| \geq A\psi_N\right)\right] \geq \max\left(\frac{1}{4}\exp\left(-\alpha\right), \frac{1-\sqrt{\frac{\alpha}{2}}}{2}\right)$$

for N large enough. Some rearrangement and the Markov Inequality then yield

$$\inf_{\hat{g}_N} \sup_{g \in \Sigma(\beta, L)} \mathbb{E}_g \left[N^{\frac{2\beta}{2\beta+d_X}} \left(g_{1N}(w) - g_{0N}(w)\right)^2\right] \geq A^2 \max\left(\frac{1}{4}\exp\left(-\alpha\right), \frac{1-\sqrt{\frac{\alpha}{2}}}{2}\right).$$

Since the constant to the right of the inequality only depends on β and L part (i) of the Theorem follows after taking the limit inferior of the expression above as $N \to \infty$.

Proof of statement (ii)

Again let P_{kN} be the probability measure of the observed data $(X_i, Y_{ij}, 1 \leq i \neq j \leq N)$ with the regression function g_{kN}. Theorem 2.5 of Tsybakov (2008) implies that part (ii) will hold if we can construct sequences of hypotheses $P_{0N}, P_{1N}, \ldots, P_{M_N N}$ such that

(a) $g_{0N}, g_{kN} \in \Sigma(\beta, L)$, $k = 1, \ldots, M_N$;

(b) $d\left(\theta_k, \theta_l\right) = ||g_{kN} - g_{lN}||_\infty \geq 2A\psi_N$, $\psi_N = \left(\frac{N}{\ln N}\right)^{-\frac{\beta}{2\beta+d}}$ and $\theta_k = g_{kN}$ and $\theta_l = g_{lN}$ for $k \neq l$ and $k, l = 1, \ldots, M_N$;

(c) $\frac{1}{M_N}\sum_{k=1}^{M_N} KL(P_{kN}, P_{0N}) \leq \alpha \ln M_N$.

Define the hypotheses:

$$\begin{aligned} g_{0N} &: (x_1, x_2) \to 0 \\ g_{kN} &: (x_1, x_2) \to L h_N^\beta \left[K\left(\frac{x_1 - x_{kN}}{h_N}\right) + K\left(\frac{x_2 - x_{kN}}{h_N}\right)\right] \end{aligned}$$

where $k \in \mathcal{I}_N = \{1, 2, \ldots, m_N\}^{d_X}$, $h_N = c_0\left(\frac{N}{\ln N}\right)^{-\frac{1}{2\beta+d_X}}$, $m_N = \lceil h_N^{-1} \rceil$, $M_N = |\mathcal{I}_N| = m_N^{d_X}$, and for $k = (k_1, k_2, \ldots, k_d)$, $x_{kN} = \left(\frac{k_1 - 1/2}{m_N}, \frac{k_2 - 1/2}{m_N}, \ldots, \frac{k_d - 1/2}{m_N}\right)$, the function $K : \mathbb{R}^{d_X} \to [0, \infty)$ satisfies (5). Notice the supports of these functions for the same N are disjoint. The results follows by verifying conditions (a), (b) and (c). We have already shown that condition

(a) holds in the proof of part (i). The condition (b) holds with $A = LK(0)c_0^\beta$ because

$$||g_{kN} - g_{lN}||_\infty \geq |g_{kN}(x_{kN}, x_{kN}) - g_{lN}(x_{kN}, x_{kN})| = 2Lh_N^\beta K(0) = 2LK(0)c_0^\beta \psi_N.$$

To verify condition (c) we evaluate the KL-divergence:

$$\begin{aligned}
\frac{1}{M_N}\sum_{k\in\mathcal{I}_N} \mathrm{KL}(P_{kN}, P_{0N}) &\leq \frac{1}{M_N}\sum_{k\in\mathcal{I}_N} \frac{1}{2}\mathbb{E}_{P_{0N}}\left[\mathbf{K}_k^\top \mathbf{K}_k\right] \\
&\leq \frac{1}{M_N}\sum_{k\in\mathcal{I}_N} 2L^2 h_N^{2\beta} K_{\max}^2 \sum_{i=1}^N \int \mathbb{1}\left(\left|\frac{x_i - x_{kN}}{h_N}\right| \leq \frac{1}{2}\right) \mathrm{d}F(x_i) \\
&= \frac{1}{M_N} 2L^2 h_N^{2\beta} K_{\max}^2 \sum_{i=1}^N \int \sum_{k\in\mathcal{I}_N} \mathbb{1}\left(\left|\frac{x_i - x_{kN}}{h_N}\right| \leq \frac{1}{2}\right) \mathrm{d}F(x_i) \\
&\leq 2L^2 h_N^{2\beta+d_X} K_{\max}^2 N \\
&= 2L^2 K_{\max}^2 c_0^{2\beta+d_X} \ln N.
\end{aligned}$$

The first and second line are proved in part (i). The fourth line use the fact that the support of functions $g_{kN}, k \in \mathcal{I}_N$ are disjoint and $\sum_{k\in\mathcal{I}_N} \mathbb{1}\left(\left|\frac{x_i - x_{kN}}{h_N}\right| \leq \frac{1}{2}\right) \leq 1$. We have $\ln M_N = \ln(m_N^{d_X}) \geq \frac{d_X}{2\beta+d_X} \ln\left(\frac{N}{\ln N}\right) - d_X \ln c_0 \geq \frac{d_X}{2\beta+d_X+1} \ln N$ for sufficiently large N. The condition is thus satisfied with sufficiently large c_0. The result follows from Theorem 2.5 of Tsybakov (2008).

Proof of Theorem 3.1

Applying the variance operator to $\hat{\Psi}(w)$ yields

$$\mathbb{V}\left(\hat{\Psi}(w)\right) = \frac{4}{N}\frac{N-2}{N-1} V_{N,1} + \binom{N}{2}^{-1} V_{N,2}$$

where, starting with the second term,

$$\begin{aligned}
V_{N,2} &= \mathbb{V}\left(\frac{1}{2}\left[Y_{12}K_{12} + Y_{21}K_{21}\right]\right) \leq \mathbb{V}\left(Y_{12}K_{12}\right) \leq \mathbb{E}\left(Y_{12}^2 K_{12}^2\right) \\
&= h_N^{-4d_X} \int \mathbb{E}\left[Y_{12}^2 | (X_1, X_2) = (x_1, x_2)\right] K^2\left(\frac{x - x_1}{h_N}, \frac{x - x_2}{h_N}\right) f(x_1) f(x_2) \mathrm{d}x_1 \mathrm{d}x_2 \\
&= h_N^{-2d_X} \int \mathbb{E}\left[Y_{12}^2 | (X_1, X_2) = (x - h_N s_1, x' - h_N s_2)\right] f(x - h_N s_1) f(x' - h_N s_2) K^2(s_1, s_2)\, \mathrm{d}s_1 \mathrm{d}s_2 \\
&\leq h_N^{-2d_X} B_4 K_{\max} B_1.
\end{aligned}$$

Next, consider the first term. We get that

$$
\begin{aligned}
V_{N,1} &= \mathbb{C}\left(\frac{1}{2}\left(Y_{12}K_{12}+Y_{21}K_{21}\right),\frac{1}{2}\left(Y_{13}K_{13}+Y_{31}K_{31}\right)\right)\\
&= \mathbb{V}\left(\mathbb{E}\left[\frac{1}{2}\left(Y_{12}K_{12}+Y_{21}K_{21}\right)\middle|X_1,U_1\right]\right)\\
&\leq \frac{1}{2}\operatorname{Var}\left(\mathbb{E}\left(Y_{12}K_{12}\middle|X_1,U_1\right)\right)+\frac{1}{2}\operatorname{Var}\left(\mathbb{E}\left(Y_{21}K_{21}\middle|X_1,U_1\right)\right)\\
&\leq \frac{1}{2}\mathbb{E}\left(Y_{12}K_{12}Y_{13}K_{13}\right)+\frac{1}{2}\mathbb{E}\left(Y_{21}K_{21}Y_{31}K_{31}\right)\\
&= \frac{1}{2}h_N^{-4d_X}\int \mathbb{E}\left(Y_{12}Y_{13}|(X_1,X_2,X_3)=(x_1,x_2,x_3)\right)\\
&\qquad\qquad \cdot K\left(\frac{x-x_1}{h_N},\frac{x'-x_2}{h_N}\right)K\left(\frac{x-x_1}{h_N},\frac{x'-x_3}{h_N}\right)f(x_1)f(x_2)f(x_3)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3\\
&\quad+\frac{1}{2}h_N^{-4d_X}\int \mathbb{E}\left(Y_{21}Y_{31}|(X_1,X_2,X_3)=(x_1,x_2,x_3)\right)\\
&\qquad\qquad \cdot K\left(\frac{x-x_2}{h_N},\frac{x'-x_1}{h_N}\right)K\left(\frac{x-x_3}{h_N},\frac{x'-x_1}{h_N}\right)f(x_1)f(x_2)f(x_3)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3\\
&= h_N^{-d_X}\frac{1}{2}\int \mathbb{E}\left(Y_{12}Y_{13}|(X_1,X_2,X_3)=(x-h_Ns_1,x'-h_Ns_2,x'-h_Ns_3)\right)\\
&\qquad\qquad \cdot f(x-h_Ns_1)f(x'-h_Ns_2)f(x'-h_Ns_3)K\left(s_1,s_2\right)K\left(s_1,s_3\right)\mathrm{d}s_1\mathrm{d}s_2\mathrm{d}s_3\\
&\quad+h_N^{-d_X}\frac{1}{2}\int \mathbb{E}\left(Y_{21}Y_{31}|(X_1,X_2,X_3)=(x'-h_Ns_1,x-h_Ns_2,x-h_Ns_3)\right)\\
&\qquad\qquad \cdot f(x'-h_Ns_1)f(x-h_Ns_2)f(x-h_Ns_3)K\left(s_1,s_2\right)K\left(s_1,s_3\right)\mathrm{d}s_1\mathrm{d}s_2\mathrm{d}s_3\\
&\leq h_N^{-d_X}B_5\int |K\left(s_1,s_2\right)||K\left(s_1,s_3\right)|\mathrm{d}s_1\mathrm{d}s_2\mathrm{d}s_3\\
&\leq h_N^{-d_X}B_5B_2B_1. \qquad (7)
\end{aligned}
$$

These two bounds imply the variance bound

$$
\begin{aligned}
\mathbb{V}\left(\hat{\Psi}(w)\right) &\leq \binom{N}{2}^{-1}h_N^{-2d_X}B_4K_{\max}B_1+\frac{4(N-2)}{N(N-1)}h_N^{-d_X}B_5B_2B_1\\
&= N^{-1}h_N^{-d_X}\left[\frac{N-2}{N-1}4B_5B_2B_1+N^{-1}h_N^{-d_X}\frac{4N}{N-1}B_4K_{\max}B_1\right],
\end{aligned}
$$

which, in turn, implies that for $M_0 = 4B_5B_2B_1+1$ and sufficiently large N, $\mathbb{V}\left(\hat{\Psi}(w)\right) \leq \frac{M_0}{Nh_N^{d_X}}$ for all $w \in \mathbb{R}^{d_W}$ as claimed.

Proof of Theorem 3.2

For τ_N a sequence of positive truncation parameters we consider the sum

$$\tilde{\Psi}_N(w) = \frac{1}{\binom{N}{2}} \sum_{1\le i<j\le N} \frac{1}{2}\left[Y_{ij}\cdot \mathbb{1}\left(|Y_{ij}|<\tau_N\right) \frac{1}{h_N^{d_W}} K\left(\frac{w-W_{ij}}{h_N}\right) \right.$$
$$\left. + Y_{ji}\cdot \mathbb{1}\left(|Y_{ji}|<\tau_N\right) \frac{1}{h_N^{d_W}} K\left(\frac{w-W_{ji}}{h_N}\right)\right].$$

We will use $\tilde{Z}_{N,ij}$ to denote the summands in the above expression in what follows. The Hoeffding decomposition of this U-like statistic is

$$\tilde{\Psi}(w) = \mathbb{E}\tilde{\Psi}(w) + \underbrace{\frac{2}{N}\sum_{i=1}^{N} \bar{Z}_{N,i}}_{T_{N,1}(w)} + \underbrace{\frac{1}{\binom{N}{2}} \sum_{1\le i<j\le N} \breve{Z}_{N,ij}}_{T_{N,2}(w)},$$

where

$$\bar{Z}_{N,i} = \mathbb{E}\left[\tilde{Z}_{N,ij}\middle| X_i, U_i\right] - \mathbb{E}\tilde{Z}_{N,ij}$$
$$\breve{Z}_{N,ij} = \tilde{Z}_{N,ij} - \mathbb{E}\left[\tilde{Z}_{N,ij}\middle| X_i, U_i\right] - \mathbb{E}\left[\tilde{Z}_{N,ij}\middle| X_j, U_j\right] + \mathbb{E}\tilde{Z}_{N,ij}.$$

Notice that $T_{N,1}(w)$ is an average of N iid mean-zero random variables while $T_{N,2}(w)$ is a degenerate second-order U-like statistic.

To proceed further we require the following Lemma.

Lemma 3.4. *Under Assumptions 3.1 and 3.2, for any $\alpha > 0$, there exists constant M_α such that*

(i) if $\tau_N \ll a_N^{-1}$, then $\sup_{w\in\mathbb{R}^{d_W}} P\left(|T_{N,1}(w)| > M_\alpha a_N\right) = O\left(N^{-\alpha}\right)$;

(ii) if $\tau_N \ll N h^{\frac{3}{2}d_X}/\ln N$ and $a_N = o(1)$, then $\sup_{w\in\mathbb{R}^{d_W}} P\left(|T_{N,2}(w)| > M_\alpha a_N\right) = O\left(N^{-\alpha}\right)$;

(iii) if for some $s>1$, $\sup_{x_1,x_2\in\mathbb{R}^{d_X}} \mathbb{E}\left[|Y_{12}|^s\middle|(X_1,X_2)=(x_1,x_2)\right]\cdot f(x_1,x_2) \le B_{4,s} < \infty$ and $\tau_N \gg a_N^{-\frac{1}{s-1}}$, then $\sup_{w\in\mathbb{R}^{d_W}} \left|\mathbb{E}\left(\hat{\Phi}(w) - \tilde{\Phi}(w)\right)\right| = o\left(a_N\right)$;

(iv) if for some $s>1$, $\mathbb{E}\left|Y_{12}\right|^s \le B_{6,s}$ and $\tau_N \gg \left(a_N h_N^{2d_X}\right)^{-\frac{1}{s-1}}$, then $\sup_{w\in\mathbb{R}^{d_W}} \left|\hat{\Phi}_N(w) - \tilde{\Phi}_N(w)\right| = o_P\left(a_N\right)$;

(v) if for some $s>2$, $\tau_N = \left(N^2\phi_N\right)^{\frac{1}{s}}$ where $\phi_N = (\ln(\ln N))^2 \ln N$, and $\mathbb{E}\left|Y_{12}\right|^s \le B_{6,s}$, then $P(\hat{\Phi}_N = \tilde{\Phi}_N) = P\left(\hat{\Phi}_N(w) = \tilde{\Phi}_N(w), \forall w\in\mathbb{R}^{2d_X}\right) \to 1$ as $N\to\infty$.

The proof of the above Lemma may be found below. The bandwidth conditions stated in the hypotheses of Theorem 3.2 ensure that we can pick truncation thresholds τ_N which satisfy the following conditions

1. $\tau_N \ll a_N^{-1}$;
2. $\tau_N \ll \frac{N}{\ln N} h_N^{\frac{3}{2}d_X}$;
3. $\tau_N \gg a_N^{-\frac{1}{s-1}}$;
4. $\tau_N \gg (N^2\phi_N)^{\frac{1}{s}}$ or $\tau_N \gg (a_N h_N^{2d_X})^{-\frac{1}{s-1}}$.

These conditions allow for the application of Lemma 3.4. Denote $R_N(w) := \hat{\Psi}_N(w) - \tilde{\Psi}_N(w)$. For any set $\mathcal{C}_N \subset \mathbb{R}^{2d}$,

$$\begin{aligned}
&P\left(\sup_{w\in\mathcal{C}_N}\left|\hat{\Psi}_N(w) - \mathbb{E}\hat{\Psi}_N(w)\right| > 8Ma_N\right) \\
&= P\left(\sup_{w\in\mathcal{C}_N}\left|\tilde{\Psi}_N(w) - \mathbb{E}\tilde{\Psi}_N(w) + R_N(w) - \mathbb{E}R_N(w)\right| > 8Ma_N\right) \\
&\leq P\left(\sup_{w\in\mathcal{C}_N}\left|\tilde{\Psi}_N(w) - \mathbb{E}\tilde{\Psi}_N(w)\right| > 6Ma_N\right) + P\left(\sup_{w\in\mathcal{C}_N}|R_N(w) - \mathbb{E}R_N(w)| > 2Ma_N\right). \quad (8)
\end{aligned}$$

The second term in inequality (8) converges to zero because

$$\begin{aligned}
&P\left(\sup_{w\in\mathcal{C}_N}|R_N(w) - \mathbb{E}R_N(w)| > 2Ma_N\right) \\
&\leq P\left(\sup_{w\in\mathbb{R}^{d_W}}|R_N(w) - \mathbb{E}R_N(w)| > 2Ma_N\right) \\
&\leq P\left(\sup_{w\in\mathbb{R}^{d_W}}|R_N(w)| > Ma_N\right) + \mathbb{1}\left(\sup_{w\in\mathbb{R}^{d_W}}|\mathbb{E}R_N(w)| > Ma_N\right) \quad (9) \\
&= o(1).
\end{aligned}$$

The last line holds because

$$\mathbb{1}\left(\sup_{w\in\mathbb{R}^{d_W}}|\mathbb{E}R_N(w)| > Ma_N\right) = 0 \qquad \text{for large } N \quad (10)$$

$$P\left(\sup_{w\in\mathbb{R}^{d_W}}|R_N(w)| > Ma_N\right) = o_P(1). \quad (11)$$

To see (10), notice part (iii) of Lemma 3.4 implies that $\sup_{w\in\mathbb{R}^{d_W}}|\mathbb{E}R_N(w)| = o(a_N)$.

Hence $\mathbb{1}\left(\sup_{w\in\mathbb{R}^{2d}}|\mathbb{E}R_N(w)|>Ma_N\right)=0$ for large N. To see (11), notice the inequality

$$P\left(\sup_{w\in\mathbb{R}^{d_W}}|R_N(w)|>Ma_N\right)\leq\min\left\{1-P(\hat{\Phi}_N=\tilde{\Phi}_N),\frac{\mathbb{E}\sup_{w\in\mathbb{R}^{d_W}}|R_N(w)|}{Ma_N}\right\},$$

suggests we can bound either term on the right-hand side to bound the term on the left-hand side. The threshold we pick meets the conditions of both parts (iv) and (v) of Lemma 3.4, which ensures either $1-P(\hat{\Phi}_N=\tilde{\Phi}_N)=o(1)$ or $\frac{\mathbb{E}\sup_{w\in\mathbb{R}^{d_W}}|R_N(w)|}{Ma_N}=o(1)$. This implies (11).

To show the first term in inequality (8) converges to zero, we will use a covering argument to reduce finding the supremum over an infinite number points to finding the maximum over a finite number of points. We then invoke point-wise concentration bounds. This part closely follows the argument in Hansen (2008). Cover any compact region $\mathcal{C}_N\subset\mathbb{R}^{d_W}$ by finite number of balls of radius a_Nh_N centered at grid points in the set $L_N=\{w_{N,1},w_{N,2},\ldots,w_{N,L_N}\}$ (Here we abuse the notation a bit: L_N is used to refer to both the set and its cardinality). Denote the ball $A_{N,j}=\{w\in\mathbb{R}^{d_W}:||w-w_{N,j}||\leq a_Nh_N\}$. For N large enough such that $a_N<L$ (L is the constant appearing in Assumption 3.3), for any point $w\in A_{N,j}$ within the ball, assumption 3.3 (ii) implies

$$\left|K\left(\frac{w-W_{ij}}{h}\right)-K\left(\frac{w_{N,j}-W_{ij}}{h}\right)\right|\leq a_NK^*\left(\frac{w_{N,j}-W_{ij}}{h}\right). \tag{12}$$

Define

$$\breve{\Phi}_N(w):=\frac{1}{N(N-1)}\sum_{1\leq i\neq j\leq N}Y_{ij}\cdot\mathbb{1}\left(|Y_{ij}|<\tau_N\right)\frac{1}{h^{d_W}}K^*\left(\frac{w-W_{ij}}{h}\right),$$

which is a version of $\tilde{\Phi}(w)$ with K replaced by K^*. The bound (12) implies

$$\left|\tilde{\Psi}_N(w)-\tilde{\Psi}_N(w_{N,j})\right|\leq a_N\breve{\Phi}_N(w_{N,j}),$$

with $|\mathbb{E}\breve{\Phi}_N(w_{N,j})|\leq B_4^{1/2}B_3^{1/2}\int|K^*(w)|\mathrm{d}w<\infty$. Next bound the sup within the ball by a

value at the center and the sup discrepancy

$$
\begin{aligned}
&\sup_{w\in A_{N,j}}\left|\tilde{\Psi}_N(w)-\mathbb{E}\tilde{\Psi}_N(w)\right| \\
&\leq \left|\tilde{\Psi}_N(w_{N,j})-\mathbb{E}\tilde{\Psi}_N(w_{N,j})\right| + \sup_{w\in A_{N,j}}\left|\tilde{\Psi}_N(w)-\tilde{\Psi}_N(w_{N,j})\right| + \sup_{w\in A_{N,j}}\left|\mathbb{E}\left(\tilde{\Psi}_N(w)-\tilde{\Psi}_N(w_{N,j})\right)\right| \\
&\leq \left|\tilde{\Psi}_N(w_{N,j})-\mathbb{E}\tilde{\Psi}_N(w_{N,j})\right| + a_N\left[\breve{\Phi}_N(w_{N,j})+\mathbb{E}\breve{\Phi}_N(w_{N,j})\right] \\
&\leq \left|\tilde{\Psi}_N(w_{N,j})-\mathbb{E}\tilde{\Psi}_N(w_{N,j})\right| + a_N\left|\breve{\Phi}_N(w_{N,j})-\mathbb{E}\breve{\Phi}_N(w_{N,j})\right| + 2a_N\mathbb{E}\breve{\Phi}_N(w_{N,j}) \\
&\leq \left|\tilde{\Psi}_N(w_{N,j})-\mathbb{E}\tilde{\Psi}_N(w_{N,j})\right| + \left|\breve{\Phi}_N(w_{N,j})-\mathbb{E}\breve{\Phi}_N(w_{N,j})\right| + 2a_N\mathbb{E}\breve{\Phi}_N(w_{N,j}).
\end{aligned}
$$

The last inequality follow because $a_N \leq 1$ for N large enough. For any constant $M \geq B_4^{1/2}B_3^{1/2}\int|K^*(w)|\mathrm{d}w \geq \mathbb{E}\breve{\Phi}_N(w_{N,j})$,

$$
\begin{aligned}
&P\left(\sup_{w\in A_{N,j}}\left|\tilde{\Psi}_N(w)-\mathbb{E}\tilde{\Psi}_N(w)\right| > 6Ma_N\right) \\
&\leq P\left(\left|\tilde{\Psi}_N(w_{N,j})-\mathbb{E}\tilde{\Psi}_N(w_{N,j})\right| + \left|\breve{\Phi}_N(w)-\mathbb{E}\breve{\Phi}_N(w)\right| + 2a_N\mathbb{E}\breve{\Phi}_N(w) > 6Ma_N\right) \\
&\leq P\left(\left|\tilde{\Psi}_N(w_{N,j})-\mathbb{E}\tilde{\Psi}_N(w_{N,j})\right| > 2Ma_N\right) + P\left(\left|\breve{\Phi}_N(w)-\mathbb{E}\breve{\Phi}_N(w)\right| > 2Ma_N\right),
\end{aligned}
$$

as well as

$$
\begin{aligned}
&P\left(\sup_{w\in\mathcal{C}_N}\left|\tilde{\Psi}_N(w)-\mathbb{E}\tilde{\Psi}_N(w)\right| > 6Ma_N\right) \\
&\leq \sum_{j=1}^{L_N} P\left(\sup_{w\in A_{N,j}}\left|\tilde{\Psi}_N(w)-\mathbb{E}\tilde{\Psi}_N(w)\right| > 6Ma_N\right) \\
&\leq L_N \max_{j\in\{1,2,\ldots,L_N\}} P\left(\sup_{w\in A_{N,j}}\left|\tilde{\Psi}_N(w)-\mathbb{E}\tilde{\Psi}_N(w)\right| > 6Ma_N\right) \\
&\leq L_N \max_{j\in\{1,2,\ldots,L_N\}} P\left(\left|\tilde{\Psi}_N(w_{N,j})-\mathbb{E}\tilde{\Psi}_N(w_{N,j})\right| > 2Ma_N\right) \\
&\qquad + L_N \max_{j\in\{1,2,\ldots,L_N\}} P\left(\left|\breve{\Phi}_N(w)-\mathbb{E}\breve{\Phi}_N(w)\right| > 2Ma_N\right). \qquad (13)
\end{aligned}
$$

We now bound the two terms in (13) using the same argument, as both K and K^* satisfy Assumption 3.1, and this is the only property of the function K or K^* we will use. For any

$\alpha > 0$ and M_α as in Lemma 3.4, for any $w \in \mathbb{R}^{d_W}$

$$\begin{aligned}\sup_{w\in\mathbb{R}^{d_W}} P\left(\left|\tilde{\Psi}_N(w) - \mathbb{E}\tilde{\Psi}_N(w)\right| > 2M_\alpha a_N\right) &= \sup_{w\in\mathbb{R}^{d_W}} P\left(|T_{N,1}(w) + T_{N,2}(w)| > 2M_\alpha a_N\right) \\ &\leq \sup_{w\in\mathbb{R}^{d_W}} P\left(|T_{N,1}(w)| > M_\alpha a_N\right) \\ &\quad + \sup_{w\in\mathbb{R}^{d_W}} P\left(|T_{N,2}(w)| > M_\alpha a_N\right) \\ &= O\left(N^{-\alpha}\right).\end{aligned}$$

Hence

$$P\left(\sup_{w\in\mathcal{C}_N}\left|\tilde{\Psi}_N(w) - \mathbb{E}\tilde{\Psi}_N(w)\right| > 6Ma_N\right) \leq O\left(L_N N^{-\alpha}\right).$$

If we take $\mathcal{C}_N = \{w \in \mathbb{R}^{d_W} : ||w|| < c_N\}$ where $c_N = N^q$, then $\mathcal{C}_N$ can be covered by $L_N = 2\left(\frac{c_N}{a_N h_N}\right)^{d_W}$ number of balls with radius $a_N h_N$. Hence we can take α large enough, e.g. $\alpha = (q + \frac{1}{2})d_W + 3$, so that $O\left(L_N N^{-\alpha}\right) = O\left(\left(\frac{c_N}{a_N h_N}\right)^{d_W} N^{-\alpha}\right) = O\left(N^{(q+\frac{1}{2})d_W + 2 - \alpha}\right) = O\left(N^{-1}\right) = o(1)$. We have therefore shown that

$$P\left(\sup_{w\in\mathcal{C}_N}\left|\tilde{\Psi}_N(w) - \mathbb{E}\tilde{\Psi}_N(w)\right| > 6Ma_N\right) = o(1). \tag{14}$$

Together the two bounds (9) and (14) imply that the right-hand side of equality (8) is $o(1)$. This is saying for sufficiently large $M < \infty$, we have $P\left(\sup_{w\in\mathcal{C}_N}\left|\hat{\Psi}_N(w) - \mathbb{E}\hat{\Psi}_N(w)\right| > 8Ma_N\right) = o(1)$, which is sufficient for

$$\sup_{||w||\leq c_N}\left|\hat{\Psi}_N(w) - \mathbb{E}\hat{\Psi}_N(w)\right| = O(a_N).$$

as required.

Proof of Lemma 3.4

Proof of claim (i)

To prove the first claim of the Lemma we will apply the classic Bernstein's inequality (see equation 2.10 on p. 36 of the textbook Boucheron et al. (2013)).

> Let $Q_1, \ldots, Q_N$ be independent random variables with finite variance such that $Q_i \leq b$ for some $b > 0$ almost surely for all $i < N$. Let $S = \sum_{i=1}^N (Q_i - \mathbb{E}Q_i)$ and

$v = \sum_{i=1}^{N} \mathbb{E}\left[Q_i^2\right]$. Then for any $t > 0$,

$$P(S \geq t) \leq \exp\left(-\frac{t^2}{2(v + bt/3)}\right).$$

In order to invoke the inequality, we first show that $Q_i(w) := \tau_N^{-1} h_N^{d_X} \bar{Z}_{N,i}(w)$ is bounded. In the following we will use the abbreviated notation Q_i for $Q_i(w)$. Remember $\bar{Z}_{N,i}$ is the mean-normalized version of $\mathbb{E}\left[\tilde{Z}_{N,ij} \middle| X_i, U_i\right]$. Since

$$\begin{aligned}
\tau_N^{-1} h^{d_X} \left|\mathbb{E}\left[\tilde{Z}_{N,ij} \middle| X_i, U_i\right]\right| =& \tau_N^{-1} h^{d_X} \left|\mathbb{E}\left[\frac{1}{2}\left[Y_{ij} \cdot \mathbb{1}\left(|Y_{ij}| < \tau_N\right) K_{ij}\right.\right.\right. \\
&\left.\left.\left. + Y_{ji} \cdot \mathbb{1}\left(|Y_{ji}| < \tau_N\right) K_{ji}\right] \middle| X_i, U_i\right]\right| \\
\leq& h_N^{d_X} \frac{1}{2} \mathbb{E}\left[|K_{ij}| + |K_{ji}| \middle| X_i, U_i\right] \\
=& h_N^{-d_X} \frac{1}{2} \mathbb{E}\left[\left|K\left(\frac{w - W_{ij}}{h_N}\right)\right| + \left|K\left(\frac{w - W_{ji}}{h_N}\right)\right| \middle| X_i\right] \\
=& h_N^{-d_X} \frac{1}{2} \int \left[\left|K\left(\frac{x - x_i}{h_N}, \frac{x' - x_j}{h_N}\right)\right| + \left|K\left(\frac{x - x_j}{h_N}, \frac{x' - x_i}{h_N}\right)\right|\right] f(x_j) \mathrm{d}x_j \\
=& \frac{1}{2} \int \left|K\left(\frac{x - x_i}{h_N}, s\right)\right| f(x' - h_N s) + \left|K\left(s, \frac{x' - x_i}{h_N}\right)\right| f(x - h_N s) \mathrm{d}s \\
\leq& B_2 B_3,
\end{aligned}$$

we have $|Q_i| = |\tau_N^{-1} h_N^{d_X} \bar{Z}_{N,i}| < 2B_2B_3$. Write $P\left(T_{N,1}(w) > M a_N\right)$ in the form suitable for applying Bernstein's inequality

$$\begin{aligned}
P\left(T_{N,1}(w) > M a_N\right) &= P\left(\frac{2}{N} \sum_{i=1}^{N} \bar{Z}_{N,i} > M a_N\right) \\
&= P\left(\sum_{i=1}^{N} \tau_N^{-1} h_N^{d_X} \bar{Z}_{N,i} > \frac{M}{2} N h_N^{d_X} a_N \tau_N^{-1}\right) \\
&= P(S \geq t),
\end{aligned}$$

in which $S = \sum_{i=1}^{N} Q_i$ and $t = \frac{M}{2} N h_N^{d_X} a_N \tau_N^{-1}$. Applying Bernstein's inequality gives us $P(S \geq t) \leq \exp\left(-\frac{t^2}{2(v + bt/3)}\right)$ where $v := \sum_{i=1}^{N} \mathbb{E}\left[Q_i^2\right]$ and $b = 2B_2B_3$. Since the function $\exp\left(-\frac{t^2}{2(v+bt/3)}\right)$ is increasing in v, we have for any $v' > v$

$$P(S \geq t) \leq \exp\left(-\frac{t^2}{2(v' + bt/3)}\right). \tag{15}$$

The upper bound v' we are going to use is the following one

$$v = \sum_{i=1}^{N} \mathbb{E}\left[Q_i^2\right] = \sum_{i=1}^{N} \mathbb{E}\left[\left(\tau_N^{-1} h_N^{d_X} \bar{Z}_{N,i}\right)^2\right] = \tau_N^{-2} h_N^{2d_X} N V_{N,1} \leq \tau_N^{-2} N h_N^{d_X} B_5 B_2 B_1 := v',$$

in which the inequality is an implication of (7). Plugging the expression of v', t, b, and a_N into the RHS of (15) gives us

$$\exp\left(-\frac{t^2}{2(v' + bt/3)}\right) = \exp\left(-\frac{M^2}{8B_5 B_2 B_1 + 8B_2 B_3 M a_N \tau_N / 3} \ln N\right).$$

By assumption $a_N \tau_N \to 0$ as $N \to \infty$, we can pick N_0 such that $8B_2 B_3 a_N \tau_N / 3 \leq 1$ for any $N > N_0$. For any $\alpha > 0$, we can pick M large enough so that $\frac{M^2}{8B_5 B_2 B_1 + M} \geq \alpha$ and $\exp\left(-\frac{M^2}{8B_5 B_2 B_1 + M} \ln N\right) < N^{\alpha}$. In particular, $M_\alpha = \frac{\alpha + \sqrt{\alpha^2 + 32 B_5 B_2 B_1 \alpha}}{2}$ will work. This means we have proved

$$P\left(T_{N,1}(w) > M_\alpha a_N\right) = O\left(N^{-\alpha}\right).$$

We get the two-sided bound by applying the same argument twice for $T_{N,1}(w)$ and $-T_{N,1}(w)$. Moreover, because the derivation of the bound and the value of M_α doesn't depend on the specific point w, we have also proved our desired result

$$\sup_{w \in \mathbb{R}^{d_W}} P\left(|T_{N,1}(w)| > M_\alpha a_N\right) = O\left(N^{-\alpha}\right).$$

Proof of claim (ii)

We will use Propsition 2.3(c), a concentration inequality, from Arcones and Gine (1993)[2] to prove the second claim.

> Let $\{X_i, i \in \mathbb{N}\}$ and $\{V_{i_1,\ldots,i_m}, (i_1,\ldots,i_m) \in I_m^{\mathbb{N}}\}$ be independent random samples; $||f||_\infty \leq c$, $\mathbb{E}]f(X_1,\ldots,X_m,V_{1,\ldots,m})] = 0$, $\sigma^2 = \mathbb{E}[f^2(X_1,\ldots,X_m,V_{1,\ldots,m})]$; f is P-canonical, then there are constants c_i depending only on m such that for any

[2]There is a small modification compared to the original proposition. Since our statistic is not exactly a U-statistic as there are the iid V_{ij} variables in our setup, we include this additional term in the statement of inequality. The proof of the inequality in our setup could follow the same steps of the original Arcones and Gine (1993) one. The reason this works is that the V_{ij} terms are iid and won't affect the randomization inequality, decoupling inequality, and the hypercontractivity inequality used in the proof.

$t > 0$,

$$P\left(\left|N^{-m/2}\sum_{(i_1,\ldots,i_m)\in I_m^N} f(X_{i_1},\ldots,X_{i_m},V_{i_1,\ldots,i_m})\right| > t\right) \leq c_1 \exp\left(-\frac{c_2 t^{2/m}}{\sigma^{2/m} + \left(ct^{1/m}N^{-1/2}\right)^{2/(m+1)}}\right).$$

In order to apply the inequality, we first show that $\tau_N^{-1}h_N^{2d_X}\breve{Z}_{N,ij}$ is bounded. Decompose

$$\breve{Z}_{N,ij} = \tilde{Z}_{N,ij} - \mathbb{E}\left[\tilde{Z}_{N,ij}\middle|X_i,U_i\right] - \mathbb{E}\left[\tilde{Z}_{N,ij}\middle|X_j,U_j\right] + \mathbb{E}\tilde{Z}_{N,ij}.$$

The last three terms on the right-hand side are bounded because $\tau_N^{-1}\mathbb{E}\tilde{Z}_{N,ij} = O(1)$ and $\tau_N^{-1}\mathbb{E}\left[\tilde{Z}_{N,ij}\middle|X_i,U_i\right] = O(h_N^{-d_X})$. Moreover, $|\tau_N^{-1}h_N^{2d_X}\tilde{Z}_{N,ij}| = \frac{1}{2}|\tau_N^{-1}Y_{ij}\cdot \mathbb{1}\left(|Y_{ij}|<\tau_N\right)K\left(\frac{w-W_{ij}}{h_N}\right)| + \frac{1}{2}|\tau_N^{-1}Y_{ji}\cdot\mathbb{1}\left(|Y_{ji}|<\tau_N\right)K\left(\frac{w-W_{ji}}{h_N}\right)| \leq K_{\max}$. Hence, there exists constant $c>0$ s.t. $|\tau_N^{-1}h_N^{2d_X}\breve{Z}_{N,ij}| < c$. Applying the concentration inequality to $T_{N,2}(w)$ then gives us

$$\begin{aligned}
P\left(|T_{N,2}(w)| > Ma_N\right) &= P\left(\left|\frac{1}{\binom{N}{2}}\sum_{1\leq i<j\leq N}\breve{Z}_{N,ij}\right| > Ma_N\right) \\
&= P\left(\left|N^{-1}\sum_{1\leq i<j\leq N}\tau_N^{-1}h_N^{2d_X}\breve{Z}_{N,ij}\right| > M\frac{N-1}{2}h_N^{2d_X}a_N\tau_N^{-1}\right) \\
&\leq c_1\exp\left(-\frac{c_2 t}{\sigma + \left(ct^{1/2}N^{-1/2}\right)^{2/3}}\right) \\
&= c_1\exp\left(-\frac{c_2 t}{\sigma\ln N + \left(ct^{1/2}N^{-1/2}\right)^{2/3}\ln N}\cdot\ln N\right)
\end{aligned}$$

where $t = M\frac{N-1}{2}h_N^{2d_X}a_N\tau_N^{-1}$ and $\sigma^2 = \mathrm{Var}\left(\tau_N^{-1}h_N^{2d_X}\breve{Z}_{N,ij}\right)$. We will show that $\frac{c_2 t}{\sigma\ln N + \left(ct^{1/2}N^{-1/2}\right)^{2/3}\ln N} \to \infty$ as $N\to\infty$ by showing both $\frac{t}{\sigma\ln N}\to\infty$ and $\frac{t}{\left(ct^{1/2}N^{-1/2}\right)^{2/3}\ln N}\to\infty$ as $N\to\infty$.

Beginning with the former claim:

$$\begin{aligned}
\frac{t}{\sigma \ln N} &= \frac{M\frac{N-1}{2}h_N^{2d_X}a_N\tau_N^{-1}}{\tau_N^{-1}h_N^{2d_X}\operatorname{Var}\left(\breve{Z}_{N,ij}\right)^{1/2}\ln N} = \frac{M(N-1)a_N}{2\operatorname{Var}\left(\breve{Z}_{N,ij}\right)^{1/2}\ln N} \geq \frac{M(N-1)a_N}{2V_{N,2}^{1/2}\ln N} \\
&\geq \frac{MNa_N}{4\left(h_N^{-2d_X}B_4K_{max}B_1\right)^{1/2}\ln N} = \frac{M}{4\left(B_4K_{max}B_1\right)^{1/2}}a_N\left(\frac{\ln N}{Nh_N^{d_X}}\right)^{-1} \\
&= \frac{M}{4\left(B_4K_{max}B_1\right)^{1/2}}a_N^{-1} \\
&\to \infty, \text{ as } N \to \infty.
\end{aligned}$$

The latter claim follows because:

$$\begin{aligned}
\frac{t}{\left(ct^{1/2}N^{-1/2}\right)^{2/3}\ln N} &= \left(\frac{t^2N}{c^2(\ln N)^3}\right)^{1/3} = \left(\frac{(M\frac{N-1}{2}h_N^{2d_X}a_N\tau_N^{-1})^2N}{c^2(\ln N)^3}\right)^{1/3} \\
&\geq \left(\frac{M^2}{16c^2}N^3(\ln N)^{-3}h_N^{4d_X}a_N^2\tau_N^{-2}\right)^{1/3} \\
&= \left(\frac{M^2}{16c^2}\right)^{1/3}\left(Nh_N^{\frac{3}{2}d_X}(\ln N)^{-1}\tau_N^{-1}\right)^{2/3} \\
&\to \infty, \text{ as } N \to \infty.
\end{aligned}$$

The last line above is an implication of the condition $\tau_N \ll Nh_N^{\frac{3}{2}d_X}/\ln N$. Combining these two limit results gives us $\frac{c_2t}{\sigma \ln N + \left(ct^{1/2}N^{-1/2}\right)^{2/3}\ln N} \to \infty$ as $N \to \infty$. Notice the bound again doesn't depend on w and the inequality still holds when we take the sup over $w \in \mathbb{R}^{d_W}$ on the left-hand side. Hence for any $M > 0$ and any $\alpha > 0$, $\sup_{w\in\mathbb{R}^{d_W}} P\left(|T_{N,2}(w)| > Ma_N\right) = O\left(N^{-\alpha}\right)$.

Proof of claim (iii)

Direct evaluation yields

$$\begin{aligned}
\left|\mathbb{E}\left(\hat{\Phi}(w)-\tilde{\Phi}(w)\right)\right| &= \left|\mathbb{E}\left[Y_{ij}\mathbb{1}\left(|Y_{ij}|>\tau_N\right)\frac{1}{h_N^{2d_X}}K\left(\frac{w-W_{ij}}{h}\right)\right]\right| \\
&\leq \mathbb{E}\left[|Y_{ij}|\left|\tau_N^{-1}Y_{ij}\right|^{s-1}\mathbb{1}\left(|Y_{ij}|>\tau_N\right)\frac{1}{h_N^{2d_X}}\left|K\left(\frac{w-W_{ij}}{h_N}\right)\right|\right] \\
&\leq \tau_N^{-(s-1)}\mathbb{E}\left[|Y_{ij}|^s\frac{1}{h_N^{2d_X}}\left|K\left(\frac{w-W_{ij}}{h}\right)\right|\right] \\
&= \tau_N^{-(s-1)}\int\mathbb{E}\left[|Y_{12}|^s\left|(X_1,X_2)=(x_1,x_2)\right.\right]\frac{1}{h_N^{2d_X}} \\
&\left|K\left(\frac{x-x_1}{h_N},\frac{x-x_2}{h_N}\right)\right|f(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2 \\
&= \tau_N^{-(s-1)}\int\mathbb{E}\left[|Y_{12}|^s\left|(X_1,X_2)=(x-h_Ns_1,x'-h_Ns_2)\right.\right] \\
&\times f(x-h_Ns_1,x'-h_Ns_2)\left|K\left(s_1,s_2\right)\right|\mathrm{d}s_1\mathrm{d}s_2 \\
&\leq \tau_N^{-(s-1)}B_{4,s}B_1.
\end{aligned}$$

Since the last expression doesn't depend on w, we have $\sup_{w\in\mathbb{R}^{d_W}}\left|\mathbb{E}\left(\hat{\Phi}(w)-\tilde{\Phi}(w)\right)\right| = o(a_N)$.

Proof of claim (iv)

First, we eliminate the sup by upper bounding the terms involving K by $K_{\max}$.

$$\begin{aligned}
\sup_{w\in\mathbb{R}^{d_W}}\left|\hat{\Phi}_N(w)-\tilde{\Phi}_N(w)\right| &= \sup_{w\in\mathbb{R}^{d_W}}\left|\frac{1}{N(N-1)}\sum_{1\leq i\neq j\leq N}Y_{ij}\mathbb{1}\left(|Y_{ij}|>\tau_N\right)\frac{1}{h_N^{d_W}}K\left(\frac{w-W_{ij}}{h_N}\right)\right| \\
&\leq \frac{1}{N(N-1)}\sum_{1\leq i\neq j\leq N}|Y_{ij}|\,\mathbb{1}\left(|Y_{ij}|>\tau_N\right)\frac{1}{h_N^{2d_X}}\sup_{w\in\mathbb{R}^{d_W}}\left|K\left(\frac{w-W_{ij}}{h_N}\right)\right| \\
&\leq K_{\max}h_N^{-2d_X}\tau_N^{-(s-1)}\frac{1}{N(N-1)}\sum_{1\leq i\neq j\leq N}|Y_{ij}|^s.
\end{aligned}$$

Then, taking expectation on both sides yields

$$\mathbb{E}\left(\sup_{w\in\mathbb{R}^{d_W}}\left|\hat{\Phi}_N(w)-\tilde{\Phi}_N(w)\right|\right)\leq K_{\max}h_N^{-2d_X}\tau_N^{-(s-1)}\mathbb{E}\left(|Y_{ij}|^s\right)\leq K_{\max}B_{6,s}h_N^{-2d_X}\tau_N^{-(s-1)}=o(a_N).$$

Proof of claim (v)

If all the $|Y_{ij}|, 1 \leq i \neq j \leq N$ are smaller than the truncation threshold τ_N, then $\hat{\Phi}_N = \tilde{\Phi}_N$,

$$P\left(\hat{\Phi}_N = \tilde{\Phi}_N\right) \geq P\left(\max_{1\leq i<j\leq N} |Y_{ij}| \leq \tau_N\right).$$

We now show that the RHS converges to 1. Observe

$$\begin{aligned}
\sum_{N=2}^{\infty}\sum_{i=1}^{N-1}[P\left(|Y_{iN}| > \tau_N\right) + P\left(|Y_{Ni}| > \tau_N\right)] &\leq \sum_{N=2}^{\infty}\sum_{i=1}^{N-1}\left[\mathbb{E}\left(|Y_{iN}|^s \tau_N^{-s}\right) + \mathbb{E}\left(|Y_{Ni}|^s \tau_N^{-s}\right)\right] \\
&= \mathbb{E}\left(|Y_{iN}|^s\right)\sum_{N=2}^{\infty}\sum_{i=1}^{N-1} 2N^{-2}\phi_N^{-1} \\
&\leq \mathbb{E}\left(|Y_{iN}|^s\right)\sum_{N=2}^{\infty}\frac{2}{N(\ln\ln N)^2 \ln N} \\
&< \infty,
\end{aligned}$$

The Borel-Cantelli lemma implies $P(A_{ij}, i \neq j, i.o.) = 0$ where the set $A_{ij} = \{\omega : Y_{ij}(\omega) > \tau_{\max\{i,j\}}\}$. This means, except for a null set $\mathcal{N}$, for any $\omega \in \mathcal{N}^c$, there exists a $N(\omega)$ s.t. for all $N \geq N(\omega)$, $Y_{iN}(\omega) \leq \tau_N$. Since $\tau_N \uparrow \infty$ as $N \to \infty$, we can take $N^*(\omega) \geq N(\omega)$ such that $\tau_{N^*(w)} > \max_{i,j\leq N(\omega)} |Y_{ij}(\omega)|$. Then for any $N \geq N^*(\omega)$, we have $\max_{1\leq i<j\leq N} |Y_{ij}(\omega)| \leq \tau_N$ and hence $\hat{\Phi}_N = \tilde{\Phi}_N$. Define the set $E_N := \{\omega : N^*(\omega) \leq N\} \subset \{\omega : \hat{\Phi}_N = \tilde{\Phi}_N\}$. Since $E_N \uparrow \mathcal{N}^c$ and $P(\mathcal{N}^c) = 1$, we have $P(\hat{\Phi}_N = \tilde{\Phi}_N) \geq P(E_N) \to 1$ as $N \to \infty$.

Proof of Theorem 3.3

The proof follows the general approach used in Hansen (2008). Denote $\hat{f}_{W,N}(w) = \frac{1}{N(N-1)}\sum_{1\leq i\neq j\leq N} K_{ij,N}(w)$. We can write

$$\hat{g}_N(w) = \frac{\hat{\Psi}_N(w)}{\hat{f}_{W,N}(w)}.$$

We examine the numerator and denominator separately. An application of Theorem 3.2 yields

$$\begin{aligned}
\sup_{||w||\leq C_N} |\hat{\Psi}_N(w) - \mathbb{E}\hat{\Psi}_N(w)| &= O_p(a_N) \\
\sup_{||w||\leq C_N} |\hat{f}_{W,N}(w) - \mathbb{E}\hat{f}_{W,N}(w)| &= O_p(a_N).
\end{aligned}$$

Standard bias calculations give

$$\sup_{||w||\leq C_N} |\mathbb{E}\hat{\Psi}_N(w) - \Psi(w)| = O(h_N^\beta)$$
$$\sup_{||w||\leq C_N} |\mathbb{E}\hat{f}_{W,N}(w) - f_W(w)| = O(h_N^\beta).$$

Combining these results we get

$$\sup_{||w||\leq C_N} |\hat{\Psi}_N(w) - \Psi(w)| = O_p(a_N) + O(h_N^\beta) = O(a_N^*)$$
$$\sup_{||w||\leq C_N} |\hat{f}_{W,N}(w) - f_W(w)| = O_p(a_N) + O(h_N^\beta) = O(a_N^*).$$

Uniformly over $||w|| \leq C_N$ we have

$$\frac{\hat{\Psi}_N(w)}{\hat{f}_{W,N}(w)} = \frac{\hat{\Psi}_N(w)/f_W(w)}{\hat{f}_{W,N}(w)/f_W(w)} = \frac{g(w) + (\hat{\Psi}_N(w) - \Psi(w))/f_W(w)}{1 + (\hat{f}_{W,N}(w) - f_W(w))/f_W(w)} = \frac{g(w) + O_p(\delta_N^{-1}a_N^*)}{1 + O_p(\delta_N^{-1}a_N^*)}$$
$$= g(w) + O_p(\delta_N^{-1}a_N^*)$$

as claimed. The optimal rate is obtained by setting $h_N \asymp \left(\frac{\ln N}{N}\right)^{\frac{1}{2\beta+d_X}}$.