Proceedings of Machine Learning Research vol 140:1–50, 2022 1st Conference on Causal Learning and Reasoning

Non-parametric Inference Adaptive to Intrinsic Dimension

Author names withheld

Editor: Under Review for CLeaR 2022

Abstract

We consider non-parametric estimation and inference of conditional moment models in high dimensions. We show that even when the dimension D of the conditioning variable is larger than the sample size n, estimation and inference is feasible as long as the distribution of the conditioning variable has small intrinsic dimension d, as measured by locally low doubling measures. Our estimation is based on a sub-sampled ensemble of the k-nearest neighbors (k-NN) Z-estimator. We show that if the intrinsic dimension of the covariate distribution is equal to d, then the finite sample estimation error of our estimator is of order $n^{-1/(d+2)}$ and our estimate is $n^{1/(d+2)}$-asymptotically normal, irrespective of D. The sub-sampling size required for achieving these results depends on the unknown intrinsic dimension d. We propose an adaptive data-driven approach for choosing this parameter and prove that it achieves the desired rates. We discuss extensions and applications to heterogeneous treatment effect estimation.

Keywords: non-parametric statistics, inference, intrinsic dimension, conditional moment equation

1. Introduction

Many non-parametric estimation problems in econometrics and causal inference can be formulated as finding a parameter vector $\theta(x) \in \mathbb{R}^p$ that is a solution to a set of conditional moment equations:

$$\mathbf{E}[\psi(Z;\theta(x))|X=x] = 0\,, \tag{1}$$

when given n i.i.d. samples $(Z_1,\ldots,Z_n)$ from the distribution of Z, where $\psi : \mathcal{Z} \times \mathbb{R}^p \to \mathbb{R}^p$ is a known vector valued moment function, $\mathcal{Z}$ is an arbitrary data space, $X \in \mathcal{X} \subset \mathbb{R}^D$ is the feature vector that is included Z. Examples include non-parametric regression[1], quantile regression[2], heterogeneous treatment effect estimation[3], instrumental variable regression[4], local maximum likelihood estimation[5] and estimation of structural econometric models (see e.g., Reiss and Wolak (2007) and examples in Chernozhukov et al. (2016); Chernozhukov et al. (2018b)). The study of such conditional moment restriction problems has a long history in econometrics (see e.g., Newey (1993); Ai and Chen (2003); Chen and Pouzo (2009); Chernozhukov et al. (2015); Chen et al. (2016)). However, the majority of the literature assumes that the conditioning variable X is low dimensional, i.e. D is a constant as the sample size n grows (see e.g., Athey et al. (2019)).[6]

1. $Z = (X, Y)$, where $Y \in \mathbb{R}^p$ is the dependent variable, and $\psi(Z;\theta(x)) = Y - \theta(x)$.
2. $Z = (X, Y)$ and $\psi(Z;\theta(x)) = 1\{Y \le \theta(x)\} - \alpha$, for some $\alpha \in [0,1]$ that denotes the target quantile.
3. $Z = (X, T, Y)$, where $T \in \mathbb{R}^p$ is a vector of treatments, and $\psi(Z;\theta(x)) = (Y - \langle\theta(x), T\rangle)\,T$.
4. $Z = (X, T, W, Y)$, where $T \in \mathbb{R}$ is a treatment, $W \in \mathbb{R}$ an instrument and $\psi(Z;\theta(x)) = (Y - \theta(x)\,T)\,W$.
5. Where the distribution of Z admits a known density $f(z;\theta(x))$ and $\psi(Z;\theta(x)) = \nabla_\theta \log(f(Z;\theta(x))$.
6. Notable exceptions include high dimensional models under parametric assumptions on $\theta(x)$, such as sparse linear forms (see e.g., Chernozhukov et al. (2018a)). There is also work that addresses the fully non-parametric setup (see e.g., Lafferty and Wasserman (2008); Dasgupta and Freund (2008); Kpotufe (2011); Biau (2012); Scornet et al. (2015)) but those are focused on the estimation problem, and do not address inference (i.e., constructing asymptotically valid confidence intervals).

Recent studies demonstrate the success of non-parametric methods (see e.g., Lewis and Syrgkanis (2018)) for solving conditional moment equations even in the high-dimensional settings. Yet, there are limited theoretical results that explain why these methods work well. Indeed without any further structural assumptions on the problem, the exponential in dimension rates of approximately $n^{1/D}$ (see e.g., Stone (1982)) cannot be avoided. Thereby estimation is in-feasible even if D grows very slowly with n.

One hypothesis is that the *intrinsic dimension* of the conditioning variables is low (i.e. even though X is high dimensional, its coordinates are highly correlated), and that causal machine learning estimators are adaptive to this hidden low dimensional structure in the data.[7] *Our work makes this argument, establishing estimation and asymptotic normality results for the general conditional moment problem, with rates that only depend on the intrinsic dimension, independent of the explicit dimension of the conditioning variable.*

We build on two literatures. The statistical machine learning literature introduces the notion of intrinsic dimension, which is defined by saying that the distribution of X has a small doubling measure around the target point x. Under assumptions of low intrinsic dimension papers in this literature establish fast estimation rates in high-dimensional kernel regression settings (Dasgupta and Freund, 2008; Kpotufe, 2011; Kpotufe and Garg, 2013; Xue and Kpotufe, 2018; Chen and Shah, 2018; Kim et al., 2018; Jiang, 2017). However these results do not apply to the conditional moment problems we study here. In the econometrics literature, the pioneering work of Wager and Athey (2018); Athey et al. (2019) does address estimation and inference of conditional moment models, but only in the low dimensional regime.[8] Relative to these literatures, our contributions are as follows:

- We extend the asymptotic normality results of Wager and Athey (2018); Athey et al. (2019) to general sub-sampled kernel estimators and for vector valued parameters $\theta(x)$. Our analysis also allows us to establish rates in the high-dimensional low intrinsic dimension regime. Given samples $S = (Z_1, \ldots, Z_n)$, our estimator solves a locally weighted empirical conditional moment equation
$$\hat{\theta}(x) \text{ solves} : \sum_{i=1}^{n} K(x, X_i, S)\, \psi(Z_i; \theta) = 0\,, \tag{2}$$
where $K(x, X_i, S)$ is a *kernel* capturing the proximity of X_i to the target point x. We consider weights $K(x, X_i, S)$ that take the form of an average over B base weights: $K(x, X_i, S) = \frac{1}{B}\sum_{b=1}^{B} K(x, X_i, S_b)\, 1\{i \in S_b\}$, where each $K(x, X_i, S_b)$ is calculated based on a randomly drawn sub-sample S_b of size $s < n$ from the original sample.
- Our main estimation and asymptotic normality results (see Theorems 6 and 7), are stated in terms of two high-level quantities, specifically an upper bound $\epsilon(s)$ on the rate at which the

7. This observation builds on a long line of work in machine learning (Dasgupta and Freund, 2008; Kpotufe, 2011; Kpotufe and Garg, 2013).

8. These results have been extended in multiple directions, such as improved rates through local linear smoothing Friedberg et al. (2018), robustness to nuisance parameter estimation error Oprescu et al. (2018) and improved bias analysis via sub-sampled nearest neighbor estimation Fan et al. (2018). However, they all require low dimensional setting and the rate provided by the theoretical analysis is roughly $n^{-1/D}$, translating to $\Omega(\epsilon^{-D})$ samples for getting a confidence interval of size ϵ, which is prohibitive in most target applications of machine learning based econometrics. In particular, Wager and Athey (2018) consider regression and heterogeneous treatment effect estimation with a scalar $\theta(x)$ and prove $n^{1/D}$-asymptotic normality of a sub-sampled random forest based estimator and Athey et al. (2019) extend it to the general conditional moment settings.

kernel "shrinks" and a lower bound $\eta(s)$ on the "incrementality" of the kernel. Notably, the explicit dimension of the conditioning variable D does not enter the theorem, so it suffices in what follows to show that $\epsilon(s)$ and $\eta(s)$ depend only on d. The shrinkage rate $\epsilon(s)$ is defined as the ℓ_2-distance between the target point x and the farthest point on which the kernel places positive weight X_i, when trained on a data set of s samplesIncrementality of a kernel describes how much information is revealed about the weight of a sample i solely by knowledge of X_i, and is captured by the second moment of the conditional expected weight The sub-sampling size s can be used to control both shrinkage and incrementality and for trading-off correctly between bias and variance. We also prove that incrementality can be lower bounded as a function of kernel shrinkage, so that having a sufficiently low shrinkage rate enables both estimation and inference. Corollary 11 and Lemma 13), rather than the explicit dimension D. In particular, we show that $\epsilon(s) = O(s^{-1/d})$ and $\eta(s) = \Theta(1/s)$, which lead to our main theorem *that the sub-sampled k-NN estimate achieves an estimation rate of order $n^{1/(d+2)}$ and is also $n^{1/(d+2)}$-asymptotically normal (Theorems 12 and 15).*

- We provide a closed form characterization of the asymptotic variance of the sub-sampled k-NN estimate, based on the conditional variance moments defined as $\sigma^2(x) = \mathrm{Var}\left(\psi(Z;\theta) \mid X = x\right)$ (Theorem 14 and Eq. (14)). For example, for the 1-NN kernel, the asymptotic variance is $\mathrm{Var}(\hat{\theta}(x)) = \frac{\sigma^2(x)s^2}{n(2s-1)}$. This strengthens prior results of Fan et al. (2018) and Wager and Athey (2018), which only proved the existence of an asymptotic variance without providing an explicit form (and thereby relied on bootstrap approaches for the construction of confidence intervals). Our Monte Carlo study shows that our constructed confidence intervals provide great finite sample coverage in a high dimensional regression setup (see Figure 1)[9].
- The sub-sampling size required to achieve optimal rates depends on the intrinsic dimension which is unknown. We discuss an adaptive data-driven approach for picking the sub-sample size s so as to achieve near-optimal estimation or asymptotic normality rates, adapting to the unknown intrinsic dimension of data (see Propositions 16 and 17). Figure 2 depicts the performance of our adaptive approach compared to two benchmarks, one constructed based on theory for intrinsic dimension d which may be unknown, and the other one constructed naïvely based on the known but sub-optimal extrinsic dimension D. As can be observed, our adaptive approach selects s close to the value suggested by the theory and therefore leads to a compelling finite sample coverage[10].

Structure of the paper. The rest of the paper is organized as follows. In §2, we provide preliminary definitions, in §2.1 and §2.2 we explain our algorithms, in §2.3 we explain doubling dimension (see Appendix B for examples). In §3 we state our assumptions, in §4 we provide general estimation and inference results for kernels that satisfy shrinkage and incrementality conditions, and in §5 we apply such results to the k-NN kernel and prove estimation and inference rates for such kernels that only depend on intrinsic dimension. We discuss the extension to heterogeneous treatment effect estimation in §6 and and defer technical proofs to Appendices.

2. Preliminaries

Suppose we have a data set M of n observations $Z_1, Z_2, \ldots, Z_n$ drawn independently from some distribution $\mathcal{D}$ over the observation domain $\mathcal{Z}$. We focus on the case that $Z_i = (X_i, Y_i)$, where X_i

9. See Appendix C for detailed explanation of our simulations
10. Code is available via https://anonymous.4open.science/r/inference-intrinsic-dimension-E037

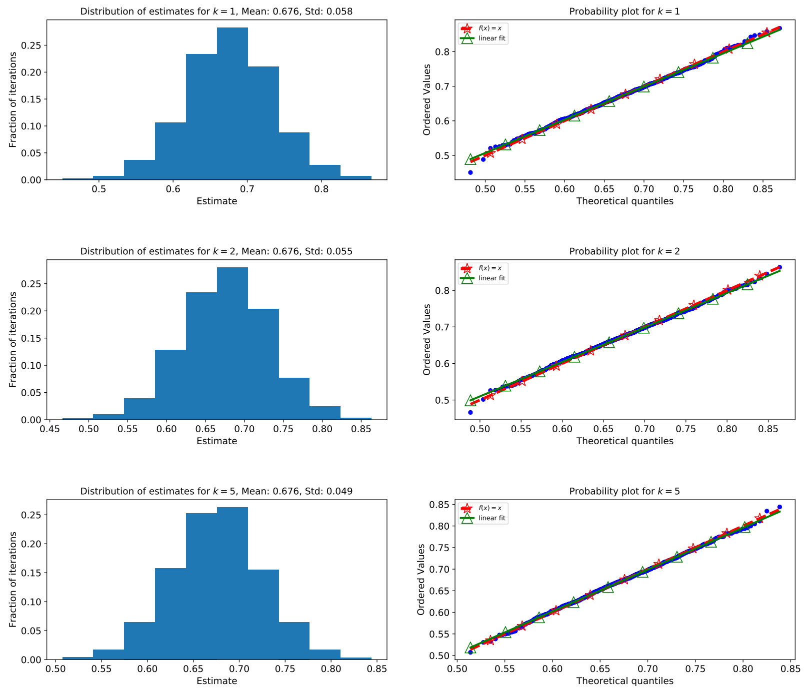

Figure 1: Left: distribution of estimates over 1000 Monte Carlo runs for $k = 1, 2, 5$. Right: the quantile-quantile plot comparison with theoretical asymptotic normal distribution of estimates stemming from our characterization. Means are $0.676, 0.676, 0.676$ and standard deviations are $0.058, 0.055, 0.049$, respectively. $n = 20000$, $D = 20$, $d = 2$, $\mathbf{E}[Y|X] = \frac{1}{1+\exp\{-3X[0]\}}$, $\sigma = 1$. Test point: $x[0] \approx 0.245$, $\mathbf{E}[Y|X = x] \approx 0.676$.

is the vector of covariates and Y_i is the outcome. In Appendix 6, we briefly discuss how our results can be extended to the setting where nuisance parameters and treatments are included in the model.

Suppose that the covariates space $\mathcal{X} \subset \mathbb{R}^D$ is contained in a ball with unknown diameter $\Delta_{\mathcal{X}}$. Denote the marginal distribution of X by μ and the empirical distribution of X on n sample points by μ_n. Let $B(x, r) = \left\{z \in \mathbb{R}^D : \|x - z\|_2 < r\right\}$ be the ℓ_2-ball centered at x with radius r and denote the standard basis for $\mathbb{R}^p$ by $\{e_1, e_2, \ldots, e_p\}$.

Let $\psi : \mathcal{Z} \times \mathbb{R}^p \to \mathbb{R}^p$ be a score function that maps observation Z and parameter $\theta \in \mathbb{R}^p$ to a p-dimensional score $\psi(Z; \theta)$. For $x \in \mathcal{X}$ and $\theta \in \mathbb{R}^p$ define the expected score as $m(x; \theta) = \mathbf{E}[\psi(Z; \theta) \mid X = x]$. The goal is to estimate the quantity $\theta(x)$ via local moment condition, i.e.

$$\theta(x) \text{ solves: } m(x; \theta) = \mathbf{E}[\psi(Z; \theta) \mid X = x] = 0.$$

2.1. Sub-Sampled Kernel Estimation

Base Kernel Learner. Our learner $\mathcal{L}_k$ takes a data set S containing m observations as input and a realization of internal randomness ω, and outputs a kernel weighting function $K_\omega : \mathcal{X} \times \mathcal{X} \times \mathcal{Z}^m \to [0, 1]$. In particular, given any target feature x and the set S, the weight of each observation Z_i in S with feature vector X_i is $K_\omega(x, X_i, S)$. Define the weighted score on a set S with internal randomness ω as $\Psi_S(x; \theta) = \sum_{i \in S} K_\omega(x, X_i, S)\psi(Z_i; \theta)$. When it is clear from context we will omit ω from our notation for succinctness and essentially treat K as a random function. For the rest of the paper, we are going to use notations $\alpha_{S,\omega}(X_i) = K_\omega(x, X_i, S)$ interchangeably.

Averaging over B sub-samples of size s. Suppose that we consider B random and independent draws from all $\binom{n}{s}$ possible subsets of size s and internal randomness variables ω and look at their average. Index these draws by $b = 1, 2, \ldots, B$ where S_b contains samples in bth draw and ω_b is the corresponding draw of internal randomness. We can define the weighted score as

$$\Psi(x;\theta) = \frac{1}{B}\sum_{b=1}^{B} \Psi_{S_b,\omega_b}(x;\theta) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i\in S_b} \alpha_{S_b,\omega_b}(X_i)\psi(Z_i;\theta)\,. \tag{3}$$

Estimating $\theta(x)$. We estimate $\theta(x)$ as a vanishing point of $\Psi(x;\theta)$. Letting $\hat\theta$ be this point, then $\Psi(x;\hat\theta) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} \alpha_{S_b,\omega_b}(X_i)\psi(Z_i;\hat\theta) = 0$. This procedure is explained in Algorithm 1.

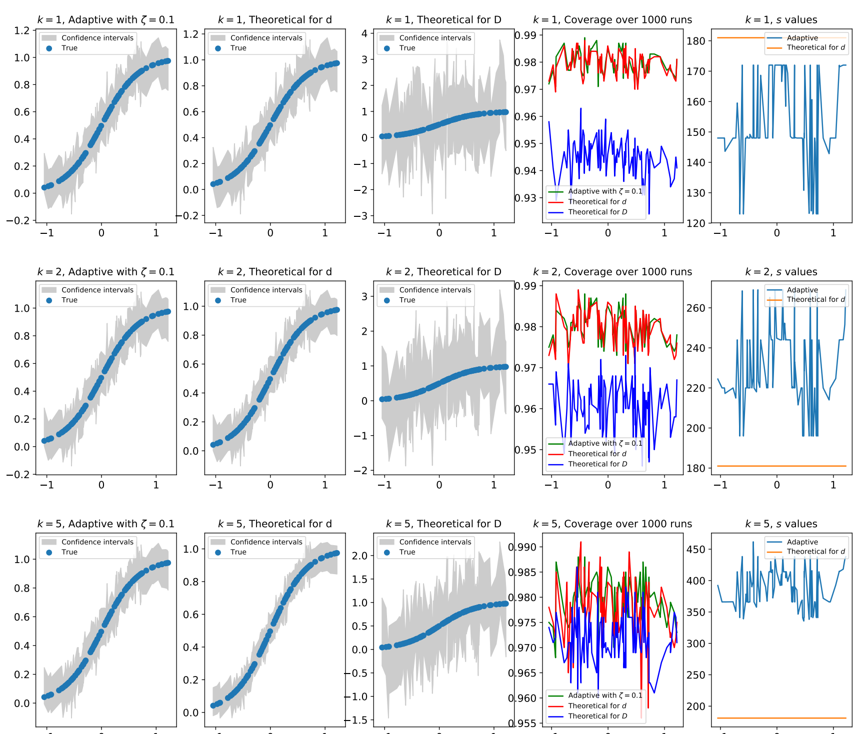

Figure 2: Confidence interval and true values for 100 randomly sampled test points on a single run for $k = 1, 2, 5$ and when (1) left: $s = s_\zeta$ is selected via Proposition 17 with $\zeta = 0.1$, (2) second from the left: $s = n^{1.05d/(d+2)}$, and (3) middle: $s = n^{1.05D/(D+2)}$. Second from the right: coverage over 1000 runs for methods considered. Right: average value of s_ζ selected via Proposition 17 for $\zeta = 0.1$ for different test points compared to the theoretical value $s = n^{1.05d/(d+2)}$. Here $n = 20000$, $D = 20$, $d = 2$, $\mathbf{E}[Y|X] = \frac{1}{1+\exp\{-3X[0]\}}$, $\sigma = 1$. Nominal coverage: 0.98.

2.2. Sub-Sampled k-NN Estimation

We especially focus on the case that the weights are distributed across the k-NN of x. In other words, given a data set S, the weights are given according to $K_\omega(x, X_i, S) = 1\{X_i \in H_k(x, S)\}/k$, where $H_k(x, S)$ are k-NN of x in the set S. The pseudo-code for this can be found in Algorithm 2.

Complete U-statistic. The expression in Equation (3) is an incomplete U-statistic. Complete U-statistic is obtained if we allow each subset of size s from n samples to be included in the model exactly once. In other words, this is achieved if $B = \binom{n}{s}$, all subsets $S_1, S_2, \ldots, S_B$ are distinct, and we also take expectation over the internal randomness ω. Denoting this by $\Psi_0(x;\theta)$, we have

$$\Psi_0(x;\theta) = \binom{n}{s}^{-1} \sum_{S \in [n]:|S|=s} \mathbf{E}_\omega \left[\sum_{i \in S} \alpha_{S,\omega}(X_i)\psi(Z_i;\theta) \right] . \tag{4}$$

Note in the case of k-NN estimator we can also represent Ψ_0 in terms of order statistics, i.e., Ψ_0 is an L-statistics (see e.g., Serfling (2009)). By sorting samples in $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ based on their distance with x as $\|X_{(1)} - x\| \le \|X_{(2)} - x\| \le \cdots \le \|X_{(n)} - x\|$, we can write $\Psi_0(x;\theta) = \sum_{i=1}^n \alpha(X_{(i)})\, \psi(Z_{(i)};\theta)$ where the weights are given by

$$\alpha(X_{(i)}) = \begin{cases} \frac{1}{k} \binom{n}{s}^{-1} \binom{n-i}{s-1} & \text{if } i \le k \\ \frac{1}{k} \binom{n}{s}^{-1} \sum_{j=0}^{k-1} \binom{i-1}{j} \binom{n-i}{s-1-j} & \text{if } i \ge k+1 \,. \end{cases}$$

2.3. Local intrinsic dimension

We are interested in settings that the distribution of X has some low dimensional structure on a ball around the target point x. The following notions are adapted from Kpotufe (2011), which we present here for completeness.

Definition 1 *The marginal μ is called* **doubling measure** *if there exists a constant $C_{db} > 0$ such that for any $x \in \mathcal{X}$ and any $r > 0$ we have $\mu(B(x,r)) \le C_{db}\mu(B(x,r/2))$.*

An equivalent definition of this notion is that, the measure μ is doubling measure if there exist $C, d > 0$ such that for any $x \in \mathcal{X}$, $r > 0$, and $\theta \in (0,1)$ we have $\mu(B(x,r)) \le C\theta^{-d}\mu(B(x,\theta r))$.

One example is given by Lebesgue measure on the Euclidean space $\mathbb{R}^d$, where for any $r > 0, \theta \in (0,1)$ we have $\mathrm{vol}(B(x,\theta r)) = \mathrm{vol}(B(x,r))\theta^d$. Building upon this, let $\mathcal{X} \in \mathbb{R}^D$ be a subset of d-dimensional hyperplane and suppose that for any ball $B(x,r)$ in $\mathcal{X}$ we have $\mathrm{vol}(B(x,r) \cap \mathcal{X}) = \Theta(r^d)$. If μ is almost uniform, then we also have $\mu(B(x,\theta r))/\mu(B(x,r)) = \Theta(\theta^d)$.

Unfortunately, this global notion of doubling measure is restrictive and most probability measures are globally complex. Rather, once restricted to local neighborhoods, they become lower dimensional and intrinsically less complex. The following definition captures this intuition better.

Definition 2 *Fix $x \in \mathcal{X}$ and $r > 0$. The marginal μ is (C,d)-***homogeneous on** *$B(x,r)$ if for any $\theta \in (0,1)$ we have $\mu(B(x,r)) \le C\theta^{-d}\mu(B(x,\theta r))$.*

Intuitively, this definition requires the marginal μ to have a local support that is intrinsically d-dimensional. This definition covers low-dimensional manifolds, mixture distributions, d-sparse data, and also any combination of these examples. These examples are explained in Appendix B.

Algorithm 1 Sub-Sampled Kernel Estimation

Input. Data $\{Z_i = (X_i, Y_i)\}_{i=1}^n$, moment ψ, kernel K, sub-sampling size s, number of iterations B

Initialize. $\alpha(X_i) = 0, 1 \leq i \leq n$ **for** $b \leftarrow 1, B$ **do**

end

Sub-sampling. Draw set S_b by sampling s points from $Z_1, Z_2, \ldots, Z_n$ without replacement.

Weight Updates. $\alpha(X_i) \leftarrow \alpha(X_i) + K_{\omega_b}(x, X_i, S_b)$

Weight Normalization. $\alpha(X_i) \leftarrow \alpha(X_i)/B$

Estimation. Denote $\hat{\theta}$ as a solution of $\Psi(x;\theta) = \sum_{i=1}^n \alpha(X_i)\psi(Z_i;\theta) = 0$

Algorithm 2 Sub-Sampled k-NN Estimation

Input. Data $\{Z_i = (X_i, Y_i)\}_{i=1}^n$, moment ψ, sub-sampling size s, number of iterations B, number of neighbors k

Initialize. $\alpha(X_i) \leftarrow 0, 1 \leq i \leq n$ **for** $b \leftarrow 1, B$ **do**

end

Sub-sampling. Draw set S_b by sampling s points from $Z_1, Z_2, \ldots, Z_n$ without replacement

Weight Updates. $\alpha(X_i) \leftarrow \alpha(X_i) + 1\{X_i \in H_k(x, S_b)\}/k$

Weight Normalization. $\alpha(X_i) \leftarrow \alpha(X_i)/B$

Estimation. Denote $\hat{\theta}$ as a solution of $\Psi(x;\theta) = \sum_{i=1}^n \alpha(X_i)\psi(Z_i;\theta) = 0$

3. Assumptions

For non-parametric sub-sampled estimators, the bias and asymptotic variance are tightly connected to the kernel shrinkage and incrementality, formally defined below.

Definition 3 (Kernel Shrinkage in Expectation) *The function $\epsilon(s)$ defines a kernel shrinkage in expectation if given a set S containing s i.i.d. observations drawn from distribution $\mathcal{D}$, it satisfies*

$$\epsilon(s) := \mathbf{E}\left[\sup\{\|x - X_i\|_2 : K(x, X_i, S) > 0\}\right]. \tag{5}$$

Definition 4 (Kernel Shrinkage in Probability) *The function $\epsilon(s,\delta)$ defines a kernel shrinkage in probability if given a set S containing s i.i.d. observations drawn from distribution $\mathcal{D}$ w.p. $1-\delta$ it satisfies*

$$\sup\{\|x - X_i\|_2 : K(x, X_i, S) > 0\} \leq \epsilon(s,\delta). \tag{6}$$

Definition 5 (Incrementality of Kernel) *The incrementality of kernel K when provided with s i.i.d. observations from distribution $\mathcal{D}$ is defined as*

$$\eta(s) = \mathbf{E}\left[\mathbf{E}\left[K(x, X_i, S)|X_i\right]^2\right]. \tag{7}$$

As shown in Wager and Athey (2018), for trees that satisfy some regularity condition, $\epsilon(s) \leq s^{-c/D}$ for a constant c. We are interested in shrinkage rates that scale as $s^{-c/d}$, where d is the local intrinsic dimension of μ on $B(x,r)$. Similar to Oprescu et al. (2018); Athey et al. (2019), we rely on the following assumptions on the moment and score functions.

Assumption 1 *The moment and score functions satisfy the following:*

1. *The moment $m(x;\theta)$ corresponds to the gradient w.r.t. θ of a λ-strongly convex loss $L(x;\theta)$. This also means that the Jacobian $M_0 = \nabla_\theta m(x;\theta(x))$ has minimum eigenvalue at least λ.*
2. *For any fixed parameters θ, $m(x;\theta)$ is a L_m-Lipschitz function in x for some constant L_m.*
3. *There exists a bound $\psi_{\max}$ such that for any observation z and any θ, $\|\psi(z;\theta)\|_\infty \leq \psi_{\max}$.*
4. *The bracketing number $N_{[]}(\mathcal{F},\epsilon,L_2)$ of the function class: $\mathcal{F} = \{\psi(\cdot;\theta) : \theta \in \Theta\}$, satisfies $\log(N_{[]}(\mathcal{F},\epsilon,L_2)) = O(1/\epsilon)$.*

Assumption 2 *The moment and score functions satisfy the following:*

1. *For any coordinate j of the moment vector m, the Hessian $H_j(x;\theta) = \nabla^2_{\theta\theta} m_j(x;\theta)$ has eigenvalues bounded above by a constant L_H for all θ.*
2. *Maximum eigenvalue of M_0 is upper bounded by L_J.*
3. *Second moment of $\psi(x;\theta)$ defined as $\mathrm{Var}\left(\psi(Z;\theta) \mid X = x\right)$ is L_{mm}-Lipschitz in x, i.e.,*

$$\| \mathrm{Var}\left(\psi(Z;\theta) \mid X = x\right) - \mathrm{Var}\left(\psi(Z;\theta) \mid X = x'\right) \|_F \leq L_{mm} \|x - x'\|_2 \,.$$

4. *Variogram is Lipschitz: $\sup_{x\in\mathcal{X}} \|\mathrm{Var}(\psi(Z;\theta) - \psi(Z;\theta') \mid X = x)\|_F \leq L_\psi \|\theta - \theta'\|_2$.*

The condition on variogram always holds for a ψ that is Lipschitz in θ. This larger class of functions ψ allows estimation in more general settings such as α-quantile regression that involves a ψ which is non-Lipschitz in θ. Similar to Athey and Imbens (2016); Athey et al. (2019), we require kernel K to be *honest* and *symmetric*.

Assumption 3 *The kernel K, built using samples $\{Z_1, Z_2, \ldots, Z_s\}$, is* **honest** *if the weight of sample i given by $K(x, X_i, \{Z_j\}_{j=1}^s)$ is independent of Y_j conditional on X_j for any $j \in [s]$.*

Assumption 4 *The kernel K, built using samples $\{Z_1, Z_2, \ldots, Z_s\}$, is* **symmetric** *if for any permutation $\pi : [s] \to [s]$, the distribution of $K(x, X_i, \{Z_j\}_{j=1}^s)$ and $K(x, X_{\pi(i)}, \{Z_{\pi(j)}\}_{j=1}^s)$ are equal. In other words, the kernel weighting distribution remains unchanged under permutations.*

For a deterministic kernel K, the above condition implies that $K(x, X_i, \{Z_j\}_{j=1}^s) = K(x, X_i, \{Z_{\pi(j)}\}_{j=1}^s)$, for any $i \in [s]$. In the next section, we provide general estimation and inference results for a general kernel based on the its shrinkage and incrementality rates.

4. Guarantees for sub-sampled kernel estimators

Our first result establishes estimation rates, both in expectation and high probability, for kernels based on their shrinkage rates. The proof of this theorem is deferred to Appendix D.

Theorem 6 (Finite Sample Estimation Rate) *Let Assumptions 1 and 3 hold. Suppose that Algorithm 1 is executed with $B \geq n/s$. If the base kernel K satisfies kernel shrinkage in expectation, with rate $\epsilon(s)$, then w.p. $1-\delta$*

$$\|\hat\theta - \theta(x)\|_2 \leq \frac{2}{\lambda}\left(L_m \epsilon(s) + O\left(\psi_{\max}\sqrt{\frac{p\,s}{n}\left(\log\log(n/s) + \log(p/\delta)\right)}\right)\right) \,. \tag{8}$$

Moreover,

$$\sqrt{\mathbf{E}\left[\|\hat\theta - \theta(x)\|_2^2\right]} \leq \frac{2}{\lambda}\left(L_m \epsilon(s) + O\left(\psi_{\max}\sqrt{\frac{p\,s}{n}\log\log(p\,n/s)}\right)\right) \,. \tag{9}$$

The next result establishes asymptotic normality of sub-sampled kernel estimators. In particular, it provides coordinate-wise asymptotic normality of our estimate $\hat{\theta}$ around its true underlying value $\theta(x)$. For this result, in addition to the shrinkage, we require the incrementality of the kernel to satisfy some conditions. The proof of this theorem is deferred to Appendix E.

Theorem 7 (Asymptotic Normality) *Let Assumptions 1, 2, 3, and 4 hold. Suppose that Algorithm 1 is executed with $B \geq (n/s)^{5/4}$ and the base kernel K satisfies kernel shrinkage, with rate $\epsilon(s,\delta)$ in probability and $\epsilon(s)$ in expectation. Let $\eta(s)$ be the incrementality of kernel K defined in Equation (7) and s grow at a rate such that $s \to \infty$, $n\eta(s) \to \infty$, and $\epsilon(s, \eta(s)^2) \to 0$. Consider any fixed coefficient $\beta \in \mathbb{R}^p$ with $\|\beta\| \leq 1$ and define the variance as*

$$\sigma_{n,\beta}^2(x) = \frac{s^2}{n} \operatorname{Var}\left[\mathbf{E}\left[\sum_{i=1}^{s} K(x, X_i, \{Z_j\}_{j=1}^s) \left\langle \beta, M_0^{-1} \psi(Z_i; \theta(x)) \right\rangle \mid Z_1\right]\right].$$

Then it holds that $\sigma_{n,\beta}(x) = \Omega\left(s\sqrt{\eta(s)/n}\right)$. Moreover, suppose that

$$\max\left(\epsilon(s), \epsilon(s)^{1/4}\left(\frac{s}{n}\log\log(n/s)\right)^{1/2}, \left(\frac{s}{n}\log\log(n/s)\right)^{5/8}\right) = o(\sigma_{n,\beta}(x)). \tag{10}$$

Then,

$$\frac{\left\langle \beta, \hat{\theta} - \theta(x) \right\rangle}{\sigma_{n,\beta}(x)} \to_d \mathsf{N}(0,1).$$

Theorems 6 and 7 generalize existing estimation and asymptotic normality results of Athey et al. (2019); Wager and Athey (2018); Fan et al. (2018) to an arbitrary kernel that satisfies appropriate shrinkage and incrementality rates (see Remark 24 in Appendix E). The following lemma relates these two and provides a lower bound on the incrementality in terms of kernel shrinkage. The proof uses the Paley-Zygmund inequality and is left to Appendix F.

Lemma 8 *For any symmetric kernel K (Assumption 4) and for any $\delta \in [0,1]$:*

$$\eta(s) = \mathbf{E}\left[\mathbf{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1\right]^2\right] \geq \frac{(1-\delta)^2\,(1/s)^2}{\inf_{\rho>0}\left(\mu(B(x, \epsilon(s,\rho))) + \rho\, s/\delta\right)}.$$

Thus if $\mu(B(x, \epsilon(s, 1/(2s^2)))) = O(\log(s)/s)$, then picking $\rho = 1/(2s^2)$ and $\delta = 1/2$ implies that $\mathbf{E}[\mathbf{E}[K(x, X_1, \{Z_j\}_{j=1}^s)|X_1]^2] = \Omega(1/s\log(s))$.

Corollary 9 *If $\epsilon(s,\delta) = O((\log(1/\delta)/s)^{1/d})$ and μ satisfies a two-sided version of the doubling measure property on $B(x,r)$, defined in Definition 2, i.e., $c\theta^d\mu(B(x,r)) \leq \mu(B(x,\theta r)) \leq C\theta^d\mu(B(x,r))$ for any $\theta \in (0,1)$. Then, $\mathbf{E}[\mathbf{E}[K(x, X_1, \{Z_j\}_{j=1}^s)|X_1]^2] = \Omega(1/(s\log(s)))$.*

Even without this extra assumption, we can still characterize the incrementality rate of the k-NN estimator, as we observe in the next section.

5. Main theorem: adaptivity of k-NN estimator

In this section, we provide estimation guarantees and asymptotic normality of the k-NN estimator by using Theorems 6 and 7. We first establish shrinkage and incrementality rates for this kernel.

5.1. Estimation guarantees for the k-NN estimator

In this section we provide shrinkage results for the k-NN kernel. As observed in Theorem 6, shrinkage rates are sufficient for bounding the estimation error. The shrinkage result that we present in the following would only depend on the local intrinsic dimension of μ on $B(x,r)$.

Lemma 10 (High probability shrinkage for the k-NN kernel) *Suppose that the measure μ is (C,d)-homogeneous on $B(x,r)$. Then, for any δ satisfying $2\exp\left(-\mu(B(x,r))s/(8C)\right) \leq \delta \leq \frac{1}{2}\exp(-k/2)$, w.p. at least $1-\delta$ we have $\|x - X_{(k)}\|_2 \leq \epsilon_k(s,\delta) = O\left(\frac{\log(1/\delta)}{s}\right)^{1/d}$.*

We can turn this result into a shrinkage rate in expectation as follows. In fact, by the very convenient choice of $\delta = s^{-1/d}$ combined with the fact that $\mathcal{X}$ has diameter $\Delta_{\mathcal{X}}$, we can establish $O\left((\log(s)/s)^{1/d}\right)$ rate on expected kernel shrinkage. However, a more careful analysis would help us to remove the $\log(s)$ dependency in the bound and is stated in the following corollary:

Corollary 11 (Expected shrinkage for the k-NN kernel) *Suppose that the conditions of Lemma 10 hold. Let k be a constant and $\epsilon_k(s)$ be the expected shrinkage for the k-NN kernel. Then, for any s larger than some constant we have $\epsilon_k(s) = \mathbf{E}\left[\|x - X_{(k)}\|_2\right] = O\left(\frac{1}{s}\right)^{1/d}$.*

We are now ready to state our estimation result for the k-NN kernel, which is honest and symmetric. Therefore, we can substitute the expected shrinkage rate established in Corollary 11 in Theorem 6 to derive estimation rates for this kernel.

Theorem 12 (Estimation Guarantees for the k-NN Kernel) *Suppose that μ is (C,d)-homogeneous on $B(x,r)$, Assumption 1 holds and that Algorithm 2 is executed with $B \geq n/s$. Then, w.p. $1-\delta$:*

$$\|\hat{\theta} - \theta(x)\|_2 \leq \frac{2}{\lambda}\left(O\left(s^{-1/d}\right) + O\left(\psi_{\max}\sqrt{\frac{p\,s}{n}\left(\log\log(n/s) + \log(p/\delta)\right)}\right)\right), \tag{11}$$

and

$$\sqrt{\mathbf{E}\left[\|\hat{\theta} - \theta(x)\|_2^2\right]} \leq \frac{2}{\lambda}\left(O\left(s^{-1/d}\right) + O\left(\psi_{\max}\sqrt{\frac{s\,p\,\log\log(p\,n/s)}{n}}\right)\right). \tag{12}$$

By picking $s = \Theta\left(n^{d/(d+2)}\right)$ and $B = \Omega\left(n^{2/(d+2)}\right)$ we get $\sqrt{\mathbf{E}\left[\|\hat{\theta} - \theta(x)\|_2^2\right]} = \tilde{O}\left(n^{-1/(d+2)}\right)$.

5.2. Asymptotic normality of the k-NN estimator

In this section we prove asymptotic normality of k-NN estimator. We first provide a bound on the incrementality of the k-NN kernel.

Lemma 13 (k-NN Incrementality) *Let K be the k-NN kernel and let $\eta_k(s)$ denote the incrementality rate of this kernel. Then, the following holds:*

$$\eta_k(s) = \mathbf{E}\left[\mathbf{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1\right]^2\right] = \frac{1}{(2s-1)\,k^2}\left(\sum_{t=0}^{2k-2}\frac{a_t}{b_t}\right),$$

where sequences $\{a_t\}_{t=0}^{2k-2}$ and $\{b_t\}_{t=0}^{2k-2}$ are defined as

$$a_t = \sum_{i=\max\{0,t-(k-1)\}}^{\min\{t,k-1\}}\binom{s-1}{i}\binom{s-1}{t-i} \qquad \text{and} \qquad b_t = \sum_{i=0}^{t}\binom{s-1}{i}\binom{s-1}{t-i}$$

We can substitute $\eta_k(s)$ in Theorem 7 to prove asymptotic normality of the k-NN estimator. The following theorem takes a step further and derives the asymptotic variance of this estimator $\sigma_{n,j}(x)$.

Theorem 14 (Asymptotic Variance of k-NN) *Let $j \in [p]$ be one of coordinates. Suppose that k is constant while $s \to \infty$. Then, for the k-NN kernel*

$$\sigma_{n,j}^2(x) = \frac{s^2}{n} \frac{\sigma_j^2(x)}{k^2\,(2s-1)}\, \zeta_k + o(s/n)\,, \tag{13}$$

where $\sigma_j^2(x) = \mathrm{Var}\left[\langle e_j, M_0^{-1}\psi(Z;\theta(x))\rangle \mid X = x\right]$ and $\zeta_k = k + \sum_{t=k}^{2k-2} 2^{-t} \sum_{i=t-k+1}^{k-1} \binom{t}{i}$.

Combining results of Theorem 7, Theorem 14, Corollary 11, and Lemma 13 we have:

Theorem 15 (Asymptotic Normality of k-NN Estimator) *Suppose that μ is (C,d)-homogeneous on $B(x,r)$. Let Assumptions 1, 2 hold and suppose that Algorithm 2 is executed with $B \geq (n/s)^{5/4}$ iterations. Suppose that s grows at a rate such that $s \to \infty$, $n/s \to \infty$, and also $s^{-1/d}(n/s)^{1/2} \to 0$. Let $j \in [p]$ be one of coordinates and $\sigma_{n,j}^2(x)$ be defined in Equation (13). Then,*

$$\frac{\hat{\theta}_j(x) - \theta_j(x)}{\sigma_{n,j}(x)} \to \mathsf{N}(0,1)\,.$$

Finally, if $s = n^\beta$ and $B \geq n^{\frac{5}{4}(1-\beta)}$ with $\beta \in (d/(d+2),1)$. Then, $\frac{\hat{\theta}_j(x) - \theta_j(x)}{\sigma_{n,j}(x)} \to \mathsf{N}(0,1)\,.$

Plug-in confidence intervals. Observe that the Theorem 14 implies that if we define $\tilde{\sigma}_{n,j}^2(x) = \frac{s^2}{n}\frac{\sigma_j^2(x)}{2s-1}\frac{\zeta_k}{k^2}$ as the leading term in the variance, then $\frac{\sigma_{n,j}^2(x)}{\tilde{\sigma}_{n,j}^2(x)} \to_p 1$. Thus, due to Slutsky's theorem

$$\frac{\hat{\theta}_j - \theta_j}{\tilde{\sigma}_{n,j}^2(x)} = \frac{\hat{\theta}_j - \theta_j}{\sigma_{n,j}^2(x)} \frac{\sigma_{n,j}^2(x)}{\tilde{\sigma}_{n,j}^2(x)} \to_d \mathsf{N}(0,1)\,. \tag{14}$$

Hence, we have a closed form solution to the variance in our asymptotic normality theorem. If we have an estimate $\hat{\sigma}_j^2(x)$ of the variance of the conditional moment around x, then we can build plug-in confidence intervals based on the normal distribution with variance $\frac{s^2}{n}\frac{\hat{\sigma}_j^2(x)}{2s-1}\frac{\zeta_k}{k^2}$. Note that ζ_k can be calculated easily for desired values of k. For instance, we have $\zeta_1 = 1, \zeta_2 = \frac{5}{2}$, and $\zeta_3 = \frac{33}{8}$ and for $k = 1,2,3$ the asymptotic variance becomes $\frac{s^2}{n}\frac{\hat{\sigma}_j^2(x)}{2s-1}$, $\frac{5}{8}\frac{s^2}{n}\frac{\hat{\sigma}_j^2(x)}{2s-1}$, and $\frac{11}{24}\frac{s^2}{n}\frac{\hat{\sigma}_j^2(x)}{2s-1}$ respectively.

5.3. Selecting s adaptively

According to Theorem 12, $s = \Theta(n^{d/(d+2)})$ would trade-off between bias and variance terms. Also, according to Theorem 15, picking $s = n^\beta$ with $d/(d+2) < \beta < 1$ would result in asymptotic normality of the estimator. However, both choices depend on the unknown intrinsic dimension of μ on the ball $B(x,r)$. Inspired by Kpotufe (2011), we explain a data-driven way for estimating s.

Suppose that $\delta > 0$ is given. Let $C_{n,p,\delta} = 2\log(2pn/\delta)$ and pick $\Delta \geq \Delta_{\mathcal{X}}$. For any $k \leq s \leq n$, let $H(s)$ be the U-statistic estimator for $\epsilon(s)$ defined as $H(s) = \sum_{S\in[n]:|S|=s} \max_{X_i \in H_k(x,S)} \|x - X_i\|_2 / \binom{n}{s}$. Each term in the summation computes the distance of x to its k-nearest neighbor on S and $H(s)$ is the average of these numbers over all $\binom{n}{s}$ possible subsets S (see Remark 34 in Appendix G regarding to efficient computation of $H(s)$). Define $G_\delta(s) = \Delta\sqrt{C_{n,p,\delta}ps/n}$. Iterate over $s = n, \cdots, k$. Let s_2 be the smallest s for which we have $H(s) > 2G_\delta(s)$ and let $s_1 = s_2 + 1$. Note that $\epsilon_k(s)$ is decreasing in s and $G_\delta(s)$ is increasing in s. Therefore, there exists a unique $1 \le s^* \le n$ such that $\epsilon_k(s^*) \le G_\delta(s^*)$ and $\epsilon_k(s^*-1) > G_\delta(s^*-1)$. We have following results.

Proposition 16 (Adaptive Estimation) *Let Assumptions of Theorem 12 hold. Suppose that s_1 is the output of the above process. Let $s_* = 9s_1 + 1$ and suppose that Algorithm 2 is executed with $s = s_*$ and $B \ge n/s_*$. Then w.p. at least $1 - 2\delta$ we have $\|\hat{\theta} - \theta(x)\|_2 = O(G_\delta(s^*)) = O\left(\left(\frac{n}{p\log(2pn/\delta)}\right)^{-1/(d+2)}\right)$. Further, for $\delta = 1/n$ we have $\sqrt{\mathbf{E}\left[\|\hat{\theta} - \theta(x)\|_2^2\right]} = \tilde{O}\left(n^{-1/(d+2)}\right)$.*

Proposition 17 (Adaptive Asymptotic Normality) *Let Assumptions of Theorem 15 hold. Let s_1 be the output of the above process when $\delta = 1/n$ and $s_* = 9s_1 + 1$. For any $\zeta \in (0, (\log(n) - \log(s_1) - \log\log^2(n))/\log(n)))$ define $s_\zeta = s_* n^\zeta$. Suppose that Algorithm 2 is executed with $s = s_\zeta$ and $B \ge (n/s_\zeta)^{5/4}$, then for any coordinate $j \in [p]$, we have $\frac{\hat{\theta}_j(x) - \theta_j(x)}{\sigma_{n,j}(x)} \to \mathsf{N}(0,1)$.*

6. Nuisance parameters and heterogeneous treatment effects

Using the techniques of Oprescu et al. (2018), our work easily extends to the case where the moments depend on, potentially infinite dimensional, nuisance components h_0, that also need to be estimated, i.e.,

$$\theta(x) \text{ solves: } m(x;\theta,h_0) = \mathbf{E}[\psi(Z;\theta,h_0) \mid x] = 0. \tag{15}$$

If the moment m is orthogonal with respect to h and assuming that h_0 can be estimated on a separate sample with a conditional MSE rate of

$$\mathbf{E}[(\hat{h}(z) - h_0(z))^2 | X = x] = o_p(\epsilon(s) + \sqrt{s/n}), \tag{16}$$

then using the techniques of Oprescu et al. (2018), we can argue that both our finite sample estimation rate and our asymptotic normality rate, remain unchanged, as the estimation error only impacts lower order terms. This extension allows us to capture settings like heterogeneous treatment effects, where the treatment model also needs to be estimated when using the orthogonal moment as

$$\psi(z;\theta,h_0) = (y - q_0(x,w) - \theta(t - p_0(x,w)))\,(t - p_0(x,w)), \tag{17}$$

where y is the outcome of interest, t is a treatment, x, w are confounding variables, $q_0(x,w) = \mathbf{E}[Y|X = x, W = w]$ and $p_0(x,w) = E[T|X = x, W = w]$. The latter two nuisance functions can be estimated via separate non-parametric regressions. In particular, if we assume that these functions are sparse linear in w, i.e.:

$$q_0(x,w) = \langle \beta(x), w\rangle\ , \qquad\qquad p_0(x,w) = \langle \gamma(x), w\rangle\ . \tag{18}$$

Then we can achieve a conditional mean-squared-error rate of the required order by using the kernel lasso estimator of Oprescu et al. (2018), where the kernel is the sub-sampled k-NN kernel, assuming the sparsity does not grow fast with n.

Conclusion

In this work we studied non-parametric inference when solving general conditional moment equations in high-dimensions and provided estimation and inference guarantees that only depend on the local intrinsic dimension of the covariate space. We confirmed our theoretical findings via numerical simulations.

References

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

Alexandr Andoni, Thijs Laarhoven, Ilya Razenshteyn, and Erik Waingarten. Optimal hashing-based time-space trade-offs for approximate near neighbors. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 47–66. Society for Industrial and Applied Mathematics, 2017.

Alexandr Andoni, Piotr Indyk, and Ilya Razenshteyn. Approximate nearest neighbor search in high dimensions. *arXiv preprint arXiv:1806.09823*, 2018.

Patrice Assouad. Plongements lipschitziens dans $\mathbb{R}^n$. *Bull. Soc. Math. France*, 111:429–448, 1983.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey and Guido W Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017.

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

RE Bellman. Adaptive control processes princeton. *Press, Princeton, NJ*, 1961.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50, 2014a.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014b.

Alexandre Belloni, Victor Chernozhukov, Ivan Fernández-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

Thomas B Berrett, Richard J Samworth, Ming Yuan, et al. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47(1):288–318, 2019.

Gérard Biau. Analysis of a random forests model. *J. Mach. Learn. Res.*, 13(1):1063–1095, April 2012. ISSN 1532-4435.

Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*. Springer, 2015.

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

A.A. Borovkov. *Probability Theory*. Springer London, 2013. ISBN 9781447152002.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Manfredo Perdigão do Carmo. *Riemannian geometry*. Birkhäuser, 1992.

B. Chen, J. Pearl, and E. Bareinboim. Incorporating knowledge into structural equation models using auxiliary variables. In S. Kambhampati, editor, *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 3577–3583. International Joint Conferences on Artificial Intelligence Organization, 2016.

George H. Chen and Devavrat Shah. Explaining the success of nearest neighbor methods in prediction. *Foundations and Trends in Machine Learning*, 10(5-6):337–588, 2018. ISSN 1935-8237.

Xiaohong Chen and Demian Pouzo. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, 152(1):46 – 60, 2009. ISSN 0304-4076. Recent Adavances in Nonparametric and Semiparametric Econometrics: A Volume Honouring Peter M. Robinson.

Victor Chernozhukov, Whitney K. Newey, and Andres Santos. Constrained Conditional Moment Restriction Models. *arXiv e-prints*, art. arXiv:1509.06311, September 2015.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K. Newey, and James M. Robins. Locally Robust Semiparametric Estimation. *arXiv e-prints*, art. arXiv:1608.00033, July 2016.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018a.

Victor Chernozhukov, Denis Nekipelov, Vira Semenova, and Vasilis Syrgkanis. Plug-in regularized estimation of high-dimensional parameters in nonlinear semiparametric models. *arXiv preprint arXiv:1806.04823*, 2018b.

Colleen D Cutler. A review of the theory and estimation of fractal dimension. In *Dimension estimation and models*, pages 1–107. World Scientific, 1993.

Sanjoy Dasgupta and Yoav Freund. Random projection trees and low dimensional manifolds. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 537–546. ACM, 2008.

Bradley Efron. *The jackknife, the bootstrap, and other resampling plans*, volume 38. Siam, 1982.

Jianqing Fan, Mark Farmen, and Irene Gijbels. Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):591–608, 1998.

Yingying Fan, Jinchi Lv, and Jingbo Wang. Dnn: A two-scale distributional tale of heterogeneous treatment effect inference. *arXiv preprint arXiv:1808.08469*, 2018.

Rina Friedberg, Julie Tibshirani, Susan Athey, and Stefan Wager. Local linear forests. *arXiv preprint arXiv:1807.11408*, 2018.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Heinrich Jiang. Rates of uniform consistency for k-nn regression. *arXiv preprint arXiv:1707.06261*, 2017.

Jisu Kim, Jaehyeok Shin, Alessandro Rinaldo, and Larry Wasserman. Uniform Convergence Rate of the Kernel Density Estimator Adaptive to Intrinsic Dimension. *arXiv e-prints*, art. arXiv:1810.05935, October 2018.

Samory Kpotufe. k-nn regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems*, pages 729–737, 2011.

Samory Kpotufe and Sanjoy Dasgupta. A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences*, 78(5):1496–1515, 2012.

Samory Kpotufe and Vikas Garg. Adaptivity to local smoothness and dimension in kernel regression. In *Advances in neural information processing systems*, pages 3075–3083, 2013.

John Lafferty and Larry Wasserman. Rodeo: Sparse, greedy nonparametric regression. *Ann. Statist.*, 36(1):28–63, 02 2008.

Arthur Lewbel. A local generalized method of moments estimator. *Economics Letters*, 94(1):124–128, 2007.

Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018.

Yue-Pok Mack. Local properties of k-nn regression estimates. *SIAM Journal on Algebraic Discrete Methods*, 2(3):311–323, 1981.

Lester Mackey, Vasilis Syrgkanis, and Ilias Zadik. Orthogonal machine learning: Power and limitations. *arXiv preprint arXiv:1711.00342*, 2017.

Sendhil Mullainathan and Jann Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

Whitney K. Newey. 16 efficient estimation of models with conditional moment restrictions. In *Econometrics*, volume 11 of *Handbook of Statistics*, pages 419 – 454. Elsevier, 1993.

Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for heterogeneous treatment effect estimation. *arXiv preprint arXiv:1806.03467*, 2018.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Thomas Peel, Sandrine Anthoine, and Liva Ralaivola. Empirical bernstein inequalities for u-statistics. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1903–1911. Curran Associates, Inc., 2010.

Peter C Reiss and Frank A Wolak. Structural econometric modeling: Rationales and examples from industrial organization. *Handbook of econometrics*, 6:4277–4415, 2007.

James M Robins and Ya'acov Ritov. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in medicine*, 16(3):285–319, 1997.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Richard J Samworth et al. Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763, 2012.

Erwan Scornet, Grard Biau, and Jean-Philippe Vert. Consistency of random forests. *Ann. Statist.*, 43(4):1716–1741, 08 2015.

Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

Joan G Staniswalis. The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84(405):276–283, 1989.

Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 12 1982.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Robert Tibshirani and Trevor Hastie. Local likelihood estimation. *Journal of the American Statistical Association*, 82(398):559–567, 1987.

Nakul Verma, Samory Kpotufe, and Sanjoy Dasgupta. Which spatial partition trees are adaptive to intrinsic dimension? In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 565–574. AUAI Press, 2009.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Lirong Xue and Samory Kpotufe. Achieving the time of 1-nn, but the accuracy of k-nn. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1628–1636, 09–11 Apr 2018.

Appendix A. Related work

There exist a very large body of work in causal inference. We refer the reader to Pearl (2009) and Imbens and Rubin (2015) for a thorough review.

There exists a vast literature on average treatment effect estimation in high-dimensional settings. The key challenge in such settings is the problem of overfitting which is usually handled by adding regularization terms. However, this leads to a shrinked estimate for the average treatment effect and therefore not desirable. The literature has taken various approaches to solve this issue. For instance, Belloni et al. (2014a,b) used a two-step method for estimating average treatment effect where in the first step feature-selection is accomplished via a lasso and then treatment effect is estimated using selected features. Athey et al. (2018) studied approximate residual balancing where a combination of weight balancing and regression adjustment is used for removing undesired bias and for achieving a double robust estimator. Chernozhukov et al. (2018a,a) considered a more general semi-parametric framework and studied debiased/double machine learning methods via first order Neyman orthogonality condition. Mackey et al. (2017) extended this result to higher order moments. Please refer to Athey and Imbens (2017); Mullainathan and Spiess (2017); Belloni et al. (2017) for a review on this literature.

However, in many applications, researchers are interested in estimating heterogeneous treatment effect on various sub-populations. One effective solution is to use one of the methods described in previous paragraph to estimate problem parameters and then project such estimations onto the sub-population of interest. However, these approaches usually perform poorly when there is a model mis-specification, i.e., when the true underlying model does not belong to the parametric search space. Consequently, researchers have studied non-parametric estimators such as k-NN estimators, kernel estimators, and random forests. While these non-parametric estimators are very robust to model mis-specification and work well under mild assumptions on the function of interest, they suffer from the curse of dimensionality (see e.g., Bellman (1961); Robins and Ritov (1997); Friedman et al. (2001)). Therefore, for applying these estimators in high-dimensional settings it is necessary to design and study non-parametric estimators that are able to overcome curse of dimensionality when possible.

The seminal work of Wager and Athey (2018) utilized random forests originally introduced by Breiman (2001) and adapted them nicely for estimating heterogeneous treatment effect. In particular, the authors demonstrated how the recursive partitioning idea, explained in Athey and Imbens (2016) for estimating heterogeneity in causal settings, can be further analyzed to establish asymptotic properties of such estimators. The main premise of random forests is that they are able to adaptively select nearest neighbors and that is very desirable in high-dimensional settings where discarding uninformative features is necessary for combating the curse of dimensionality. In a follow-up work, they extended these results and introduced Generalized Random Forests for more general setting of solving generalized method of moment (GMM) equations Athey et al. (2019). There has been some interesting developments of such ideas to other settings. Fan et al. (2018) introduced Distributional Nearest Neighbor (DNN) where they used 1-NN estimators together with sub-sampling and explained that by precisely combining two of these estimators for different sub-sampling sizes, the first order bias term can be efficiently removed. Friedberg et al. (2018) paired this idea with a local linear regression adjustment and introduced Local Linear Forests in order to improve forest estimations for smooth functions. Oprescu et al. (2018) incorporated the double machine learning methods of Chernozhukov et al. (2018a) into GMM framework of Athey et al. (2019)

and studied Orthogonal Random Forests in partially linear regression models with high-dimensional controls. Although forest kernels studied in Wager and Athey (2018) and Athey et al. (2019) seem to work well in high-dimensional applications, to the best of our knowledge, there still does not exists a theoretical result supporting it. In fact, all existing theoretical results suffer from the curse of dimensionality as they depend on the dimension of problem D.

The literature on machine learning and non-parametric statistics has recently studied how these worst-case performances can be avoided when the intrinsic dimension of problem is smaller than D. Please refer to Cutler (1993) for different notions of intrinsic dimension in metric spaces. Dasgupta and Freund (2008) studied random projection trees and showed that the structure of these trees do not depend on the actual dimension D, but rather on the intrinsic dimension d. They used the notion of Assouad Dimension, introduced by Assouad (1983), and proved that using random directions for splitting, the number of levels required for halving the diameter of a leaf scales as $O(d \log d)$. The follow-up work Verma et al. (2009) generalized these results for some other notions of dimension. Kpotufe and Dasgupta (2012) extended this idea to the regression setting and proved integrated risk bounds for random projection trees that were only dependent on intrinsic dimension. Kpotufe (2011); Kpotufe and Garg (2013) studied this in the context of k-NN and kernel estimations and established uniform point-wise risk bounds only depending on the local intrinsic dimension.

Our work is deeply rooted in the literature on intrinsic dimension explained above, literature on k-NN estimators (see e.g, Mack (1981); Samworth et al. (2012); Györfi et al. (2006); Biau and Devroye (2015); Berrett et al. (2019); Fan et al. (2018)), and generalized method of moments (see e.g., (Tibshirani and Hastie, 1987; Staniswalis, 1989; Fan et al., 1998; Hansen, 1982; Stone, 1977; Lewbel, 2007; Mackey et al., 2017)). We adapt the framework of Athey et al. (2019) and Oprescu et al. (2018) and solve a generalized moment problem using a DNN estimator, originally introduced and studied by Fan et al. (2018). We establish consistency and inference properties of this estimator and prove that these properties only depend on the local intrinsic dimension of problem. In particular, we prove that the finite sample estimation error of order $n^{-1/(d+2)}$ together with $n^{1/(d+2)}$-asymptotically normality result of DNN estimator for solving the generalized moment problem regardless of how big the actual dimension D is.

Our result differs from existing literature on intrinsic dimension (e.g., Kpotufe (2011); Kpotufe and Garg (2013)) since in addition to estimation guarantees for the regression setting, we also allow valid inference in solving conditional moment equations. Our asymptotic normality result is different from existing results for k-NN (see e.g., Mack (1981)), generalized method of moments (see e.g., Lewbel (2007)). This paper complements the work of Fan et al. (2018) and extends it to the generalized method of moment setting. Furthermore, we relax the common assumption on the existence of density for covariates and prove that DNN estimators are adaptive to intrinsic dimension.

We also provide the exact expression for the asymptotic variance of DNN estimator built using a k-NN kernel, which enables plug-in construction of confidence intervals, rather than the bootstrap method of (Efron, 1982) which was used by (Wager and Athey, 2018; Athey et al., 2019; Fan et al., 2018). While establishing consistency and asymptotic normality of our estimator, we also provide more general bounds on kernel shrinkage rate and also incrementality which can be useful for establishing asymptotic properties in other applications. One such application is given in high-dimensional settings where the exact nearest neighbor search is computationally expensive and Approximate Nearest Neighbor (ANN) search is often replaced in order to reduce this cost. Our flexible result allows us to use the state-of-the-art ANN algorithms (see e.g., Andoni et al. (2017, 2018)) while maintaining consistency and asymptotic normality.

Appendix B. Examples of spaces with small intrinsic dimension

In this section we provide examples of metric spaces that have small local intrinsic dimension. Our first example covers the setting where the distribution of data lies on a low-dimensional manifold (see e.g., Roweis and Saul (2000); Tenenbaum et al. (2000); Belkin and Niyogi (2003)). For instance, this happens for image inputs. Even though images are often high-dimensional (e.g., 4096 in the case of 64 by 64 images), all these images belong intrinsically to a 3-dimensional manifold.

Example 1 (Low dimensional manifold (adapted from Kpotufe (2011))) *Consider a d-dimensional submanifold $\mathcal{X} \subset \mathbb{R}^D$ and let μ have lower and upper bounded density on $\mathcal{X}$. The local intrinsic dimension of μ on $B(x, r)$ is d, provided that r is chosen small enough and some conditions on curvature hold. In fact, Bishop-Gromov theorem (see e.g., Carmo (1992)) implies that under such conditions, the volume of ball $B(x, r) \cap \mathcal{X}$ is $\Theta(r^d)$. This together with the lower and upper bound on the density implies that $\mu(B(x,r)\cap\mathcal{X})/\mu(B(x,\theta r)\cap\mathcal{X}) = \Theta(\theta^d)$, i.e. μ is (C, d)-homogeneous on $B(x, r)$ for some $C > 0$.*

Another example which happens in many applications, is sparse data. For example, in the bag of words representation of text documents, we usually have a vocabulary consisting of D words. Although D is usually large, each text document contains only a small number of these words. In this application, we expect our data (and measure) to have smaller intrinsic dimension. Before stating this example, let us discuss a more general example about mixture distributions.

Example 2 (Mixture distributions (adapted from Kpotufe (2011))) *Consider any mixture distribution $\mu = \sum_i \pi_i \mu_i$, with each μ_i defined on $\mathcal{X}$ with potentially different supports. Consider a point x and note that if $x \notin \operatorname{supp}(\mu_i)$, then there exists a ball $B(x, r_i)$ such that $\mu_i(B(x, r_i)) = 0$. This is true since the support of any probability measure is always closed, meaning that its complement is an open set. Now suppose that r is chosen small enough such that for any i satisfying $x \in \operatorname{supp}(\mu_i)$, μ_i is (C_i, d_i)-homogeneous on $B(x, r)$, while for any i satisfying $x \notin \operatorname{supp}(\mu_i)$ we have $\mu_i(B(x, r)) = 0$. Then,*

$$\begin{aligned}\mu(B(x,r)) = &\sum_i \pi_i\mu_i(B(x,r)) = \sum_{i:\mu_i(B(x,r))=0} \pi_i\mu_i(B(x,r)) + \sum_{i:\mu_i(B(x,r))>0} \pi_i\mu_i(B(x,r)) \\ &\le C\theta^{-d} \sum_{i:\mu_i(B(x,r))>0} \pi_i\mu_i(B(x,\theta r)) = C\theta^{-d}\sum_i \pi_i\mu_i(B(x,\theta r) = C\theta^{-d}\mu(B(x,\theta r))\,,\end{aligned}$$

where $C = \max_{i:\mu_i(B(x,r))>0} C_i$ and $d = \max_{i:\mu_i(B(x,r))>0} d_i$ and we used the fact that if $\mu_i(B(x,r)) = 0$ then $\mu_i(B(x,\theta r)) = 0$. Therefore, μ is (C, d)-homogeneous on $B(x, r)$.

This result applies to the case of d-sparse data and is explained in the following example.

Example 3 (d-sparse data) *Suppose that $\mathcal{X} \subset \mathbb{R}^D$ is defined as*

$$\mathcal{X} = \left\{ (x_1, x_2, \ldots, x_D) \in \mathbb{R}^D : \sum_{i=1}^D 1\{x_i \neq 0\} \le d \right\}.$$

Let μ be a probability measure on $\mathcal{X}$. In this case, we can write $\mathcal{X}$ as the union of $k = \binom{D}{d}$, d-dimensonal hyperplanes in $\mathbb{R}^D$. In fact,

$$\mathcal{X} = \cup_{1\le i_1 < i_2 < \cdots i_d \le D} \left\{ (x_1, x_2, \cdots, x_D) \in \mathbb{R}^D : x_j = 0,\ j \notin \{i_1, i_2, \ldots, i_d\} \right\}.$$

Letting $\mu_{i_1,i_2,\ldots,i_d}$ be the probability measure restricted to the hyperplane defined by $x_j = 0, j \notin \{i_1, i_2, \ldots, i_d\}$, we can express $\mu = \sum_{1\le i_1<i_2<\cdots i_d\le D} \pi_{i_1,i_2,\ldots,i_d}\mu_{i_1,i_2,\ldots,i_d}$. Therefore, the result of Example 2 implies that for any $x \in \mathcal{X}$, for r that is small enough μ is (C, d)-homogeneous on $B(x, r)$.

Our final example is about the product measure. This allows us to prove that any concatenation of spaces with small intrinsic dimension has a small intrinsic dimension as well.

Example 4 (Concatenation under the product measure) *Suppose that μ_i is a probability measure on $\mathcal{X}_i \subset \mathbb{R}^{D_1}$. Define $\mathcal{X} = \{(z_1, z_2) \mid z_1 \in \mathcal{X}_1, z_2 \in \mathcal{X}_2\}$ and let $\mu = \mu_1 \times \mu_2$ be the product measure on $\mathcal{X}$, i.e., $\mu(E_1 \times E_2) = \mu_1(E_1) \times \mu_2(E_2)$ for E_i that is μ_i-measurable, $i = 1, 2$. Suppose that μ_i is (C_i, d_i)-homogeneous on $B(x_i, r_i)$ and let $x = (x_1, x_2)$. Then, μ is (C, d)-homogeneous on $B(x, r)$, where $d = d_1 + d_2$, $r = \min\{r_1, r_2\}$ and $C = (C_1\, C_2\, r^{-(d_1+d_2)}\, 2^{(d_1+d_2)/2})/(r_1^{-d_1}\, r_2^{-d_2})$. To establish this, let $r = \min\{r_1, r_2\}$ and note that for any $\theta \in (0, 1)$ we have*

$$
\begin{aligned}
\mu(B(x,r)) \le\ & \mu(B(x_1,r)\times B(x_2,r)) = \mu_1(B(x_1,r))\times\mu_2(B(x_2,r)) \\
& \le \mu_1(B(x_1,r_1))\times\mu_2(B(x_2,r_2)) \\
\le\ & \left[C_1\left(\frac{r\theta}{r_1\sqrt{2}}\right)^{-d_1}\mu_1\left(B\left(x_1,\frac{r\theta}{\sqrt{2}}\right)\right)\right]\times\left[C_2\left(\frac{r\theta}{r_2\sqrt{2}}\right)^{-d_2}\mu_2\left(B\left(x_2,\frac{r\theta}{\sqrt{2}}\right)\right)\right] \\
=\ & \frac{C_1\,C_2\,r^{-(d_1+d_2)}}{r_1^{-d_1}\,r_2^{-d_2}\,\sqrt{2}^{-(d_1+d_2)}}\theta^{-d_1-d_2}\mu\left(B(x_1,r\theta/\sqrt{2})\times B(x_2,r\theta/\sqrt{2})\right) \\
\le\ & \frac{C_1\,C_2\,r^{-(d_1+d_2)\,2^{(d_1+d_2)/2}}}{r_1^{-d_1}\,r_2^{-d_2}}\theta^{-(d_1+d_2)}\mu(B(x,r\theta))\ ,
\end{aligned}
$$

where we used two simple inequalities that $\|(z_1, z_2) - (x_1, x_2)\|_2 \le r$ implies $\|z_i - x_i\|_2 \le r,\ i = 1, 2$, and further $\|z_i - x_i\|_2 \le r/\sqrt{2},\ i = 1, 2$, implies $\|(z_1, z_2) - (x_1, x_2)\|_2 \le r$.

Appendix C. Simulation Setting

Here we explain the settings for simulations shown in Figures 1 and 2.

C.1. Single test point

The data for single test point simulation, shown in Figure 1, has been generated as follows. Here $p = 1$, $D = 20$ and $d = 2$. All the points are generated using $X_i = AX_i^{\text{low}}$, where $A \in \mathbb{R}^{D\times d}$ and entries of A are independently sampled from $U[-1, 1]$. Components of each X_i^{low} are also generated independently from $U[-1, 1]$. We generate a fix test point $x_{\text{test}} = Ax_{\text{test}}^{\text{low}}$ and keep the matrix A throughout all Monte-Carlo iterations fixed. In each Monte-Carlo iteration, we generate $n = 20000$ training points as mentioned before. The values of Y_i are generated according to $Y_i = f(X_i) + \varepsilon_i$, where $f(X) = \frac{1}{1+\exp(-3X[0])}$, and $\varepsilon_i \sim \mathsf{N}(0, \sigma_e^2)$ with $\sigma_e = 1$. We are interested in estimate and inference for $f(x_{\text{test}})$ which is equivalent to solving for $\mathbf{E}[\psi(Z; \theta(x)) \mid X = x] = 0$ with $\psi(Z; \theta(x)) = Y - \theta(x)$ at $x = x_{\text{test}}$. We run DNN (Algorithm 2) for $k = 1, 2$ and 5 with parameter $s = s_\zeta$ chosen using Proposition 17 with $\zeta = 0.1$ over 1000 Monte-Carlo iterations and report the histogram and quantile-quantile plot of estimates compared to theoretical asymptotic normal distribution of estimates stemming from our characterization. In our simulations, we considered the complete U-statistic case, i.e., $B = \binom{n}{s}$.

C.2. Multiple test points

The data for the multiple test points simulation, shown in Figure 2, has been generated very similarly to the single test point setting. The only difference is that instead of generating a single test point we generate 100 test points. These test points together with matrix A are kept fixed throughout all 1000 Monte-Carlo iterations. We compare the performance of DNN (Algorithm 2) with parameter $s = s_\zeta$ chosen using Proposition 17 with $\zeta = 0.1$ with two benchmarks that set $s_d = n^{1.05d/(d+2)}$ and $s_D = n^{1.05D/(D+2)}$. This process has been repeated for $k = 1, 2$ and 5 and the coverage over a single run for all test points, the empirical coverage over 1000 runs, and chosen s_ζ versus s_d are depicted.

Appendix D. Proof of Theorem 6

Lemma 18 *For any $\theta \in \Theta$:*

$$\|\theta - \theta(x)\|_2 \leq \frac{2}{\lambda}\|m(x;\theta)\|_2 \,. \tag{19}$$

Proof By strong convexity of the loss $L(x;\theta)$ and the fact that $m(x;\theta(x)) = 0$, we have:

$$L(x;\theta) - L(x;\theta(x)) \geq \langle m(x;\theta(x)), \theta - \theta(x)\rangle + \frac{\lambda}{2}\cdot\|\theta - \theta(x)\|_2^2 = \frac{\lambda}{2}\cdot\|\theta - \theta(x)\|_2^2 \,.$$

By convexity of the loss $L(x;\theta)$ we have:

$$L(x;\theta(x)) - L(x;\theta) \geq \langle m(x;\theta), \theta(x) - \theta\rangle \,.$$

Combining the latter two inequalities we get:

$$\frac{\lambda}{2}\cdot\|\theta - \theta(x)\|_2^2 \leq \langle m(x;\theta), \theta - \theta(x)\rangle \leq \|m(x;\theta)\|_2 \cdot \|\theta - \theta(x)\|_2 \,.$$

Note that if $\|\theta - \theta(x)\|_2 = 0$, then the result is obvious. Otherwise, dividing over by $\|\theta - \theta(x)\|_2$ completes the proof of the lemma. ∎

Lemma 19 *Let $\Lambda(x;\theta) = m(x;\theta) - \Psi(x;\theta)$. Then the estimate $\hat{\theta}$ satisfies:*

$$\|m(x;\hat{\theta})\|_2 \leq \sup_{\theta\in\Theta}\|\Lambda(x;\theta)\|_2 \,. \tag{20}$$

Proof Observe that $\hat{\theta}$, by definition, satisfies $\Psi(x;\hat{\theta}) = 0$. Thus:

$$\|m(x;\hat{\theta})\|_2 = \|m(x;\hat{\theta}) - \Psi(x;\hat{\theta})\|_2 = \|\Lambda(x;\hat{\theta})\|_2 \leq \sup_{\theta\in\Theta}\|\Lambda(x;\theta)\|_2 \,.$$

∎

Lemma 20 *Suppose that the kernel is built with sub-sampling at rate s, in an honest manner (Assumption 3) and with at least $B \geq n/s$ sub-samples. If the base kernel satisfies kernel shrinkage in expectation, with rate $\epsilon(s)$, then w.p. $1-\delta$:*

$$\sup_{\theta\in\Theta}\|\Lambda(x;\theta)\|_2 \leq L_m\epsilon(s) + O\left(\psi_{\max}\sqrt{\frac{p\,s}{n}\left(\log\log(n/s) + \log(p/\delta)\right)}\right) \,. \tag{21}$$

Proof Define

$$\mu_0(x;\theta) = \mathbf{E}\left[\Psi_0(x;\theta)\right] ,$$

where we remind that Ψ_0 denotes the complete U-statistic:

$$\Psi_0(x;\theta) = \binom{n}{s}^{-1} \sum_{S_b \subset [n]:|S_b|=s} \mathbf{E}_{\omega_b}\left[\sum_{i \in S_b} \alpha_{S_b,\omega_b}(X_i)\psi(Z_i;\theta)\right] .$$

Here the expectation is taken with respect to the random draws of n samples. Then, the following result which is due to Oprescu et al. (2018) holds.

Lemma 21 (Adapted from Oprescu et al. (2018)) *For any θ and target x*

$$\mu_0(x;\theta) = \binom{n}{s}^{-1} \sum_{S_b \subset [n]:|S_b|=s} \mathbf{E}\left[\sum_{i \in S_b} \alpha_{S_b,\omega_b}(X_i) m(X_i;\theta)\right] .$$

In other words, Lemma 21 states that, in the expression for μ_0 we can simply replace $\psi(Z_i;\theta)$ with its expectation which is $m(X_i;\theta)$. We can then express $\Lambda(x;\theta)$ as sum of kernel error, sampling error, and sub-sampling error, by adding and subtracting appropriate terms, as follows:

$$\begin{aligned}
\Lambda(x;\theta) &= m(x;\theta) - \Psi(x;\theta) \\
&= \underbrace{m(x;\theta) - \mu_0(x;\theta)}_{\Gamma(x,\theta)=\text{Kernel error}} + \underbrace{\mu_0(x;\theta) - \Psi_0(x;\theta)}_{\Delta(x,\theta)=\text{Sampling error}} + \underbrace{\Psi_0(x;\theta) - \Psi(x;\theta)}_{\Upsilon(x,\theta)=\text{Sub-sampling error}}
\end{aligned}$$

The parameters should be chosen to trade-off these error terms nicely. We will now bound each of these three terms separately and then combine them to get the final bound.

Bounding the Kernel error. By Lipschitzness of m with respect to x and triangle inequality, we have:

$$\begin{aligned}
\|\Gamma(x;\theta)\|_2 &\le \binom{n}{s}^{-1} \sum_{S_b \subset [n]:|S_b|=s} \mathbf{E}\left[\sum_{i \in S_b} \alpha_{S_b,\omega_b}(X_i)\|m(x;\theta) - m(X_i;\theta)\|\right] \\
&\le L_m \binom{n}{s}^{-1} \sum_{S_b \subset [n]:|S_b|=s} \mathbf{E}\left[\sum_{i \in S_b} \alpha_{S_b,\omega_b}(X_i)\|x - X_i\|\right] \\
&\le L_m \binom{n}{s}^{-1} \sum_{S_b \subset [n]:|S_b|=s} \mathbf{E}\left[\sup\{\|x - X_i\| : \alpha_{S_b,\omega_b}(X_i) > 0\}\right] \\
&\le L_m\, \epsilon(s) ,
\end{aligned}$$

where the second to last inequality follows from the fact that $\sum_i |\alpha_{S_b}(X_i)| = 1$.

Bounding the Sampling error. For bounding the sampling error we rely on Lemma 35 and in particular Corollary 36. Observe that for each $j \in \{1, \ldots, p\}$, $\Psi_{0j}(x; \theta)$ is a complete U-statistic for each θ. Thus the sampling error defines a U-process over the class of symmetric functions $\text{conv}(\mathcal{F}_j) = \{f_j(\cdot; \theta) : \theta \in \Theta\}$, with $f_j(Z_1, \ldots, Z_s; \theta) = \mathbf{E}_\omega \left[\sum_{i=1}^s \alpha_{Z_{1:s},\omega}(X_i)\psi_j(Z_i; \theta)\right]$. Observe that since $f_j \in \text{conv}(\mathcal{F}_j)$ is a convex combination of functions in $\mathcal{F}_j = \{\psi_j(\cdot; \theta) : \theta \in \Theta\}$, the bracketing number of functions in $\text{conv}(\mathcal{F}_j)$ is upper bounded by the bracketing number of $\mathcal{F}_j$, which by our assumption, satisfies $\log(N_{[]}(\mathcal{F}_j, \epsilon, L_2)) = O(1/\epsilon)$. Moreover, by our assumptions on the upper bound $\psi_{\max}$ of $\psi_j(z; \theta)$, we have that $\sup_{f_j \in \text{conv}(\mathcal{F}_j)} \|f_j\|_2, \sup_{f_j \in \text{conv}(\mathcal{F}_j)} \|f_j\|_\infty \leq \psi_{\max}$. Thus all conditions of Corollary 36 are satisfied, with $\eta = G = \psi_{\max}$ and we get that w.p. $1 - \delta/2p$:

$$\sup_{\theta \in \Theta} |\Delta_j(x, \theta)| = O\left(\psi_{\max}\sqrt{\frac{s}{n}\left(\log\log(n/s) + \log(2p/\delta)\right)}\right). \tag{22}$$

By a union bound over j, we get that w.p. $1 - \delta/2$:

$$\sup_{\theta \in \Theta} \|\Delta_j(x, \theta)\|_2 \leq \sqrt{p} \max_{j \in [p]} \sup_{\theta \in \Theta} |\Delta_j(x, \theta)| = O\left(\psi_{\max}\sqrt{\frac{p\, s}{n}\left(\log\log(n/s) + \log(p/\delta)\right)}\right). \tag{23}$$

Bounding the Sub-sampling error. Sub-sampling error decays as B is increased. Note that for a fixed set of samples $\{Z_1, Z_2, \ldots, Z_n\}$, for a set S_b randomly chosen among all $\binom{n}{s}$ subsets of size s from the n samples, we have:

$$\mathbf{E}_{S_b,\omega_b}\left[\sum_{i \in S_b} \alpha_{S_b,\omega_b}(X_i)\psi(Z_i; \theta)\right] = \Psi_0(x; \theta).$$

Therefore, $\Psi(x; \theta)$ can be thought as the sum of B i.i.d. random variables each with expectation equal to $\Psi_0(x; \theta)$, where expectation is taken over B draws of sub-samples, each with size s. Thus one can invoke standard results on empirical processes for function classes as a function of the bracketing entropy. For simplicity, we can simply invoke Corollary 36 in the appendix for the case of a trivial U-process, with $s = 1$ and $n = B$ to get that w.p. $1 - \delta/2$:

$$\sup_{\theta \in \Theta} |\Upsilon(x; \theta)| = O\left(\psi_{\max}\sqrt{\frac{\log\log(B) + \log(2/\delta)}{B}}\right)$$

Thus for $B \geq n/s$, the sub-sampling error is of lower order than the sampling error and can be asymptotically ignored. Putting together the upper bounds on sampling, sub-sampling and kernel error finishes the proof of the Lemma. ∎

The probabilistic statement of the proof follows by combining the inequalities in the above three lemmas. The in expectation statement follows by simply integrating the exponential tail bound of the probabilistic statement.

Appendix E. Proof of Theorem 7

We will show asymptotic normality of $\hat{\alpha} = \left\langle \beta, \hat{\theta} \right\rangle$ for some arbitrary direction $\beta \in \mathbb{R}^p$, with $\|\beta\|_2 \leq R$. Consider the complete multi-dimensional U-statistic:

$$\Psi_0(x;\theta) = \binom{n}{s}^{-1} \sum_{S_b \subset [n]: |S_b| = s} \mathbf{E}_{\omega_b} \left[\sum_{i \in S_b} \alpha_{S_b, \omega_b}(X_i) \psi(Z_i; \theta) \right] . \tag{24}$$

Let

$$\Delta(x;\theta) = \Psi_0(x;\theta) - \mu_0(x;\theta) \tag{25}$$

where $\mu_0(x;\theta) = \mathbf{E}\left[\Psi_0(x;\theta)\right]$ (as in the proof of Theorem 6) and

$$\tilde{\theta} = \theta(x) - M_0^{-1} \Delta(x; \theta(x)) \tag{26}$$

Finally, let

$$\tilde{\alpha} \triangleq \left\langle \beta, \tilde{\theta} \right\rangle = \langle \beta, \theta(x) \rangle - \left\langle \beta, M_0^{-1}\, \Delta(x; \theta(x)) \right\rangle \tag{27}$$

For shorthand notation let $\alpha_0 = \langle \beta, \theta(x) \rangle$, $\psi_\beta(Z;\theta) = \left\langle \beta, M_0^{-1}(\psi(Z;\theta) - m(X;\theta)) \right\rangle$ and

$$\begin{aligned} \Psi_{0,\beta}(x;\theta) = &\ \left\langle \beta, M_0^{-1}\, \Delta(x;\theta(x)) \right\rangle \\ = &\ \binom{n}{s}^{-1} \sum_{S_b \subset [n]: |S_b| = s} \mathbf{E}_{\omega_b} \left[\sum_{i \in S_b} \alpha_{S_b, \omega_b}(X_i) \psi_\beta(Z_i; \theta) \right] \end{aligned}$$

be a single dimensional complete U-statistic. Thus we can re-write:

$$\tilde{\alpha} = \alpha_0 - \Psi_{0,\beta}(x; \theta(x))$$

We then have the following lemma which its proof is provided in Appendix I:

Lemma 22 *Under the conditions of Theorem 7:*

$$\frac{\Psi_{0,\beta}(x;\theta(x))}{\sigma_n(x)} \to \mathsf{N}(0,1)\,,$$

for $\sigma_n^2(x) = \frac{s^2}{n} \operatorname{Var}\left[\mathbf{E}\left[\sum_{i=1}^s K(x, X_i, \{X_j\}_{j=1}^s) \psi_\beta(Z_i;\theta) \mid X_1 \right] \right] = \Omega(\frac{s^2}{n}\eta(s))$.

Invoking Lemma 22 and using our assumptions on the kernel, we conclude that:

$$\frac{\tilde{\alpha} - \alpha_0(x)}{\sigma_n(x)} \to \mathsf{N}(0,1). \tag{28}$$

For some sequence σ_n^2 which decays at least as slow as $s^2\eta(s)/n$. Hence, since

$$\frac{\hat{\alpha} - \alpha_0}{\sigma_n(x)} = \frac{\tilde{\alpha} - \theta(x)}{\sigma_n(x)} + \frac{\hat{\alpha} - \tilde{\alpha}}{\sigma_n(x)},$$

if we show that $\frac{\hat{\alpha}-\tilde{\alpha}}{\sigma_n(x)} \to_p 0$, then by Slutsky's theorem we also have that:

$$\frac{\hat{\alpha}-\alpha_0}{\sigma_n(x)} \to \mathsf{N}(0,1), \tag{29}$$

as desired. Thus, it suffices to show that:

$$\frac{\|\hat{\alpha}-\tilde{\alpha}\|_2}{\sigma_n(x)} \to_p 0. \tag{30}$$

Observe that since $\|\beta\|_2 \leq R$, we have $\|\hat{\alpha}-\tilde{\alpha}\|_2 \leq R\|\hat{\theta}-\tilde{\theta}\|_2$. Thus it suffices to show that:

$$\frac{\|\hat{\theta}-\tilde{\theta}\|}{\sigma_n(x)} \to_p 0.$$

Lemma 23 *Under the conditions of Theorem 7, for* $\sigma_n^2(x) = \Omega\left(\frac{s^2}{n}\eta(s)\right)$*:*

$$\frac{\|\hat{\theta}-\tilde{\theta}\|}{\sigma_n(x)} \to_p 0. \tag{31}$$

Proof Performing a second-order Taylor expansion of $m_j(x;\theta)$ around $\theta(x)$ and observing that $m_j(x;\theta(x)) = 0$, we have that for some $\bar{\theta}_j \in \Theta$:

$$m_j(x;\hat{\theta}) = \left\langle \nabla_\theta m_j(x;\theta(x)), \hat{\theta}-\theta(x)\right\rangle + \underbrace{(\hat{\theta}-\theta(x))^\top H_j(x;\bar{\theta}_j)(\hat{\theta}-\theta(x))^\top}_{\rho_j}.$$

Letting $\rho = (\rho_1,\ldots,\rho_p)$, writing the latter set of equalities for each j in matrix form, multiplying both sides by M_0^{-1} and re-arranging, we get that:

$$\hat{\theta} = \theta(x) + M_0^{-1}m(x;\hat{\theta}) - M_0^{-1}\rho.$$

Thus by the definition of $\tilde{\theta}$ we have:

$$\hat{\theta}-\tilde{\theta} = M_0^{-1}\cdot(m(x;\hat{\theta}) + \Delta(x;\theta(x))) - M_0^{-1}\rho.$$

By the bounds on the eigenvalues of $H_j(x;\theta)$ and M_0^{-1}, we have that:

$$\|M_0^{-1}\rho\|_2 \leq \frac{L_H}{\lambda}\|\hat{\theta}-\theta(x)\|_2^2. \tag{32}$$

Thus we have:

$$\|\hat{\theta}-\tilde{\theta}\|_2 = \frac{1}{\lambda}\|m(x;\hat{\theta}) + \Delta(x;\theta(x))\|_2 + \frac{L_H}{\lambda}\|\hat{\theta}-\theta(x)\|_2^2.$$

By our estimation error Theorem 6, we have that the expected value of the second term on the right hand side is of order $O\left(\epsilon(s)^2, \frac{s}{n}\log\log(n/s)\right)$. Thus by the assumptions of the theorem, both are $o(\sigma_n)$. Hence, the second term is $o_p(\sigma_n)$.

We now argue about the convergence rate of the first term on the right hand side. Similar to the proof of Theorem 6, since $\Psi(x;\hat{\theta}) = 0$ we have:

$$m(x;\hat{\theta}) = m(x;\hat{\theta}) - \Psi(x;\hat{\theta}) = m(x;\hat{\theta}) - \Psi_0(x;\hat{\theta}) + \underbrace{\Psi_0(x;\hat{\theta}) - \Psi(x;\hat{\theta})}_{\text{Sub-sampling error}} .$$

We can further add and subtract μ_0 from $m(x;\hat{\theta})$.

$$\begin{aligned} m(x;\hat{\theta}) = &\ m(x;\hat{\theta}) - \mu_0(x;\hat{\theta}) + \mu_0(x;\hat{\theta}) - \Psi_0(x;\hat{\theta}) + \Psi_0(x;\hat{\theta}) - \Psi(x;\hat{\theta}) \\ = &\ m(x;\hat{\theta}) - \mu_0(x;\hat{\theta}) - \Delta(x;\hat{\theta}) + \Psi_0(x;\hat{\theta}) - \Psi(x;\hat{\theta}) . \end{aligned}$$

Combining we have:

$$m(x;\hat{\theta}) + \Delta(x;\theta(x)) = \underbrace{m(x;\hat{\theta}) - \mu_0(x;\hat{\theta})}_{C=\text{Kernel error}} + \underbrace{\Delta(x;\theta(x)) - \Delta(x;\hat{\theta})}_{F=\text{Stochastic equicontinuity term}} + \underbrace{\Psi_0(x;\hat{\theta}) - \Psi(x;\hat{\theta})}_{E=\text{Sub-sampling error}} .$$

Now similar to proof of Theorem 6 we bound different terms separately and combine the results.

Kernel Error. Term C is a kernel error and hence is upper bounded by $\epsilon(s)$ in expectation. Since, by assumption s is chosen such that $\epsilon(s) = o(\sigma_n(x))$, we ge that $\|C\|_2/\sigma_n(x) \to_p 0$.

Sub-sampling Error. Term E is a sub-sampling error, which can be made arbitrarily small if the number of drawn sub-samples is large enough and hence $\|E\|_2/\sigma_n(x) \to_p 0$. In fact, similar to the part about bounding sub-sampling error in Lemma 20 we have that that:

$$\mathbf{E}_{S_b}\left[\sum_{i\in S_b} \alpha_{S_b}(X_i)\psi(Z_i;\theta)\right] = \Psi_0(x;\theta) ,$$

Therefore, $\Psi(x;\theta)$ can be thought as the sum of B independent random variables each with expectation equal to $\Psi_0(x;\theta)$. Now we can invoke Corollary 36 in the appendix for the trivial U-process, with $s = 1, n = B$ to get that w.p. $1 - \delta_1$:

$$\sup_{\theta\in\Theta} \|\Psi_0(x;\theta) - \Psi(x;\theta)\| \le O\left(\Psi_{\max}\sqrt{\frac{\log\log(B) + \log(1/\delta_1)}{B}}\right) .$$

Hence, for $B \ge (n/s)^{5/4}$, due to our assumption that $(s/n \log\log(n/s))^{5/8} = o(\sigma_n(x))$ we get $\|E\|_2/\sigma_n(x) \to_p 0$.

Sampling Error. Thus it suffices that show that $\|F\|_2/\sigma_n(x) \to_p 0$, in order to conclude that $\frac{\|m(x;\hat{\theta}) + \Psi_0(x;\theta(x))\|_2}{\sigma_n(x)} \to_p 0$. Term F can be re-written as:

$$F = \Psi_0(x;\theta(x)) - \Psi_0(x;\hat{\theta}) - \mathbf{E}\left[\Psi_0(x;\theta(x)) - \Psi_0(x;\hat{\theta})\right] . \tag{33}$$

Observe that each coordinate j of F, is a stochastic equicontinuity term for U-processes over the class of symmetric functions $\text{conv}(\mathcal{F}_j) = \{f_j(\cdot;\theta) : \theta \in \Theta\}$, with $f_j(Z_1,\ldots,Z_s;\theta) = \mathbf{E}_\omega\left[\sum_{i=1}^s \alpha_{Z_{1:s},\omega}(X_i)(\psi_j(Z_i;\theta(x)) - \psi_j(Z_i;\theta))\right]$. Observe that since $f_j \in \text{conv}(\mathcal{F}_j)$ is a convex combination of functions in $\mathcal{F}_j = \{\psi_j(\cdot;\theta(x)) - \psi_j(\cdot;\theta) : \theta \in \Theta\}$, the bracketing number of functions in $\text{conv}(\mathcal{F}_j)$ is upper bounded by the bracketing number of $\mathcal{F}_j$, which in turn is upper bounded by the bracketing number of the function class $\{\psi_j(\cdot;\theta) : \theta \in \Theta\}$, which by our assumption, satisfies $\log(N_{[]}(\mathcal{F}_j, \epsilon, L_2)) = O(1/\epsilon)$. Moreover, under the variogram assumption and the lipschitz moment assumption we have that if $\|\theta - \theta(x)\| \leq r \leq 1$, then:

$$\begin{aligned}
\|f_j(\cdot;\theta)\|_{P,2}^2 = &\ \mathbf{E}\left[\left(\sum_{i=1}^{s} \alpha_{Z_{1:s}}(X_i)(\psi_j(Z_i;\theta(x)) - \psi_j(Z_i;\theta)\right)^2\right] \\
\leq &\ \mathbf{E}\left[\sum_{i=1}^{s} \alpha_{Z_{1:s}}(X_i)\left(\psi_j(Z_i;\theta(x)) - \psi_j(Z_i;\theta)\right)^2\right] && \text{(Jensen's inequality)} \\
= &\ \mathbf{E}\left[\sum_{i=1}^{s} \alpha_{Z_{1:s}}(X_i)\mathbf{E}\left[\psi_j(Z_i;\theta(x)) - \psi_j(Z_i;\theta)\right]^2 |X_i]\right] && \text{(honesty of kernel)} \\
= &\ \mathbf{E}\left[\sum_{i=1}^{s} \alpha_{Z_{1:s}}(X_i)\left(\text{Var}(\psi(Z;\theta(x)) - \psi(Z;\theta)|X_i) + (m(X_i;\theta(x)) - m(X_i;\theta))^2\right)\right] \\
\leq &\ L_\psi\|\theta - \theta(x)\| + L_J^2\|\theta - \theta(x)\|^2 \leq L_\psi r + L_J^2 r^2 = O(r)\,.
\end{aligned}$$

Moreover, $\|f_j\|_\infty \leq 2\psi_{\max}$. Thus we can apply Corollary 36, with $\eta = \sqrt{L_\psi r + L_J^2 r^2} = O(\sqrt{r})$ and $G = 2\psi_{\max}$ to get that if $\|\hat{\theta} - \theta(x)\| \leq r$, then w.p. $1 - \delta/p$:

$$\begin{aligned}
|F_j| \leq &\ \sup_{\theta:\|\theta-\theta(x)\|\leq r}\left|\Psi_0(x;\theta(x)) - \Psi_0(x;\hat{\theta}) - \mathbf{E}\left[\Psi_0(x;\theta(x)) - \Psi_0(x;\hat{\theta})\right]\right| \\
= &\ O\left(\left(r^{1/4} + \sqrt{r}\sqrt{\log(p/\delta) + \log\log(n/(s\,r))}\right)\sqrt{\frac{s}{n}}\right) \\
= &\ O\left(\left(r^{1/4}\sqrt{\log(p/\delta) + \log\log(n/s)}\right)\sqrt{\frac{s}{n}}\right) \triangleq \kappa(r,s,n,\delta)\,.
\end{aligned}$$

Using a union bound this implies that w.p. $1 - \delta$ we have

$$\max_j |F_j| \leq \kappa(r,s,n,\delta)\,.$$

By our MSE theorem and also Markov's inequality, w.p. $1 - \delta'$: $\|\hat{\theta} - \theta(x)\| \leq \nu(s)/\delta'$, where:

$$\nu(s) = \frac{1}{\lambda}\left(L_m\epsilon(s) + O\left(\psi_{\max}\sqrt{\frac{p\,s}{n}\log\log(p\,s/n)}\right)\right)$$

Thus using a union bound w.p. $1 - \delta - \delta'$, we have:

$$\max_j |F_j| = O\left(\kappa(\nu(s)/\delta', s, n, \delta)\right)$$

To improve readability from here we ignore all the constants in our analysis, while we keep all terms (even $\log$ or $\log\log$ terms) that depend on s and n. Note that we can even ignore δ and δ', because they can go to zero at very slow rate such that terms $\log(1/\delta)$ or even $\delta'^{1/4}$ appearing in the analysis grow slower than $\log\log$ terms. Now, by the definition of $\nu(s)$ and $\kappa(r, s, n, \delta')$, as well as invoking the inequality $(a+b)^{1/4} \leq a^{1/4} + b^{1/4}$ for $a, b > 0$ we have:

$$\max_j |F_j| \leq O(\kappa(\nu(s)/\delta', s, n, \delta)) \leq O\left(\epsilon(s)^{1/4}\left(\frac{s}{n}\log\log(n/s)\right)^{1/2} + \left(\frac{s}{n}\log\log(n/s)\right)^{5/8}\right), \tag{34}$$

Hence, using our Assumption on the rates in the statement of Theorem 7 we get that both of the terms above are $o(\sigma_n(x))$. Therefore, $\|F\|_2/\sigma_n(x) \to_p 0$. Thus, combining all of the above, we get that:

$$\frac{\|\tilde{\theta} - \hat{\theta}\|}{\sigma_n(x)} = o_p(1)$$

as desired. ■

Remark 24 *Our notion of incrementality is slightly different from that of Wager and Athey (2018), as there the incrementality is defined as* $\mathrm{Var}\left[\mathbf{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1\right]\right]$. *However, using the tower law of expectation*

$$\begin{aligned} &\mathbf{E}\left[\mathbf{E}[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1]^2\right] - \mathrm{Var}\left[\mathbf{E}[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1]\right] \\ =\ &\mathbf{E}\left[\mathbf{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1\right]\right]^2 = \mathbf{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s)\right]^2 . \end{aligned}$$

For a symmetric kernel the term $\mathbf{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s)\right]^2$ *is equal to* $1/s^2$ *and is asymptotically negligible compared to* $\mathrm{Var}\left[\mathbf{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1\right]\right]$, *which usually decays at a slower rate.*

Appendix F. Lower Bound on Incrementality as Function of Kernel Shrinkage

We give a generic lower bound on the quantity $\mathbf{E}[\mathbf{E}[K(x, X_1, \{Z_j\}_{j=1}^s)|X_1]^2]$ that depends only on the Kernel shrinkage. The bound essentially implies that if we know that the probability that the distribution of x's assigns to a ball of radius $\epsilon(s, 1/2s)$ around the target x is of order $1/s$, i.e. we should expect at most a constant number of samples to fall in the kernel shrinkage ball, then the main condition on incrementality of the kernel, required for asymptotic normality, holds. In some sense, this property states that the kernel shrinkage behavior is tight in the following sense. Suppose that the kernel was assigning positive weight to at most a constant number of k samples. Then kernel shrinkage property states that with high probability we expect to see at least k samples in a ball of radius $\epsilon(s, \delta)$ around x. The above assumption says that we should also not expect to see too many samples in that radius, i.e. we should also expect to see at most a constant number $K > k$ of samples in that radius. Typically, the latter should hold, if the characterization of $\epsilon(s, \delta)$ is tight, in the sense that if we expected to see too many samples in the radius, then most probably we could have improved our analysis on Kernel shrinkage and given a better bound that shrinks faster.

F.1. Proof of Lemma 8

By the Paley-Zygmund inequality, for any random variable $Z \geq 0$ and for any $\delta \in [0, 1]$:

$$\mathbf{E}[Z^2] \geq (1-\delta)^2 \frac{\mathbf{E}[Z]^2}{\Pr[Z \geq \delta \mathbf{E}[Z]]}$$

Let $W_1 = K(x, X_1, \{Z_j\}_{j=1}^s)$. Then, applying the latter to the random variable $Z = \mathbf{E}[W_1|X_1]$ and observing that by symmetry $\mathbf{E}[Z] = \mathbf{E}[W_1] = 1/s$, yields:

$$\mathbf{E}\left[\mathbf{E}[W_1|X_1]^2\right] \geq \frac{(1-\delta)^2\,\mathbf{E}[W_1]^2}{\Pr[\mathbf{E}[W_1|X_1] > \delta\mathbf{E}[W_1]]} = \frac{(1-\delta)^2\,(1/s)^2}{\Pr[\mathbf{E}[W_1|X_1] > \delta/s]}$$

Moreover, observe that by the definition of $\epsilon(s,\rho)$ for some $\rho > 0$:

$$\Pr[W_1 > 0 \wedge \|X_1 - x\| \geq \epsilon(s,\rho)] \leq \rho$$

This means that at most a mass $\rho\, s/\delta$ of the support of X_1 in the region $\|X_1 - x\| \geq \epsilon(s,\rho)$ can have $\Pr[W_1 > 0|X_1] \geq \delta/s$. Otherwise the overall probability that $W_1 > 0$ in the region of $\|X_1 - x\| \geq \epsilon(s,\rho)$ would be more than ρ. Thus we have that except for a region of mass $\rho\, s/\delta$, for each X_1 in the region $\|X_1 - x\| \geq \epsilon(s,\rho)$: $\mathbf{E}[W_1|X_1] \leq \delta/s$. Combining the above we get:

$$\Pr[\mathbf{E}[W_1|X_1] \leq \delta/s] \geq \Pr[\|X_1 - x\| \geq \epsilon(s,\rho)] - \rho\, s/\delta$$

Thus:

$$\Pr[\mathbf{E}[W_1|X_1] > \delta/s] \leq \Pr[\|X_1 - x\| \leq \epsilon(s,\rho)] + \rho\, s/\delta = \mu(B(x, \epsilon(s,\delta))) + \rho\, s/\delta$$

Since ρ was arbitrarily chosen, the latter upper bound holds for any ρ, which yields the result.

F.2. Proof of Corollary 9

Thus applying Lemma 8 with $\delta = 1/2$ yields:

$$\mathbf{E}[\mathbf{E}[K(x, X_1, \{Z_j\}_{j=1}^s)|X_1]^2] \geq \frac{(1/2s)^2}{\inf_{\rho>0}\left(\mu(B(x,\epsilon(s,\rho))) + 2\rho\, s\right)}$$

Observe that:

$$\mu(B(x,\epsilon(s,\rho))) \leq C\epsilon(s,\rho)^d \mu(B(x,r)) = O\left(\frac{\log(1/\rho)}{s}\right)$$

Hence:

$$\inf_{\rho>0}\left(\mu(B(x,\epsilon(s,\rho))) + 2\rho\, s\right) = O\left(\inf_{\rho>0}\left(\frac{\log(1/\rho)}{s} + 2\rho\, s\right)\right) = O\left(\frac{\log(s)}{s}\right)$$

where the last follows by choosing $\rho = 1/s^2$. Combining all the above yields:

$$\mathbf{E}[\mathbf{E}[K(x, X_1, \{Z_j\}_{j=1}^s)|X_1]^2] = \Omega\left(\frac{1}{s\log(s)}\right)$$

Appendix G. Proofs of Section 5

G.1. Proof of Lemma 10

For proving this result, we rely on Bernstein's inequality which is stated below:

Proposition 25 (Bernstein's Inequality) *Suppose that random variables $Z_1, Z_2, \ldots, Z_n$ are i.i.d., belong to $[-c, c]$ and $\mathbf{E}[Z_i] = \mu$. Let $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$ and $\sigma^2 = \mathrm{Var}(Z_i)$. Then, for any $\theta > 0$,*

$$\Pr\left(|\bar{Z}_n - \mu| > \theta\right) \le 2\exp\left(\frac{-n\theta^2}{2\sigma^2 + 2c\theta/3}\right).$$

This also implies that w.p. at least $1-\delta$ the following holds:

$$|\bar{Z}_n - \mu| \le \sqrt{\frac{2\sigma^2\log(2/\delta)}{n}} + \frac{2c\log(2/\delta)}{3n}. \tag{35}$$

Let A be any μ-measurable set. An immediate application of Bernstein's inequality to random variables $Z_i = 1\{X_i \in A\}$, implies that w.p. $1-\delta$ over the choice of covariates $(X_i)_{i=1}^s$, we have:

$$|\mu_s(A) - \mu(A)| \le \sqrt{\frac{2\mu(A)\log(2/\delta)}{s}} + \frac{2\log(2/\delta)}{3s}.$$

In above, we used the fact that $\mathrm{Var}(Z_i) = \mu(A)(1-\mu(A)) \le \mu(A)$. This result has the following corollary.

Corollary 26 *Define $U = 2\log(2/\delta)/s$ and let A be an arbitrary μ-measurable set. Then, w.p. $1-\delta$ over the choice of training samples, $\mu(A) \ge 4U$ implies $\mu_s(A) \ge U$.*

Proof Define $U = 2\log(2/\delta)/s$. Then, Bernstein's inequality in Proposition 25 implies that w.p. $1-\delta$ we have

$$|\mu_s(A) - \mu(A)| \le \sqrt{U\mu(A)} + \frac{U}{3}.$$

Assume that $\mu(A) \ge 4U$, we want to prove that $\mu_s(A) \ge U$. Suppose, the contrary, i.e., $\mu_s(A) < U$. Then, by dividing the above equation by $\mu(A)$ we get

$$\left|\frac{\mu_s(A)}{\mu(A)} - 1\right| \le \sqrt{\frac{U}{\mu(A)}} + \frac{1}{3}\frac{U}{\mu(A)},.$$

Note that since $\mu_s(A) < U < \mu(A)$, by letting $z = U/\mu(A) \le 1/4$ the above implies that

$$1 - z \le \sqrt{z} + \frac{z}{3} \Rightarrow \frac{4}{3}z + \sqrt{z} - 1 \ge 0,$$

which as $z > 0$ only holds for

$$\sqrt{z} \ge \frac{-3+\sqrt{57}}{8} \Rightarrow z \ge 0.3234.$$

This contradicts with $z \le 1/4$, implying the result. ■

Now we are ready to finish the proof of Lemma 10. First, note that using the definition of (C, d)-homogeneous measure. Note that for any $\theta \in (0,1)$ we have $\mu(B(x, \theta r)) \ge (1/C)\theta^d \mu(B(x, r))$. Replace $\theta r = \epsilon$ in above. It implies that for any $\epsilon \in (0, r)$

$$\mu(B(x,\epsilon)) \ge \frac{1}{C\,r^d}\epsilon^d \mu(B(x,r)). \tag{36}$$

Pick $\epsilon_k(s,\delta)$ according to

$$\epsilon_k(s,\delta) = r\left(\frac{8C\,\log(2/\delta)}{\mu(B(x,r))s}\right)^{1/d}.$$

Note that for having $\epsilon_k(s,\delta) \in (0,r)$ we need

$$\log(2/\delta) \le \frac{1}{8\,C}\mu(B(x,r))s \Rightarrow \delta \ge 2\exp\left(-\frac{1}{8\,C}\mu(B(x,r)s\right).$$

Therefore, replacing this choice of $\epsilon_k(s,\delta)$ in Equation (36) implies that $\mu(B(x,\epsilon_k(s,\delta))) \ge \frac{8\log(2/\delta)}{s}$. Now we can use the result of Corollary 26 for the choice $A = B(x,\epsilon_k(s,\delta))$. It implies that w.p. $1-\delta$ over the choice of s training samples, we have

$$\mu_s(B(x,\epsilon_k(s,\delta))) \ge \frac{2\log(2/\delta)}{s}.$$

Note that whenever $\delta \le \exp(-k/2)/2$ we have

$$\frac{2\log(2/\delta)}{s} \ge \frac{k}{s}.$$

Therefore, w.p. $1-\delta$ we have

$$\|x - X_{(k)}\| \le \epsilon_k(s,\delta) = O\left(\frac{\log(1/\delta)}{s}\right)^{1/d}.$$

G.2. Proof of Corollary 11

Lemma 10 shows that for any $t = \epsilon_k(s,\delta) = r\left(\frac{8C\log(2/\delta)}{\mu(B(x,r))s}\right)^{1/d}$, such that $t \le r$ and $t \ge r\left(\frac{4k\,C}{\mu(B(x,r))s}\right)^{1/d}$, we have that:

$$\Pr[\|x - X_{(k)}\|_2 \ge \epsilon_k(s,\delta)] \le \delta.$$

Let $\rho = \frac{1}{r}\left(\frac{\mu(B(x,r))}{8C}\right)^{1/d}$, which is a constant. Solving for δ in terms of t we get:

$$\Pr[\|x - X_{(k)}\|_2 \ge t] \le 2\exp\left(-\rho^d\, s\, t^d\right),$$

for any $t \in \left[\frac{(s/k)^{-1/d}}{\rho}, r\right]$. Thus, noting that X_i's and target x both belong to $\mathcal{X}$ that has diameter $\Delta_{\mathcal{X}}$, we can upper bound the expected value of $[\|x - X_{(k)}\|_2$ as:

$$\begin{aligned}
\mathbf{E}\left[\|x - X_{(k)}\|_2\right] &= \int_0^{\Delta_{\mathcal{X}}} \Pr\left[\|x - X_{(k)}\|_2 \ge t\right] dt \\
&\le \frac{(s/k)^{-1/d}}{\rho} + \int_{\rho\,(s/k)^{-1/d}}^{r} \Pr\left[\|x - X_{(k)}\|_2 \ge t\right] dt + \Pr\left[\|x - X_{(k)}\|_2 \ge r\right](\Delta_{\mathcal{X}} - r) \\
&\le \frac{(s/k)^{-1/d}}{\rho} + \int_{\rho\,(s/k)^{-1/d}}^{r} 2\exp\left\{-\rho^d\, s\, t^d\right\} dt + 2\exp\left\{-\rho^d\, r^d\, s\right\}(\Delta_{\mathcal{X}} - r).
\end{aligned}$$

Note that for s larger than some constant, we have $\exp\left\{-\rho^d\, r^d\, s\right\} \le s^{-1/d}$. Thus the first and last terms in the latter summation are of order $\left(\frac{1}{s}\right)^{1/d}$. We now show that the same holds for the middle term, which would complete the proof. By setting $u = \rho^d\, s\, t^d$ and doing a change of variables in the integral we get:

$$\int_{\rho\,(s/k)^{1/d}}^{r} 2\exp\left\{-\rho^d\, s\, t^d\right\} dt \le \int_0^\infty 2\exp\left\{-\rho^d\, s\, t^d\right\} dt$$
$$= \frac{1}{d\,\rho\, s^{1/d}} \int_0^\infty u^{1/d-1} \exp\left\{-u\right\} du = \frac{s^{-1/d}}{\rho} \frac{1}{d} \Gamma(1/d)\,.$$

where Γ is the Gamma function. Since by the properties of the Gamma function $z\Gamma(z) = \Gamma(z+1)$, the latter evaluates to: $\frac{s^{-1/d}}{\rho}\Gamma((d+1)/d)$. Since $(d+1)/d \in [1,2]$, we have that $\Gamma((d+1)/d) \le 2$. Thus the middle term is upper bounded by $\frac{2s^{-1/d}}{\rho}$, which is also of order $\left(\frac{1}{s}\right)^{1/d}$.

G.3. Proof of Lemma 13

Before proving this lemma we state and prove and auxiliary lemma which comes in handy in our proof.

Lemma 27 *Let P_1 denote the mass that the density of the distribution of X_i puts on the ball around x with radius $\|x - X_1\|_2$, which is a random variable as it depends on X_1. Then, for any $s \ge k$ the following holds:*

$$\mathbf{E}\left[\sum_{i=0}^{k-1} \binom{s-1}{i} (1-P_1)^{s-1-i} P_1^i\right] = \mathbf{E}\left[\mathbf{E}\left[S_1 \mid X_1\right]\right] = \frac{k}{s}\,.$$

Proof The proof is an easy consequence of symmetry. Let $S_1 = 1\{\text{sample 1 is among } k \text{ nearest neighbors}\}$, then we can write

$$\mathbf{E}\left[\mathbf{E}[S_1 \mid X_1]\right] = \mathbf{E}\left[\sum_{i=0}^{k-1} \binom{s-1}{i} (1-P_1)^{s-1-i} P_1^i\right]\,,$$

which simply computes the probability that there are at most $k-1$ other points in the ball with radius $\|x - X_1\|$. Now, by using the tower law

$$\mathbf{E}\left[\mathbf{E}[S_1 \mid X_1]\right] = \mathbf{E}[S_1] = \frac{k}{s}\,,$$

which holds because of the symmetry. In other words, the probability that sample 1 is among the k-NN is equal to k/s. Hence, the conclusion follows. ■

We can finish the proof of Lemma 13. Define $S_1 = 1\{\text{sample 1 is among } k \text{ nearest neighbors}\}$, then we can write

$$\mathbf{E}\left[\mathbf{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1\right]^2\right] = \frac{1}{k^2}\mathbf{E}\left[\mathbf{E}\left[S_1 \mid X_1\right]^2\right]\,.$$

Recall that if P_1 denotes the mass that the density of the distribution of X_i puts on the ball around x with radius $\|x - X_1\|_2$, which is a random variable depending on X_1. Therefore,

$$\mathbf{E}\left[S_1 \mid X_1\right] = \sum_{i=0}^{k-1} \binom{s-1}{i} (1-P_1)^{s-1-i} P_1^i .$$

Now we can write

$$\begin{aligned}
\mathbf{E}\left[\mathbf{E}\left[S_1 \mid X_1\right]^2\right] &= \mathbf{E}\left[\left(\sum_{i=0}^{k-1} \binom{s-1}{i} (1-P_1)^{s-1-i} P_1^i\right)^2\right] \\
&= \mathbf{E}\left[\sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \binom{s-1}{i}\binom{s-1}{j} (1-P_1)^{2s-2-i-j} P_1^{i+j}\right] \\
&= \mathbf{E}\left[\sum_{t=0}^{2k-2} (1-P_1)^{2s-2-t} P_1^t \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \binom{s-1}{i}\binom{s-1}{j} 1\left\{i+j=t\right\}\right] \\
&= \mathbf{E}\left[\sum_{t=0}^{2k-2} (1-P_1)^{2s-2-t} P_1^t \sum_{i=\max\{0,t-(k-1)\}}^{\min\{t,k-1\}} \binom{s-1}{i}\binom{s-1}{t-i}\right] \\
&= \mathbf{E}\left[\sum_{t=0}^{2k-2} a_t \left(1-P_1\right)^{2s-2-t} P_1^t\right]
\end{aligned}$$

Now using Lemma 27 (where s is replaced by $2s-1$) we know that for any value of $0 \le r \le 2s-2$ we have

$$\mathbf{E}\left[\sum_{t=0}^{r} b_t \left(1-P_1\right)^{2s-2-r} P_1^r\right] = \mathbf{E}\left[\sum_{t=0}^{r} \binom{2s-2}{t} \left(1-P_1\right)^{2s-2-t} P_1^t\right] = \frac{r+1}{2s-1} . \tag{37}$$

This implies that for any value of r we have $\mathbf{E}\left[b_r(1-P_1)^{2s-2-r} P_1^r\right] = 1/(2s-1)$. The reason is simple. Note that the above is obvious for $r = 0$ using Equation (37). For other values of $r \ge 1$, we can write Equation (37) for values r and $r-1$. Taking their difference implies the result. Note that this further implies that $\mathbf{E}\left[(1-P_1)^{2s-2-r} P_1^r\right] = 1/(b_r\,(2s-1))$, as b_r is a constant. Therefore, by plugging this back into the expression of $\mathbf{E}[\mathbf{E}[S_1 \mid X_1]^2]$ we have

$$\mathbf{E}[\mathbf{E}[S_1|X_1]^2] = \mathbf{E}\left[\sum_{t=0}^{2k-2} a_t \left(1-P_1\right)^{2s-2-t} P_1^t\right] = \frac{1}{2s-1}\left(\sum_{t=0}^{2k-2} \frac{a_t}{b_t}\right) ,$$

which implies the desired result.

Remark 28 *Note that $b_t = \binom{2s-2}{t}$ since we can view b_t as follows: how many different subsets of size t can we create from a set of $2s-2$ elements if we pick a number $i = \{0, \ldots, t\}$ and then choose i elements from the first half of these elements and $t-i$ elements from the second half. Observe that this process creates all possible sets of size t from among the $2s-2$ elements, which is equal to $\binom{2s-2}{t}$.*

Furthermore, $a_t = b_t$ for $0 \le t \le k-1$ and for any $k \le t \le 2k-2$, after some little algebra, we have

$$\frac{2k-1-t}{t+1} \le \frac{a_t}{b_t} \le 1\,.$$

This implies that the summation appeared in Lemma 29 satisfies

$$k + \sum_{t=k}^{2k-2} \frac{2k-1-t}{t+1} \le \sum_{t=0}^{2k-2} \frac{a_t}{b_t} \le 2k-1\,.$$

G.4. Proof of Theorem 14

Note that according to Lemma 38, the asymptotic variance $\sigma_{n,j}^2(x) = \frac{s^2}{n} \operatorname{Var}\left[\Phi_1(Z_1)\right]$, where $\Phi_1(Z_1) = \frac{1}{k}\mathbf{E}[\sum_{i\in H_k(x,s)} \left\langle e_j, M_0^{-1}\psi(Z_i;\theta(x))\right\rangle \mid Z_1]$. Therefore, once we establish an expression for $\operatorname{Var}\left[\Phi_1(Z_1)\right]$ we can finish the proof of this theorem. The following lemma provides such result.

Lemma 29 *Suppose that the kernel K is the k-NN kernel and let $\sigma_j^2(x) = \operatorname{Var}\left(\left\langle e_j, M_0^{-1}\psi(z;\theta(x))\right\rangle \mid X = x\right)$. Moreover, suppose that $\epsilon_k(s, 1/s^2) \to 0$ for any constant k. Then:*

$$\operatorname{Var}\left[\Phi_1(Z_1)\right] = \sigma_j(x)^2\, \mathbf{E}\left[\mathbf{E}\left[K(x, X_1, \{Z_j\}_{j=1}^s) \mid X_1\right]^2\right] + o(1/s) = \frac{\sigma_j^2(x)}{(2s-1)k^2}\left(\sum_{t=0}^{2k-2} \frac{a_t}{b_t}\right) + o(1/s)$$

where the second equality above holds due to Lemma 13 and sequences a_t and b_t, for $0 \le t \le 2k-2$, are defined in Lemma 13.

Proof In this proof for simplicity we let $Y_i = \left\langle e_j, M_0^{-1}\psi(Z_i;\theta(x))\right\rangle$ and $\mu(X_i) = \mathbf{E}[Y_i] = \left\langle e_j, M_0^{-1}m(X_i;\theta(x))\right\rangle$. Let $Z_{(i)}$ denote the random variable of the i-th closest sample to x. For the case of k-NN we have that:

$$k\,\Phi_1(Z_1) = \mathbf{E}\left[\sum_{i=1}^k Y_{(i)} \mid Z_1\right].$$

Let $S_1 = \mathbf{1}\{\text{sample 1 is among } k \text{ nearest neighbors}\}$. Then we have:

$$k\,\Phi_1(Z_1) = \mathbf{E}\left[S_1 \sum_{i=1}^k Y_{(i)} \mid Z_1\right] + \mathbf{E}\left[(1-S_1)\sum_{i=1}^k Y_{(i)} \mid Z_1\right].$$

Let $\tilde{Y}_{(i)}$ denote the label of the i-th closest point to x, excluding sample 1. Then:

$$\begin{aligned} k\,\Phi_1(Z_1) =\ & \mathbf{E}\left[S_1 \sum_{i=1}^k Y_{(i)} \mid Z_1\right] + \mathbf{E}\left[(1-S_1)\sum_{i=1}^k \tilde{Y}_{(i)} \mid Z_1\right] \\ =\ & \mathbf{E}\left[S_1 \sum_{i=1}^k \left(Y_{(i)} - \tilde{Y}_{(i)}\right) \mid Z_1\right] + \mathbf{E}\left[\sum_{i=1}^k \tilde{Y}_{(i)} \mid Z_1\right]. \end{aligned}$$

Observe that $\tilde{Y}_{(i)}$ are all independent of Z_1. Hence:

$$k\,\Phi_1(Z_1) = \mathbf{E}\left[S_1 \sum_{i=1}^{k}\left(Y_{(i)} - \tilde{Y}_{(i)}\right) \mid Z_1\right] + \mathbf{E}\left[\sum_{i=1}^{k} \tilde{Y}_{(i)}\right].$$

Therefore the variance of $\Phi(Z_1)$ is equal to the variance of the first term on the right hand side. Hence:

$$\begin{aligned} k^2\ \mathrm{Var}\left[\Phi_1(Z_1)\right] &= \mathbf{E}\left[\mathbf{E}\left[S_1 \sum_{i=1}^{k}\left(Y_{(i)} - \tilde{Y}_{(i)}\right) \mid Z_1\right]^2\right] - \mathbf{E}\left[S_1 \sum_{i=1}^{k}\left(Y_{(i)} - \tilde{Y}_{(i)}\right)\right]^2 \\ &= \mathbf{E}\left[\mathbf{E}\left[S_1 \sum_{i=1}^{k}\left(Y_{(i)} - \tilde{Y}_{(i)}\right) \mid Z_1\right]^2\right] + o(1/s). \end{aligned}$$

Where we used the fact that:

$$\left|\mathbf{E}\left[S_1 \sum_{i=1}^{k}\left(Y_{(i)} - \tilde{Y}_{(i)}\right)\right]\right| \leq \mathbf{E}\left[S_1\right] 2k\psi_{\max} = \frac{2k^2\psi_{\max}}{s}. \tag{38}$$

Moreover, observe that under the event that $S_1 = 1$, we know that the difference between the closest k values and the closest k values excluding 1 is equal to the difference between the Y_1 and $Y_{(k+1)}$. Hence:

$$\mathbf{E}\left[S_1 \sum_{i=1}^{k}\left(Y_{(i)} - \tilde{Y}_{(i)}\right) \mid Z_1\right] = \mathbf{E}\left[S_1\left(Y_1 - Y_{(k+1)}\right) \mid Z_1\right] = \mathbf{E}\left[S_1\left(Y_1 - \mu(X_{(k+1)})\right) \mid Z_1\right].$$

where the last equation holds from the fact that for any $j \neq 1$, conditional on X_j, the random variable Y_j is independent of Z_1 and is equal to $\mu(X_j)$ in expectation. Under the event $S_1 = 1$, we know that the $(k+1)$-th closest point is different from sample 1. We now argue that up to lower order terms, we can replace $\mu(X_{(k+1)})$ with $\mu(X_1)$ in the last equality:

$$\mathbf{E}\left[S_1\left(Y_1 - \mu(X_{(k+1)})\right) \mid Z_1\right] = \underbrace{\mathbf{E}\left[S_1\left(Y_1 - \mu(X_1)\right) \mid Z_1\right]}_{A} + \underbrace{\mathbf{E}\left[S_1\left(\mu(X_1) - \mu(X_{(k+1)})\right) \mid Z_1\right]}_{\rho}.$$

Observe that:

$$\mathbf{E}\left[\mathbf{E}\left[S_1\left(Y_1 - \mu(X_{(k+1)})\right) \mid Z_1\right]^2\right] = \mathbf{E}[A^2] + \mathbf{E}[\rho^2] + 2\mathbf{E}[A\rho].$$

Moreover, by Jensen's inequality, Lipschitzness of the first moments and kernel shrinkage:

$$\begin{aligned} \left|\mathbf{E}[\rho^2]\right| &= \mathbf{E}\left[\mathbf{E}\left[S_1\left(\mu(X_1) - \mu(X_{(k+1)})\right) \mid Z_1\right]^2\right] \leq \mathbf{E}\left[S_1\left(\mu(X_1) - \mu(X_{(k+1)})\right)^2\right] \\ &\leq 4L_m^2 \epsilon_{k+1}(s,\delta)^2 \mathbf{E}[\mathbf{E}[S_1|X_1]] + 4\delta\psi_{\max}^2 \leq 4L_m^2\epsilon_{k+1}(s,\delta)^2\frac{k}{s} + 4\delta\psi_{\max}^2. \end{aligned}$$

Hence, for $\delta = 1/s^2$, the latter is $o(1/s)$. Similarly:

$$\begin{aligned} |\mathbf{E}[A\rho]| \leq \mathbf{E}[|A|\,|\rho|] &\leq \psi_{\max}\mathbf{E}\left[\mathbf{E}\left[S_1\left|\mu(X_1) - \mu(X_{(k+1)})\right| \mid Z_1\right]\right] = \psi_{\max}\mathbf{E}\left[S_1\left|\mu(X_1) - \mu(X_{(k+1)})\right|\right] \\ &\leq \psi_{\max}\mathbf{E}[S_1]\epsilon_{k+1}(s,\delta) + 2\delta\psi_{\max} = \psi_{\max}\epsilon_{k+1}(s,\delta)\frac{k}{s} + 2\delta\psi_{\max}. \end{aligned}$$

which for $\delta = 1/s^2$ is also of order $o(1/s)$. Combining all the above we thus have:

$$\begin{aligned} k^2 \ \mathrm{Var}\left[\Phi_1(Z_1)\right] = & \ \mathbf{E}\left[\mathbf{E}\left[S_1\left(Y_1 - \mu(X_1)\right) \mid Z_1\right]^2\right] + o(1/s) \\ = & \ \mathbf{E}\left[\mathbf{E}\left[S_1 \mid X_1\right]^2 \left(Y_1 - \mu(X_1)\right)^2\right] + o(1/s)\,. \end{aligned}$$

We now work with the first term on the right hand side. By the tower law of expectations:

$$\begin{aligned} \mathbf{E}\left[\mathbf{E}[S_1 \mid X_1]^2 \left(Y_1 - \mu(X_1)\right)^2\right] = & \ \mathbf{E}\left[\mathbf{E}[S_1 \mid X_1]^2 \mathbf{E}\left[Y_1 - \mu(X_1)^2 \mid X_1\right]\right] = \mathbf{E}\left[\mathbf{E}[S_1 \mid X_1]^2 \sigma_j^2(X_1)\right] \\ = & \ \mathbf{E}\left[\mathbf{E}[S_1 \mid X_1]^2 \sigma_j^2(x)\right] + \mathbf{E}\left[\mathbf{E}[S_1 \mid X_1]^2 \left(\sigma_j^2(X_1) - \sigma_j^2(x)\right)\right]\ . \end{aligned}$$

By Lipschitzness of the second moments, we know that the second part is upper bounded as:

$$\begin{aligned} \left|\mathbf{E}\left[\mathbf{E}[S_1 \mid X_1]^2 \left(\sigma_j^2(X_1) - \sigma_j^2(x)\right)\right]\right| \leq & \ \left|\mathbf{E}\left[\mathbf{E}[S_1 \mid X_1] \left(\sigma_j^2(X_1) - \sigma_j^2(x)\right)\right]\right| \\ \leq & \ \left|\mathbf{E}\left[S_1 \left(\sigma_j^2(X_1) - \sigma_j^2(x)\right)\right]\right| \\ = & \ \left|\mathbf{E}\left[S_1 \left(\sigma_j^2(X_{(k)}) - \sigma_j^2(x)\right)\right]\right| \\ \leq & \ L_{mm} \mathbf{E}\left[S_1\right] \epsilon_k(s,\delta) + \delta \psi_{\max}^2 \\ = & \ \frac{L_{mm}\, \epsilon_k(s,\delta)\, k}{s} + \delta \psi_{\max}^2\,. \end{aligned}$$

For $\delta = 1/s^2$ it is of $o(1/s)$. Thus:

$$k^2 \ \mathrm{Var}\left[\Phi_1(Z_1)\right] = \mathbf{E}\left[\mathbf{E}[S_1 \mid X_1]^2\right] \sigma_j^2(x) + o(1/s)\,.$$

Note that Lemma 13 provides an expression for $\mathbf{E}\left[\mathbf{E}[S_1 \mid X_1]^2\right]$ which finishes the proof. ■

For finishing proof of Theorem 14 we need to prove that $\sum_{t=0}^{2k-2} \frac{a_t}{b_t}$ is equal to ζ_k plus lower order terms. This is proved in the following lemma.

Lemma 30 *Suppose that $s \to \infty$ and k is fixed. Then*

$$\sum_{t=0}^{2k-2} \frac{a_t}{b_t} = \zeta_k + O(1/s)\,.$$

Proof *Note that for any* $0 \leq t \leq k-1$ *we have* $a_t = b_t$ *according to Remark 28. For any* $k \leq t \leq 2k-2$ *we have*

$$\begin{aligned}
\frac{a_t}{b_t} &= \sum_{i=t-k+1}^{k-1} \frac{\binom{s-1}{i}\binom{s-1}{t-i}}{\binom{2s-2}{t}} = \sum_{i=t-k+1}^{k-1} \frac{\frac{(s-1)(s-2)\dots(s-i)}{i\,!}\,\frac{(s-1)(s-2)\dots(s-t+i)}{(t-i)\,!}}{\frac{(2s-2)(2s-3)\dots(2s-1-t)}{t\,!}} \\
&= \sum_{i=t-k+1}^{k-1} \binom{t}{i} \frac{(s-1)(s-2)\dots(s-i)\ (s-1)(s-2)\dots(s-t+i)}{(2s-2)(2s-3)\dots(2s-1-t)} \\
&= \sum_{i=t-k+1}^{k-1} \binom{t}{i} \frac{s-1}{2s-2}\frac{s-2}{2s-3}\cdots\frac{s-i}{2s-1-i}\ \frac{s-1}{2s-i}\frac{s-2}{2s-i-1}\cdots\frac{s-t+i}{2s-1-t} \\
&= \sum_{i=t-k+1}^{k-1} 2^{-t}\binom{t}{i}\left(1-\frac{1}{2s-3}\right)\cdots\left(1-\frac{i-1}{2s-1-i}\right)\left(1+\frac{i-2}{2s-i}\right)\cdots\left(1+\frac{i-(i-t+1)}{2s-1-t}\right) \\
&= 2^{-t}\sum_{i=t-k+1}^{k-1}\binom{t}{i}(1+O(1/s)) \\
&= 2^{-t}\sum_{i=t-k+1}^{k-1}\binom{t}{i} + O(1/s)\,,
\end{aligned}$$

where we used the fact that t *and* i *are both bounded above by* $2k-2$ *which is a constant. Hence,*

$$\sum_{t=0}^{2k-2}\frac{a_t}{b_t} = k + \sum_{t=k}^{2k-2} 2^{-t}\sum_{i=t-k+1}^{k-1}\binom{t}{i} + O(1/s) = \zeta_k + O(1/s)\,,$$

as desired. ■

G.5. Proof of Theorem 15

The goal is to apply Theorem 7. Note that k-NN kernel is both honest and symmetric. According to Lemma 10, we have that $\epsilon_k(s,\delta) = O\left((\log(1/\delta)/s)^{1/d}\right)$ for $\exp(-Cs) \leq \delta \leq D$, where C and D are constants. Corollary 11 also implies that $\epsilon_k(s) = O((1/s)^{1/d})$. Furthermore, according to Lemma 13, the incrementality $\eta_k(s)$ is $\Theta(1/s)$. Therefore, as s goes to ∞ we have $\epsilon_k(s,\eta_k(s)) = O\left((\log(s)/s)^{1/d}\right) \to 0$. Moreover, as $\eta_k(s) = \Theta(1/s)$, we also get that $n\eta_k(s) = O(n/s) \to \infty$. We only need to ensure that Equation (10) is satisfied. Note that $\sigma_{n,j}(x) = \Theta(\sqrt{s/n})$. Therefore, by dividing terms in Equation (10) it suffices that

$$\max\left(s^{-1/d}\left(\frac{n}{s}\right)^{1/2}, s^{-1/4d}\left(\log\log(n/s)\right)^{1/2}, \left(\frac{n}{s}\right)^{-1/8}\left(\log\log(n/s)\right)^{5/8}\right) = o(1)\,.$$

Note that due to our Assumption $n/s \to \infty$, the last term obviously goes to zero. Also, because of the assumption made in the statement of theorem, the first term also goes to zero. We claim that if the first term goes to zero, the same also holds for the second term. Note that we can write

$$s^{-1/4d}\left(\log\log(n/s)\right)^{1/2} = \left(s^{-1/d}\left(\frac{n}{s}\right)^{1/2}\right)^{1/4}\cdot\left[\left(\frac{n}{s}\right)^{-1/8}\left(\log\log(n/s)\right)^{1/2}\right],$$

and since $n/s \to \infty$, our claim follows. Therefore, all the conditions of Theorem 7 are satisfied and the result follows.

The second part of result is implied by the first part since if $s = n^{\beta}$ and $\beta \in (d/(d+2), 1)$ then $s^{-1/d}\sqrt{\frac{n}{s}} \to 0$.

G.6. Proof of Proposition 16

For proving this lemma, we need two following auxiliary results. Before that we state the Hoeffding's inequality for U-statistics Hoeffding (1994).

Proposition 31 *Suppose that $\mathbf{X} = (X_1, X_2, \dots, X_n)$ are i.i.d. and q is a function that has range $[0,1]$. Define $U_s = \binom{n}{s}^{-1}\sum_{i_1<i_2<\ldots<i_s} q(X_{i_1}, X_{i_2}, \dots, X_{i_s})$. Then, for any $\epsilon > 0$*

$$\Pr\left[|U_s - \mathbf{E}[U_s]| \geq \epsilon\right] \leq 2\exp\left(-\lfloor n/s \rfloor \epsilon^2\right) .$$

Furthermore, for any $\delta > 0$, w.p. $1-\delta$ we have

$$|U_s - \mathbf{E}[U_s]| \leq \sqrt{\frac{1}{\lfloor n/s \rfloor}\log(2/\delta)} .$$

Lemma 32 *Consider the function $H(s)$ defined in Section 5.3 and $G_\delta(s) = \Delta\sqrt{2ps/n\log(2np/\delta)}$. Then, w.p. $1-\delta$, for all values of $k \leq s \leq n$ we have*

$$|H(s) - \epsilon_k(s)| \leq G_\delta(s) .$$

Proof *Note that $H(s)$ is the complete U-statistic estimator for $\epsilon_k(s)$. For each subset S of size s from $[n]$ we have*

$$\mathbf{E}[\max_{X_i \in H_k(x,S)} \|x - X_i\|_2] = \epsilon_k(s) .$$

Further, $\|x - x'\|_2 \leq \Delta_{\mathcal{X}} \leq \Delta$ holds for any $x' \in \mathcal{X}$. Therefore, using Hoeffding's inequality for U-statistics stated in Proposition 31, for any fixed s, w.p. $1-\delta$ we have

$$|H(s) - \epsilon_k(s)| \leq \Delta\sqrt{\frac{1}{\lfloor n/s \rfloor}\log(2/\delta)} .$$

Note that $\lfloor z \rfloor \geq z/2$ for $z \geq 1$ and therefore the above translates to

$$|H(s) - \epsilon_k(s)| \leq \Delta\sqrt{\frac{2s}{n}\log(2/\delta)} .$$

Taking a union bound over $s = k, k+1, \dots, n$, replacing $\delta = \delta/n$, and using $p \geq 1$, implies the result. ∎

Lemma 33 *Consider the selection process mentioned in Section 5.3 and let s_1 be the output of this process. Then, w.p. $1-\delta$ we have*

$$\frac{s^*-1}{9} \leq s_1 \leq s^* .$$

Proof Note that using Lemma 32, w.p. $1-\delta$, for all values of s we have $|H(s)-\epsilon_k(s)| \leq G_\delta(s)$. Now consider three different cases:

- $s_1 \geq s_2 \geq s^*$: Note that based on the choice of s_1, s_2, we have $H(s_2) > 2G_\delta(s_2)$. However, $H(s_2) \leq \epsilon_k(s_2) + G_\delta(s_2)$. Hence, $\epsilon_k(s_2) > G_\delta(s_2)$ which contradicts with the assumption that $s_2 \geq s^*$. Note that this is true since $\epsilon_k(s) - G_\delta(s)$ is non-positive for $s \geq s^*$.
- $s_1 = s^*, s_2 = s^* - 1$: Obviously $s_1 \leq s^*$.
- $s_2 \leq s_1 \leq s^* - 1$: Note that we have

$$\epsilon_k(s_1) - G_\delta(s_1) \leq H(s_1) \leq 2G_\delta(s_1)\,.$$

Hence, $G_\delta(s^*-1) < \epsilon_k(s^*-1) \leq \epsilon_k(s_1) \leq 3G_\delta(s_1)$. This means that $G_\delta(s^*-1)/G_\delta(s_1) \leq 3$ which implies $\sqrt{(s^*-1)/s_1} \leq 3$. Therefore, $s_1 \geq (s^*-1)/9$.

This completes the proof. ■

Now we are ready to finalize the proof of Theorem 16. Note that using the result of Lemma 33, w.p. $1-\delta$, we have

$$\frac{s^*-1}{9} \leq s_1 \leq s^*\,.$$

This basically means that if $s_* = 9s_1 + 1$, then s_* belongs to $[s^*, 10s^*]$. Hence, we have $\epsilon_k(s_*) \leq \epsilon_k(s^*) \leq G_\delta(s^*)$ and $G_\delta(s_*) \leq G_\delta(10s^*) = \sqrt{10}G_\delta(s^*)$. Now using Theorem 6, for $B \geq n/s_*$ w.p. $1-\delta$ we have

$$\|\hat{\theta} - \theta(x)\|_2 \leq \frac{2}{\lambda}\left(L_m\epsilon(s_*) + O\left(\psi_{\max}\sqrt{\frac{p\,s_*}{n}\left(\log\log(n/s_*) + \log(p/\delta)\right)}\right)\right)\,.$$

Note that $G_\delta(s_*) = \Delta\sqrt{\frac{2ps_*}{n}\log(2pn/\delta)}$. Therefore,

$$\sqrt{\frac{p\,s_*}{n}\left(\log\log(n/s_*) + \log(p/\delta)\right)} \leq G_\delta(s_*) \leq \sqrt{10}G_\delta(s^*)\,.$$

Replacing this in above equation together with a union bound implies that w.p. at least $1-2\delta$ we have

$$\|\hat{\theta} - \theta(x)\|_2 = O(G_\delta(s^*))\,,$$

which finishes the first part of the proof. For the second part, note that according to Corollary 11, for the k-NN kernel $\epsilon(s) \leq Cs^{-1/d}$, for a constant C. Note that at $s = s^* - 1$ we have

$$\Delta\sqrt{\frac{2ps}{n}\log(2np/\delta)} = \epsilon_k(s) \leq Cs^{-1/d}\,,$$

for a constant C. The above implies that

$$s^* \leq 1 + \left(\frac{C}{\Delta}\right)^{2d/(d+2)}\left(\frac{n}{2p\log(2np/\delta)}\right)^{d/(d+2)} \leq 2\left(\frac{C}{\Delta}\right)^{2d/(d+2)}\left(\frac{n}{2p\log(2np/\delta)}\right)^{d/(d+2)}\,.$$

Hence,

$$G_\delta(s^*) \leq \sqrt{2}\Delta^{2/(d+2)}C^{d/(d+2)}\left(\frac{n}{2p\log(2np/\delta)}\right)^{-1/(d+2)}\,.$$

Remark 34 *Note that although computation of $H(s)$ may look complex as it involves calculation of distance to k-nearest neighbor of x on all $\binom{n}{s}$ subsets, there is a closed form representation for $H(s)$ according to its representation based on L-statistic. In fact, by sorting samples $(X_1, X_2, \ldots, X_n)$ based on their distance to x, i.e, $\|x - X_{(1)}\|_2 \le \|x - X_{(2)}\|_2 \le \ldots \le \|x - X_{(n)}\|_2$, we have*

$$H(s) = \binom{n}{s}^{-1} \sum_{i=k}^{n-s+k} \binom{i-1}{k-1}\binom{n-i}{s-k} \|x - X_{(i)}\|_2 \,.$$

Therefore, after sorting training samples, we can compute values of $H(s)$ very efficient and fast.

G.7. Proof of Proposition 17

Note that according to Lemma 33, w.p. $1 - 1/n$, the output of process, s_1 satisfies

$$\frac{s^* - 1}{9} \le s_1 \le s^* \,,$$

where s^* is the point for which we have $\epsilon_k(s^*) = G_{1/n}(s^*)$. This basically means that $s_* = 9s_1 + 1 \ge s^*$. Note that for the k-NN kernel we have $\eta_k(s) = \Theta(1/s)$. As $s_\zeta \ge n^\zeta$, this also implies that $\epsilon_k(s_\zeta, \eta_k(s_\zeta)) = O((\log(s_\zeta)/s_\zeta)^{1/d}) \to 0$. Also, according to the inequality $\zeta < \frac{\log(n) - \log(s_*) - \log\log^2(n)}{\log(n)}$ we have $1 - \zeta > (\log(s_*) + \log\log^2(n))/\log(n)$ and therefore

$$n^{1-\zeta} \ge s_\zeta \log^2(n) \to \frac{s_\zeta}{n} \le \frac{1}{\log^2(n)} \,,$$

and hence $n\eta_k(s_\zeta) \to 0$. Finally, note that $\sigma_{n,j}(x) = \Theta(\sqrt{s/n})$ and according to Theorem 7 it suffices that

$$\max\left(\epsilon_k(s_\zeta)\left(\frac{s_\zeta}{n}\right)^{-1/2}, \epsilon_k(s_\zeta)^{1/4}\left(\log\log(n/s_\zeta)\right)^{1/2}, \left(\frac{s_\zeta}{n}\right)^{1/8}\left(\log\log(n/s_\zeta)\right)^{5/8}\right) = o(1) \,.$$

Note that for any $\zeta > 0$, $s_\zeta \ge s^*$ and therefore $\epsilon_k(s_\zeta) \le \epsilon_k(s^*) = G_{1/n}(s^*)$. For the first term,

$$\begin{aligned}
\epsilon_k(s_\zeta)\left(\frac{s_\zeta}{n}\right)^{-1/2} &\le G_{1/n}(s^*)\left(\frac{s_\zeta}{n}\right)^{-1/2} \\
&= \Delta\sqrt{\frac{2p\, s^*}{n}\log(2n^2/p)}\;\left(\frac{s_\zeta}{n}\right)^{-1/2} \\
&= O\left(\sqrt{\frac{s^*}{s_\zeta}}\log(n)\right).
\end{aligned}$$

Now note that $s_\zeta = s_* n^\zeta \ge s^* n^\zeta$ and hence $\sqrt{s^*/s_\zeta}\log(n) = O(n^{-\zeta/2}\log(n)) \to 0$. For the second term, note that again $s_\zeta \ge s^*$ and therefore $\epsilon_k(s_\zeta) \le \epsilon_k(s^*) = G_{1/n}(s^*) \le G_{1/n}(s_\zeta)$. Now note that since $s_\zeta/n \le 1/\log^2(n)$ hence

$$\epsilon_k(s_\zeta)^{1/4}\log\log(n/s_\zeta)^{1/2} \le G_{1/n}(s_\zeta)\log\log(n) = O\left(\left(\frac{\log(n)}{\log^2(n)}\right)^{1/8}\log\log(n)\right) \to 0 \,.$$

Finally, for the last term we have $s_\zeta/n \leq 1/\log^2(n)$ and hence

$$\left(\frac{s_\zeta}{n}\right)^{1/8} (\log\log(n/s_\zeta))^{5/8} \leq \left(\frac{1}{\log(n)}\right)^{1/4} \log\log(n) \to 0.$$

This basically means w.p. $1 - 1/n$, s_ζ belongs to the interval for which the asymptotic normality result in Theorem 7 holds. As $n \to \infty$, the conclusion follows.

Appendix H. Stochastic Equicontinuity of U-statistics via Bracketing

We define here some standard terminology on bracketing numbers in empirical process theory. Consider an arbitrary function space $\mathcal{F}$ of functions from a data space $\mathcal{Z}$ to $\mathbb{R}$, equipped with some norm $\|\cdot\|$. A *bracket* $[a, b] \subseteq \mathcal{F}$, where $a, b : \mathcal{Z} \to \mathbb{R}$ consists of all functions $f \in \mathcal{F}$, such that $a \leq f \leq b$. An *ϵ-bracket* is a bracket $[a, b]$ such that $\|a - b\| \leq \epsilon$. The *bracketing number* $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$. The functions $[a, b]$ used in the definition of the brackets need not belong to $\mathcal{F}$ but satisfy the same norm constraints as functions in $\mathcal{F}$. Finally, for an arbitrary measure P on $\mathcal{Z}$, let

$$\|f\|_{P,2} = \sqrt{\mathbf{E}_{Z\sim P}[f(Z)^2]} \qquad\qquad \|f\|_{P,\infty} = \sup_{z\,\in\,\text{support}(P)} |f(z)| \tag{39}$$

Lemma 35 (Stochastic Equicontinuity for U-statistics via Bracketing) *Consider a function space $\mathcal{F}$ of symmetric functions from some data space $\mathcal{Z}^s$ to $\mathbb{R}$ and consider the U-statistic of order s, with kernel f over n samples:*

$$\Psi_s(f, z_{1:n}) = \binom{n}{s}^{-1} \sum_{1\leq i_1\leq\ldots\leq i_s\leq n} f(z_{i_1}, \ldots, z_{i_s}) \tag{40}$$

Suppose $\sup_{f\in\mathcal{F}} \|f\|_{P,2} \leq \eta$, $\sup_{f\in\mathcal{F}} \|f\|_{P,\infty} \leq G$ *and let* $\kappa = n/s$. *Then for* $\kappa \geq \frac{G^2}{\log N_{[]}(1/2, \mathcal{F}, \|\cdot\|_{P,2})}$, *w.p.* $1 - \delta$*:*

$$\sup_{f\in\mathcal{F}} |\Psi_s(f, Z_{1:n}) - \mathbf{E}[f(Z_{1:s})]| \\ = O\left(\inf_{\rho>0} \frac{1}{\sqrt{\kappa}} \int_\rho^{2\eta} \sqrt{\log(N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_{P,2})} + \eta\sqrt{\frac{\log(1/\delta) + \log\log(\eta/\rho)}{\kappa}} + \rho\right)$$

Proof Let $\kappa = n/s$. Moreover, wlog we will assume that $\mathcal{F}$ contains the zero function, as we can always augment $\mathcal{F}$ with the zero function without changing the order of its bracketing number. For $q = 1, \ldots, M$, let $\mathcal{F}_q = \cup_{i=1}^{N_q} \mathcal{F}_{qi}$ be a partition of $\mathcal{F}$ into brackets of diameter at most $\epsilon_q = 2\eta/2^q$, with $\mathcal{F}_0$ containing a single partition of all the functions. Moreover, we assume that $\mathcal{F}_q$ are nested partitions. We can achieve the latter as follows: i) consider a minimal bracketing cover of $\mathcal{F}$ of diameter ϵ_q, ii) assign each $f \in \mathcal{F}$ to one of the brackets that it is contained arbitrarily and define the partition $\bar{\mathcal{F}}_q$ of size $\bar{N}_q = N_{[]}(\epsilon_q, \mathcal{F}, \|\cdot\|_{P,2})$, by taking $\mathcal{F}_{qi}$ to be the functions assigned to bracket i, iii) let $\mathcal{F}_q$ be the common refinement of all partitions $\bar{\mathcal{F}}_0, \ldots, \bar{\mathcal{F}}_q$. The latter will have size at most $N_q \leq \prod_{q=0}^{M} \bar{N}_q$. Moreover, assign a representative function f_{qi} to each partition $\mathcal{F}_{qi}$, with the representative for the single partition at level $q = 0$ is the zero function.

Chaining definitions. Consider the following random variables, where the dependence on the random input Z is hidden:

$$\begin{aligned} \pi_q f =&\ f_{qi}, \quad \text{if } f \in \mathcal{F}_{qi} \\ \Delta_q f =&\ \sup_{g,h \in \mathcal{F}_{qi}} |g-h|, \quad \text{if } f \in \mathcal{F}_{qi} \\ B_q f =&\ \{\Delta_0 f \leq \alpha_0, \ldots, \Delta_{q-1} f \leq \alpha_{q-1}, \Delta_q f > \alpha_q\} \\ A_q f =&\ \{\Delta_0 f \leq \alpha_0, \ldots, \Delta_q f \leq \alpha_q\}, \end{aligned}$$

for some sequence of numbers $\alpha_0, \ldots, \alpha_M$, to be chosen later. By noting that $A_{q-1}f = A_q f + B_q f$ and continuously expanding terms by adding and subtracting finer approximations to f, we can write the telescoping sum:

$$\begin{aligned} f - \pi_0 f =&\ (f - \pi_0 f)\, B_0 f + (f - \pi_0 f) A_0 f \\ =&\ (f - \pi_0 f)\, B_0 f + (f - \pi_1 f) A_0 f + (\pi_1 f - \pi_0 f) A_0 f \\ =&\ (f - \pi_0 f)\, B_0 f + (f - \pi_1 f) B_1 f + (f - \pi_1 f) A_1 f + (\pi_1 f - \pi_0 f) A_0 f \\ \ldots& \\ =&\ \sum_{q=0}^{M} (f - \pi_q f) B_q f + \sum_{q=1}^{M} (\pi_q f - \pi_{q-1} f) A_{q-1} f + (f - \pi_M f) A_M f\,. \end{aligned}$$

For simplicity let $\mathbb{P}_{s,n} f = \Psi(f, Z_{1:n})$, $\mathbb{P} f = E[f(Z_{1:s})]$ and $\mathbb{G}_{s,n}$ denote the U-process:

$$\mathbb{G}_{s,n} f = \mathbb{P}_{s,n} f - \mathbb{P} f\,. \tag{41}$$

Our goal is to bound $\|\mathbb{P}_{s,n} f\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{P}_{s,n} f|$, with high probability. Observe that since $\mathcal{F}_0$ contains only the zero function, then $\mathbb{G}_{s,n} f_0 = 0$. Moreover, the operator $\mathbb{G}_{s,n}$ is linear. Thus:

$$\mathbb{G}_{s,n} f = \mathbb{G}_{s,n}(f - \pi_0 f) = \sum_{q=0}^{M} \mathbb{G}_{s,n}(f - \pi_q f) B_q f + \sum_{q=1}^{M} \mathbb{G}_{s,n}(\pi_q f - \pi_{q-1} f) A_{q-1} f + \mathbb{G}_{s,n}(f - \pi_M f) A_M f\,.$$

Moreover, by triangle inequality:

$$\|\mathbb{G}_{s,n} f\|_{\mathcal{F}} \leq \sum_{q=0}^{M} \|\mathbb{G}_{s,n}(f - \pi_q f) B_q f\|_{\mathcal{F}} + \sum_{q=1}^{M} \|\mathbb{G}_{s,n}(\pi_q f - \pi_{q-1} f) A_{q-1} f\|_{\mathcal{F}} + \|\mathbb{G}_{s,n}(f - \pi_M f) A_M f\|_{\mathcal{F}}\,.$$

We will bound each term in each summand separately.

Edge cases. The final term we will simply bound it by $2\alpha_M$, since $|(f - \pi_M f) A_M f| \leq \alpha_M$, almost surely. Moreover, the summand in the first term for $q = 0$, we bound as follows. Observe that $B_0 f = 1\{\sup_f |f| > \alpha_0\}$. But we know that $\sup_f |f| \leq G$, hence: $B_0 f \leq 1\{G > \alpha_0\}$.

$$\mathbb{G}_{s,n}(f - \pi_0 f) B_0 f = \mathbb{G}_{s,n} f B_0 f \leq |\mathbb{P}_{s,n} f B_0 f| + |\mathbb{P} f B_0 f| \leq 2G\, 1\{G > \alpha_0\}\,.$$

Hence, if we assume that α_0 is large enough such that $\alpha_0 > G$, then the latter term is zero. By the setting of α_0 that we will describe at the end, the latter would be satisfied if $\kappa \geq \frac{G^2}{\log N_{[]}(1/2, \mathcal{F}, \|\cdot\|_{P,2})}$.

B_q **terms.** For the terms in the first summand we have by triangle inequality:

$$
\begin{aligned}
|\mathbb{G}_{s,n}(f-\pi_q f)B_q f| \leq\ & \mathbb{P}_{s,n}|f-\pi_q f|B_q f + \mathbb{P}|f-\pi_q f|B_q f\\
\leq\ & \mathbb{P}_{s,n}\Delta_q f B_q f + \mathbb{P}\Delta_q f B_q f\\
\leq\ & \mathbb{G}_{s,n}\Delta_q f B_q f + 2\mathbb{P}\Delta_q f B_q f\,.
\end{aligned}
$$

Moreover, observe that:

$$
\mathbb{P}\Delta_q f B_q f \leq \mathbb{P}\Delta_q f 1\{\Delta_q f > \alpha_q\} \leq \frac{1}{\alpha_q}\mathbb{P}(\Delta_q f)^2 1\{\Delta_q f > \alpha_q\} \leq \frac{1}{\alpha_q}\mathbb{P}(\Delta_q f)^2 = \frac{1}{\alpha_q}\|\Delta_q f\|_{P,2}^2 \leq \frac{\epsilon_q^2}{\alpha_q}\,,
$$

where we used the fact that the partitions in $\mathcal{F}_q$, have diameter at most ϵ_q, with respect to the $\|\cdot\|_{P,2}$ norm. Now observe that because the partitions $\mathcal{F}_q$ are nested, $\Delta_q f \leq \Delta_{q-1} f$. Therefore, $\Delta_q f B_q f \leq \Delta_{q-1} f B_q f \leq \alpha_{q-1}$, almost surely. Moreover, $\|\Delta_q f B_q f\|_{P,2} \leq \|\Delta_q f\|_{P,2} \leq \epsilon_q$. By Bernstein's inequality for U statistics (see e.g. Peel et al. (2010)) for any fixed f, w.p. $1-\delta$:

$$
|\mathbb{G}_{s,n}\Delta_q f B_q f| \leq\ \epsilon_q\sqrt{\frac{2\log(2/\delta)}{\kappa}} + \alpha_{q-1}\frac{2\log(2/\delta)}{3\kappa}\,.
$$

Taking a union bound over the N_q members of the partition, and combining with the bound on $\mathbb{P}\Delta_q f B_q f$, we have w.p. $1-\delta$:

$$
\|\mathbb{G}_{s,n}(f-\pi_q f)B_q f\|_{\mathcal{F}} \leq \epsilon_q\sqrt{\frac{2\log(2N_q/\delta)}{\kappa}} + \alpha_{q-1}\frac{2\log(2N_q/\delta)}{3\kappa} + \frac{2\epsilon_q^2}{\alpha_q}\,. \tag{42}
$$

A_q **terms.** For the terms in the second summand, we have that since the partitions are nested, $|(\pi_q f - \pi_{q-1} f)A_{q-1} f| \leq \Delta_{q-1} f A_{q-1} f \leq \alpha_{q-1}$. Moreover, $\|(\pi_q f - \pi_{q-1} f)A_{q-1} f\|_{P,2} \leq \|\Delta_{q-1} f\|_{P,2} \leq \epsilon_{q-1} \leq 2\epsilon_q$. Thus, by similar application of Bernstein's inequality for U-statistics, we have for a fixed f, w.p. $1-\delta$:

$$
|\mathbb{G}_{s,n}(\pi_q f - \pi_{q-1} f)A_{q-1} f| \leq \epsilon_q\sqrt{\frac{8\log(2/\delta)}{\kappa}} + \alpha_{q-1}\frac{2\log(2/\delta)}{3\kappa}\,.
$$

As f ranges there are at most $N_{q-1}\, N_q \leq N_q^2$ different functions $(\pi_q f - \pi_{q-1} f)A_{q-1} f$. Thus taking a union bound, we have that w.p. $1-\delta$:

$$
\|\mathbb{G}_{s,n}(\pi_q f - \pi_{q-1} f)A_{q-1} f\|_{\mathcal{F}} \leq \epsilon_q\sqrt{\frac{16\log(2N_q/\delta)}{\kappa}} + \alpha_{q-1}\frac{4\log(2N_q/\delta)}{3\kappa}\,.
$$

Taking also a union bound over the $2M$ summands and combining all the above inequalities, we have that w.p. $1-\delta$:

$$
\|\mathbb{G}_{s,n} f\|_{\mathcal{F}} \leq \sum_{q=1}^{M} \epsilon_q\sqrt{\frac{32\log(2N_q M/\delta)}{\kappa}} + \alpha_{q-1}\frac{6\log(2N_q M/\delta)}{3\kappa} + \frac{2\epsilon_q^2}{\alpha_q}\,.
$$

Choosing $\alpha_q = \epsilon_q\sqrt{\kappa}/\sqrt{\log(2N_{q+1}M/\delta)}$ for $q < M$ and $\alpha_M = \epsilon_M$, we have for some constant C:

$$\begin{aligned}
\|\mathbb{G}_{s,n} f\|_{\mathcal{F}} \leq&~ C\sum_{q=1}^{M} \epsilon_q \sqrt{\frac{\log(2N_q M/\delta)}{\kappa}} + 3\epsilon_M\\
\leq&~ C\sum_{q=1}^{M} \epsilon_q \sqrt{\frac{\log(N_q)}{\kappa}} + C\sum_{q=1}^{M} \epsilon_q \sqrt{\frac{\log(2M/\delta)}{\kappa}} + 3\epsilon_M\\
\leq&~ C\sum_{q=1}^{M} \epsilon_q \sqrt{\frac{\log(N_q)}{\kappa}} + 2C\eta\sqrt{\frac{\log(2M/\delta)}{\kappa}} + 3\epsilon_M\,.
\end{aligned}$$

Moreover, since $\log(N_q) \leq \sum_{t=0}^{q} \log(N_{[]}(\epsilon_q, \mathcal{F}, \|\cdot\|_{P,2})$, we have:

$$\begin{aligned}
\sum_{q=1}^{M} \epsilon_q \sqrt{\log(N_q)} \leq&~ \sum_{q=1}^{M} \epsilon_q \sum_{t=0}^{q} \sqrt{\log(N_{[]}(\epsilon_t, \mathcal{F}, \|\cdot\|_{P,2})} = \sum_{t=0}^{M} \sqrt{\log(N_{[]}(\epsilon_t, \mathcal{F}, \|\cdot\|_{P,2})} \sum_{q=t}^{M} \epsilon_q\\
\leq&~ 2\sum_{t=0}^{M} \epsilon_t \sqrt{\log(N_{[]}(\epsilon_t, \mathcal{F}, \|\cdot\|_{P,2})}\\
\leq&~ 4\sum_{t=0}^{M} (\epsilon_t - \epsilon_{t+1}) \sqrt{\log(N_{[]}(\epsilon_t, \mathcal{F}, \|\cdot\|_{P,2})}\\
\leq&~ 4\int_{\epsilon_M}^{\epsilon_0} \sqrt{\log(N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_{P,2})}\,.
\end{aligned}$$

Combining all the above yields the result. ■

Corollary 36 *Consider a function space $\mathcal{F}$ of symmetric functions. Suppose that $\sup_{f\in\mathcal{F}} \|f\|_{P,2} \leq \eta$ and $\log(N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_{P,2}) = O(1/\epsilon)$. Then for $\kappa \geq O(G^2)$, w.p. $1-\delta$:*

$$\sup_{f\in\mathcal{F}} |\Psi_s(f, Z_{1:n}) - \mathbf{E}[f(Z)]| = O\left(\sqrt{\frac{\eta}{\kappa}} + \eta\sqrt{\frac{\log(1/\delta) + \log\log(\kappa/\eta)}{\kappa}}\right)\,. \tag{43}$$

Proof Applying Lemma 35, we get for every $\rho > 0$, the desired quantity is upper bounded by:

$$\begin{aligned}
&O\left(\frac{1}{\sqrt{\kappa}}\int_{\rho}^{\eta} \frac{1}{\sqrt{\epsilon}} + \eta\sqrt{\frac{\log(1/\delta) + \log\log(\eta/\rho)}{\kappa}} + \rho\right)\\
&= O\left(\frac{\sqrt{\eta} - \sqrt{\rho}}{\sqrt{\kappa}} + \eta\sqrt{\frac{\log(1/\delta) + \log\log(\eta/\rho)}{\kappa}} + \rho\right)\,.
\end{aligned}$$

Choosing $\rho = \sqrt{\eta}/\sqrt{\kappa}$, yields the desired bound. ■

Appendix I. Proof of Lemma 22

We will argue asymptotic normality of the U-statistic defined as:

$$\Psi_{0,\beta}(x;\theta(x)) = \binom{n}{s}^{-1} \sum_{b\subset[n]:|b|=s} \mathbf{E}_{\omega_b}\left[\sum_{i\in S_b} \alpha_{S_b,\omega_b}(X_i)\psi_\beta(Z_i;\theta(x))\right]$$

under the assumption that for any subset of indices S_b of size s: $\mathbf{E}\left[\mathbf{E}[\alpha_{S_b,\omega_b}(X_1)|X_1]^2\right] = \eta(s)$ and that the kernel satisfies shrinkage in probability with rate $\epsilon(s,\delta)$ such that $\epsilon(s,\eta(s)^2)\to 0$ and $n\eta(s)\to\infty$. For simplicity of notation we let:

$$Y_i = \psi_\beta(Z_i;\theta(x)) \tag{44}$$

and we then denote:

$$\Phi(Z_1,\ldots,Z_s) = \mathbf{E}_\omega\left[\sum_{i=1}^{s} K_\omega(x,X_i,\{Z_j\}_{j=1}^s)\, Y_i\right]. \tag{45}$$

Observe that we can then re-write our U-statistic as:

$$\Psi_{0,\beta}(x;\theta(x)) = \binom{n}{s}^{-1} \sum_{1\le i_1\le\ldots\le i_s\le n} \Phi(Z_{i_1},\ldots,Z_{i_s}).$$

Moreover, observe that by the definition of Y_i, $\mathbf{E}[Y_i \mid X_i] = 0$ and also

$$|Y_i| \le \|\beta\|_2 \|M_0^{-1}(\psi(Z_i;\theta(x)) - m(X_i;\theta(x))\|_2 2 \le \frac{R}{\lambda}\|\psi(Z_i;\theta(x))\|_2 \le 2\frac{R\sqrt{p}}{\lambda}\psi_{\max} \triangleq y_{\max}.$$

Invoking Lemma 38, it suffices to show that: $\mathrm{Var}\left[\Phi_1(Z_1)\right] = \Omega(\eta(s))$, where $\Phi_1(z_1) = \mathbf{E}[\Phi(z_1,Z_2,\ldots,Z_s)]$. The following lemma shows that under our conditions on the kernel, the latter property holds.

Lemma 37 *Suppose that the kernel K is symmetric (Assumption 4), has been built in an honest manner (Assumption 3) and satisfies:*

$$\mathbf{E}\left[\mathbf{E}\left[K(x,X_1,\{Z_j\}_{j=1}^s) \mid X_1\right]^2\right] = \eta(s) \le 1 \qquad \textit{and} \qquad \epsilon(s,\eta(s)^2)\to 0.$$

Then, the following holds

$$\mathrm{Var}\left[\Phi_1(Z_1)\right] \ge \mathrm{Var}(Y \mid X = x)\,\eta(s) + o(\eta(s)) = \Omega\left(\eta(s)\right).$$

Proof Note we can write

$$\Phi_1(Z_1) = \underbrace{\mathbf{E}\left[\Phi(Z_1,\ldots,Z_s) \mid X_1\right]}_{A} + \underbrace{\mathbf{E}\left[\Phi(Z_1,\ldots,Z_s) \mid X_1, Y_1\right] - \mathbf{E}\left[\Phi(Z_1,\ldots,Z_s) \mid X_1\right]}_{B}.$$

Here, B is zero mean conditional on X_1 and also A and B are uncorrelated, i.e., $\mathbf{E}[AB] = \mathbf{E}[A]\mathbf{E}[B] = 0$. Therefore:

$$\mathrm{Var}\left[\Phi_1(Z_1)\right] \ge \mathrm{Var}\left[B\right] = \mathrm{Var}\left[\sum_{i=1}^{s}\left(\mathbf{E}\left[K(x,X_i,\{Z_j\}_{j=1}^s)Y_i \mid X_1, Y_1\right] - \mathbf{E}[K(x,X_i,\{Z_j\}_{j=1}^s)Y_i \mid X_1]\right)\right].$$

For simplicity of notation let $W_i = K(x, X_1, \{Z_j\}_{j=1}^s)$ denote the random variable which corresponds to the weight of sample i. Note that thanks to the honesty of kernel defined in Assumption 3, W_i is independent of Y_1 conditional on X_1, for $i \geq 2$. Hence all the corresponding terms in the summation are zero. Therefore, the expression inside the variance above simplifies to

$$\mathbf{E}[W_1 Y_1 \mid X_1, Y_1] - \mathbf{E}[W_1 Y_1 \mid X_1] .$$

Moreover, by honesty W_1 is independent of Y_1 conditional on X_1. Thus, the above further simplifies to:

$$\mathbf{E}[W_1 \mid X_1] \left(Y_1 - \mathbf{E}[Y_1 \mid X_1]\right) .$$

Using $\mathrm{Var}(G) = \mathbf{E}[G^2] - \mathbf{E}[G]^2$, this can be further rewritten as

$$\mathrm{Var}\left[\Phi_1(Z_1)\right] \geq \mathbf{E}\left[\mathbf{E}[W_1 \mid X_1]^2 \left(Y_1 - \mathbf{E}[Y_1 \mid X_1]\right)^2\right] - \mathbf{E}\left[\mathbf{E}[W_1 \mid X_1](Y_1 - \mathbf{E}[Y_1 \mid X_1])\right]^2 .$$

Note that $Y_1 - \mathbf{E}[Y_1 \mid X_1]$ is uniformly upper bounded by some $\psi_{\max}$. Furthermore, by the symmetry of the kernel we have $\mathbf{E}\left[\mathbf{E}[W_1 \mid X_1]\right] = \mathbf{E}[W_1] = 1/s$.[11] Thus the second term in the latter is of order $1/s^2$. Hence:

$$\mathrm{Var}\left[\Phi(Z_1)\right] \geq \mathbf{E}\left[\mathbf{E}[\alpha_b(X_1) \mid X_1]^2 \left(Y_1 - \mathbf{E}[Y_1 \mid X_1]\right)^2\right] + o(1/s) .$$

Focusing at the first term and letting $\sigma^2(x) = \mathrm{Var}(Y|X = x)$, we have:

$$\begin{aligned}
\mathbf{E}\left[\mathbf{E}[W_1 \mid X_1]^2 \left(Y_1 - \mathbf{E}[Y_1 \mid X_1]\right)^2\right] &= \mathbf{E}\left[\mathbf{E}[W_1 \mid X_1]^2 \sigma^2(X_1)\right] \\
&= \mathbf{E}\left[\mathbf{E}[W_1 \mid X_1]^2\right] \sigma^2(x) + \mathbf{E}\left[\mathbf{E}[W_1 \mid X_1]^2 \left(\sigma^2(X_1) - \sigma^2(x)\right)\right] .
\end{aligned}$$

The goal is to prove that the second term is $o(1/s)$. For ease of notation let $V_1 = \mathbf{E}\left[W_1 \mid X_1\right]$. Then we can bound the second term as:

$$\begin{aligned}
\left|\mathbf{E}\left[V_1^2 \left(\sigma^2(X_1) - \sigma^2(x)\right)\right]\right| &\leq L_{mm}\epsilon(s,\delta)\, \mathbf{E}\left[V_1^2\, 1\left\{\|x - X_1\|_2 \leq \epsilon(s,\delta)\right\}\right] \\
&\qquad + 2y_{\max}^2 \mathbf{E}\left[V_1^2\, 1\left\{\|x - X_1\|_2 > \epsilon(s,\delta)\right\}\right] \\
&\leq L_{mm}\epsilon(s,\delta)\, \mathbf{E}\left[V_1^2\right] + 2y_{\max}^2 \mathbf{E}\left[V_1^2\, 1\left\{\|x - X_1\|_2 > \epsilon(s,\delta)\right\}\right] \\
&\leq L_{mm}\epsilon(s,\delta)\eta(s) + 2y_{\max}^2 \mathbf{E}\left[V_1\, 1\left\{\|x - X_1\|_2 > \epsilon(s,\delta)\right\}\right] \\
&\leq L_{mm}\epsilon(s,\delta)\eta(s) + 2y_{\max}^2 \mathbf{E}\left[W_1\, 1\left\{\|x - X_1\|_2 > \epsilon(s,\delta)\right\}\right] ,
\end{aligned}$$

where we used the fact that $V_1 \leq 1$, the assumption that $\sigma^2(\cdot)$ is L_{mm}-Lipschitz, the tower rule and the definition of $\eta(s)$. Furthermore,

$$\begin{aligned}
\mathbf{E}\left[W_1\, 1\left\{\|x - X_1\|_2 > \epsilon(s,\delta)\right\}\right] &\leq \Pr\left[\|x - X_1\|_2 \geq \epsilon(s,\delta) \text{ and } W_1 > 0\right] \\
&\leq \Pr\left[\sup_i \left\{\|x - X_i\|_2 : W_i > 0\right\} \geq \epsilon(s,\delta)\right] ,
\end{aligned}$$

which by definition is at most δ. By putting $\delta = \eta(s)^2$ we obtain

$$\left|\mathbf{E}\left[\mathbf{E}[W_1 \mid X_1]^2 \left(\sigma^2(X_1) - \sigma^2(x)\right)\right]\right| \leq L_{mm}\epsilon(s, \eta(s)^2)\eta(s) + 2y_{\max}^2 \eta(s)^2 = o(\eta(s)) ,$$

11. Since $\mathbf{E}[W_i]$ are all equal to the same value κ and $\sum_i \mathbf{E}[W_i] = 1$, we get $\kappa = 1/s$.

where we invoked our assumption that $\epsilon(s, \eta(s)^2) \to 0$. Thus we have obtained that:

$$\mathrm{Var}\left[\Phi_1(Z_1)\right] \geq \mathbf{E}\left[\mathbf{E}[W_1 \mid X_1]^2\right] \sigma^2(x) + o(\eta(s)),$$

which is exactly the form of the lower bound claimed in the statement of the lemma. This concludes the proof. ■

I.1. Hájek Projection Lemma for Infinite Order U-statistics

The following is a small adaptation of Theorem 2 of Fan et al. (2018), which we present here for completeness.

Lemma 38 (Fan et al. (2018)) *Consider a U-statistic defined via a symmetric kernel Φ:*

$$U(Z_1, \ldots, Z_n) = \binom{n}{s}^{-1} \sum_{1 \leq i_1 \leq \ldots \leq i_s \leq n} \Phi(Z_{i_1}, \ldots, Z_{i_s}), \tag{46}$$

where Z_i are i.i.d. random vectors and s can be a function of n. Let $\Phi_1(z_1) = \mathbf{E}[\Phi(z_1, Z_2, \ldots, Z_s)]$ and $\eta_1(s) = \mathrm{Var}_{z_1}[\Phi_1(z_1)]$. Suppose that $\mathrm{Var}\,\Phi$ is bounded, $n\,\eta_1(s) \to \infty$. Then:

$$\frac{U(Z_1, \ldots, Z_n) - \mathbf{E}[U]}{\sigma_n} \to_d \mathsf{N}(0, 1), \tag{47}$$

where $\sigma_n^2 = \frac{s^2}{n}\eta_1(s)$.

Proof The proof follows identical steps as the one in Fan et al. (2018). We argue about the asymptotic normality of a U-statistic:

$$U(Z_1, \ldots, Z_n) = \binom{n}{s}^{-1} \sum_{1 \leq i_1 \leq \ldots \leq i_s \leq n} \Phi(Z_{i_1}, \ldots, Z_{i_s}). \tag{48}$$

Consider the following projection functions:

$$\begin{aligned}
\Phi_1(z_1) &= \mathbf{E}[\Phi(z_1, Z_2, \ldots, Z_s)], & \tilde{\Phi}_1(z_1) &= \Phi_1(z_1) - \mathbf{E}[\Phi], \\
\Phi_2(z_1, z_2) &= \mathbf{E}[\Phi(z_1, z_2, Z_3, \ldots, Z_s), & \tilde{\Phi}_2(z_1, z_2) &= \Phi_2(z_1, z_2) - \mathbf{E}[\Phi], \\
&\vdots \\
\Phi_s(z_1, z_2, \ldots, z_s) &= \mathbf{E}[\Phi(z_1, z_2, Z_3, \ldots, Z_s), & \tilde{\Phi}_s(z_1, z_2, \ldots, z_s) &= \Phi_s(z_1, z_2, \ldots, z_s) - \mathbf{E}[\Phi],
\end{aligned}$$

where $\mathbf{E}[\Phi] = \mathbf{E}[\Phi(Z_1, \ldots, Z_s)]$. Then we define the canonical terms of Hoeffding's U-statistic decomposition as:

$$\begin{aligned}
g_1(z_1) &= \tilde{\Phi}_1(z_1)\,, \\
g_2(z_1, z_2) &= \tilde{\Phi}_2(z_1, z_2) - g_1(z_1) - g_2(z_2)\,, \\
g_3(z_1, z_2, z_3) &= \tilde{\Phi}_2(z_1, z_2, Z_3) - \sum_{i=1}^{3} g_1(z_i) - \sum_{1 \leq i < j \leq 3} g_2(z_i, z_j)\,, \\
&\vdots \\
g_s(z_1, z_2, \ldots, z_s) &= \tilde{\Phi}_s(z_1, z_2, \ldots, z_s) - \sum_{i=1}^{s} g_1(z_i) - \sum_{1 \leq i < j \leq s} g_2(z_i, z_j) - \ldots \\
&\quad \ldots - \sum_{1 \leq i_1 < i_2 < \ldots < i_{s-1} \leq s} g_{s-1}(z_{i_1}, z_{i_2}, \ldots, z_{i_{s-1}})\,.
\end{aligned}$$

Subsequently the kernel of the U-statistic can be re-written as a function of the canonical terms:

$$\tilde{\Phi}(z_1, \ldots, z_s) = \Phi(z_1, \ldots, z_s) - \mathbf{E}[\Phi] = \sum_{i=1}^{s} g_1(z_i) + \sum_{1 \leq i < j \leq s} g_2(z_i, z_j) + \ldots + g_s(z_1, \ldots, z_s)\,. \tag{49}$$

Moreover, observe that all the canonical terms in the latter expression are un-correlated. Hence, we have:

$$\mathrm{Var}\left[\Phi(Z_1, \ldots, Z_n)\right] = \binom{s}{1} \mathbf{E}\left[g_1^2\right] + \binom{s}{2} \mathbf{E}\left[g_2^2\right] + \ldots + \binom{s}{s} \mathbf{E}\left[g_s^2\right]\,. \tag{50}$$

We can now re-write the U statistic also as a function of canonical terms:

$$\begin{aligned}
U(Z_1, \ldots, Z_n) - \mathbf{E}[U] &= \binom{n}{s}^{-1} \sum_{1 \leq i_1 < i_2 < \ldots < i_s \leq n} \tilde{\Phi}(Z_{i_1}, \ldots, Z_{i_s}) \\
&= \binom{n}{s}^{-1} \left(\binom{n-1}{s-1} \sum_{i=1}^{n} g_1(Z_i) + \binom{n-2}{s-2} \sum_{1 \leq i < j \leq n} g_2(Z_i, Z_j) + \ldots \right. \\
&\quad \left. + \binom{n-s}{s-s} \sum_{1 \leq i_1 < i_2 < \ldots < i_s \leq n} g_s(Z_{i_1}, \ldots, Z_{i_s}) \right).
\end{aligned}$$

Now we define the Hájek projection to be the leading term in the latter decomposition:

$$\hat{U}(Z_1, \ldots, Z_n) = \binom{n}{s}^{-1} \binom{n-1}{s-1} \sum_{i=1}^{n} g_1(Z_i)\,. \tag{51}$$

The variance of the Hajek projection is:

$$\sigma_n^2 = \mathrm{Var}\left[\hat{U}(Z_1, \ldots, Z_n)\right] = \frac{s^2}{n} \mathrm{Var}\left[\Phi_1(z_1)\right] = \frac{s^2}{n} \eta_1(s)\,. \tag{52}$$

The Hájek projection is the sum of independent and identically distributed terms and hence by the Lindeberg-Levy Central Limit Theorem (see e.g., Billingsley (2008); Borovkov (2013)):

$$\frac{\hat{U}(Z_1, \ldots, Z_n)}{\sigma_n} \to_d \mathsf{N}(0, 1)\,. \tag{53}$$

We now argue that the remainder term: $\frac{U-\mathbf{E}[U]-\hat{U}}{\sigma_n}$ vanishes to zero in probability. The latter then implies that $\frac{U-\mathbf{E}[U]}{\sigma_n} \rightarrow_d \mathsf{N}(0,1)$ as desired. We will show the sufficient condition of convergence in mean square: $\frac{\mathbf{E}\left[\left(U-\mathbf{E}[U]-\hat{U}\right)^2\right]}{\sigma_n^2} \rightarrow 0$. From an inequality due to Wager and Athey (2018):

$$
\begin{aligned}
\mathbf{E}\left[\left(U-\mathbf{E}\left[U\right]-\hat{U}\right)^2\right] &= \binom{n}{s}^{-2}\left\{\binom{n-2}{s-2}^2\binom{n}{2}\mathbf{E}[g_2^2]+\ldots+\binom{n-s}{s-s}^2\binom{n}{s}\mathbf{E}[g_s^2]\right\} \\
&= \sum_{r=2}^{s}\left\{\binom{n}{s}^{-2}\binom{n-r}{s-r}^2\binom{n}{r}\mathbf{E}[g_r^2]\right\} \\
&= \sum_{r=2}^{s}\left\{\frac{s!(n-r)!}{n!(s-r)!}\binom{s}{r}\mathbf{E}[g_r^2]\right\} \\
&\leq \frac{s(s-1)}{n(n-1)}\sum_{r=2}^{s}\binom{s}{r}\mathbf{E}[g_r^2] \\
&\leq \frac{s^2}{n^2}\,\mathrm{Var}\left[\Phi(Z_1,\ldots,Z_s)\right]\,.
\end{aligned}
$$

Since $\mathrm{Var}\left[\Phi(Z_1,\ldots,Z_n)\right]$ is bounded by a constant V^* and $n\,\eta_1(s)\rightarrow\infty$, by our assumption, we have:

$$
\frac{\mathbf{E}\left[\left(U-\mathbf{E}\left[U\right]-\hat{U}\right)^2\right]}{\sigma_n^2} \leq \frac{\frac{s^2}{n^2}V^*}{\frac{s^2}{n}\eta_1} = \frac{V^*}{n\,\eta_1(s)} \rightarrow 0\,. \tag{54}
$$

■