Uniform Inference for the Two-Scale Distributional Nearest Neighbor Estimator

Jakob R. Juergens
University of Wisconsin - Madison

Last edited: June 19, 2024

Abstract

Recent advances in the literature on non-parametric regression and uniform inference for incomplete 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, I develop uniform inference procedures based on recent work by Ritzwoller and Syrgkanis (2024) for the TDNN estimator. ...

Supplementary material is available at: LOREM IPSUM

Contents

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 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 will be 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 extends results from Ritzwoller and Syrgkanis (2024) to the TDNN estimator and thus introduces methods to construct uniform confidence bands for the TDNN estimator. LOREM IPSUM

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, making explicit the local approach of this procedure.

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. To simplify further expressions, let $[n] = \{1, \ldots, n\}$ and define $L_{n,r}$ and $I_{n,r}$ as follows.

$$L_{n,s} = \{(l_1, \ldots, l_r) \in [n]^s \, : \text{ entries of vector are distinct}\} \tag{2.5}$$

$$I_{n,s} = \{(i_1, \ldots, i_r) \in L_{n,s} \, : \, i_1 < \ldots < i_s\} \tag{2.6}$$

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_s. This enables us to define a data-driven kernel function κ following the notation of Ritzwoller and Syrgkanis (2024).

$$\kappa(\mathbf{x}; \mathbf{Z}_i, D_\ell, \xi) = \mathbb{1}\left(\text{rk}(\mathbf{x}; \mathbf{Z}_i, D_\ell) = 1\right) \tag{2.7}$$

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_s 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 2.1 (*Symmetry and Honesty - Adapted from Ritzwoller and Syrgkanis (2024)*)

1. *The kernel $\kappa(\cdot, \cdot, D_\text{s})$ is Honest in the sense that*

$$\kappa(x, X_i, D_\mathbf{s}) \perp\!\!\!\perp m(Z_i; \mu) \mid X_i, D_{\mathbf{s}_{-i}},$$

   where $\perp\!\!\!\perp$ denotes conditional independence and s_{-i} denotes the set $\text{s}\backslash\{i\}$.

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

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

$$K(\mathbf{x}, \mathbf{X}_i) = \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} \mathbb{1}(i \in \ell) \frac{\kappa(\mathbf{x}; \mathbf{Z}_i, D_\ell)}{s!} = \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} \frac{\mathbb{1}\left(\text{rk}(\mathbf{x}; \mathbf{Z}_i, D_\ell) = 1\right)}{s!} \tag{2.8}$$

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

$$\begin{aligned} M_n(\mathbf{x}; \mu, \mathbf{D}_n) &= \sum_{i=1}^{n} K(\mathbf{x}, \mathbf{X}_i) m(\mathbf{Z}_i; \mu) \\ &= \sum_{i=1}^{n} \left(\binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} \frac{\mathbb{1}\left(\mathrm{rk}(\mathbf{x}; \mathbf{Z}_i, D_\ell) = 1\right)}{s!} \right) (Y_i - \mu(\mathbf{X}_i)) = 0 \end{aligned} \tag{2.9}$$

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 $\mu(\mathbf{x}) = \mathbb{E}\left[Y \mid \mathbf{X} = \mathbf{x}\right]$. After rearranging the terms, it is given by the following U-statistic.

$$D_n^s(\mathbf{x}) = \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} \frac{1}{s!} Y_{(1)}(\mathbf{x}; D_\ell) =: \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} \psi^s(\mathbf{x}; D_\ell) \tag{2.10}$$

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 $O_p(n^{-\frac{2}{d+4}})$.

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 entirely through a clever combination of two DNN estimators.

Theorem 2.2 (*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* $\mathbf{x} \in \mathrm{supp}(\mathbf{X})$ *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 $\mathbf{x} \in \mathrm{supp}(\mathbf{X}) \subset \mathbb{R}^d$*, we have that as* $s \to \infty$

$$\mathbb{E}\left[D_n^s(\mathbf{x})\right] = \mu(\mathbf{x}) + B(s) \tag{2.11}$$

with

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

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

where. . .

- $V_d = \frac{d^{d/2}}{\Gamma(1+d/2)}$
- $\Gamma(\cdot)$ *is the gamma function*
- $\mathrm{tr}(\cdot)$ *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

$$w_1^*(s_1, s_2) = \frac{1}{1 - (s_1/s_2)^{-2/d}} \quad \text{and} \quad w_2^*(s_1, s_2) = 1 - w_1^*(s_1, s_2) \tag{2.12}$$

they define the corresponding TDNN estimator as follows.

$$D_n^{s_1,s_2}(\mathbf{x}) = w_1^*(s_1, s_2) D_n^{s_1}(\mathbf{x}) + w_2^*(s_1, s_2) D_n^{s_2}(\mathbf{x}) \tag{2.13}$$

As a direct consequence of Theorem 2.2, 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 on the quality of the normal approximation.

Theorem 2.3 (*Demirkaya et al. (2024) - Theorem 3*)

Assume that Conditions (1) to (3) from Theorem 2.2 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* $\mathbf{x} \in \mathrm{supp}(\mathbf{X}) \subset \mathbb{R}^d$*, it holds that for some positive sequence* σ_n *of order* $(s_2/n)^{1/2}$*,*

$$\sigma_n^{-1} \left(D_n^{s_1,s_2}(\mathbf{x}) - \mu(\mathbf{x}) - \Lambda \right) \rightsquigarrow \mathcal{N}(0, 1) \tag{2.14}$$

as $n \to \infty$*, where*

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

Furthermore, the authors provide an upper bound for the point-wise MSE of the TDNN estimator and suitable variance estimators. This bound can be used to derive subsampling scales that optimize the bias-variance tradeoff.

As part of their asymptotic analysis of the TDNN estimator, the authors analyze its properties by considering the relevant Hoeffding decompositions. Borrowing the notational conventions from Lee (2019), we introduce the following notation.

$$\psi_c^s(\mathbf{x}; \mathbf{z}_1, \ldots, \mathbf{z}_c) = \mathbb{E}_{\mathbf{Z}} \left[\psi^s(\mathbf{x}; \mathbf{z}_1, \ldots, \mathbf{z}_c, \mathbf{Z}_{c+1}, \ldots, \mathbf{Z}_s) \right] \tag{2.15}$$

$$h_s^{(1)}(\mathbf{x}; \mathbf{z}_1) = \psi_1^s(\mathbf{x}; \mathbf{z}_1) - \mu(\mathbf{x}) \tag{2.16}$$

$$h_s^{(c)}(\mathbf{x}; \mathbf{z}_1, \ldots, \mathbf{z}_c) = \psi_c^s(\mathbf{x}; \mathbf{z}_1, \ldots, \mathbf{z}_c) - \sum_{j=1}^{c-1} \left(\sum_{\ell \in L_{n,j}} h_s^{(j)}(\mathbf{x}; \mathbf{z}_\ell) \right) - \mu(\mathbf{x}) \quad \text{for } c = 2, \ldots, s \tag{2.17}$$

In the usual fashion, these terms can be used to express the Hoeffding projections of different orders.

$$H_{n,s}^c = \binom{n}{c}^{-1} \sum_{\ell \in L_{n,c}} h_s^{(c)}(\mathbf{x}; \mathbf{z}_\ell) \tag{2.18}$$

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 a completely analogous fashion, we can define the corresponding terms for the TDNN estimator. In contrast to the DNN estimator, the kernel under consideration is of order s_2. The corresponding expressions will be denoted analogous to the terms above with "s_1, s_2" replacing "s". Thus, we find the following representations of the DNN and TDNN estimators using the Hoeffding decomposition.

$$D_n^s(\mathbf{x}) - \mu(\mathbf{x}) = \sum_{j=1}^{s} \binom{s}{j} H_{n,s}^j \quad \text{and} \quad D_n^{s_1,s_2}(\mathbf{x}) - \mu(\mathbf{x}) = \sum_{j=1}^{s_2} \binom{s}{j} H_{n,s_1,s_2}^j \tag{2.19}$$

Overall, Demirkaya et al. (2024) lay out a clear path to perform point-wise inference using their TDNN estimator.

3 Uniform Inference for TDNN Estimator

Absent from Demirkaya et al. (2024) is a way to construct uniformly valid confidence bands around the TDNN estimator. To consider this problem 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. Furthermore, going forward $\widehat{\mu}(\mathbf{x}; D)$ can be taken as the DNN or TDNN estimator evaluated at $\mathbf{x}$ based on the (sub-)sample D, respectively. 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 3.1 (*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

$$\widehat{\mathcal{C}}\left(\mathbf{x}^{(p)}\right) := \left\{\widehat{C}(\mathbf{x}_j^{(p)}) = \left[c_L(\mathbf{x}_j^{(p)}), c_U(\mathbf{x}_j^{(p)})\right] \,:\, j \in [p]\right\} \tag{3.1}$$

based on the observed data, such that

$$\sup_{P \in \mathbf{P}} \left| P\left(\mu(\mathbf{x}^{(d)}) \in \widehat{\mathcal{C}}\left(\mathbf{x}^{(d)}\right)\right)\right| \leq r_{n,d} \tag{3.2}$$

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

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 3.2 (*Half-Sample Bootstrap Root Approximation - Ritzwoller and Syrgkanis (2024)*)

The Half-Sample Bootstrap Root Approximation of the sampling distribution of the root

$$R\left(\mathbf{x}^{(p)}; \mathbf{D}_n\right) := \widehat{\mu}\left(\mathbf{x}^{(p)}; \mathbf{D}_n\right) - \mu(\mathbf{x}^{(p)}) \tag{3.3}$$

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

$$R^*\left(\mathbf{x}^{(p)};\mathbf{D}_n\right) := \widehat{\mu}\left(\mathbf{x}^{(p)};D_l\right) - \widehat{\mu}\left(\mathbf{x}^{(p)};\mathbf{D}_n\right) \tag{3.4}$$

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

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

$$\widehat{\lambda}_j^2\left(\mathbf{x}^{(p)};\mathbf{D}_n\right) = \text{Var}\left(\sqrt{n}R^*(\mathbf{x}_j^{(p)};\mathbf{D}_n)\,|\,\mathbf{D}_n\right) \quad \text{and} \quad \widehat{\Lambda}_n\left(\mathbf{x}^{(p)};\mathbf{D}_n\right) = \text{diag}\left(\left\{\widehat{\lambda}_j^2\left(\mathbf{x}^{(p)};\mathbf{D}_n\right)\right\}_{j=1}^{p}\right) \tag{3.5}$$

$$\widehat{S}^*\left(\mathbf{x}^{(p)};\mathbf{D}_n\right) := \sqrt{n}\left\|\left(\widehat{\Lambda}_n\left(\mathbf{x}^{(p)};\mathbf{D}_n\right)\right)^{-1/2} R^*\left(\mathbf{x}^{(p)};\mathbf{D}_n\right)\right\|_2 \tag{3.6}$$

Let $\text{cv}\left(\alpha;\mathbf{D}_n\right)$ 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 3.3 (*Uniform Confidence Region - Ritzwoller and Syrgkanis (2024)*)

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.

LOREM IPSUM

4 Simulations

4.1 Setups

4.2 Results

5 Application

6 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.

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