Pointwise and Uniform Inference for the Two-Scale Distributional Nearest Neighbor Estimator

Jakob R. Juergens
University of Wisconsin - Madison

Last edited: June 27, 2024

Abstract

Recent advances in the literature on non-parametric regression and uniform inference for infinite-order U-statistics have enabled us to consider the problem of uniform inference for a broader class of estimators. One class of such estimators are bagged nearest neighbor estimators and among them specifically the Two-Scale Distributional Nearest Neighbor Estimator (TDNN) of Demirkaya et al. (2024). In this paper, we improve on existing results for pointwise inference using the TDNN estimator and develop uniform inference procedures based on recent work by Ritzwoller and Syrgkanis (2024). As part of this work, we provide a complementary R package called *tdnnR* that implements the presented methods.

Supplementary Material and R Package available at: `https://github.com/JakobJuergens/Unif_Inf_TDNN`

1 Introduction

Nearest Neighbor Estimators and their derivatives form a flexible class of estimators for a variety of purposes including nonparametric regression. Although widely used in practice for the latter purpose, some of their properties are still elusive when it comes to performing inference. This paper contributes to improving our understanding by establishing an asymptotically valid method to construct uniform confidence bands when using the Two-Scale Distributional Nearest Neighbor (TDNN) method of Demirkaya et al. (2024). To achieve this goal, we use novel results on uniform inference for infinite-order U-statistics (IOUS) developed in Ritzwoller and Syrgkanis (2024). Due to the inherent connection of the Potential Nearest Neighbors (PNN) framework to Random Forests (RF), this work also contributes to contextualizing recent advances in inference techniques for random forests. LOREM IPSUM

The remainder of this paper is organized as follows. Section 2 introduces the TDNN estimator as developed by Demirkaya et al. (2024) and presents some of the results this paper is based on. Section 3 uses recent theoretical developments to show that the distributional approximations provided by Demirkaya et al. (2024) can be justified under weaker assumptions on the chosen subsampling scales. Section 4 extends results from Ritzwoller and Syrgkanis (2024) to the TDNN estimator and thus introduces methods to construct uniform confidence bands for the TDNN estimator. Section 5 explores the performance of the developed uniform inference methods in a number of setups and Section 6 applies them to the context of LOREM IPSUM. Section 7 concludes.

1.1 Notation

- $\rightsquigarrow$ denotes convegence in distribution
- The norm $\|\cdot\|_{\psi_1}$ denotes the ψ_1-Orlicz norm
- $[n] = \{1, \ldots, n\}$
- $L_{n,s} = \{(l_1, \ldots, l_r) \in [n]^s \, : \, \text{entries of vector are distinct}\}$
- $I_{n,s} = \{(i_1, \ldots, i_r) \in L_{n,s} \, : \, i_1 < \ldots < i_s\}$

2 Two-Scale Distributional Nearest Neighbor Estimator

As in Demirkaya et al. (2024), consider a sample of independent and identically distributed observations

$$\mathbf{D}_n = \{\mathbf{Z}_i = (\mathbf{X}_i, Y_i)\}_{i=1}^n \quad \text{from the model} \quad Y = \mu(\mathbf{X}) + \varepsilon, \tag{2.1}$$

where $Y \in \mathbb{R}$ is the response, $\mathbf{X} \in \mathbb{R}^d$ is a feature vector of fixed dimension d, ε is the unobservable model error and $\mu(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$ is the unknown mean regression function. We will denote the distribution induced by this model by P and thus $Z_i \overset{\text{iid}}{\sim} P$. As we will embed the corresponding estimation problem in the context of subsampled conditional moment regression, note that this implies a conditional moment equation of the form

$$M(\mathbf{x}; \mu) = \mathbb{E}[m(\mathbf{Z}_i; \mu) \mid \mathbf{X}_i = \mathbf{x}] = 0 \quad \text{where} \quad m(\mathbf{Z}_i; \mu) = Y_i - \mu(\mathbf{X}_i). \tag{2.2}$$

Due to the absence of nuisance parameters in the setting at hand, conditions such as local Neyman-orthogonality vacuously hold (uniformly). In practice, the non-parametric regression problem at hand can be approached by solving the corresponding empirical conditional moment equation.

$$M_n(\mathbf{x}; \mu, \mathbf{D}_n) = \sum_{i=1}^n K(\mathbf{x}, \mathbf{X}_i) m(\mathbf{Z}_i; \mu) = 0 \tag{2.3}$$

In this equation, $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a data-dependent Kernel function measuring the "distance" between the point of interest and an observation. Notationally, this makes the local and data-dependent approach of this procedure explicit.

2.1 Distributional Nearest Neighbor Estimator

With a name coined by Demirkaya et al. (2024), the Distributional Nearest Neighbor (DNN) Estimator is based on important work by Steele (2009) and Biau and Guyader (2010). While its practical properties are appealing in and of themselves, from a statistical perspective, its appeal comes in part from being easily represented as a U-statistic. Given a sample as described above and a fixed feature vector $\mathbf{x}$, consider the ordered sample $\{(\mathbf{X}_{(1)}, Y_{(1)}), \ldots, (\mathbf{X}_{(n)}, Y_{(n)})\}$ defined by

$$||\mathbf{X}_{(1)} - \mathbf{x}|| \leq ||\mathbf{X}_{(2)} - \mathbf{x}|| \leq \ldots \leq ||\mathbf{X}_{(n)} - \mathbf{x}|| \tag{2.4}$$

where draws are broken according to the natural indices of the observations in a deterministic way. Let $\text{rk}(\mathbf{x}; \mathbf{Z}_i, D)$ denote the *rank* that is assigned to observation i in a sample D relative to a point of interest $\mathbf{x}$ in this fashion. By convention, let $\text{rk}(\mathbf{x}; \mathbf{Z}_i, D) = \infty$ if $\mathbf{Z}_i \notin D$. Similarly, let $Y_{(1)}(\mathbf{x}; D)$ indicate the response value of the closest neighbor in set D.

Remark 1 (Standardization)**.** *Depending on the underlying norm that is used to calculate the distances between observations, it can be advisable to standardize the feature space before applying nearest-neighbor methods. Since this is highly context-dependent, it is impossible to give general advice on when to standardize. Thus, this decision is left to the practitioner.*

Given a subsampling scale s satisfying $1 \leq s \leq n$, a generic set of subsample indices $\ell \in L_{n,s}$ and a corresponding

generic subset of our data $D_\ell = \{\mathbf{Z}_i \,|\, i \in \ell\}$, we can consider an analogous ordering of D_ℓ. This enables us to define a data-driven kernel function κ following the notation of Ritzwoller and Syrgkanis (2024).

$$\kappa(\mathbf{x}; \mathbf{Z}_i, D_\ell, \xi) = \mathbb{1}\left(\mathrm{rk}(\mathbf{x}; \mathbf{Z}_i, D_\ell) = 1\right) \tag{2.5}$$

Here, ξ is an additional source of randomness in the construction of the base learner that comes into play when analyzing, for example, random forests as proposed by Breiman (2001) using the CART-algorithm described in Breiman et al. (2017). As the DNN estimator does not incorporate such additional randomness, the term is omitted in further considerations. Noteworthy properties of κ are its permutational symmetry in D_ℓ and that κ does not consider the response variable when assigning weights to the observations under consideration. The latter immediately implies a property that has been called *Honesty* by Wager and Athey (2018).

Definition 1 (Symmetry and Honesty - Adapted from Ritzwoller and Syrgkanis (2024)).

1. *The kernel $\kappa(\cdot, \cdot, D_\ell)$ is Honest in the sense that*
$$\kappa(x, X_i, D_\ell) \perp\!\!\!\perp m(Z_i; \mu) \mid X_i, D_{\ell_{-i}},$$
where $\perp\!\!\!\perp$ denotes conditional independence and ℓ_{-i} denotes the set $\ell \backslash \{i\}$.

2. *The kernel $\kappa(\cdot, \cdot, D_\ell)$ is positive and satisfies the restriction $\sum_{i \in s} \kappa(\cdot, X_i, D_\ell) = 1$ almost surely. Moreover, the kernel $\kappa(\cdot, X_i, D_\ell)$ is invariant to permutations of the data D_ℓ.*

Using κ, it is straightforward to find an expression for the distance function K in Equation 2.3 corresponding to the DNN estimator.

$$K(\mathbf{x}, \mathbf{X}_i) = \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} \mathbb{1}(i \in \ell) \frac{\kappa(\mathbf{x}; \mathbf{Z}_i, D_\ell)}{s!} = \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} \frac{\mathbb{1}\left(\mathrm{rk}(\mathbf{x}; \mathbf{Z}_i, D_\ell) = 1\right)}{s!} \tag{2.6}$$

Inserting into Equation 2.3, this gives us the following empirical conditional moment equation.

$$\begin{aligned} M_n(\mathbf{x}; \mu, \mathbf{D}_n) &= \sum_{i=1}^{n} K(\mathbf{x}, \mathbf{X}_i) m(\mathbf{Z}_i; \mu) \\ &= \sum_{i=1}^{n} \left(\binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} \frac{\mathbb{1}\left(\mathrm{rk}(\mathbf{x}; \mathbf{Z}_i, D_\ell) = 1\right)}{s!} \right) (Y_i - \mu(\mathbf{X}_i)) = 0 \end{aligned} \tag{2.7}$$

Solving this empirical conditional moment equation then yields the DNN estimator $D_n^s(\mathbf{x})$ with subsampling scale s estimating the conditional expectation function $\mu(\mathbf{x}) = \mathbb{E}[Y \,|\, \mathbf{X} = \mathbf{x}]$. After rearranging the terms, it is given by the following U-statistic.

$$\tilde{\mu}_s(\mathbf{x}; \mathbf{D}_n) = \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} \frac{1}{s!} Y_{(1)}(\mathbf{x}; D_\ell) =: \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} h_s(\mathbf{x}; D_\ell) \tag{2.8}$$

From the empirical conditional moment equation, it becomes apparent that the DNN estimator is a Weighted Nearest Neighbor (WNN) estimator that automatically assigns suitable weights in a distributional fashion. Weighted nearest neighbor estimators in their general form have been studied in detail by ADD REFERENCES. Among other properties, it has been shown that under smoothness assumptions and using a suitable weight vector their convergence rate is optimal with $O_p(n^{-\frac{2}{d+4}})$. Through an equivalent representation as an L-statistic developed by Steele (2009), the

estimator has a computationally simple way to be evaluated compared to the usual U-statistic approach.

2.2 Two-Scale Distributional Nearest Neighbor Esimator

Starting from this setup, Demirkaya et al. (2024) develop a novel bias-correction method for the DNN estimator that leads to appealing finite-sample properties of the resulting Two-Scale Distributional Nearest Neighbor (TDNN) estimator. Their method is based on an explicit formula for the first-order bias term of the DNN estimator, which in turn allows them to eliminate it through a clever combination of two DNN estimators.

Theorem 2.1 (Demirkaya et al. (2024) - Theorem 1)**.** *Assume that the distribution of* $\mathbf{X}$ *has a density function* $f(\cdot)$ *with respect to the Lebesgue measure* λ *on the Euclidean space* $\mathbb{R}^d$*. Let* $\mathbf{x} \in \text{supp}(\mathbf{X})$ *be a fixed feature vector. If ...*

1. *There exists some constant* $\alpha > 0$ *such that* $\mathbb{P}(\|\mathbf{X} - \mathbf{x}\| \geq R) \leq e^{-\alpha R}$ *for each* $R > 0$*.*
2. *The density* $f(\cdot)$ *is bounded away from 0 and* ∞*,* $f(\cdot)$ *and* $\mu(\cdot)$ *are four times continuously differentiable with bounded second, third, and fourth-order partial derivatives in a neighborhood of* $\mathbf{x}$*, and* $\mathbb{E}[Y^2] < \infty$*. Moreover, the model error* ε *has zero mean and finite variance* $\sigma_\epsilon^2 > 0$ *and is independent of* $\mathbf{X}$*.*
3. *We have an i.i.d. sample* $\{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)\}$ *of size* n *from the model described in Equation 2.1.*

Then, for any fixed $\mathbf{x} \in \text{supp}(\mathbf{X}) \subset \mathbb{R}^d$*, we have that as* $s \to \infty$

$$\mathbb{E}\left[\tilde{\mu}_s(\mathbf{x}; \mathbf{D}_n)\right] = \mu(\mathbf{x}) + B(s) \tag{2.9}$$

with

$$B(s) = \Gamma(2/d + 1) \frac{f(\mathbf{x}) \operatorname{tr}\left(\mu''(\mathbf{x})\right) + 2\mu'(\mathbf{x})^T f'(\mathbf{x})}{2dV_d^{2/d} f(\mathbf{x})^{1+2/d}} s^{-2/d} + R(s)$$

$$R(s) = \begin{cases} O\left(s^{-3}\right), & d = 1 \\ O\left(s^{-4/d}\right), & d \geq 2 \end{cases}$$

where...

- $V_d = \frac{d^{d/2}}{\Gamma(1+d/2)}$
- $\Gamma(\cdot)$ *is the gamma function*
- $\operatorname{tr}(\cdot)$ *stands for the trace of a matrix*
- $f'(\cdot)$ *and* $\mu'(\cdot)$ *denote the first-order gradients of* $f(\cdot)$ *and* $\mu(\cdot)$*, respectively*
- $f''(\cdot)$ *and* $\mu''(\cdot)$ *represent the* $d \times d$ *Hessian matrices of* $f(\cdot)$ *and* $\mu(\cdot)$*, respectively*

Choosing two subsampling scales $1 \leq s_1 < s_2 \leq n$ and two corresponding weights

$$w_1^*(s_1, s_2) = \frac{1}{1 - (s_1/s_2)^{-2/d}} \quad \text{and} \quad w_2^*(s_1, s_2) = 1 - w_1^*(s_1, s_2) \tag{2.10}$$

they define the corresponding TDNN estimator as follows.

$$\widehat{\mu}_{s_1,s_2}(\mathbf{x}; \mathbf{D}_n) = w_1^*(s_1, s_2)\tilde{\mu}_{s_1}(\mathbf{x}; \mathbf{D}_n) + w_2^*(s_1, s_2)\tilde{\mu}_{s_2}(\mathbf{x}; \mathbf{D}_n) \tag{2.11}$$

As a direct consequence of Theorem 2.1, this leads to the elimination of the first-order bias term leading to desirable finite-sample properties. Furthermore, the authors show that this construction improves the quality of the normal approximation.

Theorem 2.2 (Demirkaya et al. (2024) - Theorem 3)**.** *Assume that Conditions (1) to (3) from Theorem 2.1 hold. Furthermore, let $s_2 \to \infty$ with $s_2 = o(n)$ and $c_1 \leq s_1/s_2 \leq c_2$ for some constants $0 < c_1 < c_2 < 1$. Then, for any fixed $\mathbf{x} \in \text{supp}(\mathbf{X}) \subset \mathbb{R}^d$, it holds that for some positive sequence σ_n of order $(s_2/n)^{1/2}$,*

$$\sigma_n^{-1}\left(\widehat{\mu}_{s_1,s_2}(\mathbf{x}; \mathbf{D}_n) - \mu(\mathbf{x}) - \Lambda\right) \rightsquigarrow \mathcal{N}(0, 1) \tag{2.12}$$

as $n \to \infty$, where

$$\Lambda = \begin{cases} O\left(s_1^{-4/d} + s_2^{-4/d}\right) & \text{for } d \geq 2 \\ O\left(s_1^{-3} + s_2^{-3}\right) & \text{for } d = 1 \end{cases}.$$

2.3 Hoeffding Decompositions for the DNN and TDNN Estimators

As the majority of the theoretical results in the original paper rely on representations as a U-statistic, it is necessary to introduce additional notation. Recalling Equation 2.8, the DNN and TDNN estimators can be expressed in the following U-statistic form.

$$\tilde{\mu}_s(\mathbf{x}; \mathbf{D}_n) = \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} h_s(\mathbf{x}; D_\ell) \quad \text{and} \quad \widehat{\mu}_{s_1,s_2}(\mathbf{x}; \mathbf{D}_n) = \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s_2}} h_{s_1,s_2}(\mathbf{x}; D_\ell) \tag{2.13}$$

It is worth pointing out explicitly that in contrast to the DNN estimator, the kernel for the TDNN estimator is of order $s_2 > s_1$. Borrowing the notational conventions from Lee (2019), additionally, introduce the following notation.

$$\psi_c^s(\mathbf{x}; \mathbf{z}_1, \ldots, \mathbf{z}_c) = \mathbb{E}_{\mathbf{Z}}\left[h_s\left(\mathbf{x}; \mathbf{z}_1, \ldots, \mathbf{z}_c, \mathbf{Z}_{c+1}, \ldots, \mathbf{Z}_s\right)\right] \tag{2.14}$$

$$h_s^{(1)}(\mathbf{x}; \mathbf{z}_1) = \psi_1^s(\mathbf{x}; \mathbf{z}_1) - \mu(\mathbf{x}) \tag{2.15}$$

$$h_s^{(c)}(\mathbf{x}; \mathbf{z}_1, \ldots, \mathbf{z}_c) = \psi_c^s(\mathbf{x}; \mathbf{z}_1, \ldots, \mathbf{z}_c) - \sum_{j=1}^{c-1}\left(\sum_{\ell \in L_{n,j}} h_s^{(j)}(\mathbf{x}; \mathbf{z}_\ell)\right) - \mu(\mathbf{x}) \quad \text{for } c = 2, \ldots, s \tag{2.16}$$

In contrast to the notational inspiration, the subsampling size s is made explicit. Since we are dealing with an infinite-order U-statistic, s will be diverging with n. In the usual fashion, these terms can be used to express the Hoeffding

projections of different orders.

$$H_{n,s}^{c} = \binom{n}{c}^{-1} \sum_{\ell \in L_{n,c}} h_s^{(c)}(\mathbf{x}; \mathbf{z}_\ell) \tag{2.17}$$

In a completely analogous fashion to the DNN estimator, we can define the corresponding terms for the Hoeffding decomposition of the TDNN estimator. The corresponding expressions will be denoted analogous to the terms for the DNN estimator with "s_1, s_2" replacing "s" ultimately leading to the following Hoeffding decompositions.

$$\tilde{\mu}_s\left(\mathbf{x}; \mathbf{D}_n\right) = \mu(\mathbf{x}) + \sum_{j=1}^{s} \binom{s}{j} H_{n,s}^{j} \quad \text{and} \quad \widehat{\mu}_{s_1,s_2}\left(\mathbf{x}; \mathbf{D}_n\right) = \mu(\mathbf{x}) + \sum_{j=1}^{s_2} \binom{s_2}{j} H_{n,s_1,s_2}^{j} \tag{2.18}$$

Furthermore, the corresponding Hajék-Projections are as follows.

$$\begin{aligned} \tilde{\mu}_s\left(\mathbf{x}; \mathbf{D}_n\right) &= \mu(\mathbf{x}) + \frac{s}{n} \sum_{i=1}^{n} h_s^{(1)}(\mathbf{x}; \mathbf{z}_i) + \mathrm{HR}_s(\mathbf{x}; \mathbf{D}_n) \\ \widehat{\mu}_{s_1,s_2}\left(\mathbf{x}; \mathbf{D}_n\right) &= \mu(\mathbf{x}) + \frac{s_2}{n} \sum_{i=1}^{n} h_{s_1,s_2}^{(1)}(\mathbf{x}; \mathbf{z}_i) + \mathrm{HR}_{s_1,s_2}(\mathbf{x}; \mathbf{D}_n) \end{aligned} \tag{2.19}$$

Here HR stands for the Hajék residual. Bounding this residual will be integral to the theoretical arguments made in the following.

For any $1 \leq c \leq s$, let

$$\zeta_{c,s} = \mathrm{Cov}\left(\mathbf{x}; h_s\left(\mathbf{Z}_1, \ldots, \mathbf{Z}_c, \mathbf{Z}_{c+1}, \ldots, \mathbf{Z}_s\right), h_s\left(\mathbf{x}; \mathbf{Z}_1, \ldots, \mathbf{Z}_c, \mathbf{Z}'_{c+1}, \ldots, \mathbf{Z}'_s\right)\right) \tag{2.20}$$

where $\mathbf{Z}'_{c+1}, \ldots, \mathbf{Z}'_n$ are i.i.d. from P and independent of $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ and thus $\zeta_{s,s} = \mathrm{Var}\left(\mathbf{x}; h_s\left(\mathbf{Z}_1, \ldots, \mathbf{Z}_s\right)\right)$.

3 Subsampling Rate Improvements for Variance Estimation

To perform inference, the authors introduce variance estimators based on the Jackknife and Bootstrap. However, as they point out, their consistency results rely on a likely suboptimal rate condition for the subsampling scale. While Theorem 2.2 allows s_2 to be of the order $o(n)$, the variance estimators rely on the considerably stronger condition that $s_2 = o(n^{1/3})$. Establishing consistency for their variance estimators under weaker assumptions on the subsampling rates could broaden the scope of the TDNN estimator for inferential purposes considerably. At a low level, this improvement will be based on a new order-explicit bound of the remainder in a linear approximation to U-statistics established in Ritzwoller and Syrgkanis (2024). Although these new results apply to high-dimensional U-statistics, for pointwise inference, we will consider their application to a scalar argument. We will return to their high-dimensional nature in Section 4 to construct procedures for uniform inference. Going forward, all results will be stated for the TDNN estimator. However, due to the underlying similarities between the U-statistic representations of the two estimators, much of this will apply to the DNN estimator analogously.

3.1 Variance Estimators

Define the following variance we need to estimate to perform pointwise inference at a point of interest $\mathbf{x}$.

$$\omega_n^2(\mathbf{x}) = \mathrm{Var}_D\left(\widehat{\mu}_{s_1,s_2}(\mathbf{x};\mathbf{D}_n)\right) \tag{3.1}$$

We denote by $\mathbf{D}_{n,-i}$ the data set $\mathbf{D}_n$ after removing the i'th observation. Then, the proposed Jackknife variance estimator takes the following form.

$$\widehat{\omega}_{JK}^2(\mathbf{x};\mathbf{D}_n) = \frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\mu}_{s_1,s_2}(\mathbf{x};\mathbf{D}_{n,-i}) - \widehat{\mu}_{s_1,s_2}(\mathbf{x};\mathbf{D}_n)\right)^2 \tag{3.2}$$

Next, denote the chosen number of bootstrap samples by $B \geq 1$ and for all $1 \leq b \leq B$, let $\mathbf{D}_n^{(b)}$ denote a bootstrap sample drawn from $\mathbf{D}_n$ uniformly and with replacement. Then, their bootstrap variance estimator is defined as follows.

$$\widehat{\omega}_{BS}^2(\mathbf{x};\mathbf{D}_n) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\mu}_{s_1,s_2}\left(\mathbf{x};\mathbf{D}_n^{(b)}\right) - \bar{\mu}_{s_1,s_2}(\mathbf{x};\mathbf{D}_n)\right)^2 \tag{3.3}$$

Here, $\bar{\mu}_{s_1,s_2}(\mathbf{x};\mathbf{D}_n)$ is defined in the fllowing way.

$$\bar{\mu}_{s_1,s_2}(\mathbf{x};\mathbf{D}_n) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\mu}_{s_1,s_2}\left(\mathbf{x};\mathbf{D}_n^{(b)}\right) \tag{3.4}$$

3.2 Subsampling Rate Improvements

By restricting Theorem 4.2 from Ritzwoller and Syrgkanis (2024) to the univariate case, i.e. considering only a single point of interest, we obtain the following corollary.

Corollary 1. *Consider a generic order s U-statistic*

$$U_s(\mathbf{x};\mathbf{D}_n) = \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} u(\mathbf{x};D_\ell) \quad \textit{with} \quad U_s(\mathbf{x};\mathbf{D}_n) = \theta(\mathbf{x}) \tag{3.5}$$

and the Kernel of its Hajék projection

$$u^{(1)}(\mathbf{x};\mathbf{z}) = \mathbb{E}\left[u(\mathbf{x};D_{[s]}) \,|\, \mathbf{Z}_1 = \mathbf{z}\right] - \theta(\mathbf{x}). \tag{3.6}$$

Furthermore, define the following variances.

$$\nu^2 = \mathrm{Var}\left(u(\mathbf{x};D_{[s]})\right), \quad \sigma_s^2 = \mathrm{Var}\left(u^{(1)}(\mathbf{x};\mathbf{Z})\right), \quad \psi_s^2 = \nu^2 - s\sigma_s^2 \tag{3.7}$$

If the kernel function $u(\mathbf{x};D_\ell)$ satisfies the bound $\|u(\mathbf{x};D_\ell)\|_{\psi_1} \leq \phi$, then

$$\sqrt{\frac{n}{s^2\sigma_s^2}}\left(u(\mathbf{x};\mathbf{D}_n) - \theta(\mathbf{x}) - \frac{b}{n}\sum_{i=1}^{n} u^{(1)}(\mathbf{x};\mathbf{Z}_i)\right) = \sqrt{\frac{n}{s^2\sigma_s^2}}\,\mathrm{HR}(\mathbf{x};\mathbf{D}_n) \lesssim \xi_{n,s}, \quad \textit{where} \tag{3.8}$$

$$\xi_{n,s} = \left(\frac{Cs\log(n)}{n}\right)^{s/2}\left(\left(\frac{n\psi_s^2}{s^2\sigma_s^2}\right)^{1/2} + \left(\frac{\phi^2 s\log^4(n)}{\sigma_s^2}\right)^{1/2}\right), \tag{3.9}$$

with probability greater than $1 - C/n$.

We want to use this bound to improve on the subsampling bounds derived in the proofs of Theorems 5 and 6 from Demirkaya et al., 2024. These can be found in their Supplementary Material, parts D.5 and D.6. In even greater detail, we want to improve on the relation shown in Equation (A.52) in the derivation of consistency of the Jackknife variance estimator.

4 Uniform Inference for the TDNN Estimator

Absent from Demirkaya et al. (2024) is a way to construct uniformly valid confidence bands around the TDNN estimator. Luckily, as a byproduct of considering the methods from Ritzwoller and Syrgkanis (2024), procedures for uniform inference can be developed relatively easily.

To consider this problem in detail we first introduce additional notation. Instead of a single point of interest, previously denoted by $\mathbf{x}$, we will consider a vector of p points of interest denoted by $\mathbf{x}^{(p)} \in (\text{supp}(\mathbf{X}))^p$. Consequently, the j-th entry of $\mathbf{x}^{(p)}$ will be denoted by $\mathbf{x}_j^{(p)}$. In an abuse of notation, let functions (such as μ or the DNN/TDNN estimators) evaluated at $\mathbf{x}^{(p)}$ denote the vector of corresponding function values evaluated at the point, respectively. It should be pointed out that, due to the local definition of the kernel in the estimators, this does not translate to the evaluation of the same function at different points in the most immediate sense. To summarize the kind of object we want to construct, we define a uniform confidence region for the TDNN estimator in the following way following closely the notation of Ritzwoller and Syrgkanis (2024).

Definition 2 (Uniform Confidence Regions)**.** *A confidence region for the TDNN (or DNN) estimators that is uniformly valid at the rate $r_{n,d}$ is a family of random intervals*

$$\widehat{\mathcal{C}}\left(\mathbf{x}^{(p)}\right) := \left\{\widehat{C}(\mathbf{x}_j^{(p)}) = \left[c_L(\mathbf{x}_j^{(p)}), c_U(\mathbf{x}_j^{(p)})\right] \,:\, j \in [p]\right\} \tag{4.1}$$

based on the observed data, such that

$$\sup_{P \in \mathbf{P}} \left| P\left(\mu(\mathbf{x}^{(d)}) \in \widehat{\mathcal{C}}\left(\mathbf{x}^{(d)}\right)\right)\right| \leq r_{n,d} \tag{4.2}$$

for some sequence $r_{n,d}$, where $\mathbf{P}$ is some statistical family containing P.

4.1 Low-Level

In our pursuit of constructing uniform confidence regions for the TDNN estimator, we return to the results from Ritzwoller and Syrgkanis (2024) in their high-dimensional form.

Theorem 4.1 (Ritzwoller and Syrgkanis (2024) - Theorem 4.1)**.** *For any sequence of kernel orders $b = b_n$, where*

$$\frac{1}{n}\frac{\nu_j^2}{\sigma_{b,j}^2} \to 0 \quad as \quad n \to \infty, \tag{4.3}$$

we have that

$$\sqrt{\frac{n}{\sigma_{b,j}^2 b^2}}\binom{n}{b}^{-1} \sum_{\mathbf{s} \in \mathbf{S}_{n,b}} u\left(\mathbf{x}_j^{(p)}; D_{\mathbf{s}}\right) \rightsquigarrow \mathcal{N}(0,1), \quad as \quad n \to \infty. \tag{4.4}$$

Theorem 4.2 (Adapted from Ritzwoller and Syrgkanis (2024) - Theorem 4.2)**.** *Define the terms*

$$\bar{\psi}_{s_2}^2 = \max_{j \in [p]} \left\{\nu_j^2 - s_2 \sigma_{s_2,j}^2\right\} \quad and \quad \underline{\sigma}_{s_2}^2 = \min_{j \in [p]} \sigma_{s_2,j}^2. \tag{4.5}$$

If the kernel function $h_{s_1,s_2}(\mathbf{x}; D_\ell)$ *satisfies the bound*

$$\|h_{s_1,s_2}(\mathbf{x}; D_\ell)\|_{\psi_1} \leq \phi \tag{4.6}$$

for each j *in* $[d]$*, then*

$$\sqrt{\frac{n}{s_2^2 \underline{\sigma}_{s_2}^2}} \left\| \widehat{\mu}_{s_1,s_2}(\mathbf{x}^{(p)}; \mathbf{D}_n) - \mu(\mathbf{x}^{(p)}) - \frac{s_2}{n} \sum_{i=1}^{n} h^{(1)}_{s_1,s_2}(\mathbf{x}^{(p)}; \mathbf{z}_i) \right\|_\infty = \sqrt{\frac{n}{s_2^2 \underline{\sigma}_{s_2}^2}} \left\| \mathrm{HR}_{s_1,s_2}(\mathbf{x}^{(p)}; \mathbf{D}_n) \right\|_\infty \lesssim \xi_{n,s_2}, \tag{4.7}$$

where

$$\xi_{n,s_2} = \left(\frac{C s_2 \log(pn)}{n} \right)^{s_2/2} \left(\left(\frac{n \bar{\psi}_{s_2}^2}{{s_2}^2 \underline{\sigma}_{s_2}^2} \right)^{1/2} + \left(\frac{\phi^2 s_2 \log^4(pn)}{\underline{\sigma}_{s_2}^2} \right)^{1/2} \right), \tag{4.8}$$

with probability greater than $1 - C/n$*.*

4.2 High-Level

Recent advances in the field of uniform inference for infinite-order U-statistics, specifically Ritzwoller and Syrgkanis (2024), and careful analysis of the Hoeffding projections of different orders will be the cornerstones in developing uniform inference methods. The authors' approach to constructing uniform confidence regions is based on the half-sample bootstrap root.

Definition 3 (Half-Sample Bootstrap Root Approximation - Ritzwoller and Syrgkanis (2024))**.** *The Half-Sample Bootstrap Root Approximation of the sampling distribution of the root*

$$R\left(\mathbf{x}^{(p)}; \mathbf{D}_n\right) := \widehat{\mu}\left(\mathbf{x}^{(p)}; \mathbf{D}_n\right) - \mu(\mathbf{x}^{(p)}) \tag{4.9}$$

is given by the conditional distribution of the half-sample bootstrap root

$$R^*\left(\mathbf{x}^{(p)}; \mathbf{D}_n\right) := \widehat{\mu}\left(\mathbf{x}^{(p)}; D_l\right) - \widehat{\mu}\left(\mathbf{x}^{(p)}; \mathbf{D}_n\right) \tag{4.10}$$

where l *denotes a random element from* $L_{n,n/2}$*.*

Next, to standardize the relevant quantities, we introduce a corresponding studentized process.

$$\widehat{\lambda}_j^2\left(\mathbf{x}^{(p)}; \mathbf{D}_n\right) = \mathrm{Var}\left(\sqrt{n} R^*(\mathbf{x}_j^{(p)}; \mathbf{D}_n) \,|\, \mathbf{D}_n\right) \quad \text{and} \quad \widehat{\Lambda}_n\left(\mathbf{x}^{(p)}; \mathbf{D}_n\right) = \mathrm{diag}\left(\left\{\widehat{\lambda}_j^2\left(\mathbf{x}^{(p)}; \mathbf{D}_n\right)\right\}_{j=1}^p\right) \tag{4.11}$$

$$\widehat{S}^*\left(\mathbf{x}^{(p)}; \mathbf{D}_n\right) := \sqrt{n} \left\| \left(\widehat{\Lambda}_n\left(\mathbf{x}^{(p)}; \mathbf{D}_n\right)\right)^{-1/2} R^*\left(\mathbf{x}^{(p)}; \mathbf{D}_n\right) \right\|_2 \tag{4.12}$$

Let $\mathrm{cv}(\alpha; \mathbf{D}_n)$ denote the $1-\alpha$ quantile of the distribution of $\widehat{S}^*\left(\mathbf{x}^{(p)}; \mathbf{D}_n\right)$. As the authors point out specifically, and as indicated by the more explicit notation chosen in this presentation, this is a quantile of the conditional distribution given the data $\mathbf{D}_n$. Given this construction, the uniform confidence region developed in Ritzwoller and Syrgkanis (2024) adapted to the TDNN estimator takes the following form.

Theorem 4.3 (Uniform Confidence Region - Ritzwoller and Syrgkanis (2024))**.** *Define the intervals*

$$\widehat{\mathcal{C}}\left(\mathbf{x}_j^{(p)};\mathbf{D}_n\right) := \widehat{\mu}\left(\mathbf{x}_j^{(p)};\mathbf{D}_n\right) \pm n^{-1/2}\widehat{\lambda}_j\left(\mathbf{x}^{(p)};\mathbf{D}_n\right) \mathrm{cv}\left(\alpha;\mathbf{D}_n\right) \tag{4.13}$$

The α-level uniform confidence region for $\mu\left(\mathbf{x}^{(p)}\right)$ is given by $\widehat{\mathcal{C}}\left(\mathbf{x}^{(p)}\right)$.

To justify the use of this uniform confidence region, it remains to be shown if and how the other conditions for the inner workings of this procedure apply to the TDNN estimator. This is substantially simplified due to the absence of a nuisance parameter. Thus, consider the following conditions from Ritzwoller and Syrgkanis, 2024 that are simplified to fit the problem at hand.

Definition 4 (Shrinkage and Incrementality - Adapted from Ritzwoller and Syrgkanis (2024))**.** *We say that the kernel $\kappa\left(\cdot,\cdot,D_\ell\right)$ has a uniform shrinkage rate ϵ_b if*

$$\sup_{P\in\mathbf{P}}\sup_{j\in[p]}\mathbb{E}\left[\max\left\{\left\|\mathbf{X}_i-\mathbf{x}_j^{(p)}\right\|_2 : \kappa\left(\mathbf{x}_j^{(p)},\mathbf{X}_i,D_\ell\right)>0\right\}\right] \leq \epsilon_b. \tag{4.14}$$

We say that a kernel $\kappa\left(\cdot,\cdot,D_\ell\right)$ is uniformly incremental if

$$\inf_{P\in\mathbf{P}}\sup_{j\in[p]}\mathrm{Var}\left(\mathbb{E}\left[\sum_{i\in\ell}\kappa\left(\mathbf{x}_j^{(p)},\mathbf{X}_i,D_\ell\right)m\left(\mathbf{Z}_i;\mu\right) \mid l\in\ell,\mathbf{Z}_l=Z\right]\right) \gtrsim b^{-1} \tag{4.15}$$

where Z is an independent random variable with distribution P.

Translating these properties to suit the TDNN regression problem, we obtain the following conditions that need to be verified. First, to verify uniform shrinkage at a rate ϵ_b, the following remains to be shown.

$$\sup_{P\in\mathbf{P}}\sup_{j\in[p]}\mathbb{E}\left[\max\left\{\left\|\mathbf{X}_i-\mathbf{x}_j^{(p)}\right\|_2 : \mathrm{rk}(\mathbf{x}_j^{(p)};\mathbf{X}_i,D_\ell)=1\right\}\right] \leq \epsilon_b \tag{4.16}$$

Second, for uniform incrementality, we need to show the following.

$$\begin{aligned}
&\inf_{P\in\mathbf{P}}\sup_{j\in[p]}\mathrm{Var}\left(\mathbb{E}\left[\sum_{i\in\ell}\mathbb{1}\left(\mathrm{rk}(\mathbf{x}_j^{(p)};\mathbf{X}_i,D_\ell)=1\right)\left(Y_i-\mu\left(\mathbf{X}_i\right)\right) \mid l\in\ell,\mathbf{Z}_l=Z\right]\right)\\
&=\inf_{P\in\mathbf{P}}\sup_{j\in[p]}\mathrm{Var}\left(\sum_{i\in\ell}\mathbb{E}\left[\mathbb{1}\left(\mathrm{rk}(\mathbf{x}_j^{(p)};\mathbf{X}_i,D_\ell)=1\right)\varepsilon_i \mid l\in\ell,\mathbf{Z}_l=Z\right]\right)\\
&=\inf_{P\in\mathbf{P}}\sup_{j\in[p]}\mathrm{Var}\left(\sum_{i=1}^{s}\mathbb{E}\left[\mathbb{1}\left(\mathrm{rk}(\mathbf{x}_j^{(p)};\mathbf{X}_i,D_{1:s})=1\right)\varepsilon_i \mid l\in[s],\mathbf{Z}_l=Z\right]\right)\\
&=\inf_{P\in\mathbf{P}}\sup_{j\in[p]}s^2\cdot\mathrm{Var}\left(\mathbb{E}\left[\mathbb{1}\left(\mathrm{rk}(\mathbf{x}_j^{(p)};\mathbf{X}_1,D_{1:s})=1\right)\varepsilon_1 \mid l\in[s],\mathbf{Z}_l=Z\right]\right)\gtrsim b^{-1}
\end{aligned} \tag{4.17}$$

To verify these assumptions, recent theory developed in Peng, Coleman, and Mentch (2022) is of great help. Specifically, the following Proposition and its proof are helpful in showing the desired uniform incrementality property.

LOREM IPSUM

Assumption 1 (Boundedness - Adapted from Ritzwoller and Syrgkanis (2024))**.** *The absolute value of the function* $m(\cdot;\mu)$ *is bounded by the constant* $(\theta+1)\phi$ *almost surely.*

$$|m(\mathbf{Z}_i;\mu)| = |Y_i - \mu(\mathbf{X}_i)| = |\varepsilon_i| \leq (\theta+1)\phi \quad a.s. \tag{4.18}$$

To follow the notational conventions, we will further define the two functions $m^{(1)}(\mathbf{Z}_i;\mu) = -\mu(\mathbf{X}_i)$ and $m^{(2)}(\mathbf{Z}_i) = Y_i$. As the authors point out, the boundedness condition can easily be replaced by a condition on the subexponential norm. This, being more in line with the assumptions of Demirkaya et al. (2024), is a desirable substitution. Thus, we will instead consider the following assumption and fill in parts of the proofs that hinge on boundedness for ease of exposition in the original paper.

Assumption 2 (Sub-Exponential Norm Bound)**.**

LOREM IPSUM

Assumption 3 (Moment Smoothness - Adapted from Ritzwoller and Syrgkanis (2024))**.** *Define the moments*

$$M^{(1)}(\mathbf{x};\mu) = \mathbb{E}\left[m^{(1)}\left(\mathbf{Z}_i;\mu\right) \mid \mathbf{X}_i = \mathbf{x}\right] \quad \textit{and} \quad M^{(2)}(\mathbf{x}) = \mathbb{E}\left[m^{(2)}\left(\mathbf{Z}_i\right) \mid \mathbf{X}_i = \mathbf{x}\right], \tag{4.19}$$

associated with the functions $m^{(1)}(\cdot;\mu)$ *and* $m^{(2)}(\cdot)$*. Plugging in yields the following functions.*

$$M^{(1)}(\mathbf{x};\mu) = -\mu(\mathbf{x}) \quad \textit{and} \quad M^{(2)}(\mathbf{x}) = \mu(\mathbf{x}). \tag{4.20}$$

Both moments are uniformly Lipschitz in their first component, in the sense that

$$\forall \mathbf{x}, \mathbf{x}' \in \operatorname{supp}(\mathbf{X}): \quad \sup_{P\in\mathbf{P}} |\mu(\mathbf{x}) - \mu(\mathbf{x}')| \lesssim \|\mathbf{x} - \mathbf{x}'\|_2 . \tag{4.21}$$

and $M^{(1)}$ *is bounded below in the following sense*

$$\inf_{P\in\mathbf{P}} \inf_{j\in[p]} \left|M^{(1)}\left(\mathbf{x}_j^{(p)}\right)\right| = \inf_{P\in\mathbf{P}} \inf_{j\in[p]} \left|\mu\left(\mathbf{x}_j^{(p)}\right)\right| \geq c \tag{4.22}$$

for some positive constant c*.*

The Lipschitz continuity part of this assumption translates directly into a Lipschitz continuity assumption on the unknown nonparametric regression function. The boundedness assumption is LOREM IPSUM

5 Simulations

Having developed theoretical results concerning uniform inference methods for the TDNN estimator, we will proceed by testing their properties in several simulation studies.

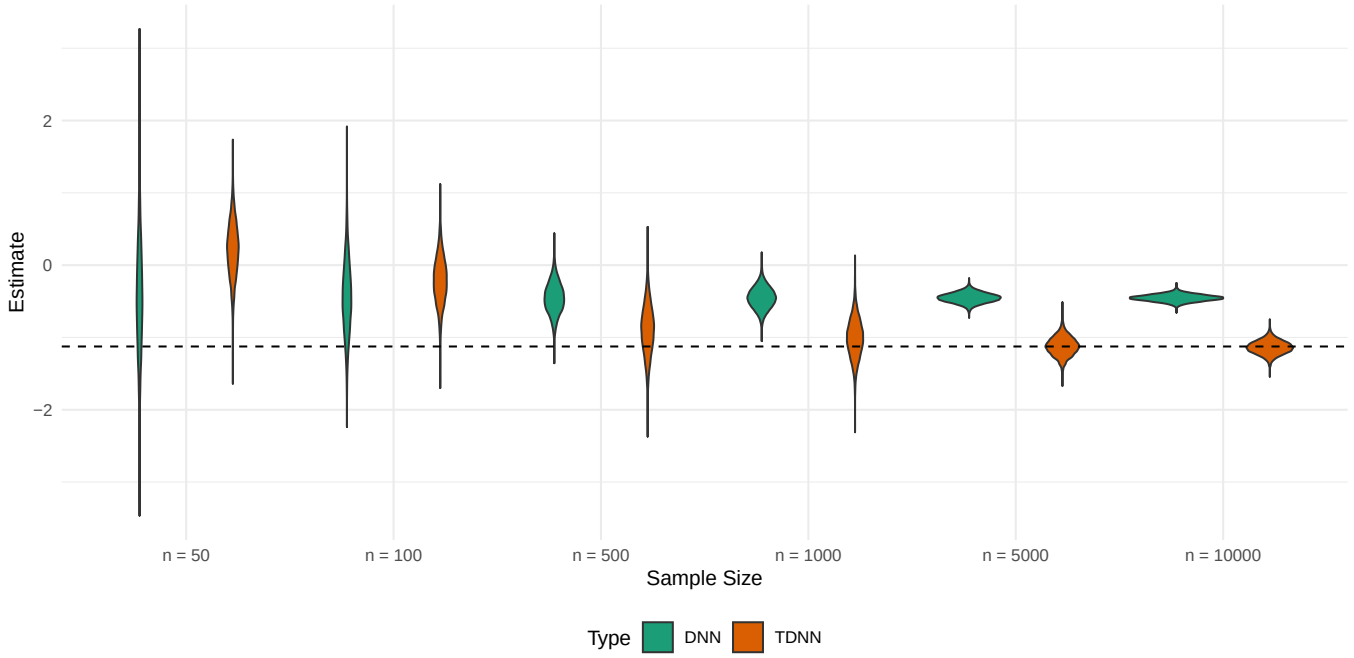

Figure 1: Comparison of the DNN ($s = 20$) and TDNN ($s_1 = 20, s_2 = 50$) Estimators for different sample sizes. The dashed line indicates the value of the unknown regression function at the point of interest. Simulation Setup replicates Setting 1 from Demirkaya et al. (2024) for 10000 Monte Carlo Replications.

LOREM IPSUM

6 Application

7 Conclusion

References

Biau, Gérard and Arnaud Guyader (Mar. 2010). "On the Rate of Convergence of the Bagged Nearest Neighbor Estimate". In: *The Journal of Machine Learning Research* 11, pp. 687–712.

Breiman, Leo (Oct. 2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

Breiman, Leo et al. (Oct. 2017). *Classification and Regression Trees*. New York: Chapman and Hall/CRC. ISBN: 978-1-315-13947-0. DOI: 10.1201/9781315139470.

Demirkaya, Emre et al. (Jan. 2024). "Optimal Nonparametric Inference with Two-Scale Distributional Nearest Neighbors". In: *Journal of the American Statistical Association* 119.545, pp. 297–307. DOI: 10.1080/01621459.2022.2115375.

Lee, A. J. (Mar. 2019). *U-Statistics: Theory and Practice*. New York: Routledge. ISBN: 978-0-203-73452-0. DOI: 10.1201/9780203734520.

Peng, Wei, Tim Coleman, and Lucas Mentch (Jan. 2022). "Rates of convergence for random forests via generalized U-statistics". In: *Electronic Journal of Statistics* 16.1, pp. 232–292. DOI: 10.1214/21-EJS1958.

Ritzwoller, David M. and Vasilis Syrgkanis (May 2024). *Uniform Inference for Subsampled Moment Regression*. DOI: 10.48550/arXiv.2405.07860.

Steele, Brian M. (Mar. 2009). "Exact bootstrap k-nearest neighbor learners". In: *Machine Learning* 74.3, pp. 235–255. DOI: 10.1007/s10994-008-5096-0.

Wager, Stefan and Susan Athey (July 2018). "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". In: *Journal of the American Statistical Association* 113.523. Publisher: Taylor & Francis, pp. 1228–1242. DOI: 10.1080/01621459.2017.1319839.

A Proofs

B PNN Calculations

B.1 1-PNN Calculations

Due to the inherent connections between random forests and the *Potential Nearest Neighbor* (PNN) framework, it is helpful to consider the properties of the k-PNN method first. Specifically, as we are considering large trees with a terminal node size of 1, we are interested in the properties of the 1-PNN framework. In this context, let $P_i(x)$ be an indicator for whether observation i is a 1-PNN of a point of interest $x \in \mathbb{R}^d$. Furthermore, let $\text{HR}(x, y)$ denote the hyperrectangle spanned by two points $x, y \in \mathbb{R}^d$ and by $x \prec y$ that $x \in \text{HR}(\mathbf{0}, y)$ and by $x \not\prec y$ the complement of that event.

Consider first the one-dimensional case and recall that given a subsample of indices $\mathcal{I}_S \subset \{1, \dots n\}$ with $i \in \mathcal{I}_S$ and $|\mathcal{I}_S| = s$, we have $P_i(x) = \mathbb{1}(i = \arg\min_{j \in \mathcal{I}_S} \|x - X_j\|)$. For simplicity and due to a symmetry of arguments, focus on the case of uniformly distributed features on $[0, 1]$ and $x = 0$ and without loss of generality consider $i = 1$. This

allows us to make the following observation

$$\begin{aligned}\mathbb{E}\left[P_1 \mid Z_1\right] = \mathbb{E}\left[P_1 \mid X_1\right] &= \mathbb{P}\left(X_1 = \min_{j \in \mathcal{I}_S} X_j \mid X_1\right) = \mathbb{P}(\forall j \in \mathcal{I}_S \backslash \{i\} : X_j > X_1 \mid X_1) \\ &= \mathbb{P}(B(s-1, X_1) = 0) = (1 - X_1)^{s-1}.\end{aligned}$$

Similarly, consider the following, concentrating without loss of generality on X_1 and X_2 for $\{1, 2\} \subset \mathcal{I}_S$ for $s \geq 2$.

$$\begin{aligned}\mathbb{E}[P_1 \mid Z_1, Z_2] = \mathbb{E}[P_1 \mid X_1, X_2] &= \mathbb{P}\left(X_1 = \min_{j \in \mathcal{I}_S} X_j \,\middle|\, X_1, X_2\right) \\ &= \mathbb{P}\left(X_1 = \min_{j \in \mathcal{I}_S} X_j \,\middle|\, X_1, X_2\right) \mathbb{1}(X_1 \geq X_2) + \mathbb{P}\left(X_1 = \min_{j \in \mathcal{I}_S} X_j \,\middle|\, X_1, X_2\right) \mathbb{1}(X_1 < X_2) \\ &= \mathbb{P}\left(B(s-2, X_1) = 0\right) \mathbb{1}(X_1 < X_2) = (1 - X_1)^{s-2} \mathbb{1}(X_1 < X_2)\end{aligned}$$

Generalizing this idea, we can find that the following holds for the case of conditioning on $p \leq s$ observations, where wlog we consider $1, \ldots, p \in \mathcal{I}_S$.

$$\mathbb{E}[P_1 \mid Z_1, \ldots, Z_p] = \mathbb{E}[P_1 \mid X_1, \ldots, X_p] = (1 - X_1)^{s-p} \mathbb{1}(\forall j = 2, \ldots, p : X_1 < X_j).$$

This is the first step to calculating the variances of the individual terms of the Hoeffding Decomposition of the weights for the 1-PNN method. Furthermore, we can recognize that the expectation of any sum of these conditional expectations has expectation equal to zero. For example, we can find the following

$$\mathbb{E}\Big[\mathbb{E}[P_1 \mid Z_1, Z_2] - \mathbb{E}[P1 \mid Z_1]\Big] = \mathbb{E}\left[\mathbb{E}[P_1 \mid Z_1, Z_2]\right] - \mathbb{E}\left[\mathbb{E}[P1 \mid Z_1]\right] = \mathbb{E}\left[P_1\right] - \mathbb{E}\left[P_1\right] = 0.$$

This, however, applies more broadly to sums of these terms, which is of interest in the upcoming calculations concerning the variances. Second, we need to consider the conditional expectations when conditioning on observations excluding i. As an illustrating example consider first the following

$$\begin{aligned}\mathbb{E}[P_1 \mid Z_2] = \mathbb{E}[P_1 \mid X_2] &= \mathbb{E}\left[\mathbb{E}[P_1 \mid X_1, X_2] \mid X_2\right] = \mathbb{E}\left[(1 - X_1)^{s-2} \mathbb{1}(X_1 < X_2) \mid X_2\right] \\ &= \int_0^{X_2} (1 - x_1)^{s-2} \, dx_1 = \frac{1 - (1 - X_2)^{s-1}}{s - 1}\end{aligned}$$

Similarly, we can consider conditioning on an arbitrary set of observations excluding $i = 1$ and make use of the law of iterated expectations analogously. Without loss of generality, consider again $i = 1$ and conditioning on observations 2 to $p < s$ and to simplify the notation, let $\underline{X}_{2:p} = \min_{i=2,\ldots,p} X_i$.

$$\begin{aligned}\mathbb{E}[P_1 \mid Z_2, \ldots, Z_p] = \mathbb{E}[P_1 \mid X_2, \ldots, X_p] &= \mathbb{E}\left[\mathbb{E}\left[P_1 \mid X_1, X_2, \ldots, X_p\right] \mid X_2, \ldots, X_p\right] \\ &= \int_0^{\underline{X}_{2:p}} (1 - x_1)^{s-p} \, dx_1 = \frac{1 - (1 - \underline{X}_{2:p})^{s+1-p}}{s + 1 - p}\end{aligned}$$

The next step is to consider the actual terms of the Hoeffding decomposition and evaluate their variance. First among

them, we can find the following

$$\begin{aligned}\operatorname{Var}(\mathbb{E}[P_1 \mid X_1]) = \operatorname{Var}\left((1-X_1)^{s-1}\right) &= \mathbb{E}\left[(1-X_1)^{2(s-1)}\right] - \mathbb{E}\left[(1-X_1)^{s-1}\right]^2 \\ &= \int_0^1 (1-x_1)^{2(s-1)} dx_1 - \left(\int_0^1 (1-x_1)^{s-1} dx_1\right)^2 = \frac{1}{2s-1} - \frac{1}{s^2} \\ &\sim s^{-1}.\end{aligned}$$

To evaluate the variance of the second term, we can proceed by combining the expressions of the first and second conditional expectations.

$$\begin{aligned}\mathbb{E}[P_1 \mid Z_1, Z_2] - \mathbb{E}[P1 \mid Z_1] &= (1-X_1)^{s-2}(\mathbb{1}(X_1 < X_2) - (1-X_1)) \\ &= (1-X_1)^{s-2}(X_1 - \mathbb{1}(X_1 \geq X_2)).\end{aligned}$$

Thus, to calculate the variance, observe that

$$\begin{aligned}\left|\mathbb{E}[P_1 \mid Z_1, Z_2] - \mathbb{E}[P_1 \mid Z_1]\right|^2 &= \left|(1-X_1)^{s-2}(X_1 - \mathbb{1}(X_1 \geq X_2))\right|^2 \\ &= (1-X_1)^{2s-4}\left|X_1 - \mathbb{1}(X_1 \geq X_2)\right|^2 \\ &= (1-X_1)^{2s-4}\left(X_1^2 + \mathbb{1}(X_1 \geq X_2)(1-2X_1)\right).\end{aligned}$$

and thus

$$\begin{aligned}&\operatorname{Var}(\mathbb{E}[P_1 \mid Z_1, Z_2] - \mathbb{E}[P1 \mid Z_1]) = \mathbb{E}\left[(1-X_1)^{2s-4}\left(X_1^2 + \mathbb{1}(X_1 \geq X_2)(1-2X_1)\right)\right] \\ =&\mathbb{E}\left[(1-X_1)^{2s-4}X_1^2\right] + \mathbb{E}\left[(1-X_1)^{2s-4}\mathbb{1}(X_1 \geq X_2)(1-2X_1)\right] \\ =&\int_0^1 (1-x_1)^{2s-4}x_1^2\,dx_1 + \int_0^1\int_0^1 (1-x_1)^{2s-4}\mathbb{1}(x_1 \geq x_2)(1-2x_1)\,dx_1\,dx_2 \\ =&\int_0^1 (1-x_1)^{2s-4}x_1^2\,dx_1 + \int_0^1\int_{x_2}^1 (1-x_1)^{2s-4}(1-2x_1)\,dx_1\,dx_2 \\ =&\frac{1}{4s^3-12s^2+11s-3} + \int_0^1 \left[\frac{(1-x_2)^{2s-3}}{2s-3}\left(1 - \frac{(2s-3)x_2+1}{s-1}\right)\right] dx_2 \\ =&\frac{1}{4s^3-12s^2+11s-3} + \frac{5-2s}{-8s^3+24s^2-22s+6} \sim s^{-2}\end{aligned}$$

Furthermore, we can find that

$$\begin{aligned}\operatorname{Var}\left(\mathbb{E}[P_1 \mid Z_2]\right) = \operatorname{Var}\left(\frac{1-(1-X_2)^{s-1}}{s-1}\right) &= \frac{1}{(s-1)^2}\operatorname{Var}\left((1-X_2)^{s-1}\right) \\ &= \frac{1}{(s-1)^2}\left(\mathbb{E}\left[(1-X_2)^{2(s-1)}\right] - \mathbb{E}\left[(1-X_2)^{s-1}\right]^2\right) \\ &= \frac{1}{(s-1)^2}\left(\frac{1}{2s-1} - \frac{1}{s^2}\right) \sim s^{-3}\end{aligned}$$

This, in turn, implies the following

$$\operatorname{Var}\left(\mathbb{E}[P_1 \mid Z_1, Z_2] - \mathbb{E}[P_1 \mid Z_1] + \mathbb{E}[P_2 \mid Z_1, Z_2] - \mathbb{E}[P_2 \mid Z_2]\right) \sim s^{-2}.$$

To approach this problem more generally, it is imperative to find an expression for the variance of the c'th term of the

Hoeffding decomposition. Now, fixing P_1 as the weight of interest, recall in analogy to the beginning of this paper the following term:

$$A_1^{(k)} := \sum_{E \in P_{s,k-1}^{\{1\}}} (-1)^{\mathbb{1}(|E| \equiv (k-1) \mod 2)} \mathbb{E}\left[P_1 \mid X_1, X_e \,:\, e \in E\right].$$

As this term forms the core of the kernel of the relevant Hoeffding decomposition, it will allow us to gain a better understanding of the relevant variances at play. First, by a simple application of the law of iterated expectations, we find that $\mathbb{E}[A_1^{(k)}] = 0$. Thus, to understand its variance, we need to investigate $\mathbb{E}\left[|A_1^{(k)}|^2\right]$.

$$|A_1^{(k)}|^2 = \sum_{E,E' \in P_{s,k-1}^{\{1\}}} (-1)^{\mathbb{1}(|E|+|E'| \equiv 0 \mod 2)} \left\{\mathbb{E}\left[P_1 \mid X_1, X_e \,:\, e \in E\right] \mathbb{E}\left[P_1 \mid X_1, X_{e'} \,:\, e' \in E'\right]\right\}$$

As this expression shows, we are interested in expressions of the following form

$$\mathbb{E}\Big[\mathbb{E}\left[P_1 \mid X_1, X_e \,:\, e \in E\right] \mathbb{E}\left[P_1 \mid X_1, X_{e'} \,:\, e' \in E'\right]\Big]$$

Now, fix $E, E' \in P_{s,k-1}^{\{1\}}$, and let $F = E \cap E'$. Given this now fixed expression, we can continue our analysis in the following way.

$$\begin{aligned}
A_{E,E'} :=& \mathbb{E}\Big[\mathbb{E}\left[P_1 \mid X_1, X_e \,:\, e \in E\right] \mathbb{E}\left[P_1 \mid X_1, X_{e'} \,:\, e' \in E'\right]\Big] \\
=& \mathbb{E}\left[(1 - X_1)^{s-|E|-1} \mathbb{1}(\forall j \in E :\, X_1 < X_j) (1 - X_1)^{s-|E'|-1} \mathbb{1}(\forall j \in E' :\, X_1 < X_j)\right] \\
=& \mathbb{E}\left[(1 - X_1)^{2s-|E|-|E'|-2} \mathbb{1}(\forall j \in E \cup E' :\, X_1 < X_j)\right] \\
=& \mathbb{E}\left[(1 - X_1)^{2s-\mathcal{E}-\mathcal{E}'-2} \mathbb{1}(\forall j \in E \cup E' :\, X_1 < X_j)\right]
\end{aligned}$$

Here, to simplify the notation, we let $\mathcal{E} = |E|$, $\mathcal{E}' = |E'|$ and $\mathcal{F} = |F|$. Recall that we are dealing with independent and identically uniformly distributed features, such that we can simplify this expectation by considering the density of the minimum of $\mathfrak{E} = \mathcal{E} + \mathcal{E}' - \mathcal{F}$ i.i.d. uniform random variables instead of the multivariate integral. Specifically, realize that defining $W = \min_{e \in E \cup E'} X_e$, W is distributed according to the following density function.

$$f_W(w) = \begin{cases} \mathfrak{E}\,(1 - w)^{\mathfrak{E}-1} & \text{if } w \in [0, 1] \\ 0 & \text{otherwise} \end{cases}.$$

Using this expression, we find that

$$\begin{aligned}
A_{E,E'} &= \int_0^1 \left(\mathfrak{E}\,(1-w)^{\mathfrak{E}-1} \int_0^w (1-x_1)^{2s-\mathcal{E}-\mathcal{E}'-2}\, dx_1 \right) dw \\
&= \int_0^1 \left(\mathfrak{E}\,(1-w)^{\mathfrak{E}-1} \frac{1}{2s-\mathcal{E}-\mathcal{E}'-1} \left(1-(1-w)^{2s-\mathcal{E}-\mathcal{E}'-1}\right) \right) dw \\
&= \frac{1}{2s-\mathcal{E}-\mathcal{E}'-1} \left(1 - \int_0^1 \mathfrak{E}\,(1-w)^{\mathfrak{E}-1} (1-w)^{2s-\mathcal{E}-\mathcal{E}'-1}\, dw \right) \\
&= \frac{1}{2s-\mathfrak{E}+\mathcal{F}-1} \left(1 - \mathfrak{E} \int_0^1 (1-w)^{2s-\mathcal{F}-2}\, dw \right) \\
&= \frac{1}{2s-\mathfrak{E}+\mathcal{F}-1} \left(1 - \frac{\mathfrak{E}}{2s-\mathcal{F}-1} \right) = \frac{2s-\mathfrak{E}-\mathcal{F}-1}{2s-\mathfrak{E}+\mathcal{F}-1} (2s-\mathcal{F}-1)^{-1} \\
&\sim s^{-1}
\end{aligned}$$

Considering this notation, we can write the expression of interest in the following way to continue investigating its behavior.

$$|A_1^{(k)}|^2 = \sum_{E,E' \in P_{s,k-1}^{\{1\}}} (-1)^{\mathbb{1}(|E|+|E'| \equiv 0 \mod 2)} A_{E,E'}$$

Considering the original test case of $k = 2$, we can check whether this approach gives us the same results.

$$\begin{aligned}
|A_1^{(2)}|^2 &= \sum_{E,E' \in P_{s,1}^{\{1\}}} (-1)^{\mathbb{1}(|E|+|E'| \equiv 0 \mod 2)} A_{E,E'} \\
&= (s-1)A_{\{2\},\{2\}} + (s-1)(s-2)A_{\{2\},\{3\}} - 2(s-1)A_{\{2\},\emptyset} + A_{\emptyset,\emptyset} \\
&= \frac{(s-1)(2s-4)}{(2s-2)(2s-2)} + \frac{(s-1)(s-2)(2s-3)}{(2s-1)(2s-1)} - \frac{2(s-1)(2s-2)}{(2s-2)(2s-1)} + \\
&= \frac{s-2}{2s-2} + \frac{(s-1)(s-2)(2s-3)}{(2s-1)(2s-1)} - \frac{2(s-1)}{2s-1}
\end{aligned}$$

LOREM IPSUM! I found a mistake. Will fix soon.

Thus, recall the following recursive Kernel definition that forms the basis to the Hoeffding decomposition.

$$\begin{aligned}
h^{(1)}(x_1) &= \psi_1(x_1) - \theta \\
h^{(c)}(x_1, x_2, \ldots, x_c) &= \psi_c(x_1, \ldots, x_c) - \sum_{j=1}^{c-1} \sum_{(c,j)} h^{(j)}(x_{i_1}, \ldots, x_{i_j}) - \theta
\end{aligned}$$

To illustrate the connection to the problem at hand, fix first P_1 as the weight of interest. Then, in an abuse of notation, consider $\psi_c(x_{i_1}, \ldots, x_{i_j}) = \mathbb{E}\left[P_1 \,|\, X_{i_1} = x_{i_1}, \ldots, X_{i_j} = x_{i_j}\right]$.

B.2 1-PNN in d Dimensions

Next, consider the case of uniformly distributed features on $[0,1]^d$ with the point of interest at the origin, i.e. $x = \mathbf{0}$. Similar to the one-dimensional case, the argument will extend to other points under consideration and other suitable distributions.

What this choice allows us to do is consider a simplified form of the probability of a point being a 1-PNN. Thus, observe that the probability of a sample point X_1 being a 1-PNN of the origin, i.e. the expectation of P_1, can be expressed in the following form

$$\begin{aligned}
\mathbb{E}[P_1 \,|\, X_1] &= \mathbb{P}(X_1 \text{ is a 1-PNN of } \mathbf{0} \,|\, X_1) \\
&= \mathbb{P}\left(\bigcap_{j=2}^{s} \{X_j \notin \mathrm{HR}(\mathbf{0}, X_1) \,|\, X_1\}\right) = \mathbb{P}\left(\bigcap_{j=2}^{s}\bigcap_{k=1}^{d} \{X_{jk} \notin [0, X_{1,k}] \,|\, X_1\}\right) \\
&= \mathbb{P}\left(B\left(s-1, \prod_{k=1}^{d} X_{1,k}\right) = 0 \,|\, X_1\right) = \mathbb{P}\left(B\left(s-1, X_{1,\bullet}\right) = 0 \,|\, X_1\right) \\
&= (1 - X_{1,\bullet})^{s-1},
\end{aligned}$$

where we let $X_{i,\bullet}$ denote $\prod_{k=1}^{d} X_{i,k}$. To calculate the variance of the conditional expectation, consider the following rewriting

$$\mathrm{Var}\left(\mathbb{E}[P_1 \,|\, X_1]\right) = \mathbb{E}[(1 - X_{1,\bullet})^{2(s-1)}] - \left(\mathbb{E}[(1 - X_{1,\bullet})^{s-1}]\right)^2.$$

Now, notice that $X_{1,\bullet}$ is distributed according to the density

$$f_{1,\bullet}(x) = \begin{cases} \frac{(-\log(x))^{d-1}}{(d-1)!} & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}.$$

Using this density, we can solve for the variance in the following way

$$\begin{aligned}
\mathrm{Var}\left(\mathbb{E}[P_1 \,|\, X_1]\right) &= \int_0^1 \frac{(-\log(x))^{d-1}}{(d-1)!}(1-x)^{2(s-1)}\, dx - \left(\int_0^1 \frac{(-\log(x))^{d-1}}{(d-1)!}(1-x)^{s-1}\, dx\right)^2 \\
&= \text{Tedious... but needs a good argument} \\
&\sim s^{-1} \quad \text{I assume...}
\end{aligned}$$

Proceeding as in the one-dimensional case, we consider next the conditional expectation of P_1 given Z_1 and Z_2, where these choices are again without loss of generality.

$$\begin{aligned}
\mathbb{E}[P_1 \,|\, Z_1, Z_2] = \mathbb{E}[P_1 \,|\, X_1, X_2] &= \mathbb{1}_{(X_2 \prec X_1)}\mathbb{E}[P_1 \,|\, X_1, X_2, X_2 \prec X_1] + \mathbb{1}_{(X_2 \not\prec X_1)}\mathbb{E}[P_1 \,|\, X_1, X_2, X_2 \not\prec X_1] \\
&= \mathbb{1}_{(X_2 \not\prec X_1)}\mathbb{E}[P_1 \,|\, X_1, X_2, X_2 \not\prec X_1] = \mathbb{1}_{(X_2 \not\prec X_1)}\mathbb{P}\left(B\left(s-2, \prod_{k=1}^{d} X_{1,k}\right) = 0 \,|\, X_1\right) \\
&= \mathbb{1}_{(X_2 \not\prec X_1)}(1 - X_{1,\bullet})^{s-2}.
\end{aligned}$$

LOREM IPSUM

C Stuff

Theorem C.1 (Ritzwoller and Syrgkanis (2024) - Corollary 4.1)**.** *Let Σ be the diagonal matrix with components $\sigma^2_{b,j}$.*

1. *Under the same conditions as Theorem 4.2, we have that*

$$\sqrt{\frac{n}{b^2}}\Sigma^{-1/2}\bar{U}_{n,b}\left(\mathbf{x}^{(p)}\right) \lesssim \log^{1/2}(dn) + \frac{\phi\log^2(dn)}{\underline{\sigma}_b n^{1/2}} + \xi_{n,b} \tag{C.1}$$

with probability greater than $1 - C/n$.

2. *Let* Z *denote a centered Gaussian random vector with covariance matrix* $\mathrm{Var}\left(\tilde{u}^{(1)}\left(\boldsymbol{x}^{(d)}, D_i\right)\right)$. *Under the same conditions as Theorem 4.2, we have that*

$$\sup_{\mathrm{R}\in\mathcal{R}}\left|P\left\{\sqrt{\frac{n}{b^2}}\Sigma^{-1/2}\bar{U}_{n,b}\left(\boldsymbol{x}^{(d)}\right)\in\mathrm{R}\right\} - P\left\{\Sigma^{-1/2}Z\in\mathrm{R}\right\}\right| \lesssim \left(\frac{\phi^2\log^5(dn)}{\sigma_b^2 n}\right)^{1/4} + \xi_{n,b}\sqrt{\log(d)}, \tag{C.2}$$

where $\mathcal{R}$ *denotes the set of hyper-rectangles in* $\mathbb{R}^d$.