# Optimal Nonparametric Inference with Two-Scale Distributional Nearest Neighbors*

Emre Demirkaya[1], Yingying Fan[2], Lan Gao[1,2], Jinchi Lv[2],
Patrick Vossler[2] and Jingbo Wang[2]

University of Tennessee Knoxville[1] and University of Southern California[2]

June 16, 2022

**Abstract**

The weighted nearest neighbors (WNN) estimator has been popularly used as a flexible and easy-to-implement nonparametric tool for mean regression estimation. The bagging technique is an elegant way to form WNN estimators with weights automatically generated to the nearest neighbors (Steele, 2009; Biau et al., 2010); we name the resulting estimator as the distributional nearest neighbors (DNN) for easy reference. Yet, there is a lack of distributional results for such estimator, limiting its application to statistical inference. Moreover, when the mean regression function has higher-order smoothness, DNN does not achieve the optimal nonparametric convergence rate, mainly because of the bias issue. In this work, we provide an in-depth technical analysis of the DNN, based on which we suggest a bias reduction approach for the DNN estimator by linearly combining two DNN estimators with different subsampling scales, resulting in the novel two-scale DNN (TDNN) estimator. The two-scale DNN estimator has an equivalent representation of WNN with weights admitting explicit forms and some being negative. We prove that, thanks to the use of negative weights, the two-scale DNN estimator enjoys the optimal nonparametric rate of convergence in estimating the regression function under the fourth-order smoothness condition. We further go beyond estimation and establish that the DNN and two-scale DNN are both asymptotically normal as the subsampling scales and sample size diverge to infinity. For the practical implementation, we also provide variance estimators and a distribution estimator using the jackknife and bootstrap

*This work was partially supported by NIH Grant 1R01GM131407. Send correspondence to Yingying Fan (*fanyingy@usc.edu*) or Lan Gao (*lgao13@utk.edu*).

techniques for the two-scale DNN. These estimators can be exploited for constructing valid confidence intervals for nonparametric inference of the regression function. The theoretical results and appealing finite-sample performance of the suggested two-scale DNN method are illustrated with several simulation examples and a real data application.

*Key words*: Nonparametric estimation and inference; $k$-nearest neighbors; Weighted nearest neighbors; Two-scale distributional nearest neighbors; Bootstrap and jackknife; Bagging

# 1 Introduction

Nonparametric regression analysis is a popular and flexible statistical tool with broad applications in various scientific fields. Among the existing nonparametric regression methods, the $k$-nearest neighbors ($k$-NN) procedure and its extensions including the weighted nearest neighbors method, have received great popularity due to their straightforward implementation and appealing theoretical properties. For existing results and some recent developments along this direction, see, for example, Mack (1980); Györfi et al. (2002); Biau and Devroye (2015); Berrett et al. (2019); Lin et al. (2021).

Despite the advantage of the weighted nearest neighbors (WNN) method over the unweighted $k$-NN, selection of the adaptive weights can be challenging in implementation. To address such issue, the bagged 1-NN estimator, an ensemble learning method, has been proposed. Specifically, Steele (2009) and Biau et al. (2010) proposed to estimate the mean regression function by averaging all 1-NN estimators constructed from randomly subsampling $s$ observations with or without replacement, where $s$ is required to diverge with the total sample size $n$. Steele (2009) showed that this procedure automatically assigns monotonic nonnegative weights to the nearest neighbors in a distributional fashion on the entire sample, motivating us to name it as the distributional nearest neighbors (DNN) in our paper for easy presentation. The bagging technique was pioneered by the seminal work of Breiman (1996) and has been employed to improve performance of the base estimators. For

instance, see Hall and Samworth (2005) for the asymptotic properties of bagged nearest neighbor classifiers.

Biau et al. (2010) proved the nice results that DNN achieves the nonparametric minimax optimal convergence rate under the Lipschitz continuity assumption of the regression function. Yet, there is a lack of asymptotic distribution results for such estimator, limiting its application to statistical inference. In addition, when the mean regression function has higher order smoothness, the DNN estimator no longer achieves the nonparametric optimal rate. In this work, we discover through thorough investigations that the non-optimality is caused by the slow convergence rate due to the bias. For further bias reduction, we establish the higher-order asymptotic expansion for the bias of DNN. Based on such a bias expansion, we propose to eliminate the leading order bias of DNN by linearly combining two DNN estimators with different subsampling scales, resulting in the novel two-scale DNN procedure for nonparametric estimation and inference.

The DNN estimator has a representation of L-statistic with weights depending only on the rank of the observations (Steele, 2009), facilitating easy and fast implementation. However, such a representation does not help with establishing the sampling properties. For the theoretical analysis, we further demonstrate that DNN estimator has an equivalent representation of U-statistic with a kernel function of diverging dimensionality equal to the subsampling scale $s$, and therefore, the two-scale DNN estimator also has a U-statistic representation with a new and carefully constructed diverging-dimensional kernel. Despite the nice U-statistic representations, the classical theory does not apply to DNN or two-scale DNN for deriving their asymptotic properties because of the diverging dimensionality of the kernel functions. To overcome such a technical challenge, we exploit Hoeffding's canonical decomposition introduced in Hoeffding (1948), and carefully collect and analyze the higher-order terms in our decomposition. Our theoretical results suggest that, when the subsampling scales are appropriately chosen, two-scale DNN achieves the nonparametric

optimal rate under the fourth-order smoothness assumption on the regression function and the density function of covariates. A larger implication of our study is that, for regression function with even higher-order smoothness, the multi-scale DNN can be constructed in the same fashion to achieve the optimal nonparametric convergence rate; we leave the detailed investigation for future study.

By construction, some weights in the two-scale DNN take negative values. The advantage of using negative weights in the weighted nearest neighbors classifiers was formally investigated in Samworth (2012). For the problem of regression, although Biau and Devroye (2015) theoretically showed that the weighted nearest neighbors estimator allowing for negative weights can improve upon that with only nonnegative weights in terms of the rate of convergence, it still remains largely unclear how to practically choose these weights. Our two-scale DNN provides an explicit and easy-to-implement way to assign negative weights which endorses the optimal nonparametric convergence rate under the higher-order smoothness assumption of the regression function.

We further show that DNN and two-scale DNN are asymptotically normal as the subsampling scales and sample size $n$ diverge to infinity. The asymptotic variance of the two-scale DNN estimator, however, does not admit a simple analytic form that is practically useful for statistical inference. We exploit two methods, the jackknife and bootstrap, for asymptotic variance estimation. We formally demonstrate that both methods yield consistent estimates of the asymptotic variance. Our proofs are more intricate than the standard technique in the literature because of the diverging subsampling scales. The key is to write the jackknife estimator as a weighted summation of a sequence of U-statistics and carefully analyze the higher-order terms. Our proof for the bootstrap estimator is built on our results for the jackknife estimator. Although both methods yield consistent variance estimates, the bootstrap estimator is much more computationally efficient. We also provide a bootstrap method to directly estimate the distribution of the two-scale DNN

estimator without estimating the asymptotic variance.

We then demonstrate the superior finite-sample performance of our method using simulation studies and a real data application. The two-scale DNN estimator has two parameters to tune – the two subsampling scales – and it is equivalent to tune the ratio between the two subsampling scales and one of the subsampling scales. We propose to jointly tune these two parameters using a two-dimensional grid, and choose the combination of the two parameters that minimizes the mean-squared estimation error (MSE). As an application, we discuss the usage of the two-scale DNN for the heterogeneous treatment effect (HTE) estimation and inference with theoretical guarantee under the setting of randomized experiments in Section B in the Supplementary Material.

The rest of the paper is organized as follows. Section 2 introduces the model setting for nonparametric regression estimation and reviews the DNN estimator. We present the two-scale distributional nearest neighbors (TDNN) procedure and its sampling properties in Section 3. Section 4 investigates the variance estimation for the TDNN estimator. We provide several simulation examples and a real data application justifying our theoretical results and illustrating the finite-sample performance of the suggested TDNN method in Sections 5 and 6, respectively. Section 7 discusses some implications and extensions of our work. In the Supplementary Material, we also provide a bootstrap estimator for the distribution of TDNN estimator, the application of TDNN in HTE estimation and inference, and all the proofs and technical details.

## 2 Model Setting

Consider a sample of independent and identically distributed (i.i.d.) observations $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ from the following nonparametric model

$$Y = \mu(\mathbf{X}) + \epsilon, \tag{1}$$

where $Y$ is the response, $\mathbf{X} \in \mathbb{R}^d$ represents the vector of covariates with fixed dimensionality $d$, $\mu(\mathbf{X})$ is the unknown mean regression function, and $\epsilon$ is the model error. The goal is to estimate and infer the underlying true mean regression function $\mu(\mathbf{x})$ at some given feature vector $\mathbf{x}$ in the support of $\mathbf{X}$.

## 2.1 Distributional nearest neighbors (DNN)

Given a fixed feature vector $\mathbf{x} \in \mathbb{R}^d$, we calculate the Euclidean distance of each observed feature vector $\mathbf{X}_i$ to the target $\mathbf{x}$ and then reorder the sample according to such distances. Denote the reordered sample as $\{(\mathbf{X}_{(1)}, Y_{(1)}), \cdots, (\mathbf{X}_{(n)}, Y_{(n)})\}$ with

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

where $\|\cdot\|$ denotes the Euclidean norm of a given vector and the ties are broken by assigning the smallest rank to the observation with the smallest natural index. Then the weighted nearest neighbors (WNN) estimate (Mack, 1980) is defined as

$$\widehat{\mu}_{\mathrm{WNN}}(\mathbf{x}) = \sum_{i=1}^{n} w_{ni} Y_{(i)}, \tag{3}$$

where $(w_{n1}, w_{n2}, \cdots, w_{nn})$ is some deterministic weight vector with all the components summing up to one. In practice, one can also use the non-Euclidean distances given by certain manifold structures.

The theoretical properties of the WNN estimator (3) have been studied extensively in Biau and Devroye (2015). In particular, it has been proved therein that, with an appropriately selected nonnegative weight vector, $\widehat{\mu}_{\mathrm{WNN}}(\mathbf{x})$ can be consistent with the optimal rate of convergence $O_P(n^{-2/(d+4)})$ when the second-order derivative exists and can have asymptotic normality. Moreover, the optimal rate of convergence can be improved by allowing for negative weights under higher-order derivatives. These existing results provide only some general sufficient conditions on the weight vector $(w_{n1}, \cdots, w_{nn})$ in order to deliver

the theoretical properties. However, identifying a practical weight vector with provably appealing properties can be highly nontrivial. Furthermore, the asymptotic variance of $\widehat{\mu}_{\mathrm{WNN}}(\mathbf{x})$ can admit a rather complicated form and depend upon some *unknown* population quantities that are very difficult to estimate in practice, hindering the applicability in statistical inference.

In contrast, the bagged 1-NN estimator proposed and studied in Steele (2009) and Biau et al. (2010) (which we refer to as the DNN estimator in this paper for the ease of presentation) automatically assigns monotonic weights to the nearest neighbors in a distributional fashion on the entire sample. Denote by $s$ with $1 \leq s \leq n$ the subsampling scale. Let $\{i_1, \cdots, i_s\}$ with $i_1 < i_2 < \cdots < i_s$ be a random subset of the full sample $\{1, \cdots, n\}$. Hereafter, we use $\mathbf{Z}_i$ as a shorthand notation for $(\mathbf{X}_i, Y_i)$ with $1 \leq i \leq n$. Let us define $\Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s})$ as the 1-NN estimator

$$\Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s}) = Y_{(1)}(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s}) \tag{4}$$

for estimating the true value $\mu(\mathbf{x})$ of the underlying mean function at the fixed point $\mathbf{x}$ based on the given subsample $\{\mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_s}\}$. Then the DNN estimator $D_n(s)(\mathbf{x})$ with subsampling scale $s$ for estimating $\mu(\mathbf{x})$ is formally defined as a U-statistic

$$D_n(s)(\mathbf{x}) = \binom{n}{s}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_s \leq n} \Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s}), \tag{5}$$

where the kernel function $\Phi(\mathbf{x}; \cdot)$ is given in (4).

The above U-statistic representation averages over all 1-NN estimators given by all possible subsamples of size $s$. For the case of $s = 1$, the DNN estimator reduces to the simple sample average $n^{-1} \sum_{i=1}^{n} Y_i$, which admits reduced variance but inflated bias. In contrast, for the case of $s = n$, the DNN estimator reduces to the simple 1-NN estimator $Y_{(1)}$ based on the full sample of size $n$, which admits the lowest bias but inflated variance. See, e.g., Hoeffding (1948); Hájek (1968); Korolyuk and Borovskich (1994) for the classical

asymptotic theory of the U-statistics. Since the computation of general U-statistics becomes more challenging when sample size $n$ grows, the following lemma in Steele (2009) shows that a different representation of the DNN estimator can be exploited for easy computation.

**Lemma 1** (Steele (2009)). *The DNN estimator $D_n(s)(\mathbf{x})$ also admits an equivalent L-statistic (Serfling, 1980) representation as*

$$D_n(s)(\mathbf{x}) = \binom{n}{s}^{-1} \sum_{i=1}^{n-s+1} \binom{n-i}{s-1} Y_{(i)}, \tag{6}$$

*where $Y_{(i)}$'s are given by the full sample of size $n$.*

One nice property of the above DNN estimator is that the distribution of weights is characterized by only two parameters of the full sample size $n$ and the subsampling scale $s$. As shown later in Section 3.2, our new higher-order asymptotic expansion for the bias reveals that the distributional weights in the DNN yield the explicit constant for the leading bias term that is free of the subsampling scale $s$, which opens the door for eliminating the first-order asymptotic bias of the DNN.

# 3 Two-scale distributional nearest neighbors

## 3.1 Two-scale DNN

We are now ready to suggest a natural extension of the single-scale DNN procedure introduced in Section 2.1. The major motivation for this extension comes from the precise higher-order asymptotic bias expansion for the single-scale DNN estimator $D_n(s)(\mathbf{x})$ unveiled in Theorem 1 to be presented in Section 3.2. In particular, we see that the explicit constant for the leading order term in the asymptotic expansion for the bias $B(s) = \mathbb{E}\, D_n(s)(\mathbf{x}) - \mu(\mathbf{x})$ is independent of the subsampling scale $s$. Such an appealing property gives us an effective way to completely remove the first-order asymptotic bias

in the order of $s^{-2/d}$, making only the second-order asymptotic bias dominating at the finite-sample level.

To achieve the aforementioned goal, let us consider a pair of single-scale DNN estimators $D_n(s_1)(\mathbf{x})$ and $D_n(s_2)(\mathbf{x})$ with different subsampling scales $1 \leq s_1 < s_2 \leq n$ as constructed in (5). Then Theorem 1 ensures that

$$\mathbb{E}\, D_n(s_1)(\mathbf{x}) = \mu(\mathbf{x}) + c\, s_1^{-2/d} + R(s_1), \tag{7}$$

$$\mathbb{E}\, D_n(s_2)(\mathbf{x}) = \mu(\mathbf{x}) + c\, s_2^{-2/d} + R(s_2), \tag{8}$$

where $c$ is some positive constant depending on the underlying distributions, but not on the subsampling scale parameter $s_1$ or $s_2$, and the higher-order remainder is given by $R(s) = O(s^{-3})$ for $d = 1$ and $R(s) = O(s^{-4/d})$ for $d \geq 2$ .

Although the specific constant $c$ in the asymptotic expansions (7) and (8) above is unknown to us, we can proceed with solving the following system of linear equations with respect to $w_1$ and $w_2$

$$w_1 + w_2 = 1, \qquad w_1\, s_1^{-2/d} + w_2\, s_2^{-2/d} = 0,$$

whose solutions are given by the specific weights

$$w_1^* = w_1^*(s_1, s_2) = 1/(1 - (s_1/s_2)^{-2/d}) \tag{9}$$

$$\text{and} \quad w_2^* = w_2^*(s_1, s_2) = -(s_1/s_2)^{-2/d}/(1 - (s_1/s_2)^{-2/d}). \tag{10}$$

Then our two-scale distributional nearest neighbors (TDNN) estimator $D_n(s_1, s_2)(\mathbf{x})$ is formally defined as

$$D_n(s_1, s_2)(\mathbf{x}) = w_1^* D_n(s_1)(\mathbf{x}) + w_2^* D_n(s_2)(\mathbf{x}). \tag{11}$$

We will impose the restriction that $s_1/s_2$ is bounded away from both 0 and 1 by some positive constants. This can avoid the undesirable cases of weights being too close to 0 or having diverging magnitude as $s_1$ and $s_2$ diverge.

Since the specific weights $w_1^*$ and $w_2^*$ depend only on subsampling scales $s_1$ and $s_2$, we see from the asymptotic expansions (7) and (8) that

$$\mathbb{E}\, D_n(s_1, s_2)(\mathbf{x}) = \mu(\mathbf{x}) + R^*(s_1), \tag{12}$$

where $R^*(s_1) = O(s_1^{-4/d})$ for $d \geq 2$, $R^*(s_1) = O(s_1^{-3})$ for $d = 1$, and we impose the constraint that $s_1 \sim s_2$ with $\sim$ representing asymptotic equivalence. The removal of the first-order asymptotic bias as shown in (12) provides the TDNN estimator appealing finite-sample performance with reduced bias and controlled variance, as demonstrated with extensive simulation examples in Section 5.

It is worth mentioning that in view of (9) and (10), weight $w_1^*$ is negative given $s_1 < s_2$. This implies that the two-scale DNN can assign negative weights to some distant nearest neighbors. In fact, the advantage of using negative weights in the $k$-NN classifier for the classification setting was discovered earlier in Samworth (2012). See also the theoretical discussions in Biau and Devroye (2015) for similar advantages in the regression setting.

TDNN is a bias-corrected version of DNN. Bias reduction techniques have been commonly used in the literature for improved mean-squared error. For example, Hall (1992) and Schucany and Sommers (1977) discussed bias correction using the bootstrap estimator and jackknife estimator, respectively. Other works have specialized bias correction in different models; for instance, see Calonico et al. (2018) and Newey et al. (2004) in kernel density estimation, Cheang and Reinsel (2000) in time series models, and Leblanc (2010) in nonparametric density estimation based on the Bernstein polynomial approximations. In such an endeavor, one always needs to estimate the bias term or find its theoretical representation. The form of the bias depends on the estimator in question, and so the bias reduction techniques can differ. For example, Calonico et al. (2018) expressed the bias of kernel density estimator in terms of the bandwidth using the Edgeworth expansion in Hall (2013). Schucany et al. (1971) established a general bias reduction method using the ratio

of bias terms of two different estimators when those estimators have different but known bias terms. In our work, Theorem 1 enables us to decompose the bias of the DNN estimator in terms of the subsampling scale. By combining two DNN estimators with different subsampling scales, we are able to eliminate the first-order bias.

## 3.2 Accuracy and asymptotic distributions of two-scale DNN

We now turn to deriving the higher-order asymptotic expansions of the DNN and TDNN estimators and their asymptotic distributions. To this end, we need to impose some necessary assumptions, which are commonly used for nonparametric regression, to facilitate our technical analysis.

Assume that the distribution of $\mathbf{X}$ has a density function $f(\cdot)$ with respect to the Lebesgue measure $\lambda$ on the Euclidean space $\mathbb{R}^d$. Let $\mathbf{x} \in \text{supp}(\mathbf{X})$ be a fixed feature vector.

**Condition 1.** *There exists some constant $\alpha > 0$ such that $\mathbb{P}(\|\mathbf{X} - \mathbf{x}\| \geq R) \leq e^{-\alpha R}$ for each $R > 0$.*

**Condition 2.** *The density $f(\cdot)$ is bounded away from 0 and $\infty$, $f(\cdot)$ 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 $\epsilon$ has zero mean and finite variance $\sigma_\epsilon^2 > 0$, and is independent of $\mathbf{X}$.*

**Condition 3.** *We have an i.i.d. sample $\{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \cdots, (\mathbf{X}_n, Y_n)\}$ of size $n$ from model (1).*

We begin with presenting an asymptotic expansion of the bias of single-scale DNN estimator in the theorem below.

**Theorem 1.** *Assume that Conditions 1–3 hold and $s \to \infty$. Then for any fixed $\mathbf{x} \in \text{supp}(\mathbf{X}) \subset \mathbb{R}^d$, we have*

$$\mathbb{E}\, D_n(s)(\mathbf{x}) = \mu(\mathbf{x}) + B(s) \tag{13}$$

*with*

$$B(s) = \Gamma(2/d+1)\frac{f(\mathbf{x})\,\mathrm{tr}(\mu''(\mathbf{x})) + 2\,\mu'(\mathbf{x})^T f'(\mathbf{x})}{2\,d\,V_d^{2/d}\,f(\mathbf{x})^{1+2/d}}\,s^{-2/d} + R(s), \tag{14}$$

$$R(s) = \begin{cases} O(s^{-3}), & d = 1, \\ O(s^{-4/d}), & d \geq 2, \end{cases} \tag{15}$$

*where* $V_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$, $\Gamma(\cdot)$ *is the gamma function,* $f'(\cdot)$ *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, and* $\mathrm{tr}(\cdot)$ *stands for the trace of a given matrix.*

Theorem 1 above shows that the first-order asymptotic bias of the single-scale DNN estimator $D_n(s)(\mathbf{x})$ is of order $s^{-2/d}$, and the second-order asymptotic bias is of order $s^{-4/d}$ for $d \geq 2$ and of order $s^{-3}$ for $d = 1$. The rate of convergence for the bias term becomes slower as the feature dimensionality $d$ grows, which is common for nonparametric estimators. It thus would be beneficial to remove the first-order asymptotic bias completely to improve the finite-sample performance.

We relate our results to the existing literature. Biau et al. (2010) showed that DNN achieves the optimal convergence rate of $n^{1/(d+2)}$ under the Lipschitz continuity assumption on the regression function when $d \geq 3$. Our Theorem 1 is proved assuming the fourth-order smoothness condition (see Condition 2). Under the fourth-order smoothness condition, DNN does not achieve the nonparametric optimal rate of $n^{4/(d+8)}$, mainly because of the bias. Our results reveal that in such a case, bias reduction is needed for improved convergence rate. The fourth-order smoothness condition is mainly used to obtain the explicit form of the coefficient in front of the first-order bias $s^{-2/d}$ and the order of the remainder $R(s)$, which are critical for successful bias reduction and also play important roles in developing our asymptotic normality theory (which until now has been absent from literature). In addition, as mentioned in the Introduction, the de-biasing idea here can be similarly

applied by constructing the multi-scale DNN to further reduce the higher-order bias under even higher-order smoothness assumption.

**Corollary 1.** *Assume that Conditions 1–3 hold and $s \to \infty$. Then for any fixed $\mathbf{x} \in \text{supp}(\mathbf{X}) \subset \mathbb{R}^d$ and the two-scale DNN estimator with weights defined in (9)–(10), we have*

$$\mathbb{E}D_n(s_1, s_2)(\mathbf{x}) = \mu(\mathbf{x}) + R(s), \tag{16}$$

*where*

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

By using the TDNN estimator, the asymptotic bias reduces to the second-order term $O(s_1^{-4/d} + s_2^{-4/d})$ for $d \geq 2$ and $O(s_1^{-3} + s_2^{-3})$ for $d = 1$. Corollary 1 is a direct consequence of Theorem 1.

We further characterize the asymptotic distribution of the single-scale DNN estimator in the following theorem, which is new to the literature.

**Theorem 2.** *Assume that Conditions 1–3 hold, $s \to \infty$, and $s = o(n)$. Then for any fixed $\mathbf{x} \in \text{supp}(\mathbf{X}) \subset \mathbb{R}^d$, it holds that for some positive sequence $\sigma_n$ of order $(s/n)^{1/2}$,*

$$\frac{D_n(s)(\mathbf{x}) - \mu(\mathbf{x}) - B(s)}{\sigma_n} \xrightarrow{\mathscr{D}} N(0, 1) \tag{17}$$

*as $n \to \infty$, where $B(s)$ is given in (14).*

Theorem 2 requires the assumptions of $s \to \infty$ and $s = o(n)$, where the former leads to vanishing bias and the latter leads to controlled variance asymptotically. The technical analysis of Theorem 2 exploits Hoeffding's canonical decomposition (Hoeffding, 1948) which is an extension of the Hájek projection.

Despite the U-statistic representation of $D_n(s)(\mathbf{x})$ given in (5), the classical U-statistic asymptotic theory (e.g., Serfling (1980); Korolyuk and Borovskich (1994)) is not readily

applicable because of the typical assumption of *fixed* subsampling scale $s$. In contrast, our method requires the opposite assumption of *diverging* subsampling scale $s$. Such a statistic is called an infinite-order U-statistic (IOUS) and has gained more interest in the recent literature; see, e.g., Borovskikh (1996); Frees (1989); Song et al. (2019); Athey et al. (2019). Unfortunately, the assumptions on the kernel functions of the U-statistics in most IOUS literature are not satisfied for the TDNN. For instance, Frees (1989) assumed that the kernels are converging as the sample size grows. However, in our case, the kernel $\Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s}) = Y_{(1)}(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s})$ becomes degenerate as $s$ tends to infinity. Another example is Borovskikh (1996) who considered scalar-valued random variables. In our case, $\mathbf{Z}_i$'s are vector-valued and thereby the results of Borovskikh (1996) are not readily applicable.

In Wager and Athey (2018), the asymptotic distribution of the random forests (Breiman, 2001, 2002; Chi et al., 2020) estimator was studied via examining the asymptotic normality of the IOUS. Both their proof and ours rely on Hoeffding's decomposition of the U-statistics (and in particular IOUS) to establish the asymptotic normality. However, the main challenge in these proofs is controlling the variance of the first-order Hájek projection. This variance term takes different forms for nearest neighbors methods and tree based methods, and thus it needs to be handled differently for each case. For instance, Theorem 3.3 and Corollary 3 in Wager and Athey (2018) demonstrate bounds for tree based methods, which are not directly extendable to the nearest neighbors methods. Instead, we use Lemma 7 in Section E.6 of the Supplementary Material to bound variance specifically for our method.

Recently, Song et al. (2019) established convergence theory similar to our Theorem 2 under more general setting and more complicated assumptions which also concern the kernel of the U-statistics and the Hájek projection of the kernel. In contrast, our Theorem 2 is developed under simpler assumptions that are more targeted to the TDNN. It might be possible to check the conditions and then employ the results of Song et al. (2019) to

prove our Theorem 2. However, the efforts on checking these assumptions can be rather significant and even comparable to the full development of our proof.

We proceed with characterizing the asymptotic distribution for the two-scale DNN estimator introduced in (11).

**Theorem 3.** *Assume that Conditions 1–3 hold,* $s_2 \to \infty$, $s_2 = o(n)$, *and there exist some constants* $0 < c_1 < c_2 < 1$ *such that* $c_1 \leq s_1/s_2 \leq c_2$. *Then for any fixed* $\mathbf{x} \in \text{supp}(\mathbf{X}) \subset \mathbb{R}^d$, *it holds that for some positive sequence* $\sigma_n$ *of order* $(s_2/n)^{1/2}$,

$$\frac{D_n(s_1, s_2)(\mathbf{x}) - \mu(\mathbf{x}) - \Lambda}{\sigma_n} \xrightarrow{\mathscr{D}} N(0, 1) \tag{18}$$

*as* $n \to \infty$, *where* $\Lambda = O(s_1^{-4/d} + s_2^{-4/d})$ *for* $d \geq 2$ *and* $\Lambda = O(s_1^{-3} + s_2^{-3})$ *for* $d = 1$.

We note that the positive sequence $\sigma_n$ in Theorem 3 is different from the sequence $\sigma_n$ in Theorem 2, with the former representing the asymptotic standard deviation of the TDNN estimator and the latter representing the asymptotic standard deviation of the single-scale DNN estimator. We use the same generic notation for the convenience of technical presentation. Since the explicit form of the asymptotic standard deviation will not be used, this should not cause any confusion. Theorem 3 requires both subsampling scales $s_1$ and $s_2$ to diverge and be of smaller orders of the full sample size $n$ in order to best trade off between the squared bias and variance. We would like to point out that Theorem 3 is not a simple consequence of Theorem 2, since marginal asymptotic normalities do not necessarily entail joint asymptotic normality. To deal with such a technical difficulty, we have to jointly analyze the two single-scale DNN estimators. A key ingredient of our technical analysis of Theorem 3 is to show that the TDNN estimator also admits a U-statistic representation, which enables us to exploit Hoeffding's decomposition and calculate the variances of the kernel and the associated first-order Hájek projection.

We also obtain the theorem below on the mean-squared error (MSE) of our TDNN

estimator. Setting $c = (s_1/s_2)^{2/d}$, the weights of the two single-scale DNN estimators are given by $w_1^* = c/(c-1)$ and $w_2^* = -1/(c-1)$ according to (9) and (10).

**Theorem 4.** *Assume that Conditions 1–3 hold, $s_2 \to \infty$, $s_2 = o(n)$, and $c$ is a constant in $(0,1)$. Then for any fixed $\mathbf{x} \in \mathrm{supp}(\mathbf{X}) \subset \mathbb{R}^d$, we have that when $d \geq 2$,*

$$\begin{aligned} &\mathbb{E}\big\{D_n(s_1,s_2)(\mathbf{x}) - \mu(\mathbf{x})\big\}^2 \\ &\leq \frac{A}{(c-1)^2}\Big\{R_1(\mathbf{x},d,f,\mu)c^{-2}s_2^{-8/d} + \sigma_\epsilon^2\frac{s_2}{n}\Big\}, \end{aligned} \tag{19}$$

*and when $d = 1$,*

$$\begin{aligned} &\mathbb{E}\big\{D_n(s_1,s_2)(\mathbf{x}) - \mu(\mathbf{x})\big\}^2 \\ &\leq \frac{A}{(c-1)^2}\Big\{R_2(\mathbf{x},d,f,\mu)c^{-1}s_2^{-6} + \sigma_\epsilon^2\frac{s_2}{n}\Big\}, \end{aligned} \tag{20}$$

*where $A$ is some positive constant, and $R_1(\mathbf{x},d,f,\mu)$ and $R_2(\mathbf{x},d,f,\mu)$ are some constants depending on the bounds of the first four derivatives of $f(\cdot)$ and $\mu(\cdot)$ in a neighborhood of $\mathbf{x}$.*

Theorem 4 provides an upper bound for the pointwise MSE and such result can be applied easily to obtain the integrated MSE under some regularity conditions. The optimal choice of subsampling scale $s_2$ in terms of achieving the best bias-variance tradeoff is given by $s_2 = O(n^{d/(8+d)})$ for $d \geq 2$ and $s_2 = O(n^{1/7})$ for $d = 1$, yielding the corresponding consistency rate at the order of $O(n^{-4/(8+d)})$ for $d \geq 2$ and $O(n^{-3/7})$ for $d = 1$. Note that such rate of convergence is minimax optimal (see, e.g., Stone (1982)) when $d \geq 2$ under the smoothness assumptions in Condition 2 . Compared to the result in Biau and Devroye (2015) where the minimax optimal convergence rate for the single-scale DNN was obtained for $d \geq 3$ under the Liptchitz continuity condition, our result still remains minimax optimal when $d \geq 2$ under different smoothness assumptions in Condition 2.

# 4 Variance estimates for two-scale DNN estimator

## 4.1 Jackknife estimator

As unveiled in Lemma 8 in Section E.7 of the Supplementary Material, the two-scale DNN estimator $D_n(s_1, s_2)(\mathbf{x})$ with $s_1 < s_2$ admits the U-statistic representation

$$D_n(s_1, s_2)(\mathbf{x}) = \binom{n}{s_2}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_{s_2} \le n} \Phi^*(\mathbf{x}; \mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_{s_2}}), \tag{21}$$

where the new kernel function is given by

$$\Phi^*(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{s_2}) = w_1^* \Phi^{(1)}(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{s_2}) + w_2^* \Phi(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{s_2})$$

with $\Phi^{(1)}(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{s_2}) = \binom{s_2}{s_1}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_{s_1} \le s_2} \Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_{s_1}})$, $\Phi(\mathbf{x}; \cdot)$ the original kernel function involved in the single-scale DNN estimator introduced in (5), and $w_1^*$ and $w_2^*$ the weights defined in equations (9) and (10). We denote by

$$\sigma_n^2 = \mathrm{Var}(D_n(s_1, s_2)(\mathbf{x})) \tag{22}$$

the variance of the two-scale DNN estimator $D_n(s_1, s_2)(\mathbf{x})$, where we drop the subscript $n$ in this population variance for notational simplicity.

For each $1 \le i \le n$, let us define the two-scale DNN estimator obtained after deleting the $i$th observation as in (21)

$$U_{n-1}^{(i)} = \binom{n-1}{s_2}^{-1} \sum_{\substack{1 \le j_1 < j_2 < \cdots < j_{s_2} \le n \\ j_1, j_2, \cdots, j_{s_2} \ne i}} \Phi^*(\mathbf{x}; \mathbf{Z}_{j_1}, \mathbf{Z}_{j_2}, \cdots, \mathbf{Z}_{j_{s_2}}). \tag{23}$$

Then the jackknife estimator (Quenouille, 1949, 1956) for $\sigma_n^2$ in (22) is given by

$$\widehat{\sigma}_J^2 = \frac{n-1}{n} \sum_{i=1}^{n} \left( U_{n-1}^{(i)} - D_n(s_1, s_2)(\mathbf{x}) \right)^2. \tag{24}$$

We formally establish the ratio consistency of the jackknife estimator $\widehat{\sigma}_J^2$ introduced in (24) in the theorem below.

**Theorem 5.** *Assume that Conditions 2–3 hold,* $\mathbb{E}[Y^4] < \infty$, $\mathbb{E}[\epsilon^4] < \infty$, $s_1 \to \infty$, *and* $s_2 \to \infty$ *with some constants* $0 < c_1 < c_2 < 1$ *such that* $c_1 \leq s_1/s_2 \leq c_2$. *Then for any fixed* $\mathbf{x} \in \text{supp}(\mathbf{X}) \subset \mathbb{R}^d$, *when* $s_2 = o(n^{1/3})$ *it holds that* $\widehat{\sigma}_J^2/\sigma_n^2 \xrightarrow{p} 1$ *as* $n \to \infty$.

The proof of Theorem 5 still builds on the U-statistic framework. Similar to the discussion after Theorem 2, the conventional technical arguments in Arvesen (1969) for the consistency of the jackknife estimator for the U-statistic are not applicable because of the diverging $s_1$ and $s_2$. As seen in Section D.5 of Supplementary Material, our technical analysis involves rather delicate calculations of the remainders. We acknowledge that the assumption of $s_2 = o(n^{1/3})$ is not necessarily optimal. Moreover, the assumption on the finite fourth moments can be relaxed to finite $(2+2\delta)$th moments with some $0 < \delta < 1$. Consequently, the bound on the order of $s_2$ will depend on parameter $\delta$ accordingly.

We point out that although the U-statistic representation plays a crucial role in obtaining our theoretical results, the computational cost of the jackknife estimator utilizing such a representation can become excessively prohibitive in practice. Instead, we should take advantage of the L-statistic representation revealed in Lemma 1 to efficiently compute the U-statistics $\{U_{n-1}^{(i)}\}_{1\leq i\leq n}$ and the two-scale DNN estimator $D_n(s_1, s_2)(\mathbf{x})$ involved in the jackknife estimator $\widehat{\sigma}_J^2$ in (24). When the sample size $n$ becomes large, one can speed up the implementation of jackknife using approximation with subsampling.

## 4.2 Bootstrap estimator

The bootstrap method (Efron, 1979) has been widely used for estimating the parameters and the distributions of statistics of interest, empowering statistical inference. We now consider the nonparametric bootstrap for estimating the variance of the two-scale DNN estimator. Given $n$ observations $\{\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n\}$, we denote by $\{\mathbf{Z}_1^*, \mathbf{Z}_2^*, \cdots, \mathbf{Z}_n^*\}$ a bootstrap sample selected independently and uniformly from the original $n$ observations with

replacement. As in (21), let us construct the two-scale DNN estimator

$$D_n^*(s_1,s_2)(\mathbf{x}) = \binom{n}{s_2}^{-1} \sum_{1\le i_1<i_2<\cdots<i_{s_2}\le n} \Phi^*(\mathbf{x};\mathbf{Z}_{i_1}^*,\mathbf{Z}_{i_2}^*,\cdots,\mathbf{Z}_{i_{s_2}}^*) \tag{25}$$

based on the bootstrap sample $\{\mathbf{Z}_1^*,\mathbf{Z}_2^*,\cdots,\mathbf{Z}_n^*\}$.

We choose the number of bootstrap samples as $B \ge 1$. For each $1 \le b \le B$, we independently select a bootstrap sample $\{\mathbf{Z}_{b,1}^*,\mathbf{Z}_{b,2}^*,\cdots,\mathbf{Z}_{b,n}^*\}$ and calculate the corresponding bootstrap version of the two-scale DNN estimator $D_n^{(b)}(s_1,s_2)(\mathbf{x})$ as in (25). Observe that given the original observations $(\mathbf{Z}_1,\mathbf{Z}_2,\cdots,\mathbf{Z}_n)$, the bootstrap samples $\{(\mathbf{Z}_{b,1}^*,\mathbf{Z}_{b,2}^*,\cdots,\mathbf{Z}_{b,n}^*)\}_{1\le b\le B}$ are independently and identically distributed as $(\mathbf{Z}_1^*,\mathbf{Z}_2^*,\cdots,\mathbf{Z}_n^*)$. Then the bootstrap estimator for $\sigma_n^2$ in (22) is given by

$$\widehat{\sigma}_{B,n}^2 = \frac{1}{B-1}\sum_{b=1}^{B}\left(D_n^{(b)}(s_1,s_2)(\mathbf{x}) - \bar{D}_{B,n}\right)^2, \tag{26}$$

where $\bar{D}_{B,n} = \frac{1}{B}\sum_{b=1}^{B} D_n^{(b)}(s_1,s_2)(\mathbf{x})$. The ratio consistency of the bootstrap estimator $\widehat{\sigma}_{B,n}^2$ introduced in (26) is shown formally in the following theorem.

**Theorem 6.** *Assume that Conditions 2–3 hold,* $\mathbb{E}[Y^4] < \infty$, $\mathbb{E}[\epsilon^4] < \infty$, $s_1 \to \infty$, *and* $s_2 \to \infty$ *with some constants* $0 < c_1 < c_2 < 1$ *such that* $c_1 \le s_1/s_2 \le c_2$. *Then for any fixed* $\mathbf{x} \in \mathrm{supp}(\mathbf{X}) \subset \mathbb{R}^d$, *when* $s_2 = o(n^{1/3})$ *and* $B \to \infty$, *it holds that* $\widehat{\sigma}_{B,n}^2/\sigma_n^2 \xrightarrow{p} 1$ *as* $n \to \infty$.

Let us gain some insights into the technical analysis for the consistency of the bootstrap estimator established in Theorem 6. First, we observe that conditional on $(\mathbf{Z}_1,\mathbf{Z}_2,\cdots,\mathbf{Z}_n)$, the bootstrap versions of the TDNN estimator $D_n^{(b)}(s_1,s_2)(\mathbf{x})$ are i.i.d. random variables and thus the law of large numbers entails that $\widehat{\sigma}_{B,n}^2$ is asymptotically close to the conditional variance $\mathrm{Var}(D_n^*(s_1,s_2)(\mathbf{x})|\mathbf{Z}_1,\mathbf{Z}_2,\cdots,\mathbf{Z}_n)$ as $B \to \infty$. Second, since the bootstrap samples are independently drawn from the empirical distribution based on $(\mathbf{Z}_1,\mathbf{Z}_2,\cdots,\mathbf{Z}_n)$ and the empirical distribution converges to the underlying distribution of $\mathbf{Z}$ asymptotically, the bootstrap version $\mathrm{Var}(D_n^*(s_1,s_2)(\mathbf{x})|\mathbf{Z}_1,\mathbf{Z}_2,\cdots,\mathbf{Z}_n)$ of the variance will converge to the

population quantity $\sigma^2$ as $n \to \infty$. It is worth mentioning that for the second part of our technical analysis, we resort to the consistency result of the jackknife estimator established in Theorem 5. In particular, we see that the jackknife and the bootstrap are asymptotically equivalent in the variance estimation for the TDNN estimator. Indeed, Efron (1979) showed that the jackknife can be viewed as a linear approximation method for the bootstrap. It was also pointed out in Efron (1979) that the jackknife can fail for certain nonsmooth functionals, while the bootstrap can still work.

# 5 Simulation studies

In this section, we investigate the finite-sample performance of the TDNN estimator for nonparametric estimation and inference in comparison to the DNN and $k$-NN. We use equations (9) and (10) to construct weights for the TDNN estimator. Specifically, we choose $w_1^* = -1/(c^{2/d} - 1)$ and $w_2^* = c^{2/d}/(c^{2/d} - 1)$ with $c = s_2/s_1$. Without loss of generality, we set $c > 1$. It is seen that $w_1^* < 0$ and hence the TDNN estimator assigns negative weights to some nearest neighbors which manages to reduce the bias for DNN. However, to control the variance of TDNN, the ratio $c = s_2/s_1$ should be chosen appropriately away from one.

With the above choice of weights, there are two parameters to tune for the TDNN: subsampling scale $s_1$ and the ratio $c = s_2/s_1$. To tune the parameters for prediction of a given feature vector $\mathbf{x}$, we perform a weighted leave-one-out cross-validation (LOOCV) procedure using each of the $B$ nearest neighbors to $\mathbf{x}$ as a single left-out observation. Specifically, we set aside each of the $B$ nearest neighbors to $\mathbf{x}$ and make prediction for it using the TDNN estimator with all the remaining $n - 1$ observations and the given combination $(c, s_1)$. Then the tuned $(c, s_1)$ is obtained by minimizing a weighted sum of the squared error over those $B$ left-out nearest neighbors, where the weights are defined by the standard Gaussian kernel distances of the nearest neighbors to the given feature vector

$\mathbf{x}$. Finally, we calculate our TDNN estimate $D_n(s_1, cs_1)(\mathbf{x})$ for the given point $\mathbf{x}$ using the $s_1$ and $c$ selected by our weighted LOOCV tuning procedure.

In our analysis, we always select the ratio $c$ from a set of values. Then for a given value of $c$, we provide our choices of the subsampling scale $s_1$ through a sign-change tuning method. Specifically, for the prediction of a feature vector $\mathbf{x}$, we compute the TDNN estimator $D_n(s_1, cs_1)(\mathbf{x})$ for each consecutive $s_1$ starting from 1. We continue this process until the difference in the absolute differences of consecutive TDNN estimators changes sign. Intuitively, the sign change represents the value of $s_1$ where the curvature of the TDNN estimator as a function of $s_1$ changes. We denote the subsampling scale chosen by the sign-change tuning process as $s_{\text{sign}}$. This process is motivated by the curve structure in Figure 1 from the simulation example in Section 5.1. One issue with the simple sign-change tuning method is that we may risk selecting a value of $s_1$ that corresponds to a local minimum for the MSE of TDNN as a function of $s_1$. To mitigate such concern, we consider a sequence of subsampling scales in the next step of our tuning process with $s_{\text{sign}}$ as our lower limit and $2s_{\text{sign}}$ as our upper limit, where the initial value $s_{\text{sign}}$ given by the sign-change tuning method provides a warm start for and specifies the order of tuning parameter $s_1$.

## 5.1 Two-scale DNN versus DNN

To illustrate the effectiveness of the two-scale framework compared to the single-scale DNN, we simulate $n = 1000$ data points from the following model.

**Setting 1.** *Assume that* $Y = \mu(\mathbf{X}) + \epsilon$, *where* $\mu(\mathbf{X}) = (x_1 - 1)^2 + (x_2 + 1)^3 - 3x_3$ *with* $\mathbf{X} = (x_1, x_2, x_3)^T$ *and* $(\mathbf{X}^T, \epsilon)^T \sim N(\mathbf{0}, I_4)$.

Our goal here is to compare the mean-squared error (MSE) of the TDNN estimator with those of the DNN and $k$-NN estimators at a fixed test point chosen to be $(0.5, -0.5, 0.5)^T$.

For the implementation of the DNN, we estimate the regression function at this test point and calculate the MSE while varying the subsampling scale $s$ from 1 to 250. For the TDNN, we estimate the regression function with fixed $c = 2$ for simplicity and $s_1$ varying from 1 to 250.

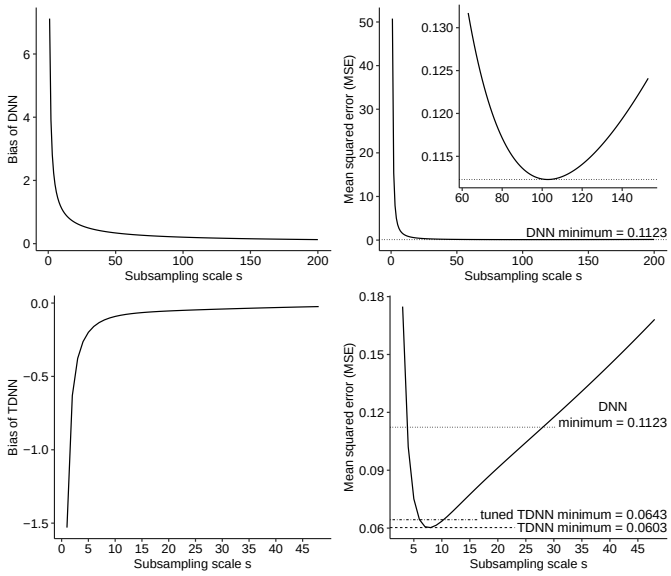


Figure 1: The results of simulation setting 1 described in Section 5.1 for DNN and TDNN. The rows show the bias and MSE as functions of the subsampling scale $s$ for DNN and TDNN, respectively. The top right panel also depicts a zoomed-in plot where the U-shaped pattern is more apparent. The dashed lines in the MSE plots are labeled with the minimum MSE value for each of the methods. The tuned TDNN MSE minimum corresponds to the weighted LOOCV tuning method described at the beginning of Section 5.

Figure 1 presents the simulation results for DNN and TDNN in terms of both the bias and the MSE. A first observation is that as the subsampling scale $s$ increases, the bias of the DNN estimator shrinks toward zero, which is intuitive from a geometric perspective since larger subsampling scale $s$ leads to the use of the information in the sample more concen-

trated around the fixed test point. From the MSE plot for DNN, we observe the classical U-shaped pattern of the bias-variance tradeoff. Thanks to the higher-order asymptotic expansions, the two-scale procedure of TDNN is completely free of the first-order asymptotic bias. The substantial difference between the dominating first-order asymptotic bias in DNN and the second-order asymptotic bias in TDNN at the finite-sample level is evident in the left panel of Figure 1.

From the MSE plot for TDNN, we also see a similar bias-variance tradeoff. An interesting phenomenon by comparing the two smooth U-shaped curves in the right panel of Figure 1 is that the minimum of the MSE for TDNN is attained at a much smaller subsampling scale $s$ than that for DNN. Furthermore, we observe that because of the reduced finite-sample bias, TDNN attains a more than 45% reduction of minimum MSE compared to the single-scale DNN. We also show in the bottom right panel of Figure 1 the MSE obtained by our weighted LOOCV tuning procedure for TDNN described at the beginning of Section 5, *without* using any knowledge of the underlying true regression function. We see that our tuning procedure provides a good approximation to the true MSE despite considering a smaller range of subsampling scales in a data-adaptive way. Finally, an additional comparison of TDNN and $k$-NN is included in Section C.1 of the Supplementary Material.

## 5.2 Comparisons with DNN and $k$-NN for nonparametric inference

We further compare TDNN with DNN and $k$-NN over two simulation examples in terms of the estimation accuracy in nonparametric regression settings.

For each simulation setting, we use a training sample size of $n = 1000$ and the summary statistics are calculated based on 1000 simulation replications. Throughout our simulations, we estimate the variance of the TDNN estimator and DNN estimator using the bootstrap

method that has been theoretically justified in Section 4.2. As for the inference by the $k$-NN estimator, we adopt the modeling strategy in Wager and Athey (2018) and model $\widehat{\mu}_{kNN}$ as Gaussian with mean $\mu(\mathbf{x})$ and variance $\widehat{\sigma}^2_{kNN}/(k-1)$, where $\widehat{\sigma}^2_{kNN}$ is the sample variance over the $k$ nearest neighbors. We tune our TDNN estimator using the weighted LOOCV tuning method by leaving out each of the $B$ nearest neighbors of a given feature vector $\mathbf{x}$ to predict, which has been described at the beginning of Section 5.1. We also adopt the same weighted LOOCV tuning strategy for the DNN estimator. We employ the `kknn` R package (Hechenbichler and Schliep, 2004) to tune the neighborhood size $k$ for the $k$-NN estimator using the leave-one-out cross-validation. In our simulation studies, $B$ for the weighted LOOCV tuning procedure is always chosen as 20, the subsampling scales $s$ for DNN varies from 1 to 250, and the neighborhood size $k$ for $k$-NN varies from 1 to 200.

The first simulation setting in this section also uses Setting 1 described in Section 5.1. We evaluate the performance of TDNN, DNN, and $k$-NN in terms of the bias, variance, and MSE at a fixed test point $(0.5, -0.5, 0.5)^T$ as well as for a set of 100 random test points drawn from the distribution of the covariates $\mathbf{X} \sim N(\mathbf{0}, I_3)$. The MSE, bias, and variance for the set of random test points are obtained by averaging over all the random test points. For the TDNN estimator, the ratio $c = s_2/s_1$ is chosen from the sequence $\{2, 4, 6, 8, 10, 15, 20, 25, 30\}$ for the random test points and we fix $c = 2$ for the fixed test point for simplicity. The subsampling scale $s_1$ is chosen from the interval $[s_{\text{sign}}, 2s_{\text{sign}}]$ for each given $c$, where $s_{\text{sign}}$ is given by the sign-change tuning process (related to the curvature) introduced at the beginning of Section 5.

We observe from Table 1 that for both fixed test point and random test points, the TDNN estimator significantly outperforms the DNN and $k$-NN estimators in terms of MSE. In addition, the improvement over the DNN is mainly due to the largely reduced bias, which is in line with our theory. In contrast, the TDNN has reduced variance compared to the $k$-NN, because TDNN is a bagged statistic and the bagging technique is known to be

| | Fixed Test Point | | | Random Test Points | | |
|---|---|---|---|---|---|---|
| Method | MSE | $\text{Bias}^2$ | Variance | MSE | $\text{Bias}^2$ | Variance |
| DNN | 0.1249 | 0.0556 | 0.0623 | 15.0989 | 14.4701 | 0.5968 |
| $k$-NN | 0.3207 | 0.0062 | 0.3114 | 9.4558 | 6.8510 | 2.2138 |
| TDNN | 0.0576 | 0.0082 | 0.0464 | 7.3142 | 5.3296 | 1.5235 |

Table 1: Comparison of DNN, $k$-NN, and TDNN in simulation setting 1 described in Section 5.1.

| | | Fixed Test Point | | | Random Test Points | | |
|---|---|---|---|---|---|---|---|
| Method | p | MSE | $\text{Bias}^2$ | Variance | MSE | $\text{Bias}^2$ | Variance |
| DNN | 3 | 0.7266 | 0.5159 | 0.2379 | 5.6775 | 4.5131 | 1.2315 |
| $k$-NN | 3 | 0.6630 | 0.1453 | 0.5158 | 5.1323 | 2.2046 | 2.8698 |
| TDNN | 3 | 0.2594 | 0.0535 | 0.2707 | 3.4788 | 1.6190 | 2.0933 |
| DNN | 5 | 0.7271 | 0.5176 | 0.2390 | 5.9057 | 4.6419 | 1.2638 |
| $k$-NN | 5 | 0.6699 | 0.1664 | 0.5272 | 5.3712 | 2.3072 | 2.9450 |
| TDNN | 5 | 0.2579 | 0.0699 | 0.2789 | 3.6883 | 1.6990 | 2.1640 |
| DNN | 10 | 0.8756 | 0.5822 | 0.2602 | 6.2705 | 4.8632 | 1.3218 |
| $k$-NN | 10 | 0.8297 | 0.1992 | 0.5643 | 5.7003 | 2.4493 | 3.0715 |
| TDNN | 10 | 0.2867 | 0.0503 | 0.2987 | 3.9243 | 1.7798 | 2.3084 |
| DNN | 15 | 0.9376 | 0.6434 | 0.2685 | 6.4583 | 4.9885 | 1.3466 |
| $k$-NN | 15 | 0.7919 | 0.2213 | 0.5693 | 5.8418 | 2.5189 | 3.1275 |
| TDNN | 15 | 0.2823 | 0.0439 | 0.3083 | 4.0509 | 1.8257 | 2.3743 |
| DNN | 20 | 0.9653 | 0.6341 | 0.2735 | 6.8909 | 5.2649 | 1.4073 |
| $k$-NN | 20 | 0.8174 | 0.1868 | 0.5729 | 6.2427 | 2.6994 | 3.2703 |
| TDNN | 20 | 0.3298 | 0.0530 | 0.3276 | 4.4064 | 1.9583 | 2.5184 |

Table 2: Comparison of DNN, $k$-NN, and TDNN in simulation setting 2 described in Section 5.2.

successful in variance reduction. We can see that the average MSE over a set of random test points is much larger than the MSE at the fixed test point (0.5, -0.5, 0.5). The main reason is that the covariate vector $\mathbf{X}$ is generated from a normal distribution and the density function at extreme values is close to zero, and thus the theoretical MSE can be very large for those extreme points. As a comparison, we also present the simulating results under the same setting except that $\mathbf{X} \sim U([0,1]^3)$ in Section C.2 of the Supplementary Material. It is seen from Table 4 in Section C.2 of Supplementary Material that under the uniform distribution setting, the MSE for random test points is only slightly larger than the MSE at the fixed test point.

For the second simulation setting, we investigate the performance of TDNN, DNN, and $k$-NN in the setting below, which is a modified version of a simulation setting first considered in Dette and Pepelyshev (2010).

**Setting 2.** *Assume that* $Y = \mu(\mathbf{X}) + \epsilon$, *where* $\mu(\mathbf{X}) = 4(4x_1 - 2 + 8x_2^2)^2 + (3 - 4x_2)^2 + 16\sqrt{x_3 + 1}(2x_3 - 1)^2$ *with* $\mathbf{X} = (x_1, \cdots, x_p)^T$, $\mathbf{X} \sim U([0,1]^p)$, *and* $\epsilon \sim N(\mathbf{0}, 1)$ *independent of* $\mathbf{X}$. *We increase the ambient dimensionality* $p$ *along the sequence* $\{3, 5, 10, 15, 20\}$.

Since the theoretical properties of TDNN established in this paper rely on the assumption of fixed dimensionality, it is natural to expect that the performance of TDNN can deteriorate as the dimensionality grows. To alleviate such difficulty, we exploit the feature screening idea (Fan and Lv, 2008; Fan and Fan, 2008; Fan and Lv, 2018) for dimension reduction to accompany the implementation of TDNN. For the screening step, we test the null hypothesis of independence between the response and each feature using the nonparametric tool of distance correlation statistic (Székely et al., 2007; Gao et al., 2021) and calculate the corresponding p-value. Then we select features with p-values less than $\alpha/p$ with some significance level $\alpha \in (0, 1)$ and make prediction by using these selected features. For our simulation studies, we fix $\alpha = 0.001$. For the TDNN estimator, the ratio $c = s_2/s_1$

is chosen from the sequence $\{2, 4, 6, 8, 10, 15, 20, 25, 30\}$ for random test points and we fix $c = 2$ for the fixed test point for simplicity. The subsampling scale $s_1$ is chosen from the interval $[s_{\text{sign}}, 2s_{\text{sign}}]$ for each given $c$, where $s_{\text{sign}}$ is given by the sign-change tuning process introduced at the beginning of Section 5.

We again evaluate the performance of the three estimators at a fixed test point chosen as $x_1 = 0.2$, $x_2 = 0.4$, $x_3 = 0.6$, and $x_j = 0.5$ for $j > 3$ as well as for a set of 100 test points randomly drawn from the hypercube $[0, 1]^p$. The simulation results in Table 2 show that the screening technique works well and the TDNN estimator has significantly reduced MSEs compared to the single-scale DNN and $k$-NN estimators. Observe that although the density function of the covariates is uniform, the average MSE for random test points is larger than the MSE for the fixed test point because the MSE also depends on the values of the regression function and its derivatives.

# 6 Real data application

In this section, we demonstrate the practical performance of the suggested TDNN procedure for nonparametric learning on the Abalone data set, which is available at the UCI repository (https://archive.ics.uci.edu/ml/datasets/abalone). The Abalone data set has been widely investigated in the literature for the illustration of various nonparametric regression methods; see, e.g., Breiman (1999, 2001) and Steele (2009). This data set contains 4177 observations on 8 input variables and a response that represents the number of rings indicating the age of an abalone. The major goal of this real data application is to predict the response based on the information of the 8 input variables. Since the first input variable is categorical and consists of three categories indicating the sex (Male, Female, and Infant), we only search nearest neighbors restrictively in each category. Consequently, there are 7 features after splitting the data set into three categories. Because the non-

| Method | $k$-NN | DNN | TDNN | RF |
|---|---|---|---|---|
| MSE | 4.99 | 4.553 | 4.512 | 4.60 |

Table 3: The MSEs of different nonparametric learning methods on the real data application in Section 6.

parametric rate of convergence for the nearest neighbors methods becomes slower as the feature dimensionality grows, we exploit the popular tool of principal component analysis (PCA) to reduce the dimensionality of the feature space and employ the first $m$ principal components for nonparametric learning. In our analysis, we choose $m = 3$ since the first three principal components account for more than 99% of the variation in the response.

Specifically, we randomly set aside 25% of the 4177 observations as a test set and train the TDNN estimator based on the remaining 75% of the observations. As mentioned in Section 5, the tuning of the two subsampling scales $s_1$ and $s_2$ is equivalent to that of the subsampling scale $s_1$ and their ratio $c = s_2/s_1$. We adopt the same strategy as described in Section 5 to tune both parameters $s_1$ and $c$ for the TDNN in a data-adaptive fashion. For a given feature vector $\mathbf{x}$ in the test set, each of the $B$ nearest neighbors to $\mathbf{x}$ is chosen as the left-out observation in the weighted LOOCV tuning procedure. Then the tuned $(c, s_1)$ is obtained by minimizing the weighted squared error over those $B$ left-out observations with the weights defined by the corresponding standard Gaussian kernel distance to the given feature vector $\mathbf{x}$. Finally, we apply the TDNN estimator constructed with the tuned $(c, s_1)$ to the test set and calculate the prediction error in terms of the MSE. The above procedure involving random data splitting is repeated 50 times and the prediction errors are averaged over those 50 random splits.

In particular, we tune $(c, s_1)$ from $c \in \{1.2, 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20\}$ and $s_1 \in [s_{\text{sign}}, 2s_{\text{sign}}]$ with $s_{\text{sign}}$ obtained by the sign-change tuning process (related to the curvature)

introduced in Section 5. The subsamping size $s$ for the DNN estimator is chosen from the sequence starting from 50 to 250 with an increment of 5. We set the neighborhood size of $B = 50$ for the implementation of the weighted LOOCV tuning procedure. We compare the prediction performance of the TDNN to that of the $k$-NN, DNN, and random forests (RF) in terms of the MSE evaluated on the test data. Table 3 summarizes the results of all the nonparametric learning methods on this real data application. In particular, the results for the $k$-NN and RF are extracted from Steele (2009). Indeed, from Table 3 we see that TDNN improves over both $k$-NN and DNN at the finite-sample level, which is in line with our theoretical results and simulation examples. Moreover, the TDNN also outperforms the RF. In contrast, there still lack optimality results for the tool of the RF.

# 7 Discussion

In this paper, we have investigated the problems of estimation and inference for nonparametric mean regression function using the two-scale DNN (TDNN), a bias reduced estimator based on the distributional nearest neighbors (DNN). Our suggested method of TDNN alleviates the finite-sample bias issue of the classical $k$-nearest neighbors and admits easy implementation with simple tuning under the assumption of the fourth-order smoothness on the mean regression function. We have provided theoretical justifications for the proposed estimator and established the asymptotic normality theory for practical use of TDNN in nonparametric statistical inference with optimality. The new TDNN tool can be exploited for the heterogeneous treatment effect (HTE) estimation and inference that is key to identifying individualized treatment effects.

Our bias reduction idea can be generalized to construct the multi-scale DNN when the mean regression function has even higher-order smoothness. In such case, DNN or TDNN no longer enjoys the nonparametric minimax optimal convergence rate. By exploiting

higher-order asymptotic bias expansion, a multi-scale DNN can be constructed in the same fashion for achieving the nonparametric optimal convergence rate. We leave the detailed investigations for future study.

It would also be interesting to extend the idea of TDNN to the settings of diverging or high feature dimensionality and consider the non-i.i.d. data settings such as time series, panel, and survival data. Since the distance function plays a natural role in identifying the nearest neighbors, it would be interesting to investigate the choice of different distance metrics, aside from the Euclidean distance, that are pertinent to specific manifold structures intrinsic to data. These problems are beyond the scope of the current paper and will be interesting topics for future research.

# References


Arvesen, J. N. (1969). Jackknifing $U$-statistics. Ann. Math. Statist. 40, 2076–2100.

Athey, S., J. Tibshirani, S. Wager, et al. (2019). Generalized random forests. Annals of Statistics 47(2), 1148–1178.

Berrett, T. B., R. J. Samworth, and M. Yuan (2019). Efficient multivariate entropy estimation via $k$-nearest neighbour distances. The Annals of Statistics 47, 288–318.

Berry, A. C. (1941). The accuracy of the Gaussian approximation to the sum of independent variates. Trans. Amer. Math. Soc. 49, 122–136.

Biau, G., F. Cérou, and A. Guyader (2010). On the rate of convergence of the bagged nearest neighbor estimate. Journal of Machine Learning Research 11, 687–712.

Biau, G. and L. Devroye (2015). Lectures on the nearest neighbor method. Springer.

Borovkov, A. A. (2013). Probability Theory. Springer.

Borovskikh, I. I. V. (1996). U-statistics in Banach Spaces. VSP.

Breiman, L. (1996). Bagging predictors. Machine learning 24(2), 123–140.

Breiman, L. (1999). Using adaptive bagging to debias regressions. Technical report, Technical Report 547, Statistics Dept. UCB.

Breiman, L. (2001). Random forests. Machine Learning 45, 5–32.

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3.1. Statistics Department University of California Berkeley, CA, USA 1, 58.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2018). On the effect of bias estimation on coverage accuracy in nonparametric inference. Journal of the American Statistical Association 113(522), 767–779.

Cheang, W.-K. and G. C. Reinsel (2000). Bias reduction of autoregressive estimates in time series regression model through restricted maximum likelihood. Journal of the American Statistical Association 95(452), 1173–1184.

Chi, C.-M., P. Vossler, Y. Fan, and J. Lv (2020). Asymptotic properties of high-dimensional random forests. arXiv preprint arXiv:2004.13953.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2008). Nonparametric tests for treatment effect heterogeneity. Review of Economics and Statistics 90, 389–405.

Dette, H. and A. Pepelyshev (2010). Generalized latin hypercube design for computer experiments. Technometrics 52(4), 421–429.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. Ann. Statist. 7, 1–26.

Fan, J. and Y. Fan (2008). High dimensional classification using features annealed independence rules. Annals of statistics 36, 2605.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70, 849–911.

Fan, J. and J. Lv (2018). Sure independence screening (invited review article). Wiley StatsRef: Statistics Reference Online.

Frees, E. W. (1989). Infinite order u-statistics. Scandinavian Journal of Statistics, 29–45.

Gao, L., Y. Fan, J. Lv, and Q. Shao (2021). Asymptotic distributions of high-dimensional distance correlation inference. The Annals of Statistics 49, 1999–2020.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002). A Distribution-Free Theory of Nonparametric Regression. Springer.

Hahn, P. R., J. S. Murray, C. M. Carvalho, et al. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. Bayesian Analysis.

Hájek, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. The Annals of Mathematical Statistics 39, 325–346.

Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. The Annals of Statistics, 675–694.

Hall, P. (2013). The bootstrap and Edgeworth expansion. Springer Science & Business Media.

Hall, P. and R. J. Samworth (2005). Properties of bagged nearest neighbour classifiers. J. R. Stat. Soc. Ser. B Stat. Methodol. 67(3), 363–379.

Hechenbichler, K. and K. Schliep (2004). Weighted k-nearest-neighbor techniques and ordinal classification.

Hitsch, G. J. and S. Misra (2018). Heterogeneous treatment effects and optimal targeting policy evaluation. Available at SSRN 3111957.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. The Annals of Mathematical Statistics 19, 293–325.

Imbens, G. W. and D. B. Rubin (2015). Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press.

Korolyuk, V. S. and Y. V. Borovskich (1994). Theory of U-statistics. Springer.

Leblanc, A. (2010). A bias-reduced approach to density estimation using bernstein polynomials. Journal of Nonparametric Statistics 22(4), 459–475.

Lee, M.-j. (2009). Non-parametric tests for distributional treatment effect for randomly censored responses. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71, 243–264.

Lin, Z., P. Ding, and F. Han (2021). Estimation based on nearest neighbor matching: from density ratio to average treatment effect. arXiv preprint arXiv:2112.13506.

Mack, Y. (1980). Local properties of k-NN regression estimates. SIAM Journal on Algebraic Discrete Methods 2, 311–323.

Newey, W. K., F. Hsieh, and J. M. Robins (2004). Twicing kernels and a small bias property of semiparametric estimators. Econometrica 72(3), 947–962.

Peng, W., T. Coleman, and L. Mentch (2019). Asymptotic distributions and rates of convergence for random forests via generalized $U$-statistics. arXiv preprint arXiv:1905.10651.

Powers, S., J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie, and R. Tibshirani (2017). Some methods for heterogeneous treatment effect estimation in high-dimensions. arXiv preprint arXiv:1707.00102.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. J. Roy. Statist. Soc. Ser. B 11, 68–84.

Quenouille, M. H. (1956). Notes on bias in estimation. Biometrika 43, 353–360.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701.

Samworth, R. J. (2012). Optimal weighted nearest neighbour classifiers. The Annals of Statistics 40, 2733–2763.

Schucany, W., H. Gray, and D. Owen (1971). On bias reduction in estimation. Journal of the American Statistical Association 66(335), 524–533.

Schucany, W. and J. P. Sommers (1977). Improvement of kernel type density estimators. Journal of the American Statistical Association 72(358), 420–423.

Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics. Wiley Series in Probability and Statistics.

Shalit, U., F. D. Johansson, and D. Sontag (2017). Estimating individual treatment effect: generalization bounds and algorithms. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 3076–3085. JMLR. org.

Song, Y., X. Chen, K. Kato, et al. (2019). Approximating high-dimensional infinite-order $u$-statistics: Statistical and computational guarantees. Electronic Journal of Statistics 13(2), 4794–4848.

Steele, B. M. (2009). Exact bootstrap k-nearest neighbor learners. Machine Learning 74(3), 235–255.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. The annals of statistics, 1040–1053.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. The Annals of Statistics 35, 2769–2794.

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

Wager, S., T. Hastie, and B. Efron (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. Journal of Machine Learning Research 15, 1625–1651.

Zaidi, A. and S. Mukherjee (2018). Gaussian process mixtures for estimating heterogeneous treatment effects. arXiv preprint arXiv:1812.07153.