Nonparametric Inference of Heterogeneous Treatment Effects with Two-Scale Distributional Nearest Neighbors

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

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

January 10, 2021

Abstract

Understanding heterogeneous treatment effects (HTE) plays a key role in many contemporary causal inference applications arising from different areas. Most of the existing works have focused on the estimation of HTE. Yet the statistical inference aspect of the problem remains relatively undeveloped. In this paper we investigate the inference of HTE in a nonparametric setting for randomized experiments. We formulate the problem as two separate nonparametric mean regressions, one for control group and the other for treatment group. For each mean regression, we extend the tool of k-nearest neighbors to the framework of distributional nearest neighbors (DNN). We show that the DNN estimator has two equivalent representations of L-statistic and U-statistic, where the former endorses easy and fast implementation, and the latter enables us to obtain higher-order asymptotic expansion of bias and establish the asymptotic normality. To reduce the finite sample bias of DNN, we further suggest a new method of two-scale distributional nearest neighbors (TDNN). Under some regularity conditions, we show through delicate higher-order asymptotic expansions that the TDNN heterogeneous treatment effect estimator is asymptotically normal. We further establish the consistency of the variance estimates of the TDNN estimator with both jackknife and bootstrap, enabling user-friendly inference tools for heterogeneous treatment effects. The theoretical results and appealing finite-sample performance of the suggested TDNN method are illustrated with several simulation examples and a children's birth weight application.

Key words: Causal inference; Heterogeneous treatment effects; Nonparametric estimation and inference; k-nearest neighbors; Two-scale distributional nearest neighbors; Bootstrap

1 Introduction

The problems of treatment effect estimation and inference in causal inference have broad applications in a wide variety of scientific areas, ranging from economics to medical studies. Examples include evaluating whether food stamps increases adult obesity, deciding whether to launch a job training program, and finding out whether a vaccine is truly effective. Under the potential outcomes framework (Rubin, 1974; Imbens and Rubin, 2015), the treatment effect usually refers to the average causal effect of a binary treatment indicator variable on some outcome of interests. Beyond regressions, the potential outcomes framework perceives outcomes as in parallel universes, one with the event of treatment happened and the other without. The interest is often the average of the difference between the two conceptual potential outcomes, the average treatment effect (ATE). It is seen that for each unit, only one potential outcome can be observed. Indeed, a common challenge in many causal inference problems is caused by the missingness of potential outcomes. There is a long line of literature on ATE estimation. See, for example, Belloni et al. (2014); Chernozhukov et al. (2015); Belloni et al. (2017); Athey et al. (2017); Mullainathan and Spiess (2017); Fan et al. (2016), among many others.

Different from ATE, heterogeneous treatment effect (HTE) focuses on the unit level effect by considering the treatment effect conditional on pre-treatment covariates. The estimation and inference of HTE have received increasing attention in recent years because of their ability to provide information that ATE cannot provide. For example, it is often possible that a drug is less/more effective for certain subgroups of patients than the general, average population. In addition, if HTE can be well estimated for each individual unit, then

ATE can be easily estimated by aggregating over all individuals in the whole population. See Crump et al. (2008) for additional examples for the importance of HTE. In fact, the unit level effect measured by HTE can be invaluable information in a wide range of modern big data applications including economics, business, and healthcare. However, despite its importance, related research about HTE is still relatively limited.

There has been some recent efforts to use machine learning methods to estimate and infer HTE. Crump et al. (2008) proposed two nonparametric tests, one aiming at testing whether the treatment has zero effect for all subpopulations identified by covariates and the second one aiming at testing the existence of heterogeneity in treatment effects by covairates. Their tests were constructed by using two separate sieve estimates of nonparametric regression functions for control group and treatment group, respectively. Lee (2009) proposed a smoothed nonparametric test for assessing heterogeneity of treatment effect when covariates are continuous and the outcome variable is randomly censored.

Besides testing, there are also recent works on HTE estimation. Wager and Athey (2018) proposed causal forests whose estimate is shown to be pointwise consistent for the true treatment effect and asymptotically Gaussian. It was further proposed in the same paper to construct the asymptotic confidence intervals using the derived asymptotic distribution with variance estimated by the infinitesimal jackknife (Wager et al., 2014). Shalit et al. (2017) proposed to estimate HTE jointly using a neural network-based algorithm with loss function motivated from a generalization error bound established in the paper. Another popular line of work is to formulate HTE estimation as nonparametric regressions. Under the unconfoundedness and overlap assumptions, the problem of HTE estimation can be formulated as two separate nonparametric mean regression problems, one for the control group, and the other for the treatment group. Then the treatment effect is estimated as the difference of these two estimated mean functions. Examples along this line include Bayesian causal forests (Hahn et al. (2020)), causal boosting and causal MARS (Powers

et al. (2017)), Gaussian process mixtures (Zaidi and Mukherjee (2018)), and causal KNN (Hitsch and Misra (2018)). Among them, causal KNN is most closely related to our approach. This method estimates the treatment effect function by taking the difference of two separate k-nearest neighbor (KNN) regression functions for the treatment group and control group. The tuning parameter of neighborhood size k was chosen by minimizing the squared difference between the estimated treatment effect function and the propensity score weighted response. However, the work does not provide theoretical justification for the causal KNN estimator

In this paper, we propose a new method for HTE estimation in completely randomized experiments and provide formal statistical estimation and inference guarantees. We adopt the previously discussed approach of fitting two separate nonparametric mean regressions and then combine to form an HTE estimate. For each mean regression, we revisit and enhance the classical tool of KNN regression with a simple yet powerful algorithm, extending it to the distributional setting. Specifically, we estimate the mean regression function as the average of all 1-NN estimators constructed from subsampling the data of size s without replacement. Here, it is important for s to diverge with total sample size n. We show that this subsampling and aggregating approach is equivalent to assigning monotonic weights to the nearest neighbors in a distributional fashion, motivating the name of distributional nearest neighbors (DNN). We show that DNN estimator has a representation of L-statistic with weights depending only on the rank of the observations. Although the L-statistic representation endorses easy and fast implementation of DNN, it does not help with establishing the sampling properties. For theoretical analysis, we further demonstrate that DNN estimator has an equivalent representation of U-statistic with a kernel function of diverging dimension equal to the subsampling scale s. Despite the nice U-statistic representation, classical theory does not apply to DNN for deriving its asymptotic properties, because of the diverging dimensionality of the kernel function. To overcome this technical

challenge, we exploit Hoeffding's canonical decomposition introduced in Hoeffding (1948), and carefully collect and analyze the higher-order terms in our decomposition. We provide higher-order asymptotic expansion for the bias of DNN in estimating the mean regression function, and show that DNN is asymptotically normal as s and n diverge to infinity. Here, d is the dimension of the pre-treatment covariate in constructing DNN estimator and is considered to be fixed.

A nice feature of DNN estimator is that its first-order bias can be eliminated by combining two DNN estimators with different subsampling scales, resulting in two-scale DNN. Utilizing the U-statistic representation of two-scale DNN with a new and carefully constructed diverging dimensional kernel, we further establish the asymptotic normality of the two-scale DNN estimator. These results naturally lead to the asymptotic normality of the resulting HTE estimator.

To further provide statistical inferential guarantees such as valid confidence intervals, we need to estimate the asymptotic variance of the two-scale DNN estimator, which does not admit simple analytic form that is practically useful. We explore two methods, jackknife and bootstrap, for such a purpose. We formally demonstrate that both methods yield consistent estimates of the 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 was built on the results we derived for jackknife estimator. Although both methods yield consistent variance estimates, the bootstrap estimator is much more computationally efficient.

We then demonstrate the superior performance of our method using Monte Carlo simulations and a real data analysis on the impact of smoking on children's birth weights. The two scale DNN estimator has two parameters to tune – the subsampling scales. We propose to prefix the ratio of the two subsampling scales at some level that is adaptive to

dimension d so that the variance of two-scale DNN can be better controlled. An additional advantage of such strategy is that it leaves us only one subsampling scale s to tune. Our simulation results show that as s increases, the mean squared estimation error (MSE) of two-scale DNN estimator follows a smooth U-shaped curve. This suggests a simple tuning strategy for two scale DNN which starts with some small value of s, and increases its value until the MSE starts increasing. This simple tuning yields attractive performance in our numerical studies.

An important contribution of our work is that we provide an easy-to-implement method with simple tuning for HTE estimation and inference with theoretical guarantee. As discussed above, most existing works focus only on the estimation and are not able to provide confidence intervals. A noticeable exception is the causal forests in Wager and Athey (2018). Compared to the causal forests, the confidence intervals provided by our method transit more smoothly as the value of pre-treatment covariates changes, as demonstrated in our numerical study. Consequently, the confidence intervals provided by our method are easier to interpret. In addition, we observe in our simulation studies that the variance estimate in causal forest can be crude in some applications, resulting in low coverage of confidence intervals.

Our paper makes a standalone contribution to the nonparametric statistics literature, going beyond applications to HTE estimation. Two-scale DNN belongs to the class of weighted nearest neighbors methods. Some weights in two-scale DNN take negative values. The advantage of using negative weights in weighted nearest neighbors classifiers was formally investigated in Samworth (2012). For classical and other recent results on nearest neighbors related methods and theory, see, for example, Mack (1980); Györfi et al. (2002); Biau and Devroye (2015); Berrett et al. (2019). Despite the similarity, two-scale DNN has at least two advantages compared to these existing literature. First, we provide an explicit and easy-to-implement way to choose weights. Although sufficient conditions on weights for

ensuring the asymptotic normality for general weighted nearest neighbors estimator have been developed (Biau and Devroye, 2015), it is usually unclear how to practically choose such weights. Second, even if weights can be successfully constructed, it is still unclear how to estimate such estimator's variance for confidence interval construction.

The rest of the paper is organized as follows. Section 2 introduces the model setting for heterogeneous treatment effect estimation and reviews the approach of weighted nearest neighbors for nonparametric regression. 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 children's birth weight 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. All the proofs and technical details are provided in the Supplementary Material.

2 Heterogeneous treatment effect estimation

2.1 Model setting

Consider the example of new drug and vaccine developments. The clinical trials for these studies involve carefully designed randomized experiments, in which each individual will be assigned randomly to either the treatment group denoted as $T = 1$ or the control group denoted as $T = 0$. Using the potential outcomes framework (Rubin, 1974), let $Y_{T=1} \in \mathbb{R}$ and $Y_{T=0} \in \mathbb{R}$ represent the potential outcomes for the treatment and control groups, respectively. Then the observed scalar response can be written as $Y = TY_{T=1} + (1 - T)Y_{T=0}$. Let $\mathbf{X} \in \mathbb{R}^d$ be the random feature vector for an individual with d some fixed positive integer representing the feature dimension. We consider the randomized

experiment setting which amounts to the choice of constant treatment propensity $\mathbb{P}(T = 1|\mathbf{X}, Y_{T=1}, Y_{T=0}) = 1/2$. Here, $1/2$ can be replaced with any other constant in $(0, 1)$. The evaluation of the effectiveness for a newly developed drug or vaccine often boils down to the average treatment effect (ATE) of treatment T on response Y defined as $\tau = \mathbb{E}[Y_{T=1} - Y_{T=0}]$. Given the scale of the costs associated with the medical developments, a practically important question even with the presence of a less significant treatment effect τ on the entire population characterized by the distribution of $\mathbf{X}$ is whether the treatment effect could be significant on certain individuals in the population.

The above illustrative example demonstrates that characterizing the treatment effects at the individual level with precision can have important real applications. Given a fixed feature vector $\mathbf{x} \in \mathbb{R}^d$, the heterogeneous treatment effect (HTE) of treatment T on response Y is defined as

$$\tau(\mathbf{x}) = \mathbb{E}[Y_{T=1} - Y_{T=0}|\mathbf{X} = \mathbf{x}], \tag{1}$$

where the expectation is taken with respect to the randomness associated with the subgroup of all individuals with specific feature vector $\mathbf{x}$. Such a subgroup can be viewed as a group of identical twins who differ only by which treatment they received. We see that the fundamental challenge of causal inference applications is the missing data problem, that is, the latency of the above *ideal* subgroup. It can be infeasible or even unethical to assign certain individuals to the treatment or control group, not to mention that having an identical twin can become a luxury in practice.

Since the setting of randomized experiments entails the unconfoundedness given by

$$(Y_{T=0}, Y_{T=1}) \perp\!\!\!\perp T \mid \mathbf{X}, \tag{2}$$

our goal of heterogeneous treatment effect estimation and inference for (1) reduces to the problem of nonparametric regression applied separately to the treatment and control

groups, giving rise to

$$\tau(\mathbf{x}) = \mathbb{E}[Y_{T=1}|\mathbf{X}=\mathbf{x}] - \mathbb{E}[Y_{T=0}|\mathbf{X}=\mathbf{x}] = \mathbb{E}[Y|\mathbf{X}=\mathbf{x}, T=1] - \mathbb{E}[Y|\mathbf{X}=\mathbf{x}, T=0]. \quad (3)$$

Specifically, let us consider the following nonparametric regression model for the treatment group $Y_{T=1} = \mu(\mathbf{X}) + \epsilon$, where $\mu(\mathbf{X}) = \mathbb{E}[Y_{T=1}|\mathbf{X}]$ denotes the true mean regression function and the model error ϵ with zero mean and finite variance is independent of the d-dimensional random feature vector $\mathbf{X}$. Similarly, we can introduce the corresponding nonparametric regression model for the control group; see Section 3.3 for more detailed technical descriptions.

To simplify the technical presentation, hereafter, unless otherwise specified, we will slightly abuse the notation by using Y to denote the generic response variable and focus our attention on the following nonparametric regression model when introducing our main ideas and major theoretical developments

$$Y = \mu(\mathbf{X}) + \epsilon. \quad (4)$$

We assume that there are independent and identically distributed (i.i.d.) observations $(\mathbf{X}_i, Y_i)$, $i = 1, \cdots, n$ observed from model (4). Our method and theory will be separately applied to the control and treatment groups and then combined together by using (3) to estimate the heterogeneous treatment effect.

2.2 Weighted nearest neighbors

Among the nonparametric regression methods, the k-nearest neighbors (KNN) procedure and its generalizations have received great popularity by many researchers and practitioners due to its simplicity of implementation and nice theoretical properties. For an overview of nearest neighbors methods, see, e.g., Biau and Devroye (2015). Given a fixed vector $\mathbf{x} \in \mathbb{R}^d$, we can calculate the Euclidean distance from each observed feature vector $\mathbf{X}_i$ in

the sample to the target $\mathbf{x}$ and then reorder the sample with such distances. This results in a reordered sample $\{(\mathbf{X}_{(1)}, Y_{(1)}), \cdots, (\mathbf{X}_{(n)}, Y_{(n)})\}$ with

$$\|\mathbf{X}_{(1)} - \mathbf{x}\| \leq \|\mathbf{X}_{(2)} - \mathbf{x}\| \leq \cdots \leq \|\mathbf{X}_{(n)} - \mathbf{x}\|, \tag{5}$$

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

$$\widehat{\mu}_{\mathrm{WNN}}(\mathbf{x}) = \sum_{i=1}^{n} w_{ni} Y_{(i)}, \tag{6}$$

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 (6) have been studied extensively in Biau and Devroye (2015). In particular, it has been proved therein that, with an appropriately selected weight vector $(w_{n1}, \cdots, w_{nn})$, the weighted nearest neighbors estimate $\widehat{\mu}_{\mathrm{WNN}}(\mathbf{x})$ can be consistent with the rate of convergence $o_P(n^{-2/(d+4)})$ and can have asymptotic normality. The existing results in the literature 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 the general weighted nearest neighbors estimator $\widehat{\mu}_{\mathrm{WNN}}(\mathbf{x})$ can admit a rather complicated form and depend upon some *unknown* population quantities that are very difficult to estimate in practice. To address these difficulties, we will introduce a new framework of the two-scale distributional nearest neighbors estimate for heterogeneous treatment effect estimation and inference.

3 Two-scale distributional nearest neighbors

3.1 Distributional nearest neighbors

The classical KNN method for nonparametric regression assigns equal weights $1/k$ to the k nearest neighbors of a fixed vector $\mathbf{x} \in \mathbb{R}^d$ and zero weights to all remaining observations in the sample, resulting in the following specific form of estimator in (6)

$$\widehat{\mu}_{\mathrm{KNN}}(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}. \tag{7}$$

In particular, the choice of $k = 1$ in (7) gives rise to the 1-nearest neighbor (1NN) estimator $Y_{(1)}$, which relies on the single closest observation for nonparametric estimation.

Our distributional nearest neighbors (DNN) procedure builds upon the ideas of both subsampling and 1NN. 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 1NN estimator

$$\Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s}) = Y_{(1)}(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s}) \tag{8}$$

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 single-scale DNN estimator $D_n(s)(\mathbf{x})$ with subsampling scale s for estimating $\mu(\mathbf{x})$ is formally defined as a U-statistic

$$D_n(s)(\mathbf{x}) = \binom{n}{s}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_s \leq n} \Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s}), \tag{9}$$

where the kernel function $\Phi(\mathbf{x}; \cdot)$ is given in (8).

The above U-statistic representation averages over all 1NN 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 1NN estimator $Y_{(1)}$ based on the full sample of size n, which admits 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, we show in the lemma below that a different representation of the DNN estimator can be exploited for easy computation.

Lemma 1. *In addition to the U-statistic representation given in (9), the single-scale DNN estimator $D_n(s)(\mathbf{x})$ also admits an equivalent L-statistic (Serfling, 1980) representation as*

$$D_n(s)(\mathbf{x}) = \binom{n}{s}^{-1} \sum_{i=1}^{n-s+1} \binom{n-i}{s-1} Y_{(i)}, \tag{10}$$

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

The above L-statistic representation relieves the computational burden greatly from the U-statistic representation. Let us gain some insights into the distributional weights unveiled in (10). It is easy to see that there are a total of $\binom{n}{s}$ subsamples of size s from the full sample of size n. Out of the $\binom{n}{s}$ options, $(\mathbf{X}_{(1)}, Y_{(1)})$ will appear in $\binom{n-1}{s-1}$ of them, and $Y_{(1)}$ will become the 1NN estimator from this particular subsample whenever $(\mathbf{X}_{(1)}, Y_{(1)})$ is included, giving rise to the specific weight of $\binom{n-1}{s-1}/\binom{n}{s}$ for $Y_{(1)}$. We can obtain similar intuitions on the remaining weights. Since the weights in (10) are assigned in a distributional fashion on the entire sample, we name the new procedure introduced in (9) as the distributional nearest neighbors.

Such a distributional view differs from the conventional idea of seeking weights for the nearest neighbors with a relatively small neighborhood size. One interesting feature is that the distribution of weights used in the single-scale DNN is characterized by only two parameters of the full sample size n and the subsampling scale s. As shown later in Section

3.3, our new higher-order asymptotic expansion for bias reveals that the distributional view yields explicit constant for the leading bias term that is free of the subsampling scale s, which opens the door for removing the first-order asymptotic bias of DNN.

3.2 Two-scale DNN

We are now ready to suggest a natural extension of the single-scale DNN procedure introduced in Section 3.1. 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 3.3. 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 remove completely 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 (9). Without loss of generality, we assume that $s_1 < s_2$. Then Theorem 1 ensures that

$$\mathbb{E}\, D_n(s_1)(\mathbf{x}) = \mu(\mathbf{x}) + c\, s_1^{-2/d} + R(s_1), \tag{11}$$

$$\mathbb{E}\, D_n(s_2)(\mathbf{x}) = \mu(\mathbf{x}) + c\, s_2^{-2/d} + R(s_2), \tag{12}$$

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 $R(s) = O(s^{-3})$ for $d = 1$ and $R(s) = O(s^{-4/d})$ for $d \geq 2$.

Although the specific constant c in the asymptotic expansions (11) and (12) 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

$$w_1^* = w_1^*(s_1, s_2) = 1/(1 - (s_1/s_2)^{-2/d}) \tag{13}$$

$$\text{and} \quad w_2^* = w_2^*(s_1, s_2) = -(s_1/s_2)^{-2/d}/(1 - (s_1/s_2)^{-2/d}). \tag{14}$$

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

$$D_n(s_1, s_2)(\mathbf{x}) = w_1^* D_n(s_1)(\mathbf{x}) + w_2^* D_n(s_2)(\mathbf{x}). \tag{15}$$

We will impose the restriction that s_1/s_2 is bounded away from both 0 and 1 by some positive constant. This can avoid the undesirable cases of weights being too close to 0 or having diverging magnitudes as s diverges. In the implementation, we can also choose s_1/s_2 adaptively to dimensionality d so that the weights w_1^* and w_2^* are free from the impact of dimensionality.

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 (11) and (12) that

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

where $R^*(s_1) = O(s_1^{-4/d})$ for $d \geq 2$ and $R^*(s_1) = O(s_1^{-3})$ for $d = 1$, and we assume that s_1 and s_2 are of the same order for simplicity. The removal of the first-order asymptotic bias as shown in (16) makes the TDNN estimator enjoy appealing finite-sample performance with reduced bias and controlled variance, as demonstrated with the extensive simulation examples in Section 5.

It is worth mentioning that in view of (13) and (14), 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 KNN classifier for the classification setting was discovered earlier in Samworth (2012). See also the discussions in Biau and Devroye (2015) for similar advantages in the regression setting.

3.3 Asymptotic distributions of two-scale DNN

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

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.

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 ∞, $f(\cdot)$ and $\mu(\cdot)$ are four-times continuously differentiable with bounded second-order, third-order, 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 $\sigma_\epsilon^2 > 0$.*

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 (4).*

Conditions 1–3 are some basic assumptions that have been employed commonly for nonparametric regression. 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*

$$\mathbb{E}\, D_n(s)(\mathbf{x}) = \mu(\mathbf{x}) + B(s) \tag{17}$$

with

$$B(s) = \Gamma(2/d+1)\frac{f(\mathbf{x})\,\mathrm{tr}(\mu''(\mathbf{x})) + 2\,\mu'(\mathbf{x})^T f'(\mathbf{x})}{2\,d\,V_d^{2/d}\,f(\mathbf{x})^{1+2/d}}\,s^{-2/d} + R(s), \tag{18}$$

$$R(s) = \begin{cases} A_1(d,f,\mathbf{x})s^{-3} + o(s^{-3}), & d = 1, \\ A_2(d,f,\mathbf{x})s^{-4/d} + o(s^{-4/d}), & d \geq 2, \end{cases} \tag{19}$$

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,* $\mathrm{tr}(\cdot)$ *stands for the trace of a given matrix, and* $A_1(d,f,\mathbf{x})$ *and* $A_2(d,f,\mathbf{x})$ *are some positive bounded quantities that depend only on* d*, the underlying density function* $f(\cdot)$ *of* $\mathbf{X}$*, and the regression function* $\mu(\cdot)$*.*

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 $d \geq 2$ (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 would be beneficial to remove the first-order asymptotic bias completely to improve the finite-sample performance. Such a goal can be achieved thanks to the interesting feature revealed in Theorem 1 that the explicit constants for the first two leading terms in the higher-order asymptotic expansion do not depend on the subsampling scale parameter s. The technical analysis of Theorem 1 exploits the idea of projecting the mean function $\mu(\mathbf{X}) = \mathbb{E}(Y|\mathbf{X})$ onto the positive half line $\mathbb{R}_+ = [0,\infty)$ given by $\|\mathbf{X} - \mathbf{x}\|$. In particular, the assumption of $s \to \infty$ plays an important role in establishing the higher-order asymptotic expansion. We further characterize the asymptotic distribution of the single-scale DNN estimator in the following theorem.

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* σ_n *of order* $(s/n)^{1/2}$,

$$\frac{D_n(s)(\mathbf{x}) - \mu(\mathbf{x}) - B(s)}{\sigma_n} \xrightarrow{\mathscr{D}} N(0,1) \tag{20}$$

as $n \to \infty$, *where* $B(s)$ *is given in (18).*

Theorem 2 establishes the asymptotic normality of the single-scale DNN estimator $D_n(s)(\mathbf{x})$ with subsampling scale s. In particular, it requires the assumptions of $s \to \infty$ and $s = o(n)$, where the former leads to reduced 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 projection idea. Despite the U-statistic representation of $D_n(s)(\mathbf{x})$ given in (9), 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 new asymptotic analysis requires the opposite assumption of *diverging* subsampling scale s. As a result, we have to conduct a more delicate and challenging technical analysis to derive the asymptotic normality. It is worth mentioning that when the subsampling scale s is chosen in the order of $n^{\frac{d}{d+4}}$, the optimal rate of convergence in terms of the mean-squared error can be obtained. The exact form of the asymptotic variance σ_n^2 can take a complicated form that depends upon the underlying distributions and fixed vector $\mathbf{x}$, and is left for future study.

We proceed with characterizing the asymptotic distribution for the two-scale DNN estimator introduced in (15).

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* $c_1 \le s_1/s_2 \le c_2$. *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 oder* $(s_2/n)^{1/2}$,

$$\frac{D_n(s_1, s_2)(\mathbf{x}) - \mu(\mathbf{x}) - \Lambda}{\sigma_n} \xrightarrow{\mathscr{D}} N(0,1) \tag{21}$$

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

We note that the positive sequence σ_n in Theorem 3 is different from the σ_n in Theorem 2, with the former representing the asymptotic standard deviation of TDNN estimator and the latter representing the asymptotic standard deviation of single-scale DNN estimator. We used the same generic notation for the convenience of presentation. Since the explicit form of the asymptotic standard deviation will not be used, this should not cause any confusion. Theorem 3 above establishes the asymptotic normality for the two-scale DNN estimator $D_n(s_1, s_2)(\mathbf{x})$ with a pair of subsampling scales s_1 and s_2. Similarly, it requires that both subsampling scales s_1 and s_2 are diverging and of a smaller order of the full sample size n asymptotically in order to balance between the bias and variance. In particular, the asymptotic bias of the two-scale DNN estimator is reduced 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 $d = 1$. The optimal choice of s_2 in terms of achieving the best bias and variance tradeoff is $s_2 = O(n^{d/(8+d)})$ for $d \geq 2$ and $s_2 = O(n^{1/7})$ for $d = 1$, yielding the corresponding consistency rate of $n^{-4/(8+d)}$ for $d \geq 2$ and $n^{-3/7}$ for $d = 1$. Note that this rate is faster than the optimal convergence rate $n^{-2/(d+4)}$ established in Theorem 14.4 of Biau and Devroye (2015) for weighted nearest neighbors estimate with nonnegative weights in regression setting. The main reason for improved convergence rate is that TDNN allows for negative weights.

We would also like to point out that the conclusion in Theorem 3 is not a simple consequence of that in Theorem 2, since the marginal asymptotic normalities do not necessarily entail the joint asymptotic normality. To deal with such a technical difficulty, we have to analyze the two single-scale DNN estimators involved in the definition of TDNN estimator in a joint fashion. A key ingredient of our technical analysis of Theorem 3 is to show that the two-scale DNN 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.

Now we are ready to apply two-scale DNN to hetergeneous treatment effect estimation

discussed in Section 2.1. We will first introduce some necessary notation. Denote by n_1 and n_0 the sizes of the i.i.d. samples from the treatment and control groups, respectively. The assumption of completely randomized experiment entails that $n_0/n_1 \xrightarrow{p} 1$ as $n \to \infty$ and the two samples for the treatment and control groups are independent of each other. Let $\mathbf{x} \in \text{supp}(\mathbf{X}_1) \cap \text{supp}(\mathbf{X}_0)$ be a fixed vector, where $\text{supp}(\mathbf{X}_1)$ and $\text{supp}(\mathbf{X}_0)$ represent the supports of the corresponding feature vector distributions for the treatment and control groups, respectively. Similarly, denote by $\mu_1(\cdot)$ and $\mu_0(\cdot)$ the mean regression functions corresponding to responses $Y_{T=1}$ and $Y_{T=0}$, respectively, and ϵ_1 and ϵ_0 the model errors, with the subscript indicating the treatment and control groups, respectively. Then we can construct two individual two-scale DNN estimators $D^{(1)}_{n_1}(s^{(1)}_1, s^{(1)}_2)(\mathbf{x})$ and $D^{(0)}_{n_0}(s^{(0)}_1, s^{(0)}_2)(\mathbf{x})$ separately based on the treatment and control samples with pairs of subsampling scales $(s^{(1)}_1, s^{(1)}_2)$ and $(s^{(0)}_1, s^{(0)}_2)$, respectively.

In view of (3), the population version of the heterogeneous treatment effect at the fixed vector $\mathbf{x}$ is given by

$$\tau(\mathbf{x}) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}). \tag{22}$$

We estimate $\tau(\mathbf{x})$ using the following TDNN heterogeneous treatment effect estimator

$$\widehat{\tau}(\mathbf{x}) = D^{(1)}_{n_1}(s^{(1)}_1, s^{(1)}_2)(\mathbf{x}) - D^{(0)}_{n_0}(s^{(0)}_1, s^{(0)}_2)(\mathbf{x}). \tag{23}$$

The theorem below characterizes the asymptotic distribution of $\widehat{\tau}(\mathbf{x})$.

Theorem 4. *Assume that Conditions 1–3 with the subscripts attached hold for both treatment and control groups. Further assume that* $s^{(i)}_2 \to \infty$, $s^{(i)}_2 = o(n)$, *and there exist some constants* $0 < c_1 < c_2 < 1$ *such that* $c_1 \le s^{(i)}_1/s^{(i)}_2 \le c_2$ *for* $i = 0, 1$. *Then for any fixed* $\mathbf{x} \in \text{supp}(\mathbf{X}_1) \cap \text{supp}(\mathbf{X}_0) \subset \mathbb{R}^d$, *it holds that for some positive sequence* σ_n *of order* $\{(s^{(1)}_2 + s^{(0)}_2)/n\}^{1/2}$,

$$\frac{[D^{(1)}_{n_1}(s^{(1)}_1, s^{(1)}_2)(\mathbf{x}) - D^{(0)}_{n_0}(s^{(0)}_1, s^{(0)}_2)(\mathbf{x})] - \tau(\mathbf{x}) - \Lambda}{\sigma_n} \xrightarrow{\mathscr{D}} N(0, 1) \tag{24}$$

as $n \to \infty$, *where* $\Lambda = O\{(s_1^{(1)})^{-4/d} + (s_2^{(1)})^{-4/d} + (s_1^{(0)})^{-4/d} + (s_2^{(0)})^{-4/d}\}$ *for* $d \geq 2$ *and* $\Lambda = O\{(s_1^{(1)})^{-3} + (s_2^{(1)})^{-3} + (s_1^{(0)})^{-3} + (s_2^{(0)})^{-3}\}$ *for* $d = 1$.

The same as explained before, σ_n in Theorem 4 is a generic notation representing the asymptotic standard deviation of the TDNN heterogeneous treatment effect estimator. We see that the subsampling scales need to satisfy that $s_2^{(i)} \to \infty$ and $s_2^{(i)} = o(n)$ for $i = 0, 1$. The asymptotic bias of estimator $\widehat{\tau}(\mathbf{x})$ is only of the second order $O\{(s_1^{(1)})^{-4/d} + (s_2^{(1)})^{-4/d} + (s_1^{(0)})^{-4/d} + (s_2^{(0)})^{-4/d}\}$ for $d \geq 2$ and $O\{(s_1^{(1)})^{-3} + (s_2^{(1)})^{-3} + (s_1^{(0)})^{-3} + (s_2^{(0)})^{-3}\}$ for $d = 1$. The asymptotic variance identified in Theorems 3 and 4 generally depends on the underlying distributions and the fixed vector $\mathbf{x}$, whose complicated form calls for a need to develop practical approaches to the estimation of the asymptotic variance for the TDNN estimator.

4 Variance estimate for two-scale DNN estimator

4.1 Jackknife estimator

Let us gain some insights into the problem of variance estimation for the TDNN estimator before presenting the major theoretical results. We have shown in Lemma 8 in Section B.8 of Supplementary Material that the two-scale DNN estimator $D_n(s_1, s_2)(\mathbf{x})$ is in fact a U-statistic. It is well known that the jackknife and bootstrap approaches are employed commonly to estimate the variance of a U-statistic. By the independence assumption of the treatment and control samples in the randomized experiment setting, the variance of the TDNN heterogeneous treatment effect estimator is naturally the sum of their individual variances. Thus, we only need to estimate the variance of the TDNN estimator constructed based on an individual i.i.d. sample $\{\mathbf{Z}_1, \cdots, \mathbf{Z}_n\}$ corresponding to either control group or treatment group.

As mentioned before, the two-scale DNN estimator $D_n(s_1, s_2)(\mathbf{x})$ admits the U-statistic

representation revealed in Lemma 8

$$D_n(s_1, s_2)(\mathbf{x}) = \binom{n}{s_2}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_{s_2} \le n} \Phi^*(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{s_2}), \tag{25}$$

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 (9), and w_1^* and w_2^* the weights defined in equations (13) and (14). We denote by

$$\sigma^2 = \mathrm{Var}(D_n(s_1, s_2)(\mathbf{x})) \tag{26}$$

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 ith observation as in (25)

$$U_{n-1}^{(i)} = \binom{n-1}{s_2}^{-1} \sum_{\substack{1 \le j_1 < j_2 < \cdots < j_{s_2} \le n \\ j_1, j_2, \cdots, j_{s_2} \ne i}} \Phi^*(\mathbf{x}; \mathbf{Z}_{j_1}, \mathbf{Z}_{j_2}, \cdots, \mathbf{Z}_{j_{s_2}}). \tag{27}$$

Then the jackknife estimator (Quenouille, 1949, 1956) for σ^2 in (26) is given by

$$\widehat{\sigma}_J^2 = \frac{n-1}{n} \sum_{i=1}^{n} \left(U_{n-1}^{(i)} - D_n(s_1, s_2)(\mathbf{x}) \right)^2. \tag{28}$$

We will formally establish the ratio consistency of the jackknife estimator $\widehat{\sigma}_J^2$ introduced in (28) 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* $c_1 \le s_1/s_2 \le c_2$. *Then for any fixed* $\mathbf{x} \in \mathrm{supp}(\mathbf{X}) \subset \mathbb{R}^d$, *when* $s_2 = o(n^{1/3})$ *it holds that* $\widehat{\sigma}_J^2/\sigma^2 \xrightarrow{p} 1$ *as* $n \to \infty$.

The proof of Theorem 5 still builds on the U-statistic framework. Similar to the discussions 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 B.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 δ accordingly.

We would like to 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 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 $\{U_{n-1}^{(i)}\}_{1\le i\le n}$ and the two-scale DNN estimator $D_n(s_1,s_2)(\mathbf{x})$ involved in the jackknife estimator $\widehat{\sigma}_J^2$ in (28). 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 used widely in many applications 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 (25), let us construct the two-scale DNN estimator

$$D_n^*(s_1,s_2)(\mathbf{x}) = \binom{n}{s_2}^{-1} \sum_{1\le i_1<i_2<\cdots<i_{s_2}\le n} \Phi^*(\mathbf{x}; \mathbf{Z}_{i_1}^*, \mathbf{Z}_{i_2}^*, \cdots, \mathbf{Z}_{i_{s_2}}^*) \tag{29}$$

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

We choose the number of bootstrap samples as $B \geq 1$. For each $1 \leq b \leq B$, we select independently a boostrap sample $\{\mathbf{Z}_{b,1}^*, \mathbf{Z}_{b,2}^*, \cdots, \mathbf{Z}_{b,n}^*\}$ and calculate the corresponding boostrap version of the two-scale DNN estimator $D_n^{(b)}(s_1, s_2)(\mathbf{x})$ as in (29). Observe that given the orginal observations $(\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n)$, the bootstrap samples $\{(\mathbf{Z}_{b,1}^*, \mathbf{Z}_{b,2}^*, \cdots, \mathbf{Z}_{b,n}^*)\}_{1\leq b\leq B}$ are independently and identically distributed as $(\mathbf{Z}_1^*, \mathbf{Z}_2^*, \cdots, \mathbf{Z}_n^*)$. Then the bootstrap estimator for σ^2 in (26) is given by

$$\widehat{\sigma}_{B,n}^2 = \frac{1}{B-1}\sum_{b=1}^{B}\left(D_n^{(b)}(s_1, s_2)(\mathbf{x}) - \bar{D}_{B,n}\right)^2, \tag{30}$$

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 (30) 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* $c_1 \leq s_1/s_2 \leq c_2$. *Then for any fixed* $\mathbf{x} \in \text{supp}(\mathbf{X}) \subset \mathbb{R}^d$, *when* $s_2 = o(n^{1/3})$ *and* $B \to \infty$, *it holds that* $\widehat{\sigma}_{B,n}^2/\sigma^2 \xrightarrow{p} 1$ *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 $\text{Var}(D_n^*(s_1, s_2)(\mathbf{x})|\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n)$ as $B \to \infty$. Second, since the bootstrap samples are drawn independently 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 $\text{Var}(D_n^*(s_1, s_2)(\mathbf{x})|\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n)$ for the variance will converge to the population quantity σ^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.

Let us elaborate more on the practical implementation of the two-scale procedure of TDNN for heterogeneous treatment effect inference. As mentioned in Section 4.1, the U-statistic representation of the TDNN estimator is key to our theoretical developments, but a different representation can be used for computation. Since the single-scale DNN estimator is in fact an L-statistic as shown in Lemma 1, the two-scale DNN estimator, which is a linear combination of a pair of single-scale DNN estimators, is still an L-statistic. The L-statistic representation of the TDNN estimator enables simple yet fast computation. For the randomized experiment setting considered in our paper, we can construct a pair of two-scale DNN estimators separately based on the treatment and control subsamples and then take a difference. As suggested by Theorem 6, we can further bootstrap such difference by resampling within each group to provide tight heterogeneous treatment effect inference. Thus, the two-scale procedure of TDNN coupled with bootstrap enjoys both theoretical justifications and computational scalability.

5 Simulation studies

In this section, we investigate the finite-sample performance of TDNN for heterogeneous treatment effect estimation and inference in comparison to the DNN, KNN, and causal forests (CF) in Wager and Athey (2018) which is a natural extension of random forests (Breiman, 2001, 2002; Chi et al., 2020) to the causal inference setting.

We construct weights for two-scale DNN estimator by resorting to equations (13) and (14). Since nonparametric methods including TDNN generally suffer from the curse of

dimensionality, we propose to first conduct dimension reduction if the ambient dimensionality p is large before applying TDNN. Let d be the reduced dimensionality. It is seen that although negative weight (as given in (13)) can ensure faster consistency rate, it may also induce inflated variance (at constant level) if w_1^* and w_2^* are not chosen appropriately. To control the variance, we propose to choose w_1^* and w_2^* adaptively to dimensionality d. Specifically, we suggest to fix $w_1^* = c/(c-1)$ and $w_2^* = -1/(c-1)$ with $c \in (0,1)$ some constant, and then solve for $s_2 = c^{-d/2}s_1$. In our numerical studies, we fix $c = 0.8$. We acknowledge that this choice may not be optimal and leave the optimal choice of c for future study.

With the above choice of weights, we have one tuning parameter which is the subsampling scale s. In implementation, we compute the TDNN estimator $D_n(s, c^{-d/2}s)(\mathbf{x})$ with s starting from 1. We continue this process until the difference in the absolute differences of consecutive MSEs changes the sign. Intuitively, it is the point when the curvature of the MSE of two-scale DNN estimator as a function of s changes, as demonstrated in the curve structure in Figure 1 from the simulation example in Section 5.1. Such a simple data-driven tuning strategy works fast and well in our numerical studies.

5.1 Comparisons with DNN and KNN

To illustrate the effectiveness of the two-scale framework compared to the single-scale DNN and the classical KNN, we simulate $n = 1000$ data points from the model $Y = \tau(\mathbf{x}) + \epsilon$, where $\tau(\mathbf{x}) = (x_1 - 1)^2 + (x_2 + 1)^3 - 3x_3$ with $\mathbf{x} = (x_1, x_2, x_3)^T$ and $(\mathbf{x}^T, \epsilon)^T \sim N(\mathbf{0}, I_4)$. Our goal is to estimate the true value of the mean response $\tau(\mathbf{x})$ at some test point chosen to be $(0.5, -0.5, 0.5)^T$. For the implementation of the single-scale DNN, we estimate the regression function at this test point while varying the subsampling scale s from 1 to 250. Since the dimension here is low we skip the step of dimension reduction and apply TDNN directly. As discussed at the beginning of this section, TDNN is implemented with the

choice of subsampling scales $s_1 = s$ and $s_2 = c^{-d/2}s_1$ with $d = 3$.

Figure 1 presents the simulation results for DNN and TDNN in terms of both the bias and the mean-squared error (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 the geometrical point of view since larger subsampling scale s leads to the use of the information in the sample concentrated more toward 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 expansions, the two-scale procedure of TDNN is free completely 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. Such a feature makes the tuning of the subsampling scale s relatively simple as discussed at the beginning of this section. Furthermore, we observe that in addition to the reduced finite-sample bias, TDNN admits substantially improved minimum MSE with over 50% reduction in comparison to the single-scale DNN. Indeed, the minimum values of the MSE attained by the single-scale DNN and TDNN are 0.1157 and 0.0488, respectively. To save space, the comparison results with KNN are summarized in the supplementary file.

5.2 Comparisons with causal forests

We further compare TDNN with causal forests (CF) over four simulation examples on heterogeneous treatment effect estimation and inference. For each simulation setting, we consider a sample size of $n = 1000$. The first three examples are about estimation accuracy

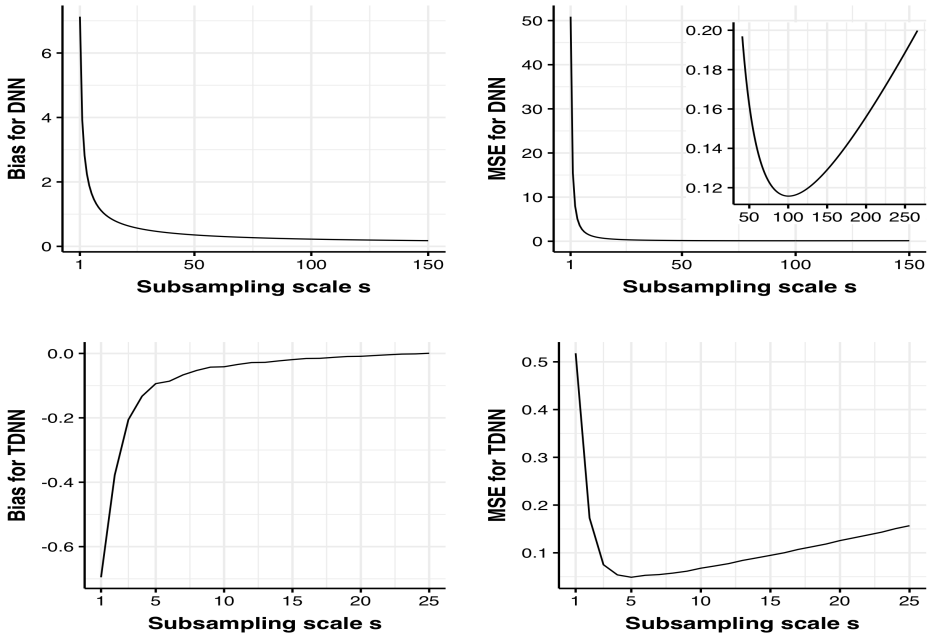

Figure 1: The bias and MSE results for DNN and TDNN in Section 5.1. The top right panel also shows a zoomed-in plot where the U-shaped pattern is more apparent.

and the last one is on confidence interval coverage. Because of the space constraint, we present the simulation results for the first three examples in supplementary file, and only include the fourth example in the current section. We adapt the second simulation setting in Wager and Athey (2018), which was introduced to demonstrate the ability of CF in adapting to the heterogeneity in the treatment effect function $\tau(\mathbf{x})$ in a randomized experiment setting. We hold the treatment propensity $e(\mathbf{x}) = 0.5$ and the main effect $m(\mathbf{x}) = 0$ for the control group fixed, but increase the support of the heterogeneous treatment effect function as $\tau(\mathbf{x}) = \varsigma(x_1)\varsigma(x_2)\varsigma(x_3)$ with $\varsigma(x) = 1 + \{1 + \exp(-20(x - \frac{1}{3}))\}^{-1}$. We set the ambient dimension $p = 20$ and simulate $\mathbf{x} \sim \text{Uniform}([0, 1]^p)$ and model error $\epsilon \sim N(0, 1)$. We will examine the finite-sample coverage of the true value of the heterogeneous treatment effect $\tau(\mathbf{x})$ at some fixed test point with a target coverage probability of 0.95. The fixed test

point is chosen as $x_1 = 0.2$, $x_2 = 0.4$, $x_3 = 0.6$, and $x_j = 0.5$ for $j > 3$.

As discussed at the beginning of this section, 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) 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., 2020) and calculate the corresponding p-value. Then we select features with p-values less than α/p with some significance level $\alpha \in (0, 1)$. We estimate the variance of the TDNN estimator using bootstrap method that has been theoretically justified in Section 4.2. The CF is implemented with the R package `grf` (Tibshirani et al., 2019), which is a generalization of the algorithm introduced in Wager and Athey (2018).

Method	True Mean	Est. Mean (SE)	Bias	MSE	Est. Variance	Coverage	CI Length
TDNN	3.80639	3.81184 (0.00892)	0.00545	0.07947	0.07180	0.93600	1.04211
CF	3.80639	0.18305 (0.00024)	-3.62334	13.12864	0.00008	0.00000	0.03299

Table 1: Comparison of the performance of TDNN and CF in the fourth simulation setting in Section 5.2.

From Table 1 we observe that due to the bias of the CF estimate, its confidence interval estimates fail to cover the true value of the heterogeneous treatment effect. In sharp contrast, TDNN with pre-screening successfully achieves a coverage probability that is very close to the target coverage probability of 0.95, which is in line with the theoretical justifications for TDNN heterogeneous treatment effect inference with bootstrap established formally in Sections 3 and 4. In particular, we see that although the variance estimates

corresponding to TDNN is larger than the one corresponding to CF in our simulation settings, they actually provide more accurate confidence interval coverage. Indeed, TDNN coupled with bootstrap provides a computationally simple yet theoretically justified tool for tight heterogeneous treatment effect inference with appealing finite-sample performance.

6 Real data application

In this section, we demonstrate the performance of TDNN for heterogeneous treatment effect inference on a children's birth weight application. The major goal of this real data application is to characterize the heterogeneity of a treatment effect across subpopulations defined by the values of some continuous covariate. In particular, we aim to study the effect of smoking on a child's birth weight across mothers' ages. We utilize the data set studied originally in Abrevaya et al. (2015) with a kernel-based estimator*.

For our analysis, the feature vector $\mathbf{x}$ includes mother's age, mother's education, father's education, gestation length in weeks, and the number of prenatal visits. We use this subset of covariates because each of these covariates has been shown to have some association with low birth weight (Silvestrin et al., 2013). The response Y is the child's birth weight. The binary treatment indicator T is whether or not the mother smoked during pregnancy. In particular, the data set consists of 591,547 observations in total, 85,976 of whom smoked during pregnancy. We estimate the heterogeneous treatment effects with all the features fixed at the average levels of the corresponding treatment group, except for mother's age which we vary from 16 to 35. This allows us to see how the effect of smoking changes with age, while controlling for all other observed confounders. Similarly as in the simulation examples in Section 5, we also conduct the same analysis using the causal forests (CF) as

*The data set used in our analysis can be found on the research web page of Robert P. Lieli, https://sites.google.com/site/robertplieli/research.

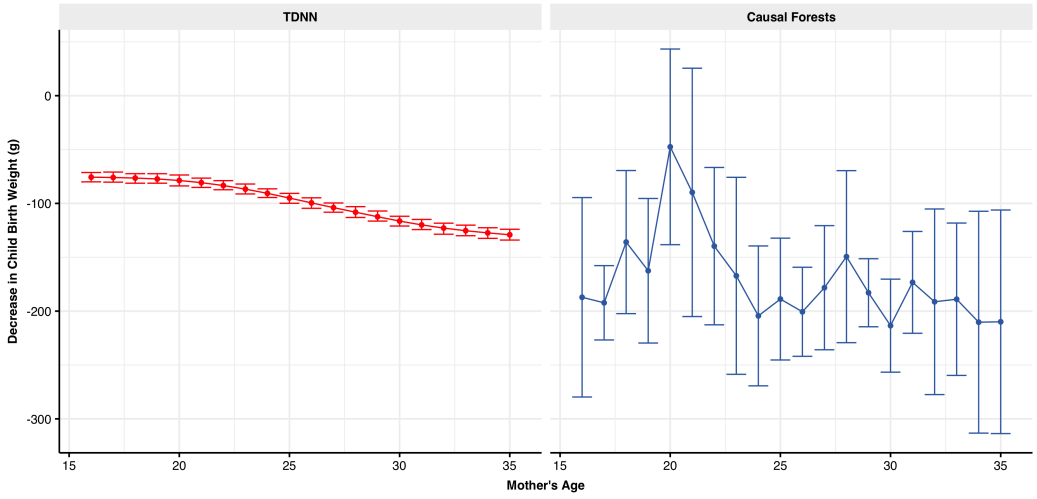

Figure 2: The effect of mother's smoking on a child's birth weight as a function of mother's age. The error bars correspond to 95% confidence intervals.

a popular benchmark for nonparametric heterogeneous treatment effect inference.

Figure 2 presents the heterogeneous treatment effect estimates for both TDNN and CF along with 95% confidence intervals. The confidence intervals for TDNN are generated by bootstrapping the difference between the treatment and control groups, while the confidence intervals for CF are generated using the variance estimation method in the R package `grf`. The results from both procedures suggest that as age increases, the decrease in a newborn's weight associated with a mother's smoking behavior becomes larger. In particular, we see from Figure 2 that the TDNN estimates exhibit a monotone decreasing relationship between mother's age and the heterogeneous treatment effect of smoking with tight confidence intervals. In contrast, the CF estimates show a much more irregular relationship with wide confidence intervals.

7 Discussions

In this paper, we have investigated the problem of heterogeneous treatment effect inference in the setting of randomized experiments. Our suggested method of TDNN alleviates the finite-sample bias issue of the classical k-nearest neighbors and admits easy implementation with simple tuning. The new TDNN tool can enable tight heterogeneous treatment effect inference that is important to identify individualized treatment effects.

Due to the complexities of our new theoretical developments, we have contented ourselves with the setting of randomized experiments and fixed feature dimensionality. It would be interesting to extend the idea of TDNN to the settings of observational studies exploiting the treatment propensity information and of diverging or high feature dimensionality. It would also be interesting to 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 distances that are beyond the Euclidean one and pertinent to specific manifold structures intrinsic to data. It would also be interesting to explore the idea of k-scale DNN with $k \geq 3$ for further bias reduction. These problems are beyond the scope of the current paper and will be interesting topics for future research.

References

Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics 33*, 485–505.

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

Athey, S., G. Imbens, T. Pham, and S. Wager (2017). Estimating average treatment effects:

Supplementary analyses and remaining challenges. *American Economic Review 107*, 278–281.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*, 29–50.

Belloni, A., V. Chernozhukov, F. I, and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica 85*, 233–298.

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.

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

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

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.

Chernozhukov, V., C. Hansen, and M. Spindler (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics 7*, 649–688.

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.

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., K. Imai, H. Liu, Y. Ning, and X. Yang (2016). Improving covariate balancing propensity score: A doubly robust and efficient approach. *Manuscript*.

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*.

Gao, L., Y. Fan, J. Lv, and Q. Shao (2020). Asymptotic distributions of high-dimensional distance correlation inference. *The Annals of Statistics, to appear*.

Györfi, L., M. Kohler, A. Krzyak, 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.

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.

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.

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

Mullainathan, S. and J. Spiess (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives 31*, 87–106.

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.

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.

Silvestrin, S., C. H. da Silva, V. N. Hirakata, A. A. Goldani, P. P. Silveira, and M. Z. Goldani (2013). Maternal education level and low birth weight: a meta-analysis. *Jornal de Pediatria (Versão em Português) 89*, 339–345.

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.

Tibshirani, J., S. Athey, and S. Wager (2019). *grf: Generalized Random Forests*. R package version 0.10.4.

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*.

Supplementary Material to "Nonparametric Inference of Heterogeneous Treatment Effects with Two-Scale Distributional Nearest Neighbors"

Emre Demirkaya, Yingying Fan, Lan Gao, Jinchi Lv,
Patrick Vossler and Jingbo Wang

This Supplementary Material contains additional simulation examples and the proofs of all main results and key lemmas, and some additional technical details.

A Additional Simulation Examples

A.1 Comparison with KNN

We repeat the same simulation study as in Section 5.1 of the main text using the KNN estimator whose performance is depicted in Figure 3. From Figure 3, we see that the finite-sample bias of KNN tends to increase with the neighborhood size k, which is sensible since moving further away from the fixed test point incurs naturally inflated bias. The MSE plot in Figure 3 shows a similar U-shaped pattern of the bias-variance tradeoff. In contrast, the minimum value of the MSE attained by KNN is 0.1250, which is outperformed by both the single-scale DNN and TDNN. It is largely unclear to us whether and how the two-scale idea can be implemented for the KNN estimator, so we opt not to explore it in the current paper.

A.2 Comparison with Causal Forest Estimation

We present three simulation examples which complement the study in Section 5.2 of the main text for heterogeneous treatment effect estimation. While the number of truly im-

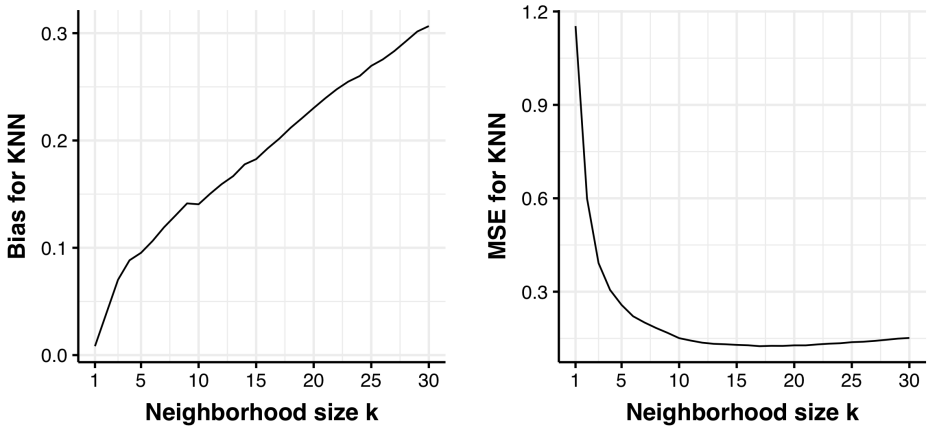

Figure 3: The bias and MSE results for KNN in Section 5.1.

portant predictors is always chosen to be 3, we explore higher ambient dimensionality p for the predictors. We use the same screening method as discussed in the main text for dimension reduction when implementing TDNN. The sample size is fixed at $n = 1000$.

The first simulation setting uses ambient dimensionality $p = 10$. The main effect $m(\mathbf{x})$ for the control group is defined as $m(\mathbf{x}) = x_1^2 + x_2$, and the treatment effect function is $\tau(\mathbf{x}) = (x_1 - 1)^2 + (x_2 + 1)^3 - 3x_3$, where $\mathbf{x} = (x_1, x_2, \cdots, x_p)^T$ and $(\mathbf{x}^T, \epsilon)^T \sim N(\mathbf{0}, I_{p+1})$. We evaluate the performance of both TDNN and CF in terms of the bias and MSE at a fixed test point $(x_1, \cdots, x_p)^T$ chosen as $x_1 = 0.5$, $x_2 = -0.5$, $x_3 = 0.5$, and $x_j = 0$ for $j > 3$.

For the second simulation setting, we investigate the performance of TDNN and CF as we increase the ambient dimensionality. Specifically, for ambient dimensionality $p = 10$, 20, 30, 40, and 50, the data generating process is $\tau(\mathbf{x}) = \left|\log\left[\sum_{j=1}^{3}(x_j^3 - 2x_j^2 + 2x_j)\right]^2\right| + \epsilon$ with $(\mathbf{x}^T, \epsilon)^T \sim N(\mathbf{0}, I_{p+1})$, and we set the main effect $m(\mathbf{x}) = 0$ for the control group. In addition to the fixed test point as chosen in example 1, we also evaluate the performance of both methods at a random test point chosen as $x_j = 0$ for $j > 3$, and $x_j \sim_{i.i.d} N(0, 1)$

for $j = 1$, 2 and 3.

The third simulation setting is a modification of the first simulation setting in Wager and Athey (2018). This simulation setting is designed to test the capability of TDNN and CF in resisting the bias due to an interaction between the treatment propensity $e(\mathbf{x}) = \frac{1}{4}(1 + \beta_{2,4}(x_3))$ and the main effect $m(\mathbf{x}) = x_1^2 + x_2 + x_3^2$ for the control group, where $\beta_{2,4}$ denotes the beta distribution with shape parameters 2 and 4. We choose $\mathbf{x} \sim \text{Uniform}([0,1]^p)$ and $\epsilon \sim N(0,1)$. As a departure from the original simulation setting, we set the population version of the heterogeneous treatment effect $\tau(\mathbf{x}) = x_2$ and ambient dimensionality $p = 20$. We fix $x_j = 0.5$ for $j > 3$, choose $x_1 = 0.2$, $x_2 = 0.4$ and $x_3 = 0.6$ for the fixed test point, and draw the first three components independently from the uniform distribution on the interval $[0,1]$ for the random test point.

The first part in Table 2 presents the comparison results from our first simulation setting, where TDNN produces a considerably more accurate estimate of the heterogeneous treatment effect than CF. The second part in Table 2 summarizes the comparison results as the ambient dimensionality p increases from 10 to 50 in the second simulation setting. It is seen that the extra screening step ensures the robustness of TDNN with respect to the increased number of noise predictors – it is still able to accurately estimate the heterogeneous treatment effect. CF gives a consistently biased estimate, regardless of the ambient dimension. The third part in Table 2 summarizes the results for the third simulation setting and highlights the better robustness of TDNN to the bias caused by an interaction between the treatment propensity and the main effect for the control group in comparison to CF.

B Proofs of main results

B.1 Proof of Theorem 1

Let us investigate the higher-order asymptotic expansion for the bias term of the single-scale distributional nearest neighbors (DNN) estimator $D_n(s)(\mathbf{x})$ introduced in (9) under the asymptotic setting when the subsampling scale $s \to \infty$ as the sample size n increases. Recall that the target point $\mathbf{x}$ is a given vector inside the domain $\mathrm{supp}(\mathbf{X}) \subset \mathbb{R}^d$ of the covariate distribution, where the feature dimensionality d is assumed to be fixed for simplifying the technical presentation of our work. The main idea of the proof is to first consider the specific case of $s = n$ in Lemma 5 in Section B.5, and then analyze the general case of $s \to \infty$ by exploiting the projection of the mean function $\mu(\mathbf{X}) = \mathbb{E}(Y|\mathbf{X})$ onto the positive half line $\mathbb{R}_+ = [0, \infty)$ given by $\|\mathbf{X} - \mathbf{x}\|$ in Lemma 6 in Section B.6.

Since $\{i_1, \cdots, i_s\}$ is a random subsample of $\{1, \cdots, n\}$ with subsampling scale s, in view of (8) and (9) we have

$$\begin{aligned} \mathbb{E}\, D_n(s)(\mathbf{x}) &= \mathbb{E}\, \Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s}) \\ &= \mathbb{E}\, [Y_{(1)}(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s})] \\ &= \mathbb{E}\, [m(r_{(1)})(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s})], \end{aligned} \tag{B.1}$$

where the kernel $\Phi(\mathbf{x}; \cdot)$ in the U-statistic representation of the DNN estimator is simply the 1-nearest neighbor (1NN) estimator $Y_{(1)}(\cdot)$ given by the response for the closest neighbor $\mathbf{X}_{i_{(1)}}$ of $\mathbf{x}$ in the random subsample $\{\mathbf{X}_{i_1}, \cdots, \mathbf{X}_{i_s}\}$ with $\mathbf{Z}_{i_j}$ denoting $(\mathbf{X}_{i_j}, Y_{i_j})$, $m(r) = \mathbb{E}(Y \,|\, \|\mathbf{X} - \mathbf{x}\| = r)$ is the projection of the mean function $\mu(\mathbf{X})$ onto the positive half line introduced in (C.97) in Lemma 6, and $r_{(1)} = \|\mathbf{X}_{i_{(1)}} - \mathbf{x}\|$. The representation in (B.1) provides a useful starting point for our technical analysis.

From (B.1) above, we see that it is necessary to first study the asymptotic behavior of the term $r_{(1)}$. Without loss of generality, for this step we can simply replace parameter

s with parameter n since both subsample size s and full sample size n are assumed to diverge simultaneously. With such a notational simplification, the 1NN $\mathbf{X}_{i_{(1)}}$ of $\mathbf{x}$ in the subsample becomes the 1NN $\mathbf{X}_{(1)}$ of $\mathbf{x}$ in the full sample and thus $r_{(1)} = \|\mathbf{X}_{(1)} - \mathbf{x}\|$. We see from Lemma 5 that $\mathbb{E}r_{(1)}^2 = \mathbb{E}\|\mathbf{X}_{(1)} - \mathbf{x}\|^2$ admits a higher-order asymptotic expansion with explicit constants provided for the first two leading orders, which are respectively $n^{-2/d}$ and $n^{-4/d}$ for $d \geq 2$ as shown in (C.83) and (C.84), and respectively n^{-2} and n^{-3} for $d = 1$ as shown in (C.84). To apply such an asymptotic expansion in Lemma 5 to the term $r_{(1)} = \|\mathbf{X}_{i_{(1)}} - \mathbf{x}\|$ in (B.1), we now need to replace parameter n back with parameter s, which also diverges by assumption.

A natural next step is to consider the expectation on the right-hand side of (B.1) by conditioning on $r_{(1)} = \|\mathbf{X}_{i_{(1)}} - \mathbf{x}\|$. Indeed, this motivates us to investigate the higher-order asymptotic expansion of the projected mean function $m(r) = \mathbb{E}(Y \mid \|\mathbf{X} - \mathbf{x}\| = r)$ in Lemma 6, where $r \to 0$ and some constants are given for the first two leading orders r^2 and r^4 in (C.99). Observe that the asymptotic regime of $r \to 0$ is reasonable since it has been shown by Lemma 2.2 in Biau and Devroye (2015) that $r_{(1)} = \|\mathbf{X}_{i_{(1)}} - \mathbf{x}\| \to 0$ almost surely as $s \to \infty$.

Based on the expansion of $\mathbb{E}\|\mathbf{X}_{(1)} - \mathbf{x}\|^2$ under different regimes of d provided in (C.82)–(C.84) in Lemma 5, we can see that there are two cases for the expansion of $\mathbb{E}\|\mathbf{X}_{(1)} - \mathbf{x}\|^2$. Specifically, the first two leading orders are n^{-2} and n^{-3} for $d = 1$, while the first two leading orders are $n^{-2/d}$ and $n^{-4/d}$ for $d \geq 2$. Thus, we calculate $\mathbb{E}\, D_n(s)(\mathbf{x})$ for $d \geq 2$ and $d = 1$, separately.

First, for the case of $d = 1$, combining the arguments above using (C.82) and Lemma

6, from (B.1) we can deduce that

$$\begin{aligned}
\mathbb{E}\, D_n(s)(\mathbf{x}) &= \mu(\mathbf{x}) + \frac{f(\mathbf{x})\mathrm{tr}(\mu''(\mathbf{x})) + 2\,\mu'(\mathbf{x})^T f'(\mathbf{x})}{2\,d\,f(\mathbf{x})}\,\mathbb{E}r_{(1)}^2 + O_4\mathbb{E}r_{(1)}^4 \\
&= \mu(\mathbf{x}) + \frac{f(\mathbf{x})\mathrm{tr}(\mu''(\mathbf{x})) + 2\,\mu'(\mathbf{x})^T f'(\mathbf{x})}{2\,d\,f(\mathbf{x})} \\
&\quad \times \left(\frac{\Gamma(2/d+1)}{(f(\mathbf{x})V_d)^{2/d}}s^{-2/d} - \left(\frac{\Gamma(2/d+2)}{d(f(\mathbf{x})V_d)^{2/d}}\right)s^{-(1+2/d)}\right) \\
&\quad + O_4\frac{\Gamma(4/d+1)}{(f(\mathbf{x})V_d)^{4/d}}s^{-4/d} + o(s^{-(1+2/d)}) \\
&= \mu(\mathbf{x}) + \Gamma(2/d+1)\frac{f(\mathbf{x})\mathrm{tr}(\mu''(\mathbf{x})) + 2\,\mu'(\mathbf{x})^T f'(\mathbf{x})}{2\,d\,V_d^{2/d}\,f(\mathbf{x})^{2/d+1}}s^{-2/d} + R(s), \qquad \text{(B.2)}
\end{aligned}$$

where $R(s) = A_1(d, f, \mathbf{x})s^{-3} + o(s^{-3})$ with $A_1(d, f, \mathbf{x})$ a bounded quantity that depends only on d, the underlying density function $f(\cdot)$, and regression function $\mu(\cdot)$. In addition, $\Gamma(\cdot)$ denotes the gamma function, $V_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$, $f'(\mathbf{x})$ and $\mu'(\mathbf{x})$ represent the first-order gradients of $f(\mathbf{x})$ and $\mu(\mathbf{x})$ at $\mathbf{x}$, respectively, $\mu''(\mathbf{x})$ denotes the Hessian matrix of $\mu(\cdot)$ at $\mathbf{x}$, O_4 is some constant given in Lemma 6, and $\mathrm{tr}(\cdot)$ stands for the trace operator.

We proceed to prove for the case of $d \geq 2$. In the same fashion of deriving (B.2), applying (C.83)–(C.84) and Lemma 6, from (B.1) we can obtain that

$$\begin{aligned}
\mathbb{E}\, D_n(s)(\mathbf{x}) &= \mu(\mathbf{x}) + \frac{f(\mathbf{x})\mathrm{tr}(\mu''(\mathbf{x})) + 2\,\mu'(\mathbf{x})^T f'(\mathbf{x})}{2\,d\,f(\mathbf{x})}\,\mathbb{E}r_{(1)}^2 + O_4\mathbb{E}r_{(1)}^4 \\
&= \mu(\mathbf{x}) + \frac{f(\mathbf{x})\mathrm{tr}(\mu''(\mathbf{x})) + 2\,\mu'(\mathbf{x})^T f'(\mathbf{x})}{2\,d\,f(\mathbf{x})} \times \left(\frac{\Gamma(2/d+1)}{(f(\mathbf{x})V_d)^{2/d}}s^{-2/d} - C(d, f, \mathbf{x})s^{-4/d}\right) \\
&\quad + O_4\frac{\Gamma(4/d+1)}{(f(\mathbf{x})V_d)^{4/d}}s^{-4/d} + o(s^{-4/d}) \\
&= \mu(\mathbf{x}) + \Gamma(2/d+1)\frac{f(\mathbf{x})\mathrm{tr}(\mu''(\mathbf{x})) + 2\,\mu'(\mathbf{x})^T f'(\mathbf{x})}{2\,d\,V_d^{2/d}\,f(\mathbf{x})^{2/d+1}}s^{-2/d} + R(s), \qquad \text{(B.3)}
\end{aligned}$$

where $R(s) = A_2(d, f, \mathbf{x})s^{-4/d} + o(s^{-4/d})$ with $C(d, f, \mathbf{x})$ and $A_2(d, f, \mathbf{x})$ two bounded quantities that depend only on d, the underlying density function $f(\cdot)$, and regression function $\mu(\cdot)$.

Therefore, combining the above results, we obtain the desired higher-order asymptotic expansion for the bias term of the single-scale DNN estimator $B(s) = \mathbb{E}\, D_n(s)(\mathbf{x}) - \mu(\mathbf{x})$. This completes the proof of Theorem 1.

B.2 Proof of Theorem 2

We now proceed to prove the asymptotic normality of the single-scale DNN estimator $D_n(s)(\mathbf{x})$. Recall that in Theorem 1, the higher-order asymptotic expansion for the bias term $B(s)$ of $D_n(s)(\mathbf{x})$ requires the assumption that the subsampling scale $s \to \infty$ as sample size n increases. As shown in the proof of Theorem 1 in Section B.1, the single-scale DNN estimator $D_n(s)(\mathbf{x})$ reduces to the 1NN estimator when we choose $s = n$, since in such a case, there is a single subsample with size $s = n$, i.e., the full sample. We immediately realize that although the choice of $s = n$ satisfies the need on the bias side, it does not make the variance shrink asymptotically. Intuitively, we would need to form the empirical average over a diverging number of such individual estimates in order to establish the desired asymptotic normality. This naturally calls for the assumption of $s = o(n)$, which entails that the total number of these individual estimates $\binom{n}{s}$ diverges as sample size n increases. Thus we will work with the asymptotic regime of subsampling scale with $s \to \infty$ and $s = o(n)$.

In view of the U-statistic representation of $D_n(s)(\mathbf{x})$ given in (9), a natural idea of the proof for the asymptotic normality of the single-scale DNN estimator is to exploit the asymptotic theory of the U-statistic framework. However, the classical U-statistic asymptotic theory is not readily applicable due to the common assumption of *fixed* subsampling scale s. In contrast, as discussed above, our asymptotic analysis needs the opposite assumption of *diverging* subsampling scale s, i.e., $s \to \infty$. Such a discrepancy causes additional technical challenges when we derive the asymptotic normality.

Let us first exploit Hoeffding's canonical decomposition introduced in (Hoeffding, 1948),

which is an extension of the projection idea. For each $1 \le i \le s$, we define the centered conditional expectation

$$\widetilde{\Phi}_i(\mathbf{x}; \mathbf{z}_1, \cdots, \mathbf{z}_i) = \mathbb{E}[\Phi(\mathbf{x}; \mathbf{z}_1, \cdots, \mathbf{z}_i, \mathbf{Z}_{i+1}, \cdots, \mathbf{Z}_s) \,|\, \mathbf{z}_1, \cdots, \mathbf{z}_i] - \mathbb{E}\Phi(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_s), \tag{B.4}$$

where $\Phi(\mathbf{x}; \cdot)$ is the kernel defined in (8) for the U-statistic representation of the single-scale DNN estimator. Then in light of (B.4), for each $1 \le i \le s$ we can successively define the canonical term

$$g_i(\mathbf{x}; \mathbf{z}_1, \cdots, \mathbf{z}_i) = \widetilde{\Phi}_i(\mathbf{x}; \mathbf{z}_1, \cdots, \mathbf{z}_i) - \sum_{j=1}^{i-1} \sum_{1 \le \alpha_1 < \cdots < \alpha_j \le i} g_j(\mathbf{x}; \mathbf{z}_{\alpha_1}, \cdots, \mathbf{z}_{\alpha_j}), \tag{B.5}$$

where $g_1(\mathbf{x}; \mathbf{z}_1) = \widetilde{\Phi}_1(\mathbf{x}; \mathbf{z}_1)$ by definition. Combining (8), (B.4), and (B.5), we see that the kernel $\Phi(\mathbf{x}; \cdot)$ can be rewritten as a sum of the canonical terms

$$\Phi(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_s) - \mathbb{E}\Phi(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_s) = \sum_{j=1}^{s} \sum_{1 \le \alpha_1 < \cdots < \alpha_j \le s} g_j(\mathbf{x}; \mathbf{Z}_{\alpha_1}, \cdots, \mathbf{Z}_{\alpha_j}). \tag{B.6}$$

Moreover, it holds that

$$\mathrm{Var}(\Phi(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_s)) = \sum_{j=1}^{s} \binom{s}{j} \mathrm{Var}(g_j(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_j)). \tag{B.7}$$

The above Hoeffding's canonical decomposition in (B.6) plays an important role in establishing the asymptotic normality.

In view of (9), (B.4), and (B.6), we can deduce that

$$\begin{aligned} D_n(s) - \mathbb{E}\, D_n(s) &= \binom{n}{s}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_s \le n} \widetilde{\Phi}_s(\mathbf{x}; \mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s}) \\ &= \binom{n}{s}^{-1} \Big\{ \binom{n-1}{s-1} \sum_{i_1=1}^{n} g_1(\mathbf{x}; \mathbf{Z}_{i_1}) + \binom{n-2}{s-2} \sum_{1 \le i_1 < i_2 \le n} g_2(\mathbf{x}; \mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}) + \cdots \\ &\quad + \binom{n-s}{s-s} \sum_{1 \le i_1 < i_2 < \cdots < i_s \le n} g_s(\mathbf{x}; \mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s}) \Big\}. \end{aligned} \tag{B.8}$$

From the above Hoeffding's canonical decomposition in (B.8) for the single-scale DNN estimator, we see that the Hájek projection introduced in (Hájek, 1968) of the centered DNN estimator $D_n(s) - \mathbb{E}D_n(s)$ is given by

$$\widehat{D}_n(s) = \binom{n}{s}^{-1} \binom{n-1}{s-1} \sum_{i=1}^{n} g_1(\mathbf{x}; \mathbf{Z}_i), \tag{B.9}$$

which is the first-order part of the decomposition in (B.8).

A useful observation is that the Hájek projection given in (B.9) involves the sum of some independent and identically distributed (i.i.d.) terms. Denote by σ_n^2 the variance of the Hájek projection. Then it follows from $g_1(\mathbf{x}; \mathbf{z}_1) = \widetilde{\Phi}_1(\mathbf{x}; \mathbf{z}_1)$ and (B.4) that

$$\begin{aligned} \sigma_n^2 = \operatorname{Var}(\widehat{D}_n(s)) &= \frac{s^2}{n} \operatorname{Var}(\widetilde{\Phi}_1(\mathbf{x}; \mathbf{Z}_1)) \\ &= \frac{s^2}{n} \operatorname{Var}(\Phi_1(\mathbf{x}; \mathbf{Z}_1)) = \frac{s^2}{n} \eta_1, \end{aligned} \tag{B.10}$$

where the non-centered conditional expectation $\Phi_1(\mathbf{x}; \mathbf{Z}_1)$ is defined later in (C.114) and η_1 is defined as the variance of $\Phi_1(\mathbf{x}; \mathbf{Z}_1)$. From (B.4), we see that each term $g_1(\mathbf{x}; \mathbf{Z}_i) = \widetilde{\Phi}_1(\mathbf{x}; \mathbf{Z}_i)$ of the i.i.d. sum in (B.9) has zero mean. Thus by (B.10), an application of the Lindeberg–Lévy central limit theorem in (Borovkov, 2013) leads to

$$\frac{\widehat{D}_n(s)}{\sigma_n} \xrightarrow{\mathscr{D}} N(0, 1), \tag{B.11}$$

which establishes the asymptotic normality of the Hájek projection $\widehat{D}_n(s)$.

Finally, we aim to show that similar asymptotic normality as above holds when the Hájek projection $\widehat{D}_n(s)$ in the numerator on the left-hand side of (B.11) is replaced with the centered single-scale DNN estimator $D_n(s)(\mathbf{x}) - \mathbb{E}\, D_n(s)(\mathbf{x}) = D_n(s)(\mathbf{x}) - \mu(\mathbf{x}) - B(s)$, where $B(s)$ is the bias term identified in Theorem 1. With the aid of the Slutsky's lemma, we see that it suffices to show that

$$\frac{D_n(s) - \mathbb{E}D_n(s) - \widehat{D}_n(s)}{\sigma_n} = o_P(1). \tag{B.12}$$

In view of (B.7) and Hoeffding's canonical decomposition in (B.8), we can deuce that

$$\begin{aligned}
\mathbb{E}[D_n(s) - \mathbb{E}D_n(s) - \widehat{D}_n(s)]^2 &= \binom{n}{s}^{-2}\Big\{\binom{n-2}{s-2}^2\binom{n}{2}\mathbb{E}\left(g_2(\mathbf{x};\mathbf{Z}_1,\mathbf{Z}_2)\right)^2 + \\
&\cdots + \binom{n-s}{s-s}^2\binom{n}{s}\mathbb{E}\left(g_s(\mathbf{x};\mathbf{Z}_1,\cdots,\mathbf{Z}_s)\right)^2\Big\} \\
&= \sum_{r=2}^{s}\Big\{\binom{n}{s}^{-2}\binom{n-r}{s-r}^2\binom{n}{r}\mathbb{E}\left(g_r(\mathbf{x};\mathbf{Z}_1,\cdots,\mathbf{Z}_r)\right)^2\Big\} \\
&= \sum_{r=2}^{s}\Big\{\frac{s!(n-r)!}{n!(s-r)!}\binom{s}{r}\mathbb{E}\left(g_r(\mathbf{x};\mathbf{Z}_1,\cdots,\mathbf{Z}_r)\right)^2\Big\} \\
&\le \frac{s(s-1)}{n(n-1)}\sum_{r=2}^{s}\binom{s}{r}\mathbb{E}\left(g_r(\mathbf{x};\mathbf{Z}_1,\cdots,\mathbf{Z}_r)\right)^2 \\
&\le \frac{s^2}{n^2}\mathrm{Var}(\Phi(\mathbf{x};\mathbf{Z}_1,\cdots,\mathbf{Z}_s)). \qquad \text{(B.13)}
\end{aligned}$$

It remains to bound the variance term $\mathrm{Var}(\Phi(\mathbf{x};\mathbf{Z}_1,\cdots,\mathbf{Z}_s))$ above.

By Lemma 7 in Section B.7, we have an important result that

$$\mathrm{Var}(\Phi(\mathbf{x};\mathbf{Z}_1,\cdots,\mathbf{Z}_s)) = o(n\eta_1). \qquad \text{(B.14)}$$

Combining (B.10), (B.13), and (B.14), it holds that

$$\begin{aligned}
\mathbb{E}\left[\frac{D_n(s) - \mathbb{E}D_n(s) - \widehat{D}_n(s))}{\sigma_n}\right]^2 &= o\left\{\frac{1}{\sigma_n^2}\frac{s^2}{n^2}(n\eta_1)\right\} \\
&= o\left\{\frac{n}{s^2\eta_1}\frac{s^2}{n^2}(n\eta_1)\right\} = o(1). \qquad \text{(B.15)}
\end{aligned}$$

Therefore, we are ready to see that (B.15) entails the desired claim (B.12). Finally, by (B.10) and (C.119) obtained in the proof of Lemma 7 in Section B.7, we see that σ_n is of order $(s/n)^{1/2}$, which concludes the proof of Theorem 2.

B.3 Proof of Theorem 3

We further prove the asymptotic normality of the two-scale DNN estimator $D_n(s_1, s_2)(\mathbf{x})$ introduced in (15). It is worth mentioning that Theorem 3 is not a simple consequence

of Theorem 2 since the marginal asymptotic normalities do not necessarily lead to the joint asymptotic normality. This means that we need to analyze the two single-scale DNN estimators involved in the definition of the two-scale DNN estimator in a joint fashion. To this end, we will exploit the ideas in the proof of Theorem 2 in Section B.2. To facilitate the technical analysis, some key technical tools are provided in Lemmas 8–10 in Sections B.8–B.10, respectively.

Without loss of generality, let us assume that $s_1 < s_2$ for the two subsampling scales. In particular, we make the assumptions that $s_1, s_2 \to \infty$, $s_1, s_2 = o(n)$, and $c_1 \le s_1/s_2 \le c_2$ for some constants $0 < c_1 < c_2 < 1$. From Lemma 8 in Section B.8, we see that the two-scale DNN estimator $D_n(s_1, s_2)(\mathbf{x})$ is also a U-statistic of order s_2 with a new kernel $\Phi^*(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{s_2})$ introduced later in (C.121). Thus Hoeffding's canonical decomposition for U-statistics can be applied to derive the asymptotic normality of the two-scale DNN estimator. For each $1 \le i \le s_2$, let us define

$$\Phi_i^*(\mathbf{x}; \mathbf{z}_1, \cdots, \mathbf{z}_i) = \mathbb{E}[\Phi^*(\mathbf{x}; \mathbf{z}_1, \cdots, \mathbf{z}_i, \mathbf{Z}_{i+1}, \cdots, \mathbf{Z}_{s_2}) \,|\, \mathbf{z}_1, \cdots, \mathbf{z}_i], \tag{B.16}$$

$$\begin{aligned} g_i^*(\mathbf{x}; \mathbf{z}_1, \cdots, \mathbf{z}_i) &= \Phi_i^*(\mathbf{x}; \mathbf{z}_1, \cdots, \mathbf{z}_i) - \mathbb{E}\Phi_i^*(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_i) \\ &\quad - \sum_{j=1}^{i-1} \sum_{1 \le \alpha_1 < \cdots < \alpha_j \le i} g_j^*(\mathbf{x}; \mathbf{z}_{\alpha_1}, \cdots, \mathbf{z}_{\alpha_j}), \end{aligned} \tag{B.17}$$

where $g_1^*(\mathbf{x}; \mathbf{z}_1) = \Phi_1^*(\mathbf{x}; \mathbf{z}_1) - \mathbb{E}\Phi_1^*(\mathbf{x}; \mathbf{Z}_1)$ by definition. We further define

$$\operatorname{Var}\Phi^* = \operatorname{Var}(\Phi^*(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{s_2})) \quad \text{and} \quad \eta_1^* = \operatorname{Var}(\Phi_1^*(\mathbf{x}; \mathbf{Z}_1)). \tag{B.18}$$

In view of (15), (B.16), and (B.17), an application of similar U-statistic and Hoeffding's canonical decomposition arguments to those in the proof of Theorem 2 in Section B.2 entails that

$$(n^{-1}s_2^2\eta_1^*)^{-1/2}\big(D_n(s_1, s_2)(\mathbf{x}) - \mathbb{E}[D_n(s_1, s_2)(\mathbf{x})]\big) \tag{B.19}$$

can be approximated by the first-order part of Hoeffding's canonical decomposition that converges to a normal distribution with the remainders asymptotically negligible, where η_1^* is given in (B.18). More specifically, denote by

$$\widehat{D}_n(s_1, s_2) = \frac{s_2}{n} \sum_{i=1}^{n} g_1^*(\mathbf{x}; \mathbf{Z}_i), \tag{B.20}$$

where $g_1^*(\mathbf{x}; \mathbf{Z}_i)$ is defined in (B.17). It follows from (B.18), (B.20), and the classical central limit theorem for i.i.d. random variables that

$$\frac{\widehat{D}_n(s_1, s_2)}{\sqrt{n^{-1} s_2^2 \eta_1^*}} \xrightarrow{\mathscr{D}} N(0, 1), \tag{B.21}$$

since it holds that $\mathrm{Var}(g_1^*(\mathbf{x}; \mathbf{Z}_1)) = \mathrm{Var}(\Phi_1^*(\mathbf{x}; \mathbf{Z}_1)) = \eta_1^*$.

Similar to (B.13), by (B.17), (B.18), and (B.20) we can deduce that

$$\begin{aligned} &\frac{\mathbb{E}\big[D_n(s_1, s_2)(\mathbf{x}) - \mathbb{E}D_n(s_1, s_2)(\mathbf{x}) - \widehat{D}_n(s_1, s_2)\big]^2}{n^{-1} s_2^2 \eta_1^*} \\ &\leq \frac{n^{-2} s_2^2 \,\mathrm{Var}\, \Phi^*}{n^{-1} s_2^2 \eta_1^*} = \frac{\mathrm{Var}\, \Phi^*}{n \eta_1^*}. \end{aligned} \tag{B.22}$$

Moreover, it follows from the upper bound on $\mathrm{Var}\, \Phi^*$ obtained in Lemma 9 in Section B.9 and the asymptotic order of η_1^* established in Lemma 10 in Section B.10 that

$$\mathrm{Var}\, \Phi^* / (n \eta_1^*) \to 0 \tag{B.23}$$

since $s_2/n \to 0$ by assumption. Therefore, combining (B.21)–(B.23), an application of the Slutsky's lemma yields the desired claim in (B.19), that is,

$$\frac{D_n(s_1, s_2)(\mathbf{x}) - \mathbb{E}D_n(s_1, s_2)(\mathbf{x})}{\sigma_n} \xrightarrow{\mathscr{D}} N(0, 1), \tag{B.24}$$

where we define $\sigma_n^2 = n^{-1} s_2^2 \eta_1^*$. Finally, we see from Lemma 10 that $\sigma_n = (n^{-1} s_2^2 \eta_1^*)^{1/2}$ is of order $(s_2/n)^{1/2}$, and from the higher-order asymptotic expansion of the bias term in Theorem 1 that

$$\Lambda = \mathbb{E}D_n(s_1, s_2)(\mathbf{x}) - \mu(\mathbf{x}) = \begin{cases} O(s_1^{-4/d} + s_2^{-4/d}), & d \geq 2, \\ O(s_1^{-3} + s_2^{-3}), & d = 1. \end{cases}$$

This together with (B.24) completes the proof of Theorem 3.

B.4 Proof of Theorem 4

We now aim to prove the asymptotic normality of the HTE estimator $\widehat{\tau}(\mathbf{x}) = D_{n_1}^{(1)}(s_1^{(1)}, s_2^{(1)})(\mathbf{x}) - D_{n_0}^{(0)}(s_1^{(0)}, s_2^{(0)})(\mathbf{x})$ introduced in (23), where $D_{n_1}^{(1)}(s_1^{(1)}, s_2^{(1)})(\mathbf{x})$ and $D_{n_0}^{(0)}(s_1^{(0)}, s_2^{(0)})(\mathbf{x})$ denote the two-scale DNN estimators constructed using the treatment sample of size n_1 and the control sample of size n_0, respectively. Denote by $n = n_0 + n_1$ the total sample size. By the assumption $P(T = 1|\mathbf{X}, Y_{T=0}, Y_{T=1}) = 1/2$, it is easy to see that $n_0/n_1 \xrightarrow{\mathrm{P}} 1$ as $n \to \infty$. For each of the treatment and control groups in the randomized experiment, by the assumptions a separate application of Theorem 3 shows that there exist some positive numbers σ_{n_1} of order $(s_2^{(1)}/n_1)^{1/2}$ and σ_{n_0} of order $(s_2^{(0)}/n_0)^{1/2}$ such that

$$\frac{D_{n_1}^{(1)}(s_1^{(1)}, s_2^{(1)})(\mathbf{x}) - \mathbb{E}[D_{n_1}^{(1)}(s_1^{(1)}, s_2^{(1)})(\mathbf{x})]}{\sigma_{n_1}} \xrightarrow{\mathscr{D}} N(0,1) \tag{B.25}$$

and

$$\frac{D_{n_0}^{(0)}(s_1^{(0)}, s_2^{(0)})(\mathbf{x}) - \mathbb{E}[D_{n_0}^{(0)}(s_1^{(0)}, s_2^{(0)})(\mathbf{x})]}{\sigma_{n_0}} \xrightarrow{\mathscr{D}} N(0,1). \tag{B.26}$$

In view of the randomized experiment assumption, the treatment sample and control sample are independent of each other, which entails that the two separate two-scale DNN estimators $D_{n_1}^{(1)}(s_1^{(1)}, s_2^{(1)})(\mathbf{x})$ and $D_{n_0}^{(0)}(s_1^{(0)}, s_2^{(0)})(\mathbf{x})$ are independent. Thus it follows from (B.25) and (B.26) that

$$\frac{D_{n_1}^{(1)}(s_1^{(1)}, s_2^{(1)})(\mathbf{x}) - D_{n_0}^{(0)}(s_1^{(0)}, s_2^{(0)})(\mathbf{x}) - \mathbb{E}[D_{n_1}^{(1)}(s_1^{(1)}, s_2^{(1)})(\mathbf{x}) - D_{n_0}^{(0)}(s_1^{(0)}, s_2^{(0)})(\mathbf{x})]}{\sigma_n}$$
$$\xrightarrow{\mathscr{D}} N(0,1), \tag{B.27}$$

where we define $\sigma_n = (\sigma_{n_1}^2 + \sigma_{n_0}^2)^{1/2}$. Moreover, from the higher-order asymptotic expansion of the bias term in Theorem 1 applied to the potential treatment and control responses,

respectively, and the definition of the heterogeneous treatment effect (HTE) $\tau(\mathbf{x})$ introduced in (22), we see that

$$\mathbb{E}[D_{n_1}^{(1)}(s_1^{(1)}, s_2^{(1)})(\mathbf{x})] - \mathbb{E}[D_{n_0}^{(0)}(s_1^{(0)}, s_2^{(0)})(\mathbf{x})] = \tau(\mathbf{x}) + \Lambda, \tag{B.28}$$

where $\Lambda = O\{(s_1^{(1)})^{-4/d} + (s_2^{(1)})^{-4/d} + (s_1^{(0)})^{-4/d} + (s_2^{(0)})^{-4/d}\}$ for $d \geq 2$ and $\Lambda = O\{(s_1^{(1)})^{-3} + (s_2^{(1)})^{-3} + (s_1^{(0)})^{-3} + (s_2^{(0)})^{-3}\}$ for $d = 1$. Therefore, combining (B.27) and (B.28) yields the desired asymptotic normality of the HTE estimator $\widehat{\tau}(\mathbf{x})$ based on the two-scale DNN estimators. This concludes the proof of Theorem 4.

B.5 Proof of Theorem 5

In view of the arguments in the proof of Theorem 4 in Section B.4, the variance of the HTE estimator $\widehat{\tau}(\mathbf{x}) = D_{n_1}^{(1)}(s_1^{(1)}, s_2^{(1)})(\mathbf{x}) - D_{n_0}^{(0)}(s_1^{(0)}, s_2^{(0)})(\mathbf{x})$ is naturally the sum of the variances of the pair of two-scale DNN estimators. Thus the variance estimate for the HTE estimator reduces to that for the two-scale DNN estimator $D_n(s_1, s_2)(\mathbf{x})$, whose asymptotic normality has been shown in Theorem 3. Now we aim to establish the consistency of the jackknife estimator $\widehat{\sigma}_J^2$ introduced in (28) for the variance σ^2 of the two-scale DNN estimator $D_n(s_1, s_2)(\mathbf{x})$ as defined in (26). We will build on the technique in Arvesen (1969) that expands and reorganizes the jackknife estimator $\widehat{\sigma}_J^2$. However, a major theoretical challenge is that instead of an application of the classical asymptotic theory for the case of fixed order, a more delicate technical analysis of the remainders is essential to proving the consistency under our current assumption of diverging order $s_2 \to \infty$.

More specifically, we will show that the jackknife estimator $\widehat{\sigma}_J^2$ can be written as a weighted sum of a sequence of U-statistics $\{U_c\}_{0 \leq c \leq s_2}$ to be introduced in (B.34) later, where U_0 and U_1 are the dominating terms and the remaining ones are asymptotically negligible under the assumption of $s_2 = o(n^{1/3})$. Since U-statistics are symmetric with

respect to the input arguments, it follows from (25) and (27) that

$$\sum_{i=1}^{n} \binom{n-1}{s_2} U_{n-1}^{(i)} = (n-s_2)\binom{n}{s_2} D_n(s_1, s_2)(\mathbf{x}),$$

which entails that

$$n^{-1} \sum_{i=1}^{n} U_{n-1}^{(i)} = D_n(s_1, s_2)(\mathbf{x}). \tag{B.29}$$

Thus, in light of the definition of the jackknife estimator $\widehat{\sigma}_J^2$ in (28) and (B.29), we can deduce that

$$\begin{aligned} n\widehat{\sigma}_J^2 &= (n-1)\Big\{ \sum_{i=1}^{n} \big(U_{n-1}^{(i)}\big)^2 - n(D_n(s_1, s_2)(\mathbf{x}))^2 \Big\} \\ &= (n-1)\Big\{ \binom{n-1}{s_2}^{-2} \sum_{i=1}^{n} \sum_{i} \Phi^*(\mathbf{x}; \mathbf{Z}_{\alpha_1^i}, \cdots, \mathbf{Z}_{\alpha_{s_2}^i}) \Phi^*(\mathbf{x}; \mathbf{Z}_{\beta_1^i}, \cdots, \mathbf{Z}_{\beta_{s_2}^i}) \\ &\quad - n\binom{n}{s_2}^{-2} \sum \Phi^*(\mathbf{x}; \mathbf{Z}_{\alpha_1}, \cdots, \mathbf{Z}_{\alpha_{s_2}}) \Phi^*(\mathbf{x}; \mathbf{Z}_{\beta_1}, \cdots, \mathbf{Z}_{\beta_{s_2}}) \Big\}, \end{aligned} \tag{B.30}$$

where we use the shorthand notation $\sum_i$ for

$$\sum_{\substack{1 \le \alpha_1^i < \alpha_2^i < \cdots < \alpha_{s_2}^i \le n \\ 1 \le \beta_1^i < \beta_2^i < \cdots < \beta_{s_2}^i \le n \\ \alpha_1^i, \alpha_2^i, \cdots, \alpha_{s_2}^i \ne i;\, \beta_1^i, \beta_2^i, \cdots, \beta_{s_2}^i \ne i}} \tag{B.31}$$

and $\sum$ for

$$\sum_{\substack{1 \le \alpha_1 < \alpha_2 < \cdots < \alpha_{s_2} \le n \\ 1 \le \beta_1 < \beta_2 < \cdots < \beta_{s_2} \le n}} \tag{B.32}$$

to simplify the technical presentation.

For each $0 \le c \le s_2$, by calculating the number of terms with c overlapping components

in $\Phi^*(\mathbf{x}; \mathbf{Z}_{\alpha_1}, \cdots, \mathbf{Z}_{\alpha_{s_2}})\Phi^*(\mathbf{x}; \mathbf{Z}_{\beta_1}, \cdots, \mathbf{Z}_{\beta_{s_2}})$, we can obtain from (B.30)–(B.32) that

$$\begin{aligned} n\widehat{\sigma}_J^2 = (n-1)\Bigg\{ & \binom{n-1}{s_2}^{-2} \sum_{c=0}^{s_2} (n - 2s_2 + c) \sum \Phi^*(\mathbf{x}; \mathbf{Z}_{\alpha_1}, \cdots, \mathbf{Z}_{\alpha_c}, \mathbf{Z}_{\beta_1}, \cdots, \mathbf{Z}_{\beta_{s_2-c}}) \\ & \cdot \Phi^*(\mathbf{x}; \mathbf{Z}_{\alpha_1}, \cdots, \mathbf{Z}_{\alpha_c}, \mathbf{Z}_{\gamma_1}, \cdots, \mathbf{Z}_{\gamma_{s_2-c}}) \\ & - n\binom{n}{s_2}^{-2} \sum_{c=0}^{s_2} \sum \Phi^*(\mathbf{x}; \mathbf{Z}_{\alpha_1}, \cdots, \mathbf{Z}_{\alpha_c}, \mathbf{Z}_{\beta_1}, \cdots, \mathbf{Z}_{\beta_{s_2-c}}) \\ & \cdot \Phi^*(\mathbf{x}; \mathbf{Z}_{\alpha_1}, \cdots, \mathbf{Z}_{\alpha_c}, \mathbf{Z}_{\gamma_1}, \cdots, \mathbf{Z}_{\gamma_{s_2-c}}) \Bigg\} \\ = & \frac{n-1}{n}\binom{n-1}{s_2}^{-2} \sum_{c=0}^{s_2} (cn - s_2^2)\binom{n}{2s_2 - c}\binom{2s_2 - c}{s_2}\binom{s_2}{c} U_c, \end{aligned} \tag{B.33}$$

where we introduce a sequence of U-statistics $\{U_c\}_{0 \le c \le s_2}$ defined as

$$\begin{aligned} U_c = & \left\{ \binom{n}{2s_2 - c}\binom{2s_2 - c}{s_2}\binom{s_2}{c} \right\}^{-1} \\ & \cdot \sum \Phi^*(\mathbf{x}; \mathbf{Z}_{\alpha_1}, \cdots, \mathbf{Z}_{\alpha_c}, \mathbf{Z}_{\beta_1}, \cdots, \mathbf{Z}_{\beta_{s_2-c}}) \Phi^*(\mathbf{x}; \mathbf{Z}_{\alpha_1}, \cdots, \mathbf{Z}_{\alpha_c}, \mathbf{Z}_{\gamma_1}, \cdots, \mathbf{Z}_{\gamma_{s_2-c}}). \end{aligned} \tag{B.34}$$

Here, with slight abuse of notation, $\sum$ is short for denoting the summation over all possible combinations of distinct $\alpha_1, \cdots, \alpha_c, \beta_1, \cdots, \beta_{s_2-c}, \gamma_1, \cdots, \gamma_{s_2-c}$ satisfying that $1 \le \alpha_1 < \cdots < \alpha_c \le n$, $1 \le \beta_1 < \cdots < \beta_{s_2-c} \le n$, and $1 \le \gamma_1 < \cdots < \gamma_{s_2-c} \le n$.

Observe that by symmetrization, U_c defined in (B.34) is indeed a U-statistic that can be represented as

$$U_c = \binom{n}{2s_2 - c}^{-1} \sum_{C_{2s_2-c}} K^{(c)}(\mathbf{x}; \mathbf{Z}_{\alpha_1}, \cdots, \mathbf{Z}_{\alpha_c}, \mathbf{Z}_{\beta_1}, \cdots, \mathbf{Z}_{\beta_{s_2-c}}, \mathbf{Z}_{\gamma_1}, \cdots, \mathbf{Z}_{\gamma_{s_2-c}}), \tag{B.35}$$

where $\sum_{C_{2s_2-c}}$ represents the summation taken over all combinations of $1 \le \alpha_1 < \cdots < \alpha_c < \beta_1 < \cdots < \beta_{s_2-c} < \gamma_1 < \cdots < \gamma_{s_2-c} \le n$, and the symmetrized kernel function $K^{(c)}$ is given

by

$$
\begin{aligned}
&K^{(c)}(\mathbf{x}; \mathbf{Z}_{\alpha_1}, \cdots, \mathbf{Z}_{\alpha_c}, \mathbf{Z}_{\beta_1}, \cdots, \mathbf{Z}_{\beta_{s_2-c}}, \mathbf{Z}_{\gamma_1}, \cdots, \mathbf{Z}_{\gamma_{s_2-c}}) \\
&= \left\{ \binom{2s_2 - c}{c} \binom{2s_2 - 2c}{s_2 - c} \right\}^{-1} \sum_{\prod_{2s_2-c}} \Phi^*(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_c}, \mathbf{Z}_{i_{c+1}}, \cdots, \mathbf{Z}_{i_{s_2}}) \\
&\quad \cdot \Phi^*(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_c}, \mathbf{Z}_{i_{s_2+1}}, \cdots, \mathbf{Z}_{i_{2s_2-c}}) \qquad \text{(B.36)}
\end{aligned}
$$

with $\sum_{\prod_{2s_2-c}}$ standing for the summation over all the $\binom{2s_2-c}{c}\binom{2s_2-2c}{s_2-c}$ possible permutations of $(\alpha_1, \cdots, \alpha_c, \beta_1, \cdots, \beta_{s_2-c}, \gamma_1, \cdots, \gamma_{s_2-c})$ that are not permuted within sets $(\alpha_1, \cdots, \alpha_c)$, $(\beta_1, \cdots, \beta_{s_2-c})$, and $(\gamma_1, \cdots, \gamma_{s_2-c})$.

From (B.33)–(B.36) above, we can further deduce that as long as $s_2 = o(\sqrt{n})$, it holds that

$$
\begin{aligned}
n\widehat{\sigma}_J^2 &= \sum_{c=0}^{s_2} (cn - s_2^2) \frac{(n - s_2 - 1)(n - s_2 - 2) \cdots (n - 2s_2 + c + 1)}{(n-2)(n-3)\cdots(n - s_2)c!} \\
&\quad \cdot [s_2(s_2 - 1) \cdots (s_2 - c + 1)]^2 U_c \\
&= -s_2^2 \Big[1 + O\big(\frac{s_2^2}{n}\big)\Big] U_0 + s_2^2 [1 + O\big(\frac{s_2^2}{n}\big)] U_1 + \sum_{c=2}^{s_2} O\big(\frac{s_2^2}{n}\big)^{c-1} \frac{s_2^2}{c!} U_c \\
&= s_2^2 (U_1 - U_0) + O\big(\frac{s_2^4}{n}\big)(U_0 + U_1) + \sum_{c=2}^{s_2} O\big(\frac{s_2^2}{n}\big)^{c-1} \frac{s_2^2}{c!} U_c,
\end{aligned}
$$

which leads to

$$
\frac{n}{s_2^2} \widehat{\sigma}_J^2 = U_1 - U_0 + O\big(\frac{s_2^2}{n}\big)(U_0 + U_1) + \sum_{c=2}^{s_2} O\big(\frac{s_2^2}{n}\big)^{c-1} \frac{U_c}{c!}. \qquad \text{(B.37)}
$$

By (B.34), for the mean we have

$$
\begin{aligned}
\mathbb{E} U_c &= \mathbb{E}\big[\Phi^*(\mathbf{x}; \mathbf{Z}_{\alpha_1}, \cdots, \mathbf{Z}_{\alpha_c}, \mathbf{Z}_{\beta_1}, \cdots, \mathbf{Z}_{\beta_{s_2-c}}) \Phi^*(\mathbf{x}; \mathbf{Z}_{\alpha_1}, \cdots, \mathbf{Z}_{\alpha_c}, \mathbf{Z}_{\gamma_1}, \cdots, \mathbf{Z}_{\gamma_{s_2-c}})\big] \\
&= \mathbb{E}\big([\Phi_c^*(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_c)]^2\big), \qquad \text{(B.38)}
\end{aligned}
$$

where $\Phi_c^*(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_c) = \mathbb{E}[\Phi^*(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2}) | \mathbf{Z}_1, \cdots, \mathbf{Z}_c]$.

As for the variance, it follows from Lemmas 2 and 3 in Sections B.2 and B.3, respectively, that for each $0 \le c \le s_2$ and fixed $\mathbf{x}$, we have

$$\mathrm{Var}(U_c) = O(s_2/n). \tag{B.39}$$

Moreover, in view of (B.38) and Jensen's inequality, it holds that for each $2 \le c \le s_2$,

$$\mathbb{E}U_c \le \mathbb{E}[(\Phi^*)^2]. \tag{B.40}$$

Consequently, it follows from (B.37)–(B.40) that

$$\begin{aligned}
&\mathbb{E}\Big(\Big[\frac{n}{s_2^2}\widehat{\sigma}_J^2 - \mathrm{Var}(\Phi_1^*(\mathbf{x};\mathbf{Z}_1))\Big]^2\Big) \\
&\le C\Big\{\mathrm{Var}(U_1) + \mathrm{Var}(U_0) + \frac{s_2^4}{n^2}\big[(\mathbb{E}\Phi^*)^4 + \big(\mathbb{E}[\Phi_1^*(\mathbf{x};\mathbf{Z}_1)]^2\big)^2\big] \\
&\quad + \sum_{j=2}^{s_2}\sum_{i=2}^{s_2}\Big(\frac{s_2^2}{n}\Big)^{i+j-2}\big[\mathrm{Var}(U_i) + (\mathbb{E}(\Phi^*)^2)^2\big]^{1/2}\big[\mathrm{Var}(U_j) + (\mathbb{E}(\Phi^*)^2)^2\big]^{1/2}\Big\},
\end{aligned} \tag{B.41}$$

where C is some positive constant. Recall the facts that $\mathbb{E}[\Phi^*] = O(1)$ and $\mathbb{E}[(\Phi^*)^2] = O(1)$, which have been shown previously in the proof of Theorem 3 in Section B.3. Combining (B.41) with these facts yields

$$\mathbb{E}\Big(\Big[\frac{n}{s_2^2}\widehat{\sigma}_J^2 - \mathrm{Var}(\Phi_1^*(\mathbf{x};\mathbf{Z}_1))\Big]\Big)^2 \le C\Big(\frac{s_2}{n} + \frac{s_2^4}{n^2}\Big). \tag{B.42}$$

Furthermore, it has been shown in the proof of Theorem 3 in Section B.3 that

$$\mathrm{Var}(\Phi_1^*(\mathbf{x};\mathbf{Z}_1)) \ge Cs_2^{-1} \tag{B.43}$$

with C some positive constant. Thus, when $s_2 = o(n^{1/3})$, we can obtain from (B.42) and (B.43) that

$$\frac{\widehat{\sigma}_J^2}{\frac{s_2^2}{n}\mathrm{Var}(\Phi_1^*(\mathbf{x};\mathbf{Z}_1))} \xrightarrow{p} 1. \tag{B.44}$$

In addition, it follows from (B.22) and the decomposition for the variance of the U-statistic that as long as $s_2 = o(n)$, we have

$$\frac{\sigma^2}{\frac{s_2^2}{n}\mathrm{Var}(\Phi_1^*(\mathbf{x};\mathbf{Z}_1))} \to 1. \tag{B.45}$$

Therefore, combining (B.44) and (B.45) results in $\widehat{\sigma}_J^2/\sigma^2 \xrightarrow{p} 1$, which establishes the desired consistency of the jackknife estimator $\widehat{\sigma}_J^2$. This completes the proof of Theorem 5.

B.6 Proof of Theorem 6

We now proceed with establishing the consistency of the bootstrap estimator $\widehat{\sigma}_{B,n}^2$ introduced in (30) for the variance σ^2 of the two-scale DNN estimator $D_n(s_1, s_2)(\mathbf{x})$ as defined in (26). Let us define the bootstrap version of the quantity σ^2 conditional on the given sample $\{\mathbf{Z}_1, \cdots, \mathbf{Z}_n\}$ as

$$\widehat{\sigma}_n^2 = \mathrm{Var}(D_n^*(s_1, s_2)(\mathbf{x})|\mathbf{Z}_1, \cdots, \mathbf{Z}_n), \tag{B.46}$$

where $D_n^*(s_1, s_2)$ defined in (29) denotes the two-scale DNN estimator constructed as in (25) using the bootstrap sample $\{\mathbf{Z}_1^*, \cdots, \mathbf{Z}_n^*\}$. In fact, the quantity introduced in (B.46) above provides a crucial bridge. The main ingredients of the proof consist of two parts. First, we will show that the bootstrap estimator $\widehat{\sigma}_{B,n}^2$ is asymptotically close to $\widehat{\sigma}_n^2$ given in (B.46) as the number of bootstrap samples $B \to \infty$. Second, we will prove that the bootstrap version $\widehat{\sigma}_n^2$ is further asymptotically close to the population quantity σ^2 under the assumption of $s_2 = o(n^{1/3})$. It is worth mentioning that the technical analysis for the second part relies on the consistency of the jackknife estimator $\widehat{\sigma}_J^2$ established in Theorem 5.

For each $1 \le b \le B$, denote by $D_n^{(b)}(s_1, s_2)(\mathbf{x})$ the two-scale DNN estimator $D_n^*(s_1, s_2)$ constructed using the bth bootstrap sample. It is easy to see from (B.46) that for each

$1 \leq b \leq B$,

$$\mathrm{Var}(D_n^{(b)}(s_1, s_2)(\mathbf{x}) | \mathbf{Z}_1, \cdots, \mathbf{Z}_n) = \widehat{\sigma}_n^2. \tag{B.47}$$

Since the sample variance defined in (30) is an unbiased estimator for the population variance, by (B.47) it holds that

$$\mathbb{E}[\widehat{\sigma}_{B,n}^2 | \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n] = \widehat{\sigma}_n^2. \tag{B.48}$$

Thus, in view of (30) and (B.48), we can obtain

$$\mathbb{E}[(\widehat{\sigma}_{B,n}^2 - \sigma^2)^2] = \mathbb{E}[(\widehat{\sigma}_{B,n}^2 - \widehat{\sigma}_n^2)^2] + \mathbb{E}[(\widehat{\sigma}_n^2 - \sigma^2)^2]. \tag{B.49}$$

Without loss of generality, let us assume that $\mathbb{E}[D_n(s_1, s_2)(\mathbf{x})] = 0$ to ease our technical presentation; otherwise we can subtract the mean first.

We begin with considering the first term $\mathbb{E}[(\widehat{\sigma}_{B,n}^2 - \widehat{\sigma}_n^2)^2]$ on the right-hand side of (B.49). Since $\{D_n^{(b)}(s_1, s_2)(\mathbf{x})\}_{1 \leq b \leq B}$ are i.i.d. random variables conditional on the given

sample $\{\mathbf{Z}_1, \cdots, \mathbf{Z}_n\}$, we can deduce that

$$
\begin{aligned}
&\mathbb{E}[(\widehat{\sigma}^2_{B,n} - \widehat{\sigma}^2_n)^2 | \mathbf{Z}_1, \cdots, \mathbf{Z}_n] \\
&= \mathbb{E}\Big[\frac{1}{(B-1)^2}\Big(\sum_{b=1}^{B}\big([D_n^{(b)}(s_1,s_2)(\mathbf{x})]^2 - \widehat{\sigma}^2_n\big) - (B\bar{D}^2_{B,n} - \widehat{\sigma}^2_n)\Big)^2 \Big| \mathbf{Z}_1, \cdots, \mathbf{Z}_n\Big] \\
&\le \frac{2}{(B-1)^2}\Big\{\mathbb{E}\Big[\Big(\sum_{b=1}^{B}\big([D_n^{(b)}(s_1,s_2)(\mathbf{x})]^2 - \widehat{\sigma}^2_n\big)\Big)^2 \Big| \mathbf{Z}_1, \cdots, \mathbf{Z}_n\Big] \\
&\quad + \mathbb{E}\big[(B\bar{D}^2_{B,n} - \widehat{\sigma}^2_n)^2 | \mathbf{Z}_1, \cdots, \mathbf{Z}_n\big]\Big\} \\
&\le \frac{2}{(B-1)^2}\Big\{B\mathbb{E}\big[\big([D_n^{(1)}(s_1,s_2)(\mathbf{x})]^2 - \widehat{\sigma}^2_n\big)^2 \big| \mathbf{Z}_1, \cdots, \mathbf{Z}_n\big] \\
&\quad + \frac{2B}{B^2}\mathbb{E}\big[\big([D_n^{(1)}(s_1,s_2)(\mathbf{x})]^2 - \widehat{\sigma}^2_n\big)^2 \big| \mathbf{Z}_1, \cdots, \mathbf{Z}_n\big] \\
&\quad + \frac{4}{B^2}\sum_{1\le i\ne j\le B}\mathbb{E}\big[\big(D_n^{(i)}(s_1,s_2)(\mathbf{x})\big)^2\big(D_n^{(j)}(s_1,s_2)(\mathbf{x})\big)^2 \big| \mathbf{Z}_1, \cdots, \mathbf{Z}_n\big]\Big\} \\
&\le \frac{C}{B}\mathbb{E}\big[\big([D_n^{(1)}(s_1,s_2)(\mathbf{x})]^2 - \widehat{\sigma}^2_n\big)^2 \big| \mathbf{Z}_1, \cdots, \mathbf{Z}_n\big] \\
&\quad + \frac{C}{B^2}\Big(\mathbb{E}\big[[D_n^{(1)}(s_1,s_2)(\mathbf{x})]^2 \big| \mathbf{Z}_1, \cdots, \mathbf{Z}_n\big]\Big)^2 \\
&\le \frac{C}{B}\mathbb{E}\big[[D_n^{(1)}(s_1,s_2)(\mathbf{x})]^4 \big| \mathbf{Z}_1, \cdots, \mathbf{Z}_n\big], \qquad \text{(B.50)}
\end{aligned}
$$

where the last inequality follows from the conditional Jensen's inequality.

Let $k = [n/s_2]$ be the integer part of the number n/s_2. We define

$$
\begin{aligned}
h(\mathbf{Z}_1, \cdots, \mathbf{Z}_n) = k^{-1}\big(&\Phi^*(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2}) + \Phi^*(\mathbf{x}; \mathbf{Z}_{s_2+1}, \cdots, \mathbf{Z}_{2s_2}) \\
&+ \cdots + \Phi^*(\mathbf{x}; \mathbf{Z}_{ks_2-s_2+1}, \cdots, \mathbf{Z}_{ks_2})\big). \qquad \text{(B.51)}
\end{aligned}
$$

Note that it has been shown in (2.1.15) in Korolyuk and Borovskich (1994) that

$$
\mathbb{E}\big[[D_n^{(1)}(s_1,s_2)(\mathbf{x})]^4 \big| \mathbf{Z}_1, \cdots, \mathbf{Z}_n\big] \le \mathbb{E}[h^4(\mathbf{Z}_1^*, \cdots, \mathbf{Z}_n^*) | \mathbf{Z}_1, \cdots, \mathbf{Z}_n] \qquad \text{(B.52)}
$$

with the functional $h(\cdot)$ given in (B.51). Moreover, with an application of Rosenthal's

inequality for independent random variables, we can obtain that

$$\begin{aligned}&\mathbb{E}[h^4(\mathbf{Z}_1^*,\cdots,\mathbf{Z}_n^*)|\mathbf{Z}_1,\cdots,\mathbf{Z}_n]\\&\leq Ck^{-4}k^2\mathbb{E}\big([\Phi^*(\mathbf{x};\mathbf{Z}_1^*,\cdots,\mathbf{Z}_{s_2}^*)]^4\big|\mathbf{Z}_1,\cdots,\mathbf{Z}_n\big),\end{aligned}\tag{B.53}$$

where C is some positive constant. Then in light of (B.53), it remains to bound the quantity $\mathbb{E}([\Phi^*(\mathbf{x};\mathbf{Z}_1^*,\cdots,\mathbf{Z}_{s_2}^*)]^4)$, which has been dealt with in Lemma 4 in Section B.4. Thus, it follows from (B.50), (B.52)–(B.53), and Lemma 4 that

$$\begin{aligned}\mathbb{E}[(\widehat{\sigma}_{B,n}^2-\widehat{\sigma}_n^2)^2]&\leq\frac{C}{B}\frac{s_2^2}{n^2}n^{-s_2}\sum_{i_1=1}^{n}\cdots\sum_{i_{s_2}=1}^{n}\mathbb{E}\big([\Phi^*(\mathbf{x};\mathbf{Z}_{i_1},\cdots,\mathbf{Z}_{i_{s_2}})]^4\big)\\&\leq\frac{CMs_2^2}{Bn^2},\end{aligned}\tag{B.54}$$

where M is some positive constant given in Lemma 4.

We next proceed with analyzing the second term $\mathbb{E}[(\widehat{\sigma}_n^2-\sigma^2)^2]$ on the right-hand side of (B.49). Recall the definition of the bootstrap version $\widehat{\sigma}_n^2$ for the population quantity σ^2 introduced in (B.46). Let us define

$$m_n=\mathbb{E}[\Phi^*(\mathbf{x};\mathbf{Z}_1^*,\cdots,\mathbf{Z}_{s_2}^*)|\mathbf{Z}_1,\cdots,\mathbf{Z}_n]\tag{B.55}$$

and

$$h_1(\mathbf{z})=\mathbb{E}[\Phi^*(\mathbf{x};\mathbf{Z}_1^*,\cdots,\mathbf{Z}_{s_2}^*)-m_n|\mathbf{Z}_1^*=\mathbf{z}].\tag{B.56}$$

Then applying similar arguments as for (B.13) in the proof of Theorem 2 in Section B.2, we can deduce that

$$\widehat{\sigma}_n^2=\frac{s_2^2}{n}\mathbb{E}[h_1^2(\mathbf{Z}_1^*)|\mathbf{Z}_1,\cdots,\mathbf{Z}_n]+\Delta_1,\tag{B.57}$$

where $0\leq\Delta_1\leq\frac{s_2^2}{n^2}\operatorname{Var}(\Phi^*(\mathbf{x};\mathbf{Z}_1^*,\cdots,\mathbf{Z}_{s_2}^*)|\mathbf{Z}_1,\cdots,\mathbf{Z}_n)$ and function $h_1(\cdot)$ is given in (B.56) and (B.55). Similarly, it holds that

$$\sigma^2=\frac{s_2^2}{n}\mathbb{E}[g_1^2(\mathbf{Z}_1)]+\Delta_2,\tag{B.58}$$

where $g_1(\mathbf{Z}_1) = \mathbb{E}[\Phi^*(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_n)|\mathbf{Z}_1]$ and $0 \leq \Delta_2 \leq \frac{s_2^2}{n^2} \operatorname{Var}(\Phi^*(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2}))$. Hence, by (B.57) and (B.58) we can obtain that

$$\begin{aligned}
&\mathbb{E}[(\widehat{\sigma}_n^2 - \sigma^2)^2] \\
&\leq C\mathbb{E}\Big(\frac{s_2^4}{n^2}\big[\mathbb{E}[h_1^2(\mathbf{Z}_1^*)|\mathbf{Z}_1, \cdots, \mathbf{Z}_n] - \mathbb{E}[g_1^2(\mathbf{Z}_1)]\big]^2 + \Delta_1^2 + \Delta_2^2\Big),
\end{aligned} \tag{B.59}$$

where C is some positive constant.

Observe that

$$\Delta_2^2 = O(\frac{s_2^4}{n^4}) \tag{B.60}$$

and

$$\begin{aligned}
\mathbb{E}(\Delta_1^2) &\leq \frac{s_2^4}{n^4}\mathbb{E}\big[\mathbb{E}\big([\Phi^*(\mathbf{x}; \mathbf{Z}_1^*, \cdots, \mathbf{Z}_{s_2}^*)]^2|\mathbf{Z}_1, \cdots, \mathbf{Z}_n\big)\big]^2 \\
&\leq \frac{s_2^4}{n^4}\mathbb{E}\big([\Phi^*(\mathbf{x}; \mathbf{Z}_1^*, \cdots, \mathbf{Z}_{s_2}^*)]^4\big) \\
&\leq \frac{Ms_2^4}{n^4},
\end{aligned} \tag{B.61}$$

where the last inequality follows from Lemma 4 with M some positive constant. In addition, it holds that

$$\mathbb{E}[h_1^2(\mathbf{Z}_1^*)|\mathbf{Z}_1, \cdots, \mathbf{Z}_n] = \frac{1}{n}\sum_{i=1}^n h_1^2(\mathbf{Z}_i) \tag{B.62}$$

and

$$\begin{aligned}
h_1(\mathbf{Z}_i) = n^{-s_2+1}&\sum_{i_2=1}^n \cdots \sum_{i_{s_2}=1}^n \Phi^*(\mathbf{x}; \mathbf{Z}_i, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_{s_2}}) \\
- n^{-s_2}&\sum_{i_1=1}^n \cdots \sum_{i_{s_2}=1}^n \Phi^*(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_{s_2}}).
\end{aligned} \tag{B.63}$$

Let us further define

$$S_i = \binom{n-1}{s_2-1}^{-1} \sum_{\substack{1 \leq j_1 < j_2 < \cdots < j_{s_2-1} \leq n \\ j_1, j_2, \cdots, j_{s_2-1} \neq i}} \Phi^*(\mathbf{x}; \mathbf{Z}_i, \mathbf{Z}_{j_1}, \cdots, \mathbf{Z}_{j_{s_2-1}}). \tag{B.64}$$

From the equality $\binom{n-1}{s_2}U_{n-1}^{(i)} + \binom{n-1}{s_2-1}S_i = \binom{n}{s_2}D_n(s_1,s_2)(\mathbf{x})$ in view of (B.64), it is easy to see that the jackknife estimator $\widehat{\sigma}_J^2$ introduced in (28) satisfies that

$$\frac{n\widehat{\sigma}_J^2}{s_2^2} = \frac{n-1}{(n-s_2)^2}\sum_{i=1}^{n}\big(S_i - D_n(s_1,s_2)(\mathbf{x})\big)^2. \tag{B.65}$$

Then the main idea of the remaining proof is to show that under the assumption of $s_2 = o(n^{1/3})$, $h_1(\mathbf{Z}_i)$ is asymptotically close to $S_i - D_n(s_1,s_2)(\mathbf{x})$ and thus $\mathbb{E}[h_1^2(\mathbf{Z}_1^*)|\mathbf{Z}_1,\cdots,\mathbf{Z}_n]$ is asymptotically close to $\frac{n\widehat{\sigma}_J^2}{s_2^2}$. Observe that

$$n^{-s_2+1}\binom{n-1}{s_2-1}(s_2-1)! = 1 + O(s_2^2/n)$$

and

$$n^{-s_2}\binom{n}{s_2}s_2! = 1 + O(s_2^2/n),$$

which entail that

$$\Big(n^{s_2-1} - \binom{n-1}{s_2-1}(s_2-1)!\Big)n^{-s_2+1} = O(s_2^2/n)$$

and

$$\Big(n^{s_2} - \binom{n}{s_2}s_2!\Big)n^{-s_2} = O(s_2^2/n).$$

Thus, it follows from (B.63) and these facts that

$$\begin{aligned} h_1(\mathbf{Z}_i) = (1+O(s_2^2/n))\big[S_i - D_n(s_1,s_2)(\mathbf{x})\big] + n^{-s_2+1}\sum_{\mathscr{D}_1}\Phi^*(\mathbf{x};\mathbf{Z}_i,\mathbf{Z}_{i_2},\cdots,\mathbf{Z}_{i_{s_2}}) \\ - n^{-s_2}\sum_{\mathscr{D}_2}\Phi^*(\mathbf{x};\mathbf{Z}_{i_1},\mathbf{Z}_{i_2},\cdots,\mathbf{Z}_{i_{s_2}}), \end{aligned} \tag{B.66}$$

where $\mathscr{D}_1 = \{(i_2,\cdots,i_{s_2}) : \text{there is at least one pair that are equal or there is a component that is equal to } i\}$ and $\mathscr{D}_2 = \{(i_1,\cdots,i_{s_2}) : \text{there is at least one pair of components that are equal}\}$.

With an application of similar arguments as in the proof of Lemma 4 in Section B.4, we can obtain that

$$\mathbb{E}\big([\Phi^*(\mathbf{x};\mathbf{Z}_{i_1},\mathbf{Z}_{i_2},\cdots,\mathbf{Z}_{i_{s_2}})]^4\big) \le M \tag{B.67}$$

with M some positive constant, regardless of how many components of $(i_1, i_2, \cdots, i_{s_2})$ are equal. As a consequence, by (B.67) it holds that

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}\Big(n^{-s_2+1}\sum_{\mathscr{D}_1}\Phi^*(\mathbf{x};\mathbf{Z}_i,\mathbf{Z}_{i_2},\cdots,\mathbf{Z}_{i_{s_2}})\Big)^2\Big)^2\Big] \\
&\le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(n^{-s_2+1}\sum_{\mathscr{D}_1}\Phi^*(\mathbf{x};\mathbf{Z}_i,\mathbf{Z}_{i_2},\cdots,\mathbf{Z}_{i_{s_2}})\Big)^4\Big] \le \frac{CMs_2^8}{n^4}
\end{aligned} \tag{B.68}$$

and similarly,

$$\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}\Big(n^{-s_2}\sum_{\mathscr{D}_2}\Phi^*(\mathbf{x};\mathbf{Z}_{i_1},\mathbf{Z}_{i_2},\cdots,\mathbf{Z}_{i_{s_2}})\Big)\Big)^2\Big] \le \frac{CMs_2^8}{n^4}, \tag{B.69}$$

where C represents some positive constant whose value may change from line to line. Hence, combining (B.62), (B.66), and (B.68)–(B.69), we can deduce that as long as $s_2 = o(n^{1/3})$, it holds that

$$\begin{aligned}
&\mathbb{E}\Big(\big[\mathbb{E}[h_1^2(\mathbf{Z}_1^*)|\mathbf{Z}_1,\cdots,\mathbf{Z}_n] - \mathbb{E}[g_1^2(\mathbf{Z}_1)]\big]^2\Big) \\
&\le C\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}(1+O(s_2^2/n))^2[S_i - D_n(s_1,s_2)(\mathbf{x})]^2 - \mathrm{Var}(\Phi_1^*(\mathbf{x};\mathbf{Z}_1))\Big)^2\Big] \\
&\quad + \frac{CMs_2^8}{n^4} \\
&\le C\mathbb{E}\Big[\Big(\frac{(n-s_2)^2}{n(n-1)}(1+O(s_2^2/n))\frac{n}{s_2^2}\widehat{\sigma}_J^2 - \mathrm{Var}(\Phi_1^*(\mathbf{x};\mathbf{Z}_1))\Big)^2\Big] + \frac{CMs_2^8}{n^4} \\
&\le C(1+O(s_2^2/n))\mathbb{E}\big([\frac{n}{s_2^2}\widehat{\sigma}_J^2 - \mathrm{Var}(\Phi_1^*(\mathbf{x};\mathbf{Z}_1))]^2\big) + \frac{Cs_2^4}{n^2}(\mathrm{Var}(\Phi_1^*(\mathbf{x};\mathbf{Z}_1)))^2 \\
&\quad + \frac{CMs_2^8}{n^4} \\
&\le \frac{Cs_2}{n} + \frac{s_2^2}{n^2} + \frac{CMs_2^8}{n^4} \le \frac{C(M+1)s_2}{n},
\end{aligned} \tag{B.70}$$

where the second to the last inequality comes from (B.42) and (C.132) in the proof of Lemma 10 in Section B.10.

Substituting the above bounds in (B.60)–(B.61) and (B.70) into (B.59) leads to

$$\mathbb{E}[(\widehat{\sigma}_n^2 - \sigma^2)^2] = O\Big(\frac{s_2^5}{n^3} + \frac{s_2^4}{n^4}\Big). \tag{B.71}$$

Thus, combining (B.54) and (B.71), we can obtain that

$$\mathbb{E}[(\widehat{\sigma}^2_{B,n} - \sigma^2)^2] = O\Big(\frac{s_2^5}{n^3} + \frac{s_2^2}{Bn^2}\Big). \tag{B.72}$$

Recall the fact that $\sigma^2 = O(\frac{s_2}{n})$ under the assumption of $s_2 = o(n)$. Consequently, such fact along with (B.72) entails that

$$\mathbb{E}\Big[\Big(\frac{\widehat{\sigma}^2_{B,n}}{\sigma^2} - 1\Big)^2\Big] = O(\frac{s_2^3}{n} + \frac{1}{B}). \tag{B.73}$$

Therefore, combining (B.73) and the assumptions of $s_2 = o(n^{1/3})$ and $B \to \infty$ yields $\widehat{\sigma}^2_{B,n}/\sigma^2 \xrightarrow{p} 1$, which establishes the desired consistency of the bootstrap estimator $\widehat{\sigma}^2_{B,n}$. This concludes the proof of Theorem 6.

C Some key lemmas and their proofs

B.1 Proof of Lemma 1

Observe that the set of possible values $Y_{(1)}(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s})$ can take is $\{Y_{(1)}, Y_{(2)}, \cdots, Y_{(n-s+1)}\}$. If $Y_{(1)}(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s}) = Y_{(1)}$, then the observation corresponding to $Y_{(1)}$ must be selected and there are $\binom{n-1}{s-1}$ options for the remaining $s-1$ places in the subsample. In general, if $Y_{(1)}(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_s}) = Y_{(j)}$ for some $1 \le j \le n-s+1$, then the observation corresponding to $Y_{(j)}$ must be selected and the observations corresponding to $Y_{(1)}, Y_{(2)}, \cdots, Y_{(j-1)}$ will not be selected, which entails that there are $\binom{n-j}{s-1}$ options for the remaining $s-1$ places in the subsample. Applying these arguments, we can obtain the representation in (10) for the single-scale DNN estimator $D_n(s)(\mathbf{x})$.

It remains to show that for any given $\mathbf{x}$, $D_n(s)(\mathbf{x})$ is in fact an L-statistic. Denote by $Y_{1:n}, Y_{2:n}, \cdots, Y_{n:n}$ the order statistics for n scalars $Y_1, Y_2, \cdots, Y_n$. Note that for each given $\mathbf{x}$, $Y_{(1)}, Y_{(2)}, \cdots, Y_{(n-s-1)}$ are fixed and there exists some set $S(\mathbf{x}) := \{i_j : 1 \le j \le n-s+1\}$

that may depend upon $\mathbf{x}$ such that

$$(Y_{(1)}, Y_{(2)}, \cdots, Y_{(n-s+1)}) = (Y_{i_1:n}, Y_{i_2:n}, \cdots, Y_{i_{n-s+1}:n}). \tag{C.74}$$

Conversely, for each $k \in S(\mathbf{x})$, there exists some $1 \leq m_k \leq n-s+1$ that may depend on $\mathbf{x}$ such that $Y_{k:n} = Y_{(m_k)}$. Consequently, combining such fact, (C.74), and the representation in (10) established above, we can obtain that

$$D_n(s)(\mathbf{x}) = \binom{n}{s}^{-1} \sum_{k \in S(\mathbf{x})} \binom{n-m_k}{s-1} Y_{k:n},$$

which shows that for any given $\mathbf{x}$, $D_n(s)(\mathbf{x})$ is an L-statistic. This completes the proof of Lemma 1.

B.2 Lemma 2 and its proof

Lemma 2. *Under the conditions of Theorem 5, we have that for each $0 \leq c \leq s_2$ and fixed $\mathbf{x}$,*

$$\mathrm{Var}(U_c) \leq \frac{2s_2 - c}{n} \mathrm{Var}(K^{(c)}), \tag{C.75}$$

where U_c is the U-statistic defined in (B.34) and $K^{(c)}$ is the symmetrized kernel function given in (B.36).

Proof. For notational simplicity, we will drop the dependence of all the functionals on the fixed vector $\mathbf{x}$ whenever there is no confusion. For each $1 \leq j \leq 2s_2 - c$, let us define

$$\begin{aligned} K_j^{(c)}(\mathbf{Z}_1, \cdots, \mathbf{Z}_j) &= \mathbb{E}[K^{(c)} | \mathbf{Z}_1, \cdots, \mathbf{Z}_j], \\ g_j^{(c)}(\mathbf{Z}_1, \cdots, \mathbf{Z}_j) &= K_j^{(c)} - \mathbb{E}[K^{(c)}] - \sum_{i=1}^{j-1} \sum_{1 \leq \alpha_1 < \cdots < \alpha_i \leq j} g_i^{(c)}(\mathbf{Z}_{\alpha_1}, \cdots, \mathbf{Z}_{\alpha_i}), \end{aligned}$$

and $V_j = \mathrm{Var}(g_j^{(c)}(\mathbf{Z}_1, \cdots, \mathbf{Z}_j))$. Then it follows from Hoeffding's decomposition that

$$U_c = \mathbb{E}[K^{(c)}] + \binom{n}{2s_2 - c}^{-1} \sum_{i=1}^{2s_2-c} \binom{n-i}{2s_2 - c - i} \sum_{1 \le \alpha_1 < \cdots < \alpha_i \le n} g_i^{(c)}(\mathbf{Z}_{\alpha_1}, \cdots, \mathbf{Z}_{\alpha_i}). \quad \text{(C.76)}$$

Observe that $\mathrm{Var}(K^{(c)}) = \sum_{i=1}^{2s_2-c} \binom{2s_2-c}{i} V_i$. Thus, in view of (C.76), we can deduce that

$$\begin{aligned}
\mathrm{Var}(U_c) &= \sum_{i=1}^{2s_2-c} \binom{n}{2s_2 - c}^{-2} \binom{n-i}{2s_2 - c - i}^2 \binom{n}{i} V_j \\
&= \sum_{i=1}^{2s_2-c} \frac{(2s_2 - c)!(n-i)!}{n!(2s_2 - c - i)!} \binom{2s_2 - c}{i} V_i \\
&\le \frac{2s_2 - c}{n} \sum_{i=1}^{2s_2-c} \binom{2s_2 - c}{i} V_i \\
&= \frac{2s_2 - c}{n} \mathrm{Var}(K^{(c)}),
\end{aligned}$$

which establishes the desired upper bound in (C.75). This concludes the proof of Lemma 2.

B.3 Lemma 3 and its proof

Lemma 3. *Under the conditions of Theorem 5, it holds that for each* $0 \le c \le s_2$ *and fixed* $\mathbf{x}$,

$$\mathrm{Var}(K^{(c)}) \le C[(w_1^*)^4 + (w_2^*)^4]\big(\mu^4(\mathbf{x}) + 6\mu^2(\mathbf{x})\sigma_\epsilon + 4\mu(\mathbf{x}) + \mathbb{E}[\epsilon_1^4]\big), \quad \text{(C.77)}$$

where $K^{(c)}$ *is the symmetrized kernel function given in (B.36) and* C *is some positive constant.*

Proof. By the Cauchy–Schwarz inequality, we can deduce that

$$\begin{aligned}
&\mathrm{Var}(K^{(c)}) \le \mathbb{E}[(K^{(c)})^2] \\
&= \left[\binom{2s_2 - c}{c}\binom{2s_2 - 2c}{s_2 - c}\right]^{-2} \sum_{\Pi_{2s_2-c}} \sum_{\Pi_{2s_2-c}} \\
&\quad \mathbb{E}\big\{\Phi^*(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_c}, \mathbf{Z}_{i_{c+1}}, \cdots, \mathbf{Z}_{i_{s_2}})\Phi^*(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_c}, \mathbf{Z}_{i_{s_2+1}}, \cdots, \mathbf{Z}_{i_{2s_2-c}}) \\
&\quad \times \Phi^*(\mathbf{x}; \mathbf{Z}_{j_1}, \cdots, \mathbf{Z}_{j_c}, \mathbf{Z}_{j_{c+1}}, \cdots, \mathbf{Z}_{j_{s_2}})\Phi^*(\mathbf{x}; \mathbf{Z}_{j_1}, \cdots, \mathbf{Z}_{j_c}, \mathbf{Z}_{j_{s_2+1}}, \cdots, \mathbf{Z}_{j_{2s_2-c}})\big\} \\
&\le \mathbb{E}\big\{\big[\Phi^*(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{s_2})\big]^4\big\}, \qquad \text{(C.78)}
\end{aligned}$$

where $\sum_{\Pi_{2s_2-c}}$ denotes the summation introduced in (B.36). In light of the definition of Φ^* in (C.121), we have

$$\begin{aligned}
\mathbb{E}\big\{\big[\Phi^*(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{s_2})\big]^4\big\} &\le 8(w_1^*)^4\mathbb{E}[\Phi^4(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_1})] \\
&\quad + 8(w_2^*)^4\mathbb{E}[\Phi^4(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2})]. \qquad \text{(C.79)}
\end{aligned}$$

Let us make some useful observations. Note that

$$\begin{aligned}
\mathbb{E}[\Phi^4(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_1})] &= \mathbb{E}\Big[\Big(\sum_{i=1}^n y_i\zeta_{i,s_1}\Big)^4\Big] \\
&= \sum_{i=1} \mathbb{E}[y_i^4\zeta_{i,s_1}] = s_1\mathbb{E}[y_1^4\zeta_{1,s_1}]
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[y_1^4\zeta_{1,s_1}] &= \mathbb{E}\big([\mu(\mathbf{X}_1) + \epsilon_1]^4\zeta_{1,s_1}\big) \\
&= \mathbb{E}[\mu^4(\mathbf{X}_1)\zeta_{1,s_1}] + 6\mathbb{E}[\mu^2(\mathbf{X}_1)\zeta_{1,s_1}]\sigma_\epsilon^2 + 4\mathbb{E}[\mu(\mathbf{X}_1)\zeta_{1,s_1}] + \mathbb{E}[\epsilon_1^4],
\end{aligned}$$

where $\zeta_{i,s}$ represents the indicator function for the event that $\mathbf{X}_i$ is the 1NN of $\mathbf{x}$ among $\mathbf{X}_1, \cdots, \mathbf{X}_s$. Moreover, it follows from Lemma 13 in section C.3 that as $s_1 \to \infty$,

$$s_1\mathbb{E}[\mu^k(\mathbf{X}_1)\zeta_{1,s_1}] \to \mu^k(\mathbf{x})$$

for $k = 1, 2, 4$. Hence, it holds that

$$\begin{aligned}\mathbb{E}[\Phi^4(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_1})] &= s_1 \mathbb{E}[y_1^4 \zeta_{1,s_1}] \\ &\to \mu^4(\mathbf{x}) + 6\mu^2(\mathbf{x})\sigma_\epsilon + 4\mu(\mathbf{x}) + \mathbb{E}[\epsilon_1^4]\end{aligned}$$

as $s_1 \to \infty$.

Using similar arguments, we can show that as $s_2 \to \infty$,

$$\mathbb{E}[\Phi^4(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2})] \to \mu^4(\mathbf{x}) + 6\mu^2(\mathbf{x})\sigma_\epsilon + 4\mu(\mathbf{x}) + \mathbb{E}[\epsilon_1^4].$$

Therefore, combining the asymptotic limits obtained above, (C.78), and (C.79) results in

$$\mathrm{Var}(K^{(c)}) \le C[(w_1^*)^4 + (w_2^*)^4]\big(\mu^4(\mathbf{x}) + 6\mu^2(\mathbf{x})\sigma_\epsilon + 4\mu(\mathbf{x}) + \mathbb{E}[\epsilon_1^4]\big),$$

where C is some positive constant. This completes the proof of Lemma 3.

B.4 Lemma 4 and its proof

Lemma 4. *Under the conditions of Theorem 6, there exists some constant $M > 0$ depending upon w_1^*, w_2^*, $\mathbf{x}$, and the distribution of ϵ such that*

$$\mathbb{E}\big([\Phi^*(\mathbf{x}; \mathbf{Z}_1^*, \cdots, \mathbf{Z}_{s_2}^*)]^4\big) \le M. \tag{C.80}$$

Proof. Since the observations in the bootstrap sample $\{\mathbf{Z}_1^*, \cdots, \mathbf{Z}_n^*\}$ are selected independently and uniformly from the original sample $\{\mathbf{Z}_1, \cdots, \mathbf{Z}_n\}$, we have

$$\begin{aligned}\mathbb{E}\big([\Phi^*(\mathbf{x}; \mathbf{Z}_1^*, \cdots, \mathbf{Z}_{s_2}^*)]^4\big) &= \mathbb{E}\Big(\mathbb{E}\big([\Phi^*(\mathbf{x}; \mathbf{Z}_1^*, \cdots, \mathbf{Z}_{s_2}^*)]^4 \big| \mathbf{Z}_1, \cdots, \mathbf{Z}_n\big)\Big) \\ &= n^{-s_2} \sum_{i_1=1}^{n} \cdots \sum_{i_{s_2}=1}^{n} \mathbb{E}\big([\Phi^*(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_{s_2}})]^4\big).\end{aligned}$$

Observe that for distinct $i_1, \cdots, i_{s_2}$, we have shown in the proof of Lemma 3 in Section B.3 that as $s_2 \to \infty$,

$$\mathbb{E}\big([\Phi^*(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2})]^4\big) \to A$$

for some positive constant A that depends upon w_1^*, w_2^*, $\mathbf{x}$, and the distribution of ϵ.

Furthermore, note that if $i_1 = i_2 = \cdots = i_c$ and the remaining arguments are distinct, then it holds that

$$\Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_{s_2}}) = \Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \mathbf{Z}_{i_{c+1}}, \cdots, \mathbf{Z}_{i_{s_2}}).$$

Therefore, there exists some positive constant M depending upon w_1^*, w_2^*, $\mathbf{x}$, and the distribution of ϵ such that

$$\mathbb{E}\big([\Phi^*(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_{s_2}})]^4\big) \leq M$$

for any $1 \leq i_1 \leq n, \cdots, 1 \leq i_{s_2} \leq n$. This concludes the proof of Lemma 4.

B.5 Lemma 5 and its proof

In Lemma 5 below, we will provide the asymptotic expansion of $\mathbb{E}\,\|\mathbf{X}_{(1)} - \mathbf{x}\|^k$ with $k \geq 1$ and its higher-order asymptotic expansion for the case of $k = 2$ as the sample size $n \to \infty$.

Lemma 5. *Assume that Conditions 1–3 hold and $\mathbf{x} \in \mathrm{supp}(\mathbf{X}) \subset \mathbb{R}^d$ is fixed. Then the 1-nearest neighbor (1NN) $\mathbf{X}_{(1)}$ of $\mathbf{x}$ in the i.i.d. sample $\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ satisfies that for any $k \geq 1$,*

$$\mathbb{E}\,\|\mathbf{X}_{(1)} - \mathbf{x}\|^k = \frac{\Gamma(k/d+1)}{(f(\mathbf{x})V_d)^{k/d}} n^{-k/d} + o(n^{-k/d}) \tag{C.81}$$

as $n \to \infty$, where $\Gamma(\cdot)$ is the gamma function and $V_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$. In particular, when $k = 2$, there are three cases. If $d = 1$, we have

$$\mathbb{E}\,\|\mathbf{X}_{(1)} - \mathbf{x}\|^2 = \frac{\Gamma(2/d+1)}{(f(\mathbf{x})V_d)^{2/d}} n^{-2/d} - \left(\frac{\Gamma(2/d+2)}{d(f(\mathbf{x})V_d)^{2/d}}\right) n^{-(1+2/d)} + o(n^{-(1+2/d)}). \tag{C.82}$$

If $d = 2$, we have

$$\mathbb{E}\,\|\mathbf{X}_{(1)} - \mathbf{x}\|^2 = \frac{\Gamma(2/d+1)}{(f(\mathbf{x})V_d)^{2/d}} n^{-2/d} - \left(\frac{\mathrm{tr}(f''(\mathbf{x}))\Gamma(4/d+1)}{f(\mathbf{x})(f(\mathbf{x})V_d)^{4/d}d(d+2)} + \frac{\Gamma(2/d+2)}{d(f(\mathbf{x})V_d)^{2/d}}\right) n^{-4/d} + o(n^{-4/d}), \tag{C.83}$$

where $f''(\cdot)$ stands for the Hessian matrix of the density function $f(\cdot)$. If $d \geq 3$, we have

$$\mathbb{E}\,\|\mathbf{X}_{(1)} - \mathbf{x}\|^2 = \frac{\Gamma(2/d+1)}{(f(\mathbf{x})V_d)^{2/d}} n^{-2/d} - \left(\frac{\mathrm{tr}(f''(\mathbf{x}))\Gamma(4/d+1)}{f(\mathbf{x})(f(\mathbf{x})V_d)^{4/d} d(d+2)}\right) n^{-4/d} + o(n^{-4/d}). \tag{C.84}$$

Proof. Denote by φ the probability measure on $\mathbb{R}^d$ given by random vector $\mathbf{X}$. We begin with obtaining an approximation of $\varphi(B(\mathbf{x}, r))$, where $B(\mathbf{x}, r)$ represents a ball in the Euclidean space $\mathbb{R}^d$ with center $\mathbf{x}$ and radius $r > 0$. Recall that by Condition 2, the density function $f(\cdot)$ of measure φ with respect to the Lebesgue measure λ is four-times continuously differentiable with bounded corresponding derivatives in a neighborhood of $\mathbf{x}$. Then using the Taylor expansion, we see that for any $\boldsymbol{\xi} \in S^{d-1}$ and $0 < \rho < r$,

$$f(\mathbf{x} + \rho\boldsymbol{\xi}) = f(\mathbf{x}) + f'(\mathbf{x})^T \boldsymbol{\xi}\rho + \frac{1}{2}\boldsymbol{\xi}^T f''(\mathbf{x})\boldsymbol{\xi}\rho^2 + o(\rho^2), \tag{C.85}$$

where S^{d-1} denotes the unit sphere in $\mathbb{R}^d$, and $f'(\cdot)$ and $f''(\cdot)$ stand for the gradient vector and the Hessian matrix, respectively, of the density function $f(\cdot)$. With the aid of the representation in (C.85), an application of the spherical integration leads to

$$\begin{aligned}
\varphi(B(\mathbf{x}, r)) &= \int_0^r \int_{S^{d-1}} f(\mathbf{x} + \rho\boldsymbol{\xi})\rho^{d-1}\nu(\mathrm{d}\boldsymbol{\xi})\,\mathrm{d}\rho \\
&= \int_0^r \int_{S^{d-1}} \left(f(\mathbf{x}) + f'(\mathbf{x})^T\boldsymbol{\xi}\rho + \frac{1}{2}\boldsymbol{\xi}^T f''(\mathbf{x})\boldsymbol{\xi}\rho^2 + o(\rho^2)\right)\rho^{d-1}\nu(\mathrm{d}\boldsymbol{\xi})\,\mathrm{d}\rho \\
&= \int_0^r \left[f(\mathbf{x})dV_d\rho^{d-1} + \frac{\mathrm{tr}(f''(\mathbf{x}))V_d}{2}\rho^{d+1} + o(\rho^{d+1})\right]\mathrm{d}\rho \\
&= f(\mathbf{x})V_d r^d + \frac{\mathrm{tr}(f''(\mathbf{x}))V_d}{2(d+2)} r^{d+2} + o(r^{d+2}),
\end{aligned} \tag{C.86}$$

where ν denotes a measure constructed on the unit sphere $\mathbb{S}^{d-1}$ as characterized in Lemma 11 in Section C.1 and $\mathrm{d}\cdot$ stands for the differential of a given variable hereafter.

We now turn our attention to the target quantity $\mathbb{E}\left\|\mathbf{X}_{(1)}-\mathbf{x}\right\|^k$ for any $k \geq 1$. It holds that

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{X}_{(1)}-\mathbf{x}\right\|^k &= \int_0^\infty \mathbb{P}\left(\left\|\mathbf{X}_{(1)}-\mathbf{x}\right\|^k > t\right) \mathrm{d}t \\
&= \int_0^\infty \mathbb{P}\left(\left\|\mathbf{X}_{(1)}-\mathbf{x}\right\| > t^{1/k}\right) \mathrm{d}t \\
&= \int_0^\infty \left[1-\varphi(B(\mathbf{x}, t^{1/k}))\right]^n \mathrm{d}t \\
&= n^{-k/d} \int_0^\infty \left[1-\varphi\left(B\left(\mathbf{x}, \frac{t^{1/k}}{n^{1/d}}\right)\right)\right]^n \mathrm{d}t. \qquad \text{(C.87)}
\end{aligned}$$

To evaluate the integration in (C.87), we need to analyze the term $\left[1-\varphi\left(B\left(\mathbf{x}, \frac{t^{1/k}}{n^{1/d}}\right)\right)\right]^n$. It follows from the asymptotic expansion of $\varphi(B(\mathbf{x}, r))$ in (C.86) that

$$\begin{aligned}
&\left[1-\varphi\left(B\left(\mathbf{x}, \frac{t^{1/k}}{n^{1/d}}\right)\right)\right]^n \\
&= \left[1-\frac{f(\mathbf{x})V_d t^{d/k}}{n} - \frac{\frac{\mathrm{tr}(f''(\mathbf{X}))V_d}{2(d+2)} t^{(d+2)/k}}{n^{1+2/d}} + o(n^{-(1+2/d)})\right]^n. \qquad \text{(C.88)}
\end{aligned}$$

From (C.88), we see that for each fixed $t > 0$,

$$\lim_{n\to\infty}\left[1-\varphi\left(B\left(\mathbf{x}, \frac{t^{1/k}}{n^{1/d}}\right)\right)\right]^n = \exp(-f(\mathbf{x})V_d t^{d/k}).$$

Moreover, by Condition 1, we have

$$\begin{aligned}
\left[1-\varphi\left(B\left(\mathbf{x}, \frac{t^{1/k}}{n^{1/d}}\right)\right)\right]^n &\leq \left[\exp\left(-\alpha\frac{t^{1/k}}{n^{1/d}}\right)\right]^n \\
&\leq \exp\left(-\alpha t^{1/k}\right).
\end{aligned}$$

Thus, an application of the dominated convergence theorem yields

$$\begin{aligned}
\lim_{n\to\infty}\int_0^\infty \left[1-\varphi\left(B\left(\mathbf{x}, \frac{t^{1/k}}{n^{1/d}}\right)\right)\right]^n \mathrm{d}t &= \int_0^\infty \lim_{n\to\infty}\left[1-\varphi\left(B\left(\mathbf{x}, \frac{t^{1/k}}{n^{1/d}}\right)\right)\right]^n \mathrm{d}t \\
&= \int_0^\infty \exp(-f(\mathbf{x})V_d t^{d/k})\, \mathrm{d}t \\
&= \frac{\Gamma(k/d+1)}{(f(\mathbf{x})V_d)^{k/d}}, \qquad \text{(C.89)}
\end{aligned}$$

which establishes the desired asymptotic expansion in (C.81) for any $k \geq 1$.

We further investigate higher-order expenasion for the case of $k = 2$. The leading term of the asymptotic expansion for $\mathbb{E}\,\|\mathbf{X}_{(1)} - \mathbf{x}\|^2$ has been identified in (C.89) with the choice of $k = 2$. But we now aim to conduct a higher-order asymptotic expansion. To do so, we will resort to the higher-order asymptotic expansion given in (C.88). In view of (C.88), we can deduce from Taylor expansion for $\log(1-x)$ around 0 that

$$\begin{aligned}
&\left[1 - \varphi\left(B\left(\mathbf{x}, \frac{t^{1/2}}{n^{1/d}}\right)\right)\right]^n - \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\} \\
&= \exp\left\{n \log\left[1 - \frac{f(\mathbf{x})V_d t^{d/2}}{n} - \frac{\frac{\mathrm{tr}(f''(\mathbf{X}))V_d}{2(d+2)} t^{(d+2)/2}}{n^{1+2/d}} + o(n^{-(1+2/d)})\right]\right\} \\
&\quad - \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\} \\
&= \exp\left\{-f(\mathbf{x})V_d t^{d/2} - \frac{\frac{\mathrm{tr}(f''(\mathbf{X}))V_d}{2(d+2)} t^{(d+2)/2}}{n^{2/d}} - \frac{f^2(\mathbf{x})V_d^2 t^d}{2n} + o(n^{-(2/d)})\right\} \\
&\quad - \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\}
\end{aligned} \tag{C.90}$$

as $n \to \infty$. To determine the order of above remainders, there are three cases, that is, $d = 1$, $d = 2$, and $d \geq 3$. First for $d = 1$, it follows from (C.90) that

$$\begin{aligned}
&\left[1 - \varphi\left(B\left(\mathbf{x}, \frac{t^{1/2}}{n^{1/d}}\right)\right)\right]^n - \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\} \\
&= \exp\left\{-f(\mathbf{x})V_d t^{d/2} - \frac{f^2(\mathbf{x})V_d^2 t^d}{2n} + o(n^{-1})\right\} - \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\} \\
&= \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\}\left(\exp\left\{-\frac{f^2(\mathbf{x})V_d^2 t^d}{2n} + o(n^{-1})\right\} - 1\right) \\
&= \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\}\left(-\frac{f^2(\mathbf{x})V_d^2 t^d}{2n} + o(n^{-1})\right)
\end{aligned} \tag{C.91}$$

as $n \to \infty$. Furthermore, it holds that

$$\int_0^\infty \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\}\left(-\frac{f^2(\mathbf{x})V_d^2 t^d}{2}\right)\,\mathrm{d}t = -\frac{\Gamma(2/d+2)}{d(f(\mathbf{x})V_d)^{2/d}}, \tag{C.92}$$

where we have used the fact that for any $a > 0$ and $b > 0$,

$$\int_0^\infty x^{a-1} \exp(-bx^p)\, \mathrm{d}x = \frac{1}{p} b^{-a/p} \Gamma(\frac{a}{p}). \tag{C.93}$$

Therefore, combining (C.87), (C.89), (C.91), and (C.92) results in the desired higher-order asymptotic expansion in (C.82) for the case of $k = 2$ and $d = 1$.

When $d = 2$, noting that $2/d = 1$, it follows from (C.90) that

$$\begin{aligned}
&\left[1 - \varphi\left(B\left(\mathbf{x}, \frac{t^{1/2}}{n^{1/d}}\right)\right)\right]^n - \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\} \\
&= \exp\left\{-f(\mathbf{x})V_d t^{d/2} - \frac{\frac{\mathrm{tr}(f''(\mathbf{X}))V_d}{2(d+2)} t^{(d+2)/2}}{n^{2/d}} - \frac{f^2(\mathbf{x})V_d^2 t^d}{2n^{2/d}} + o(n^{-(2/d)})\right\} \\
&\quad - \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\} \\
&= \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\} \left(\exp\left\{-\frac{\frac{\mathrm{tr}(f''(\mathbf{X}))V_d}{2(d+2)} t^{(d+2)/2}}{n^{2/d}} - \frac{f^2(\mathbf{x})V_d^2 t^d}{2n^{2/d}} + o(n^{-(2/d)})\right\} - 1\right) \\
&= \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\} \left(-\frac{\frac{\mathrm{tr}(f''(\mathbf{X}))V_d}{2(d+2)} t^{(d+2)/2}}{n^{2/d}} - \frac{f^2(\mathbf{x})V_d^2 t^d}{2n^{2/d}} + o(n^{-(2/d)})\right)
\end{aligned} \tag{C.94}$$

as $n \to \infty$. Applying equality (C.93) again yields

$$\begin{aligned}
&\int_0^\infty \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\} \left(-\frac{\mathrm{tr}(f''(\mathbf{x}))V_d}{2(d+2)n^{2/d}} t^{(d+2)/2}\right) \mathrm{d}t \\
&= -\left(\frac{\mathrm{tr}(f''(\mathbf{x}))\Gamma(4/d+1)}{d(d+2)f(\mathbf{x})(f(\mathbf{x})V_d)^{4/d}}\right) n^{-2/d}.
\end{aligned} \tag{C.95}$$

Hence, combining (C.87), (C.89), (C.92), (C.94), and (C.95) leads to the desired higher-order asymptotic expansion in (C.83) for the case of $k = 2$ and $d = 2$.

Finally, it remains to investigate the case of $d \geq 3$. Noticing that $n^{-1} = o(n^{-2/d})$ for

$d \geq 3$, we obtain from (C.90) that

$$\begin{aligned}
&\left[1-\varphi\left(B\left(\mathbf{x}, \frac{t^{1/2}}{n^{1/d}}\right)\right)\right]^n - \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\} \\
&= \exp\left\{-f(\mathbf{x})V_d t^{d/2} - \frac{\frac{\mathrm{tr}(f''(\mathbf{X}))V_d}{2(d+2)} t^{(d+2)/2}}{n^{2/d}} + o(n^{-(2/d)})\right\} - \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\} \\
&= \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\}\left(\exp\left\{-\frac{\frac{\mathrm{tr}(f''(\mathbf{X}))V_d}{2(d+2)} t^{(d+2)/2}}{n^{2/d}} + o(n^{-(2/d)})\right\} - 1\right) \\
&= \exp\left\{-f(\mathbf{x})V_d t^{d/2}\right\}\left(-\frac{\frac{\mathrm{tr}(f''(\mathbf{X}))V_d}{2(d+2)} t^{(d+2)/2}}{n^{2/d}} + o(n^{-(2/d)})\right). \qquad \text{(C.96)}
\end{aligned}$$

Consequently, combining (C.87), (C.89), (C.95), and (C.96) yields the desired higher-order asymptotic expansion in (C.84) for the case of $k = 2$ and $d \geq 3$. This completes the proof of Lemma 5.

B.6 Lemma 6 and its proof

As in Biau and Devroye (2015), we define the projection of the mean function $\mu(\mathbf{X}) = \mathbb{E}(Y|\mathbf{X})$ onto the positive half line $\mathbb{R}_+ = [0, \infty)$ given by $\|\mathbf{X} - \mathbf{x}\|$ as

$$m(r) = \lim_{\delta \to 0+} \mathbb{E}\left[\mu(\mathbf{X}) \mid r \leq \|\mathbf{X} - \mathbf{x}\| \leq r + \delta\right] = \mathbb{E}\left[Y \mid \|\mathbf{X} - \mathbf{x}\| = r\right] \qquad \text{(C.97)}$$

for any $r \geq 0$. Clearly, the definition in (C.97) entails that

$$m(0) = \mathbb{E}\left[Y \mid \mathbf{X} = \mathbf{x}\right] = \mu(\mathbf{x}). \qquad \text{(C.98)}$$

We will show in Lemma 6 below that the projection $m(\cdot)$ admits an explicit higher-order asymptotic expansion as the distance $r \to 0$.

Lemma 6. *For each fixed* $\mathbf{x} \in \mathrm{supp}(\mathbf{X}) \subset \mathbb{R}^d$, *we have*

$$m(r) = m(0) + \frac{f(\mathbf{x})\,\mathrm{tr}(\mu''(\mathbf{x})) + 2\,\mu'(\mathbf{x})^T f'(\mathbf{x})}{2\,d\,f(\mathbf{x})}\, r^2 + O_4 r^4 \qquad \text{(C.99)}$$

as $r \to 0$, where O_4 is some bounded quantity depending only on d and the fourth-order partial derivatives of the underlying density function $f(\cdot)$ and regression function $\mu(\cdot)$. Here $g'(\cdot)$ and $g''(\cdot)$ stand for the gradient vector and the Hessian matrix, respectively, of a given function $g(\cdot)$.

Proof. We will exploit the spherical coordinate integration in our proof. Let us first introduce some necessary notation. Denote by $B(\mathbf{0}, r)$ the ball centered at $\mathbf{0}$ and with radius r in the Euclidean space $\mathbb{R}^d$, $\mathbb{S}^{d-1}$ the unit sphere in $\mathbb{R}^d$, ν a measure constructed on the unit sphere $\mathbb{S}^{d-1}$ as in (C.86), and $\boldsymbol{\xi} = (\xi_i) \in \mathbb{S}^{d-1}$ an arbitrary point on the unit sphere. Let V_d be the volume of the unit ball in $\mathbb{R}^d$ as given in (C.81). The integration with the spherical coordinates is equivalent to the standard integration through the identity

$$\int_{B(\mathbf{0},r)} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_0^r u^{d-1} \int_{\mathbb{S}^{d-1}} f(u\,\boldsymbol{\xi})\,\nu(\mathrm{d}\boldsymbol{\xi})\,\mathrm{d}u. \tag{C.100}$$

From Lemma 11 in Section C.1, we have the following integration formulas with the spherical coordinates

$$\begin{aligned}
\int_{\mathbb{S}^{d-1}} \nu(\mathrm{d}\boldsymbol{\xi}) &= d\,V_d, && \text{(C.101)}\\
\int_{\mathbb{S}^{d-1}} \boldsymbol{\xi}\,\nu(\mathrm{d}\boldsymbol{\xi}) &= \mathbf{0}, && \text{(C.102)}\\
\int_{\mathbb{S}^{d-1}} \boldsymbol{\xi}^T A\,\boldsymbol{\xi}\,\nu(\mathrm{d}\boldsymbol{\xi}) &= \mathrm{tr}(A)\,V_d, && \text{(C.103)}\\
\int_{\mathbb{S}^{d-1}} \xi_i\xi_j\xi_k\nu(\mathrm{d}\boldsymbol{\xi}) &= 0 \quad \text{for any } 1 \le i,j,k \le d, && \text{(C.104)}
\end{aligned}$$

where A is any $d \times d$ symmetric matrices. We will make use of the identities in (C.101)–(C.104) in our technical analysis.

Let us decompose $m(r)$ into two terms that we will analyze separately

$$\begin{aligned}
m(r) &= \lim_{\delta\to 0+} \mathbb{E}\left[\mu(\mathbf{X}) \mid r \le \|\mathbf{X}-\mathbf{x}\| \le r+\delta\right]\\
&= \lim_{\delta\to 0+} \frac{\mathbb{E}\left[\mu(\mathbf{X})\mathbb{1}(r \le \|\mathbf{X}-\mathbf{x}\| \le r+\delta)\right]}{\mathbb{P}\left(r \le \|\mathbf{X}-\mathbf{x}\| \le r+\delta\right)},
\end{aligned} \tag{C.105}$$

where $\mathbb{1}(\cdot)$ stands for the indicator function. In view of (C.100), we can obtain the spherical coordinate representations for the denominator and numerator in (C.105)

$$\mathbb{P}\left(r \le \|\mathbf{X}-\mathbf{x}\| \le r+\delta\right) = \int_r^{r+\delta} u^{d-1} \int_{\mathbb{S}^{d-1}} f(\mathbf{x}+u\,\boldsymbol{\xi})\,\nu(\mathrm{d}\boldsymbol{\xi})\,\mathrm{d}u \tag{C.106}$$

and

$$\begin{aligned}&\mathbb{E}\left[\mu(\mathbf{X})\mathbb{1}(r \le \|\mathbf{X}-\mathbf{x}\| \le r+\delta)\right]\\ &= \int_r^{r+\delta} u^{d-1} \int_{\mathbb{S}^{d-1}} \mu(\mathbf{x}+u\,\boldsymbol{\xi}) f(\mathbf{x}+u\,\boldsymbol{\xi})\,\nu(\mathrm{d}\boldsymbol{\xi})\,\mathrm{d}u.\end{aligned} \tag{C.107}$$

Note that in light of (C.105)–(C.107), an application of L'Hôpital's rule leads to

$$\begin{aligned}m(r) &= \lim_{\delta\to 0+} \frac{\mathbb{E}\left[\mu(\mathbf{X})\mathbb{1}(r \le \|\mathbf{X}-\mathbf{x}\| \le r+\delta)\right]}{\mathbb{P}\left(r \le \|\mathbf{X}-\mathbf{x}\| \le r+\delta\right)}\\ &= \frac{\int_{\mathbb{S}^{d-1}} \mu(\mathbf{x}+r\,\boldsymbol{\xi}) f(\mathbf{x}+r\,\boldsymbol{\xi})\,\nu(\mathrm{d}\boldsymbol{\xi})}{\int_{\mathbb{S}^{d-1}} f(\mathbf{x}+r\,\boldsymbol{\xi})\,\nu(\mathrm{d}\boldsymbol{\xi})}.\end{aligned} \tag{C.108}$$

First let us expand the denominator. Using the spherical coordinate integration, we can deduce that

$$\begin{aligned}&\int_{\mathbb{S}^{d-1}} f(\mathbf{x}+r\boldsymbol{\xi})\,\nu(\mathrm{d}\boldsymbol{\xi})\\ &= \int_{\mathbb{S}^{d-1}} \Big(f(\mathbf{x}) + f'(\mathbf{x})^T\,\boldsymbol{\xi}\,r + \frac{1}{2}\,\boldsymbol{\xi}^T f''(\mathbf{x})\,\boldsymbol{\xi}\,r^2 + \frac{1}{6}\sum_{1\le i,j,k\le d} \frac{\partial^3 f(\mathbf{x})}{\partial \mathbf{x}_i \partial \mathbf{x}_j \partial \mathbf{x}_k} \xi_i\xi_j\xi_k r^3\\ &\quad + \frac{1}{24}\sum_{1\le i,j,k,l\le d} \frac{\partial^4 f(\mathbf{x}+\theta r\boldsymbol{\xi})}{\partial \mathbf{x}_i \partial \mathbf{x}_j \partial \mathbf{x}_k \partial \mathbf{x}_l} \xi_i\xi_j\xi_k\xi_l r^4 \Big)\,\nu(\mathrm{d}\boldsymbol{\xi}),\end{aligned} \tag{C.109}$$

where $0<\theta<1$. Note that the fourth-order partial derivatives of f are bounded in some neiborghhood of $\mathbf{x}$ by Condition 2, and

$$\begin{aligned}\int_{\mathbb{S}^{d-1}} \sum_{1\le i,j,k,l\le d} |\xi_i\xi_j\xi_k\xi_l|\,\nu(\mathrm{d}\boldsymbol{\xi}) &= \int_{\mathbb{S}^{d-1}} \Big(\sum_{i=1}^d |\xi_i|\Big)^4 \nu(\mathrm{d}\boldsymbol{\xi})\\ &\le \int_{\mathbb{S}^{d-1}} d^2 \Big(\sum_{i=1}^d \xi_i^2\Big)^2 \nu(\mathrm{d}\boldsymbol{\xi})\\ &= d^2 \int_{\mathbb{S}^{d-1}} \nu(\mathrm{d}\boldsymbol{\xi}) = d^3 V_d.\end{aligned} \tag{C.110}$$

Thus, from (C.101)–(C.104) and (C.110) we can obtain

$$\int_{\mathbb{S}^{d-1}} f(\mathbf{x}+r\boldsymbol{\xi})\,\nu(\mathrm{d}\boldsymbol{\xi}) = f(\mathbf{x})\,d\,V_d + \frac{1}{2}\,\mathrm{tr}(f''(\mathbf{x}))V_d\,r^2 + R_1(d,f,\mathbf{x})\,r^4, \tag{C.111}$$

where the coefficient $R_1(d,f,\mathbf{x})$ in the remainder term is bounded and depends only on the fourth-order partial derivatives of f and dimensionality d.

For the numerator, it holds that

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}} \mu(\mathbf{x}+r\,\boldsymbol{\xi}) f(\mathbf{x}+r\,\boldsymbol{\xi})\,\nu(\mathrm{d}\boldsymbol{\xi}) \\
&= \int_{\mathbb{S}^{d-1}} \Big[\mu(\mathbf{x}) + \mu'(\mathbf{x})^T\boldsymbol{\xi}\,r + \frac{1}{2}\,\boldsymbol{\xi}^T\mu''(\mathbf{x})\,\boldsymbol{\xi}\,r^2 \\
&\quad + \frac{1}{6}\sum_{1\le i,j,k\le d} \frac{\partial^3\mu(\mathbf{x})}{\partial\mathbf{x}_i\partial\mathbf{x}_j\partial\mathbf{x}_k}\xi_i\xi_j\xi_k r^3 + \frac{1}{24}\sum_{1\le i,j,k,l\le d} \frac{\partial^4\mu(\mathbf{x}+\theta_1 r\boldsymbol{\xi})}{\partial\mathbf{x}_i\partial\mathbf{x}_j\partial\mathbf{x}_k\partial\mathbf{x}_l}\xi_i\xi_j\xi_k\xi_l r^4\Big] \\
&\quad \times \Big[f(\mathbf{x}) + f'(\mathbf{x})^T\boldsymbol{\xi}\,r + \frac{1}{2}\boldsymbol{\xi}^T f''(\mathbf{x})\,\boldsymbol{\xi}\,r^2 \\
&\quad + \frac{1}{6}\sum_{i,j,k} \frac{\partial^3 f(\mathbf{x})}{\partial\mathbf{x}_i\partial\mathbf{x}_j\partial\mathbf{x}_k}\xi_i\xi_j\xi_k r^3 + \frac{1}{24}\sum_{1\le i,j,k,l\le d} \frac{\partial^4 f(\mathbf{x}+\theta_2 r\boldsymbol{\xi})}{\partial\mathbf{x}_i\partial\mathbf{x}_j\partial\mathbf{x}_k\partial\mathbf{x}_l}\xi_i\xi_j\xi_k\xi_l r^4\Big]\,\nu(\mathrm{d}\boldsymbol{\xi}),
\end{aligned} \tag{C.112}$$

where $0<\theta_1<1$ and $0<\theta_2<1$. In the same manner as deriving (C.110), we can bound the integrals associated with r^4 and the higher-orders r^5, r^6, r^7, and r^8 under Condition 2 that the fourth-order partial derivatives of $f(\cdot)$ and $\mu(\cdot)$ are bounded in a neighborhood of $\mathbf{x}$. Hence, we can deduce that

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}} \mu(\mathbf{x}+r\,\boldsymbol{\xi}) f(\mathbf{x}+r\,\boldsymbol{\xi})\,\nu(\mathrm{d}\boldsymbol{\xi}) \\
&= \mu(\mathbf{x})f(\mathbf{x})\int_{\mathbb{S}^{d-1}} \nu(\mathrm{d}\boldsymbol{\xi}) + \frac{\mu(\mathbf{x})r^2}{2}\int_{\mathbb{S}^{d-1}} \boldsymbol{\xi}^T f''(\mathbf{x})\boldsymbol{\xi}\,\nu(\mathrm{d}\boldsymbol{\xi}) \\
&\quad + r^2\int_{\mathbb{S}^{d-1}} \boldsymbol{\xi}^T\mu'(\mathbf{x})f'(\mathbf{x})^T\boldsymbol{\xi}\,\nu(\mathrm{d}\boldsymbol{\xi}) + \frac{f(\mathbf{x})r^2}{2}\int_{\mathbb{S}^{d-1}} \boldsymbol{\xi}^T\mu''(\mathbf{x})\boldsymbol{\xi}\,\nu(\mathrm{d}\boldsymbol{\xi}) \\
&\quad + R_2(d,f,\mathbf{x})r^4 + o(r^4) \\
&= \mu(\mathbf{x})f(\mathbf{x})dV_d + \frac{1}{2}\left[f(\mathbf{x})\,\mathrm{tr}(\mu''(\mathbf{x})) + \mu(\mathbf{x})\;\mathrm{tr}(f''(\mathbf{x}))\right]V_d\,r^2 \\
&\quad + \mu'(\mathbf{x})^T f'(\mathbf{x})V_d\,r^2 + R_2(d,f,\mathbf{x})r^4 + o(r^4),
\end{aligned} \tag{C.113}$$

where the coefficient $R_2(d, f, \mathbf{x})$ in the remainder term is bounded and depends only on the fourth-order partial derivatives of f and dimensionality d. The last equality in (C.113) follows from (C.101)–(C.104). Therefore, substituting (C.111) and (C.113) into (C.108) leads to

$$m(r) = \mu(\mathbf{x}) + \frac{f(\mathbf{x})\mathrm{tr}(\mu''(\mathbf{x})) + 2\,\mu'(\mathbf{x})^T f'(\mathbf{x})}{2\,d\,f(\mathbf{x})}\, r^2 + O_4 r^4$$

as $r \to 0$, where O_4 is a bounded quantity depending only on d and the fourth-order partial derivatives of $f(\cdot)$ and $\mu(\cdot)$. This concludes the proof of Lemma 6.

B.7 Lemma 7 and its proof

Lemma 7 below provides us with the order of the variance for the first-order Hájek projection. To simplify the technical presentation, we use $\mathbf{Z}_i$ as a shorthand notation for $(\mathbf{X}_i, Y_i)$. Given any fixed vector $\mathbf{x}$, the projection of $\Phi(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_s)$ onto $\mathbf{Z}_1$ is denoted as $\Phi_1(\mathbf{x}; \mathbf{z}_1)$ given by

$$\begin{aligned}\Phi_1(\mathbf{x}; \mathbf{z}_1) &= \mathbb{E}\left[\Phi(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_s) | \mathbf{Z}_1 = \mathbf{z}_1\right] \\ &= \mathbb{E}\left[\Phi(\mathbf{x}; \mathbf{z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_s)\right]. \end{aligned} \tag{C.114}$$

Denote by $\mathbb{E}_i$ and $\mathbb{E}_{i:s}$ the expectations with respect to $\mathbf{Z}_i$ and $\{\mathbf{Z}_i, \mathbf{Z}_{i+1}, \cdots, \mathbf{Z}_s\}$, respectively.

Lemma 7. *For any fixed* $\mathbf{x}$*, the variance* η_1 *of* $\Phi_1(\mathbf{x}; \mathbf{z}_1)$ *defined in (C.114) satisfies that when* $s \to \infty$ *and* $s = o(n)$*,*

$$\lim_{n\to\infty} \frac{\mathrm{Var}(\Phi)}{n\eta_1} = 0. \tag{C.115}$$

Proof. A main ingredient of the proof is to decompose $\mathrm{Var}(\Phi)$ and η_1 using the conditioning arguments. Denote by $\zeta_{i,s}$ the indicator function for the event that $\mathbf{X}_i$ is the 1NN of $\mathbf{x}$

among $\{\mathbf{X}_1, \cdots, \mathbf{X}_s\}$. By symmetry, we can see that $\zeta_{i,s}$ are identically distributed with mean

$$\mathbb{E}\zeta_{i,s} = s^{-1}.$$

In addition, observe that $\Phi(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_s) = \sum_{i=1}^{s} y_i \zeta_{i,s}$. Then we can obtain an upper bound of $\operatorname{Var}\Phi$ as

$$\begin{aligned}\operatorname{Var}(\Phi) \leq \mathbb{E}[\Phi^2] &= \mathbb{E}\Big[\Big(\sum_{i=1}^{s} y_i \zeta_{1,s}\Big)^2\Big] = \sum_{i=1}^{s} \mathbb{E}[y_i^2 \zeta_{1,s}] \\ &= s\mathbb{E}[y_1^2 \zeta_{1,s}],\end{aligned}$$

where we have used the fact that $\zeta_{i,s}\zeta_{j,s} = 0$ with probability one when $i \neq j$.

Since $\mathbb{E}[\epsilon|\mathbf{X}] = 0$ by assumption, it holds that

$$\begin{aligned}s\mathbb{E}[y_1^2 \zeta_{1,s}] &= s\mathbb{E}[\mu^2(\mathbf{X}_1)\zeta_{1,s}] + \sigma_\epsilon^2 s\mathbb{E}[\zeta_{1,s}] \\ &= \mathbb{E}_1[\mu^2(\mathbf{X}_1) s\mathbb{E}_{2:s}[\zeta_{1,s}]] + \sigma_\epsilon^2.\end{aligned}$$

A key observation is that $\mathbb{E}_{2:s}[\zeta_{1,s}] = \{1 - \varphi(B(\mathbf{x}, \|\mathbf{X}_1 - \mathbf{x}\|))\}^{s-1}$ and $\mathbb{E}_1[s\mathbb{E}_{2:s}[\zeta_{1,s}]] = 1$. See Lemma 12 in Section C.2 for a list of properties for the indicator functions $\zeta_{i,s}$. Thus, $s\mathbb{E}_{2:s}[\zeta_{1,s}]$ behaves like a a Dirac measure at $\mathbf{x}$ as $s \to \infty$. Such observation leads to Lemma 13 in Section C.3, which entails that

$$\operatorname{Var}(\Phi) \leq \mu^2(\mathbf{x}) + \sigma_\epsilon^2 + o(1) \tag{C.116}$$

as $s \to \infty$.

To derive a lower bound for η_1, we exploit the idea in Theorem 3 of Peng et al. (2019). Let B be the event that $\mathbf{X}_1$ is the nearest neighbor of $\mathbf{x}$ among $\{\mathbf{X}_1, \cdots, \mathbf{X}_s\}$. Denote by

$\mathbf{X}_1^*$ the nearest point to $\mathbf{x}$ and y_1^* the corresponding response. Then we can deduce that

$$\begin{aligned}\Phi_1(\mathbf{x};\mathbf{Z}_1) &= \mathbb{E}[y_1\mathbb{1}_B|\mathbf{Z}_1] + \mathbb{E}[y_1^*\mathbb{1}_{B^c}|\mathbf{Z}_1] \\ &= y_1\mathbb{E}[\mathbb{1}_B|\mathbf{Z}_1] + \mathbb{E}[y_1^*\mathbb{1}_{B^c}|\mathbf{Z}_1] \\ &= \epsilon_1\mathbb{E}[\mathbb{1}_B|\mathbf{X}_1] + \mu(\mathbf{X}_1)\mathbb{E}[\mathbb{1}_B|\mathbf{X}_1] + \mathbb{E}[\mu(\mathbf{X}_1^*)\mathbb{1}_{B^c}|\mathbf{X}_1] \\ &= \epsilon_1\mathbb{E}[\mathbb{1}_B|\mathbf{X}_1] + \mathbb{E}[\mu(\mathbf{X}_1^*)|\mathbf{X}_1].\end{aligned}$$

Since ϵ is an independent model error term with $\mathbb{E}[\epsilon|\mathbf{X}] = 0$ by assumption, it holds that

$$\begin{aligned}\eta_1 &= \mathrm{Var}(\Phi_1(\mathbf{x};\mathbf{Z}_1)) = \mathrm{Var}(\epsilon_1\mathbb{E}[\mathbb{1}_B|\mathbf{X}_1]) + \mathrm{Var}(\mathbb{E}[\mu(\mathbf{X}_1^*)|\mathbf{X}_1]) \\ &\geq \mathrm{Var}(\epsilon_1\mathbb{E}[\mathbb{1}_B|\mathbf{X}_1]) = \sigma_\epsilon^2\mathbb{E}\big[\mathbb{E}^2[\mathbb{1}_B|\mathbf{X}_1]\big] \\ &= \frac{\sigma_\epsilon^2}{2s-1},\end{aligned} \tag{C.117}$$

where we have used the fact that

$$\mathbb{E}[\mathbb{E}^2[\mathbb{1}_B|\mathbf{X}_1]] = \mathbb{E}[\mathbb{1}_{B'}|\mathbf{X}_1] = \frac{1}{2s-1}$$

with B' representing the event that $\mathbf{X}_1$ is the nearest neighbor of $\mathbf{x}$ among the i.i.d. observations $\{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_s, \mathbf{X}_2', \cdots, \mathbf{X}_s'\}$.

We now turn to the upper bound for η_1. From the variance decomposition for $\mathrm{Var}(\Phi)$ given in (B.7), we can obtain

$$\begin{aligned}\mathrm{Var}(\Phi) &= \sum_{j=1}^{s}\binom{s}{j}\mathrm{Var}(g_j(\mathbf{x};\mathbf{Z}_1,\cdots,\mathbf{Z}_j)) \\ &= s\eta_1 + \sum_{j=2}^{s}\binom{s}{j}\mathrm{Var}(g_j(\mathbf{x};\mathbf{Z}_1,\cdots,\mathbf{Z}_j)),\end{aligned}$$

which along with (C.116) entails that

$$s\eta_1 \leq \mathrm{Var}(\Phi) \leq \mu^2(\mathbf{x}) + \sigma_\epsilon^2 + o(1). \tag{C.118}$$

Consequently, combining (C.117) and (C.118) leads to

$$\eta_1 \sim s^{-1}, \tag{C.119}$$

where $\sim$ denotes the asymptotic order. Finally, recall that it has been shown that $\mathrm{Var}(\Phi) \leq C$ for some positive constant depending upon $\mu(\mathbf{x})$ and σ_ϵ. Therefore, we see that as long as $s \to \infty$ and $s = o(n)$,

$$\frac{\mathrm{Var}(\Phi)}{n\eta_1} = O(\frac{s}{n}) \to 0,$$

which yields the desired conclusion in (C.115). This completes the proof of Lemma 7.

B.8 Lemma 8 and its proof

Assume that $s_1 < s_2$ for the two subsampling scales. Let us define

$$\Phi^{(1)}(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{s_2}) = \binom{s_2}{s_1}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_{s_1} \leq s_2} \Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \cdots, \mathbf{Z}_{i_{s_1}}) \tag{C.120}$$

and

$$\Phi^*(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{s_2}) = w_1^* \Phi^{(1)}(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{s_2}) + w_2^* \Phi(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{s_2}), \tag{C.121}$$

where w_1^* and w_2^* are determined by the system of linear equations (13)–(14).

Lemma 8. *The two-scale DNN estimator $D_n(s_1, s_2)(\mathbf{x})$ admits a U-statistic representation given by*

$$D_n(s_1, s_2)(\mathbf{x}) = \binom{n}{s_2}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_{s_2} \leq n} \Phi^*(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{s_2}), \tag{C.122}$$

where the kernel function $\Phi^*(\mathbf{x}; \cdot)$ is defined in (C.121).

Proof. From the definition of the two-scale DNN estimator $D_n(s_1, s_2)(\mathbf{x})$ introduced in (15), we have

$$D_n(s_1, s_2)(\mathbf{x}) = w_1^* \binom{n}{s_1}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_{s_1} \le n} \Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_{s_1}}) + w_2^* \binom{n}{s_2}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_{s_2} \le n} \Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_{s_2}}).$$

Thus, to establish the U-statistic representation for the two-scale DNN estimator $D_n(s_1, s_2)(\mathbf{x})$, it suffices to show that

$$\begin{aligned}
&\binom{n}{s_1}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_{s_1} \le n} \Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_{s_1}}) \\
&= \binom{n}{s_2}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_{s_2} \le n} \Phi^{(1)}(\mathbf{x}; \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{s_2}) \\
&= \binom{n}{s_2}^{-1} \binom{s_2}{s_1}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_{s_2} \le n} \sum_{1 \le j_1 < j_2 < \cdots < j_{s_1} \le s_2} \Phi(\mathbf{x}; \mathbf{Z}_{i_{j_1}}, \mathbf{Z}_{i_{j_2}}, \cdots, \mathbf{Z}_{i_{j_{s_1}}}). \qquad \text{(C.123)}
\end{aligned}$$

Observe that for each given tuple $1 \le u_1 < u_2 < \cdots < u_{s_1} \le n$, it will appear a total of $\binom{n-s_1}{s_2-s_1}$ times in the summation

$$\sum_{1 \le i_1 < i_2 < \cdots < i_{s_2} \le n} \sum_{1 \le j_1 < j_2 < \cdots < j_{s_1} \le s_2} \Phi(\mathbf{x}; \mathbf{Z}_{i_{j_1}}, \mathbf{Z}_{i_{j_2}}, \cdots, \mathbf{Z}_{i_{j_{s_1}}}).$$

Indeed, if $(i_{j_1}, i_{j_2}, \cdots, i_{j_{s_1}}) = (u_1, u_2, \cdots, u_{s_1})$ are fixed, then there exist $\binom{n-s_1}{s_2-s_1}$ options for the remaining $s_2 - s_1$ places in $(i_1, i_2, \cdots, i_{s_2})$. Consequently, it holds that

$$\begin{aligned}
&\binom{n}{s_2}^{-1} \binom{s_2}{s_1}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_{s_2} \le n} \sum_{1 \le j_1 < j_2 < \cdots < j_{s_1} \le s_2} \Phi(\mathbf{x}; \mathbf{Z}_{i_{j_1}}, \mathbf{Z}_{i_{j_2}}, \cdots, \mathbf{Z}_{i_{j_{s_1}}}) \\
&= \binom{n}{s_2}^{-1} \binom{s_2}{s_1}^{-1} \binom{n-s_1}{s_2-s_1} \sum_{1 \le i_1 < i_2 < \cdots < i_{s_1} \le n} \Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_{s_1}}) \\
&= \binom{n}{s_1}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_{s_1} \le n} \Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_{s_1}}),
\end{aligned}$$

which establishes the desired claim in (C.123). This concludes the proof of Lemma 8.

B.9 Lemma 9 and its proof

We provide in Lemma 9 below the order of the variance of the kernel function Φ^* defined in (C.121) for the two-scale DNN estimator $D_n(s_1, s_2)(\mathbf{x})$, which states that the variance of the kernel function is bounded from above by some positive constant depending upon the underlying distributions. Denote by $\mathrm{Var}(\Phi^*) = \mathrm{Var}[\Phi^*(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2})]$ for simplicity.

Lemma 9. *Under the conditions of Theorem 3, there exists some positive constant C depending upon c_1 and c_2 such that*

$$\mathrm{Var}(\Phi^*) \leq C\big(\mu^2(\mathbf{x}) + \sigma_\epsilon^2 + o(1)\big) \tag{C.124}$$

as $s_1 \to \infty$ and $s_2 \to \infty$.

Proof. Since $\mathrm{Var}(\Phi^*) \leq \mathbb{E}[(\Phi^*)^2]$, it suffices to bound $\mathbb{E}[(\Phi^*)^2]$. It follows that

$$\begin{aligned}
\mathbb{E}[(\Phi^*)^2] &\leq 2(w_1^*)^2 \mathbb{E}\big\{[\Phi^{(1)}(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2})]^2\big\} \\
&\quad + 2(w_2^*)^2 \mathbb{E}[\Phi^2(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2})] \\
&\leq 2(w_1^*)^2 \mathbb{E}[\Phi^2(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_1})] \\
&\quad + 2(w_2^*)^2 \mathbb{E}[\Phi^2(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2})],
\end{aligned} \tag{C.125}$$

where the last inequality holds since

$$\begin{aligned}
&\mathbb{E}\big\{[\Phi^{(1)}(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2})]^2\big\} \\
&= \binom{s_2}{s_1}^{-2} \sum_{\substack{1 \leq i_1 < \cdots < i_{s_1} \leq s_2 \\ 1 \leq j_1 < \cdots < j_{s_1} \leq s_2}} \mathbb{E}\big\{\Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_{s_1}}) \Phi(\mathbf{x}; \mathbf{Z}_{j_1}, \cdots, \mathbf{Z}_{j_{s_1}})\big\} \\
&\leq \binom{s_2}{s_1}^{-2} \sum_{\substack{1 \leq i_1 < \cdots < i_{s_1} \leq s_2 \\ 1 \leq j_1 < \cdots < j_{s_1} \leq s_2}} \mathbb{E}[\Phi^2(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_1})] \\
&= \mathbb{E}[\Phi^2(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_1})].
\end{aligned}$$

Since $\Phi(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_1}) = \sum_{i=1}^{s_1} y_i \zeta_{i,s_1}$ and $\zeta_{i,s_1}\zeta_{j,s_1} = 0$ with probability one when $i \neq j$, we can deduce that

$$\begin{aligned}\mathbb{E}[\Phi^2(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_1})] &= \mathbb{E}\Big[\Big(\sum_{i=1}^{s_1} y_i \zeta_{i,s_1}\Big)^2\Big] = \sum_{i=1}^{s_1}\sum_{j=1}^{s_1} y_i y_j \zeta_{i,s_1}\zeta_{j,s_1} \\ &= \sum_{i=1}^{s_1} y_i^2 \zeta_{i,s_1} = s_1 \mathbb{E}[y_1^2 \zeta_{1,s_1}] \\ &= s_1 \mathbb{E}[\mu^2(\mathbf{X}_1)\zeta_{1,s_1}] + \sigma_\epsilon^2 s_1 \mathbb{E}[\zeta_{1,s_1}].\end{aligned}$$

Note that $s_1\mathbb{E}[\zeta_{1,s_1}] = \sum_{i=1}^{n} \zeta_{i,s_1}$. Furthermore, it follows from Lemma 13 in Section C.3 that

$$s_1 \mathbb{E}[\mu^2(\mathbf{X}_1)\zeta_{1,s_1}] \to \mu^2(\mathbf{x})$$

as $s_1 \to \infty$. Thus, we have that as $s_1 \to \infty$,

$$\mathbb{E}[\Phi^2(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_1})] = \mu^2(\mathbf{x}) + \sigma_\epsilon^2 + o(1). \tag{C.126}$$

Similarly, we can show that as $s_2 \to \infty$,

$$\mathbb{E}[\Phi^2(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2})] = \mu^2(\mathbf{x}) + \sigma_\epsilon^2 + o(1). \tag{C.127}$$

Consequently, combining (C.125), (C.126), and (C.127) results in

$$\mathbb{E}[(\Phi^*)^2] \leq 2\big[(w_1^*)^2 + (w_2^*)^2\big]\big[\mu^2(\mathbf{x}) + \sigma_\epsilon^2 + o(1)\big]. \tag{C.128}$$

Since $c_1 \leq s_1/s_2 \leq c_2$ by assumption, it holds that

$$(w_1^*)^2 \leq C \quad \text{and} \quad (w_2^*)^2 \leq C$$

for some absolute positive constant C depending upon c_1 and c_2, which together with (C.128) entails the desired upper bound in (C.124). This completes the proof of Lemma 9.

B.10 Lemma 10 and its proof

Lemma 10 below establishes the order of the variance for the first-order Hájek projection of the kernel function Φ^* defined in (C.121). Recall that in the proof of Theorem 3 in Section B.3, we have defined that for each $1 \le i \le s_2$,

$$\Phi_i^*(\mathbf{x}; \mathbf{z}_1, \cdots, \mathbf{z}_i) = \mathbb{E}[\Phi^*(\mathbf{x}; \mathbf{z}_1, \cdots, \mathbf{z}_i, \mathbf{Z}_{i+1}, \cdots, \mathbf{Z}_{s_2}) \,|\, \mathbf{z}_1, \cdots, \mathbf{z}_i],$$

$$\begin{aligned} g_i^*(\mathbf{z}_1, \cdots, \mathbf{z}_i) = \Phi_i^*(\mathbf{x}; \mathbf{z}_1, \cdots, \mathbf{z}_i) - \mathbb{E}\Phi^*(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_i) \\ - \sum_{j=1}^{i-1} \sum_{1 \le \alpha_1 < \cdots < \alpha_j \le i} g_j^*(\mathbf{z}_{\alpha_1}, \cdots, \mathbf{z}_{\alpha_j}), \end{aligned}$$

and $\eta_1^* = \mathrm{Var}(\Phi_1^*(\mathbf{x}; \mathbf{Z}_1))$.

Lemma 10. *Under the conditions of Theorem 3, it holds that*

$$\eta_1^* \sim s_2^{-1}, \tag{C.129}$$

where $\sim$ denotes the asymptotic order.

Proof. We begin with the lower bound for η_1^*. The proof follows the ideas used in the proof of Lemma 7 in Section B.7. By definition, it holds that

$$\begin{aligned} \Phi_1^*(\mathbf{x}; \mathbf{Z}_1) &= w_1^* \binom{s_2}{s_1}^{-1} \sum_{1 \le i_1 < \cdots < i_{s_1} \le s_2} \mathbb{E}[\Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_{s_1}}) | \mathbf{Z}_1] \\ &\quad + w_2^* \mathbb{E}[\Phi(\mathbf{x}; \mathbf{Z}_{i_1}, \cdots, \mathbf{Z}_{i_{s_2}}) | \mathbf{Z}_1] \\ &= w_1^* \frac{s_2 - s_1}{s_2} \mathbb{E}[\Phi(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_1})] + w_1^* \frac{s_1}{s_2} \mathbb{E}[\Phi(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_1}) | \mathbf{Z}_1] \\ &\quad + w_2^* \mathbb{E}[\Phi(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2}) | \mathbf{Z}_1]. \end{aligned}$$

Since the first term on the right-hand side of the above equality is a constant, we have

$$\begin{aligned} \mathrm{Var}(\Phi_1^*(\mathbf{x}; \mathbf{Z}_1)) = \mathrm{Var}\Big(& w_1^* \frac{s_1}{s_2} \mathbb{E}[\Phi(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_1}) | \mathbf{Z}_1] \\ & + w_2^* \mathbb{E}[\Phi(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2}) | \mathbf{Z}_1] \Big). \end{aligned}$$

Denote by A_1 the event that $\mathbf{X}_1$ is the nearest neighbor of $\mathbf{x}$ among $\{\mathbf{X}_1, \cdots, \mathbf{X}_{s_1}\}$ and A_2 the event that $\mathbf{X}_1$ is the nearest neighbor of $\mathbf{x}$ among $\{\mathbf{X}_1, \cdots, \mathbf{X}_{s_2}\}$. Let $\mathbf{X}_1^*$ be the nearest point to $\mathbf{x}$ among $\{\mathbf{X}_1, \cdots, \mathbf{X}_{s_1}\}$ and y_1^* the corresponding value of the response. Similarly, we define $\breve{\mathbf{X}}_1$ as the nearest point to $\mathbf{x}$ among $\{\mathbf{X}_1, \cdots, \mathbf{X}_{s_2}\}$ and $\breve{y}_1$ as the corresponding value of the response. Since $\epsilon_i \perp\!\!\!\perp \mathbf{X}_i$ and $\mathbb{E}[\epsilon_i] = 0$ by assumption, we can write

$$\begin{aligned}
\mathbb{E}[\Phi(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_1})|\mathbf{Z}_1] &= \mathbb{E}[y_1 \mathbb{1}_{A_1}|\mathbf{Z}_1] + \mathbb{E}[y_1^* \mathbb{1}_{A_1^c}|\mathbf{Z}_1] \\
&= \epsilon_1 \mathbb{E}[\mathbb{1}_{A_1}|\mathbf{X}_1] + \mathbb{E}[\mu(\mathbf{X}_1)\mathbb{1}_{A_1}|\mathbf{X}_1] + \mathbb{E}[\mu(\mathbf{X}_1^*)\mathbb{1}_{A_1^c}|\mathbf{X}_1] \\
&= \epsilon_1 \mathbb{E}[\mathbb{1}_{A_1}|\mathbf{X}_1] + \mathbb{E}[\mu(\mathbf{X}_1^*)|\mathbf{X}_1].
\end{aligned}$$

Similarly, we can show that

$$\mathbb{E}[\Phi(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2})|\mathbf{Z}_1] = \epsilon_1 \mathbb{E}[\mathbb{1}_{A_2}|\mathbf{X}_1] + \mathbb{E}[\mu(\breve{\mathbf{X}}_1)|\mathbf{X}_1].$$

Thus, we can obtain

$$\begin{aligned}
\mathrm{Var}(\Phi_1^*(\mathbf{x}; \mathbf{Z}_1)) = \mathrm{Var}\Big\{ &\epsilon_1 \Big(w_1^* \frac{s_1}{s_2} \mathbb{E}[\mathbb{1}_{A_1}|\mathbf{X}_1] + w_2^* \mathbb{E}[\mathbb{1}_{A_2}|\mathbf{X}_1] \Big) \\
&+ w_1^* \frac{s_1}{s_2} \mathbb{E}[\mu(\mathbf{X}_1^*)|\mathbf{X}_1] + w_2^* \mathbb{E}[\mu(\breve{\mathbf{X}}_1)|\mathbf{X}_1] \Big\},
\end{aligned}$$

which along with the assumption of $\epsilon_1 \perp\!\!\!\perp \mathbf{X}_1$ and $\mathbb{E}[\epsilon_1] = 0$ yields

$$\begin{aligned}
\mathrm{Var}(\Phi_1^*(\mathbf{x}; \mathbf{Z}_1)) &= \mathrm{Var}\Big\{ \epsilon_1 \Big(w_1^* \frac{s_1}{s_2} \mathbb{E}[\mathbb{1}_{A_1}|\mathbf{X}_1] + w_2^* \mathbb{E}[\mathbb{1}_{A_2}|\mathbf{X}_1] \Big) \Big\} \\
&\quad + \mathrm{Var}\Big\{ w_1^* \frac{s_1}{s_2} \mathbb{E}[\mu(\mathbf{X}_1^*)|\mathbf{X}_1] + w_2^* \mathbb{E}[\mu(\breve{\mathbf{X}}_1)|\mathbf{X}_1] \Big\} \\
&\geq \mathrm{Var}\Big\{ \epsilon_1 \Big(w_1^* \frac{s_1}{s_2} \mathbb{E}[\mathbb{1}_{A_1}|\mathbf{X}_1] + w_2^* \mathbb{E}[\mathbb{1}_{A_2}|\mathbf{X}_1] \Big) \Big\}.
\end{aligned}$$

Furthermore, we can deduce that

$$\begin{aligned}
&\mathrm{Var}\left\{\epsilon_1\Big(w_1^*\frac{s_1}{s_2}\mathbb{E}[\mathbb{1}_{A_1}|\mathbf{X}_1]+w_2^*\mathbb{E}[\mathbb{1}_{A_2}|\mathbf{X}_1]\Big)\right\}\\
&=\sigma_\epsilon^2\mathbb{E}\Big\{\Big(w_1^*\frac{s_1}{s_2}\mathbb{E}[\mathbb{1}_{A_1}|\mathbf{X}_1]+w_2^*\mathbb{E}[\mathbb{1}_{A_2}|\mathbf{X}_1]\Big)^2\Big\}\\
&=\sigma_\epsilon^2\Big\{\Big(w_1^*\frac{s_1}{s_2}\Big)^2\mathbb{E}\big[\mathbb{E}^2[\mathbb{1}_{A_1}|\mathbf{X}_1]\big]+2w_1^*w_2^*\frac{s_1}{s_2}\mathbb{E}\big[\mathbb{E}[\mathbb{1}_{A_1}|\mathbf{X}_1]\mathbb{E}[\mathbb{1}_{A_2}|\mathbf{X}_1]\big]\\
&\quad+(w_2^*)^2\mathbb{E}\big[\mathbb{E}^2[\mathbb{1}_{A_2}|\mathbf{X}_1]\big]\Big\}.
\end{aligned}$$

Let us make use of the following basic facts

$$\begin{aligned}
\mathbb{E}\big[\mathbb{E}^2[\mathbb{1}_{A_1}|\mathbf{X}_1]\big]&=\frac{1}{2s_1-1},\\
\mathbb{E}\big[\mathbb{E}[\mathbb{1}_{A_1}|\mathbf{X}_1]\mathbb{E}[\mathbb{1}_{A_2}|\mathbf{X}_1]\big]&=\frac{1}{s_1+s_2-1},\\
\mathbb{E}\big[\mathbb{E}^2[\mathbb{1}_{A_2}|\mathbf{X}_1]\big]&=\frac{1}{2s_2-1}.
\end{aligned}$$

Then it follows that

$$\begin{aligned}
\mathrm{Var}(\Phi_1^*(\mathbf{x};\mathbf{Z}_1))\geq\sigma_\epsilon^2\Big\{&\Big(w_1^*\frac{s_1}{s_2}\Big)^2\frac{1}{2s_1-1}+2w_1^*w_2^*\frac{s_1}{s_2}\frac{1}{s_1+s_2-1}\\
&+(w_2^*)^2\frac{1}{2s_2-1}\Big\}.
\end{aligned}\tag{C.130}$$

By (C.130) and the assumption of $c_1\leq s_1/s_2\leq c_2$, we can obtain

$$\mathrm{Var}(\Phi_1^*(\mathbf{x};\mathbf{Z}_1))\geq C\sigma_\epsilon^2 s_2^{-1}\tag{C.131}$$

for some positive constant C depending upon c_1 and c_2.

We next proceed to show the upper bound for $\mathrm{Var}(\Phi_1^*(\mathbf{x};\mathbf{Z}_1))$. Since

$$\Phi^*(\mathbf{x};\mathbf{Z}_1,\cdots,\mathbf{Z}_{s_2})-\mathbb{E}\Phi^*(\mathbf{x};\mathbf{Z}_1,\cdots,\mathbf{Z}_{s_2})=\sum_{j=1}^{s_2}\sum_{1\leq\alpha_1<\cdots<\alpha_j\leq s_2}g_j^*(\mathbf{Z}_{\alpha_1},\cdots,\mathbf{Z}_{\alpha_j}),$$

we see that

$$\mathrm{Var}(\Phi^*(\mathbf{x};\mathbf{Z}_1,\cdots,\mathbf{Z}_{s_2}))=\sum_{j=1}^{s_2}\binom{s_2}{j}\mathrm{Var}(g_j^*(\mathbf{Z}_1,\cdots,\mathbf{Z}_j)).$$

Then it follows that

$$\mathrm{Var}(\Phi^*(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2})) \geq s_2 \, \mathrm{Var}(\Phi_1^*(\mathbf{x}; \mathbf{Z}_1)).$$

Recall that it has been shown in Lemma 9 in Section B.9 that

$$\mathrm{Var}(\Phi^*(\mathbf{x}; \mathbf{Z}_1, \cdots, \mathbf{Z}_{s_2})) \leq C,$$

where C is some positive constant depending upon c_1, c_2, and the underlying distributions. Therefore, we can deduce that

$$\mathrm{Var}(\Phi_1^*(\mathbf{x}; \mathbf{Z}_1)) \leq C s_2^{-1}, \tag{C.132}$$

which together with (C.131) entails the desired asymptotic order in (C.129). This concludes the proof of Lemma 10.

D Additional technical details

C.1 Lemma 11 and its proof

We present in Lemma 11 below some useful spherical integration formulas.

Lemma 11. *Let $\mathbb{S}^{d-1}$ be the unit sphere in $\mathbb{R}^d$, ν some measure constructed specifically on the unit sphere $\mathbb{S}^{d-1}$, and $\boldsymbol{\xi} = (\xi_i) \in \mathbb{S}^{d-1}$ an arbitrary point on the unit sphere. Then for any $d \times d$ symmetric matrices A, it holds that*

$$\int_{\mathbb{S}^{d-1}} \nu(\mathrm{d}\boldsymbol{\xi}) = d\, V_d, \tag{D.133}$$

$$\int_{\mathbb{S}^{d-1}} \boldsymbol{\xi}\, \nu(\mathrm{d}\boldsymbol{\xi}) = \mathbf{0}, \tag{D.134}$$

$$\int_{\mathbb{S}^{d-1}} \boldsymbol{\xi}^T A\, \boldsymbol{\xi}\, \nu(\mathrm{d}\boldsymbol{\xi}) = \mathrm{tr}(A)\, V_d, \tag{D.135}$$

$$\int_{\mathbb{S}^{d-1}} \xi_i \xi_j \xi_k \nu(\mathrm{d}\boldsymbol{\xi}) = 0 \quad \textit{for any } 1 \leq i, j, k \leq d, \tag{D.136}$$

where $V_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$ *denotes the volume of the unit ball in* $\mathbb{R}^d$.

Proof. It is easy to see that identities (D.134) and (D.136) hold. This is because for each of them, the integrand is an odd function of variable $\boldsymbol{\xi}$, which entails that the integral is zero. Identity (D.133) can be derived using the iterated integral

$$\begin{aligned} V_d &= \int_0^1 \int_{\mathbb{S}^{d-1}} \rho^{d-1} \nu(\mathrm{d}\boldsymbol{\xi})\, \mathrm{d}\rho = \left(\int_0^1 \rho^{d-1}\, \mathrm{d}\rho \right) \left(\int_{\mathbb{S}^{d-1}} \nu(\mathrm{d}\boldsymbol{\xi}) \right) \\ &= \frac{1}{d} \int_{\mathbb{S}^{d-1}} \nu(\mathrm{d}\boldsymbol{\xi}). \end{aligned}$$

To prove (D.135), we first represent the integral in (D.135) as a sum of integrals by expanding the quadratic expression in the integrand

$$\int_{\mathbb{S}^{d-1}} \boldsymbol{\xi}^T A\, \boldsymbol{\xi}\ \nu(\mathrm{d}\boldsymbol{\xi}) = \sum_{1 \le i,j \le d} A_{ij} \int_{\mathbb{S}^{d-1}} \xi_i \xi_j\ \nu(\mathrm{d}\boldsymbol{\xi}). \tag{D.137}$$

For $i \neq j$, we have by symmetry that

$$\int_{\mathbb{S}^{d-1}} \xi_i \xi_j\ \nu(\mathrm{d}\boldsymbol{\xi}) = \int_{\mathbb{S}^{d-1}} -\xi_i \xi_j\ \nu(\mathrm{d}\boldsymbol{\xi}) = 0. \tag{D.138}$$

Thus, it holds that

$$\begin{aligned} \int_{\mathbb{S}^{d-1}} \boldsymbol{\xi}^T A\, \boldsymbol{\xi}\ \nu(\mathrm{d}\boldsymbol{\xi}) &= \sum_{i=1}^{d} A_{ii} \int_{\mathbb{S}^{d-1}} \xi_i^2\ \nu(\mathrm{d}\boldsymbol{\xi}) \\ &= \mathrm{tr}(A) \int_{\mathbb{S}^{d-1}} \xi_1^2\ \nu(\mathrm{d}\boldsymbol{\xi}). \end{aligned} \tag{D.139}$$

When $d = 1$, $\mathbb{S}^{d-1}$ reduces to the trivial case of two points, 1 and -1. Then we can obtain that for $d = 1$,

$$\int_{\mathbb{S}^{d-1}} \boldsymbol{\xi}^T A\, \boldsymbol{\xi}\ \nu(\mathrm{d}\boldsymbol{\xi}) = 2\mathrm{tr}(A) = \mathrm{tr}(A) V_d, \tag{D.140}$$

where the last equality comes from the fact that $V_d = 2$ for $d = 1$. When $d \geq 2$, we now use the spherical coordinates: $\xi_1 = \cos(\phi_1)$, $\xi_k = \cos(\phi_k) \prod_{i=1}^{k-1} \sin(\phi_i)$ for $1 \leq k \leq d-1$,

and $\xi_d = \prod_{i=1}^{d-1} \sin(\phi_i)$, where $0 \le \phi_{d-1} < 2\pi$ and $0 \le \phi_i < \pi$ for $1 \le i \le d-2$. Then the volume element becomes

$$\nu(\mathrm{d}\boldsymbol{\xi}) = \left(\prod_{i=1}^{d-2} \sin^{d-1-i}(\phi_i) \right) \prod_{i=1}^{d} \mathrm{d}\phi_i.$$

It follows that

$$\begin{aligned}
\int_{\mathbb{S}^{d-1}} \xi_1^2 \, \nu(\mathrm{d}\boldsymbol{\xi}) &= \int_0^{2\pi} \int_0^{\pi} \cdots \int_0^{\pi} \cos^2(\phi_1) \left(\prod_{i=1}^{d-2} \sin^{d-1-i}(\phi_i) \right) \prod_{i=1}^{d} \mathrm{d}\phi_i \\
&= \frac{\int_0^{\pi} \cos^2(\phi_1) \sin^{d-2}(\phi_1) \, \mathrm{d}\phi_i}{\int_0^{\pi} \sin^{d-2}(\phi_1) \, \mathrm{d}\phi_i} \int_0^{2\pi} \int_0^{\pi} \cdots \int_0^{\pi} \left(\prod_{i=1}^{d-2} \sin^{d-1-i}(\phi_i) \right) \prod_{i=1}^{d} \mathrm{d}\phi_i \\
&= \frac{\int_0^{\pi} \cos^2(\phi_1) \sin^{d-2}(\phi_1) \, \mathrm{d}\phi_i}{\int_0^{\pi} \sin^{d-2}(\phi_1) \, \mathrm{d}\phi_i} \int_{\mathbb{S}^{d-1}} \nu(\mathrm{d}\boldsymbol{\xi}) \\
&= \frac{\int_0^{\pi} \cos^2(\phi_1) \sin^{d-2}(\phi_1) \, \mathrm{d}\phi_i}{\int_0^{\pi} \sin^{d-2}(\phi_1) \, \mathrm{d}\phi_i} dV_d. \qquad \text{(D.141)}
\end{aligned}$$

By applying the integration by parts twice to the numerator from the above expression, we can obtain

$$\int_0^{\pi} \cos^2(\phi_1) \sin^{d-2}(\phi_1) \, \mathrm{d}\phi_1 = \frac{1}{d-1} \int_0^{\pi} \sin^d(\phi_1) \, \mathrm{d}\phi_1.$$

In addition, using the trigonometric integration formulas, we can show that

$$\frac{\int_0^{\pi} \sin^d(\phi_1) \, \mathrm{d}\phi_1}{\int_0^{\pi} \sin^{d-2}(\phi_1) \, \mathrm{d}\phi_1} = \frac{d-1}{d},$$

which along with (D.139) and (D.141) leads to

$$\int_{\mathbb{S}^{d-1}} \boldsymbol{\xi}^T A \, \boldsymbol{\xi} \, \nu(\mathrm{d}\boldsymbol{\xi}) = \mathrm{tr}(A) V_d$$

for the case of $d = 2$. Also, it is easy to see that the same formula holds for the case of $d = 1$ by (D.140). This completes the proof of Lemma 11.

C.2 Lemma 12 and its proof

Let us define $\zeta_{i,s}$ as the indicator function for the event that $\mathbf{X}_i$ is the 1NN of $\mathbf{x}$ among $\{\mathbf{X}_1, \cdots, \mathbf{X}_s\}$. We provide in Lemma 12 below a list of properties for these indicator functions $\zeta_{i,s}$.

Lemma 12. *The indicator functions $\zeta_{i,s}$ satisfy that*

1) For any $i \neq j$, we have $\zeta_{i,s}\zeta_{j,s} = 0$ with probability one;

2) $\sum_{i=1}^{s} \zeta_{i,s} = 1$;

3) $\mathbb{E}[\zeta_{i,s}] = s^{-1}$;

4) $\mathbb{E}_{2:s}[\zeta_{1,s}] = \{1-\varphi(B(\mathbf{x}, \|\mathbf{X}_1-\mathbf{x}\|))\}^{s-1}$, where $\mathbb{E}_{i:s}$ denotes the expectation with respect to $\{\mathbf{Z}_i, \mathbf{Z}_{i+1}, \cdots, \mathbf{Z}_s\}$.

The proof of Lemma 12 involves some standard calculations and thus we omit it here for simplicity. Let us make some remarks on $\mathbb{E}_{2:s}[\zeta_{1,s}]$ that can be regarded as a function of $\mathbf{X}_1$. The last property in Lemma 12 above shows that $\mathbb{E}_{2:s}[\zeta_{1,s}]$ vanishes asymptotically as s tends to infinity, unless $\mathbf{X}_1$ is equal to $\mathbf{x}$. Moreover, we see that

$$\mathbb{E}_1[\mathbb{E}_{2:s}[\zeta_{1,s}]] = s^{-1}.$$

These two facts suggest that $\mathbb{E}_{2:s}[\zeta_{1,s}]$ tends to approximate the Dirac delta function at $\mathbf{x}$, which will be established formally in Lemma 13 in Section C.3.

C.3 Lemma 13 and its proof

Lemma 13. *For any L^1 function f that is continuous at $\mathbf{x}$, it holds that*

$$\lim_{s\to\infty} \mathbb{E}_1[f(\mathbf{X}_1)s\mathbb{E}_{2:s}[\zeta_{1,s}]] = f(\mathbf{x}). \tag{D.142}$$

Proof. We will show that the absolute difference $|\mathbb{E}_1[f(\mathbf{X}_1)s\mathbb{E}_{2:s}[\zeta_{1,s}]] - f(\mathbf{x})|$ converges to zero as $s \to \infty$. By property 3) in Lemma 12 in Section C.2, we have

$$\mathbb{E}_1[s\mathbb{E}_{2:s}[\zeta_{1,s}]] = 1.$$

Thus, we can deduce that

$$\begin{aligned} |\mathbb{E}_1[f(\mathbf{X}_1)s\mathbb{E}_{2:s}[\zeta_{1,s}]] - f(\mathbf{x})| &= |\mathbb{E}_1[(f(\mathbf{X}_1) - f(\mathbf{x}))\, s\mathbb{E}_{2:s}[\zeta_{1,s}]]| \\ &\leq \mathbb{E}_1[|f(\mathbf{X}_1) - f(\mathbf{x})|s\mathbb{E}_{2:s}[\zeta_{1,s}]]. \end{aligned} \tag{D.143}$$

Let $\epsilon > 0$ be arbitrarily given. By the continuity of function f at point $\mathbf{x}$, there exists a neighborhood $B(\mathbf{x}, \delta)$ of $\mathbf{x}$ with some $\delta > 0$ such that

$$|f(\mathbf{X}_1) - f(\mathbf{x})| < \epsilon$$

for all $\mathbf{X}_1 \in B(\mathbf{x}, \delta)$. We will decompose the above expectation in (D.143) into two parts: one inside and the other outside of $B(\mathbf{x}, \delta)$ as

$$\begin{aligned} \mathbb{E}_1[|f(\mathbf{X}_1) - f(\mathbf{x})|s\mathbb{E}_{2:s}[\zeta_{1,s}]] &= \mathbb{E}_1[|f(\mathbf{X}_1) - f(\mathbf{x})|s\mathbb{E}_{2:s}[\zeta_{1,s}]\mathbb{1}_{B(\mathbf{x},\delta)}(\mathbf{X}_1)] \\ &+ \mathbb{E}_1[|f(\mathbf{X}_1) - f(\mathbf{x})|s\mathbb{E}_{2:s}[\zeta_{1,s}]\mathbb{1}_{B^c(\mathbf{x},\delta)}(\mathbf{X}_1)], \end{aligned} \tag{D.144}$$

where the superscript c stands for set complement in $\mathbb{R}^d$.

The first term on the right-hand side of (D.144) is bounded by ϵ since

$$\begin{aligned} \mathbb{E}_1[|f(\mathbf{X}_1) - f(\mathbf{x})|s\mathbb{E}_{2:s}[\zeta_{1,s}]\mathbb{1}_{B(\mathbf{x},\delta)}(\mathbf{X}_1)] &\leq \mathbb{E}_1[\epsilon s\mathbb{E}_{2:s}[\zeta_{1,s}]\mathbb{1}_{B(\mathbf{x},\delta)}(\mathbf{X}_1)] \\ &\leq \mathbb{E}_1[\epsilon s\mathbb{E}_{2:s}[\zeta_{1,s}]] = \epsilon. \end{aligned} \tag{D.145}$$

To bound the second term on the right-hand side of (D.144), observe that

$$B(\mathbf{x}, \delta) \subset B(\mathbf{x}, \|\mathbf{X}_1 - \mathbf{x}\|)$$

when $\mathbf{X}_1 \in B^c(\mathbf{x}, \delta)$. Then an application of Lemma 12 gives

$$\mathbb{E}_{2:s}[\zeta_{1,s}] \leq (1 - \varphi(B(\mathbf{x}, \delta)))^{s-1}$$

when $\mathbf{X}_1 \in B^c(\mathbf{x}, \delta)$. Thus, we can deduce that

$$\begin{aligned}
&\mathbb{E}_1[|f(\mathbf{X}_1) - f(\mathbf{x})| s\mathbb{E}_{2:s}[\zeta_{1,s}]\mathbb{1}_{B^c(\mathbf{X},\delta)}(\mathbf{X}_1)] \\
&\leq \mathbb{E}_1[|f(\mathbf{X}_1) - f(\mathbf{x})| s(1 - \varphi(B(\mathbf{x}, \delta)))^{s-1}\mathbb{1}_{B^c(\mathbf{X},\delta)}(\mathbf{X}_1)] \\
&\leq s(1 - \varphi(B(\mathbf{x}, \delta)))^{s-1}\mathbb{E}_1[|f(\mathbf{X}_1) - f(\mathbf{x})|] \\
&\leq s(1 - \varphi(B(\mathbf{x}, \delta)))^{s-1} \left(\|f\|_{L^1} + f(\mathbf{x})\right). \qquad \text{(D.146)}
\end{aligned}$$

Finally, we see that the right-hand side of the last equation in (D.146) tends to 0 as $s \to \infty$. Therefore, for large enough s,

$$\mathbb{E}_1[|f(\mathbf{X}_1) - f(\mathbf{x})| s\mathbb{E}_{2:s}[\zeta_{1,s}]\mathbb{1}_{B^c(\mathbf{X},\delta)}(\mathbf{X}_1)]$$

can be bounded from above by 2ϵ. Since the choice of $\epsilon > 0$ is arbitrary, combining such upper bound, (D.143), (D.144), and (D.145) yields the desired limit in (D.142) as $s \to \infty$. This concludes the proof of Lemma 13.

p	Method	True Mean	Est. Mean (SE)	Bias	MSE	Est. Variance	Bias (R)	MSE (R)
			Simulation setting 1					
10	TDNN	-1.12500	-1.63763 (0.01)	-0.51263	0.36274	0.09192	-	-
10	CF	-1.12500	0.06665 (0.00078)	1.19165	1.42063	0.00078	-	-
			Simulation setting 2					
10	TDNN	1.96166	1.44716 (0.00989)	-0.51449	0.36240	0.09372	0.06597	0.29096
10	CF	1.96166	0.19061 (0.00024)	-1.77105	3.13668	0.00007	-0.90673	1.05201
20	TDNN	1.96166	1.44355 (0.00977)	-0.51811	0.36378	0.09270	0.04766	0.29497
20	CF	1.96166	0.18936 (2e-04)	-1.77230	3.14110	0.00006	-0.87784	1.00680
30	TDNN	1.96166	1.44243 (0.00955)	-0.51923	0.36070	0.09403	0.00592	0.27087
30	CF	1.96166	0.18877 (2e-04)	-1.77289	3.14316	0.00005	-0.89306	1.03192
40	TDNN	1.96166	1.4347 (0.00989)	-0.52696	0.37550	0.09592	-0.01069	0.27223
40	CF	1.96166	0.18679 (0.00017)	-1.77487	3.15021	0.00005	-0.90926	1.04946
50	TDNN	1.96166	1.41684 (0.00987)	-0.54482	0.39408	0.09824	0.02465	0.28106
50	CF	1.96166	0.1856 (0.00017)	-1.77606	3.15442	0.00005	-0.91459	1.05752
			Simulation setting 3					
20	TDNN	0.40000	0.47782 (0.00781)	0.07782	0.06706	0.06077	-1.08856	1.42839
20	CF	0.40000	0.09126 (0.00052)	-0.30874	0.09559	0.00048	-1.47286	2.33523

Table 2: Comparison of TDNN and CF in the three simulation settings in Section A.2. The true mean corresponds to the data generating process evaluated at the fixed test point. The standard error of the estimated mean is included in the parentheses, while the estimated variance corresponds to the variance estimated from the fixed test point. The columns with "(R)" correspond to the bias and MSE for the random test point.