Inference for Conditional Average Treatment Effects using Distributional Nearest Neighbors

Jakob R. Juergens
University of Wisconsin - Madison

Last edited: February 22, 2025

Abstract

This paper presents a computationally simple method for estimating heterogeneous treatment effects based on a weighted nearest-neighbor-type estimator. As part of this analysis, for a class of generalized U-statistics, I improve on conditions currently available in the literature required for consistent variance estimation. Furthermore, I provide corresponding results for asymptotically valid pointwise inference in a nonparametric regression setup. I extend the ideas of this nonparametric regression estimator to the estimation of conditional average treatment effects using ideas from double / debiased machine learning including Neyman-orthogonal moments and cross-fitting. This effectively results in a weighted-nearest-nearest-based DDML variant a that is new to the literature and resembles recent Kernel-based approaches under a data-adaptive kernel.

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

The author thanks Jack Porter and Harold D. Chiang for their invaluable feedback.

1 Introduction

Nearest-Neighbor type estimators and their derivatives are a popular class of estimators that is frequently used in fields such as Computer Science or Economics. However, the development of inferential theory for these estimators is not yet up to par with their widespread adoption in practice. One such estimator with a particularly close connection to random forests (RF) is the "two-scale distributional nearest neighbor estimator" (TDNN) of Demirkaya et al. (2024). In the aforementioned paper, the authors develop a novel debiasing method that promises great improvements on the finite sample properties of the estimator and show its asymptotic normality. The main contributions of this paper are twofold. First, this paper provides extended consistency results for the variance estimators for the DNN and TDNN estimators. These results show consistency for three Jackknife-based variance estimators in a broad class of asymptotically Gaussian generalized U-statistics that extends considerably beyond the DNN-based regression approaches. Second, I propose a novel estimator for the CATE based on the ideas inherent to the DNN estimator and methods from DDML. Simulations show promising performance of the estimator, and its simple structure when compared to competing estimators motivates further research into its use for pointwise and simultaneous inference. These extensions will be at the heart of future iterations of this paper.

After short sections on notation and literature review, the remainder of this paper is organized as follows. Section 2 introduces the two setups covered in this paper: first, a relatively simple nonparametric regression setup and second, a setup that mimics the problem of estimating conditional average treatment effects (CATE). To give readers a more economic understanding of the ideas presented in this paper, I will also introduce a running example that I will refer to throughout this paper. This running example will be a simplified version of the well-known problem of estimating treatment effects in a job-training program for individuals of differing characteristics. Furthermore, this section introduces and contextualizes most of the assumptions that I refer to at later stages of the paper. Section 3 is used to define the DNN and TDNN estimators in the context of nonparametric regression and introduces a novel estimator for the CATE setup. In addition, the main results of the distributional approximations of the estimators are introduced. Section 4 further embeds the DNN estimators and their CATE derivatives in the context of U-statistics and introduces essential notation for the analysis of the estimators. Section 5 introduces consistency results for variance estimators to allow for pointwise inference using the DNN estimator and its derivatives. While purely asymptotic in nature, these results improve on currently available results for generalized U-statistics and apply to a broader context than the one presented in this paper. The future iterations of this paper will then tackle the problem of simultaneous inference using techniques developed by Ritzwoller and Syrgkanis (2024). These novel developments have the potential to significantly extend the applicability of the estimator to scenarios where the treatment effect for a large number of subgroups is of importance. Section 6 contains multiple simulation experiments that show the performance of the methods presented in this paper in a setting that mimics an economic analysis. Lastly, Section 8 concludes.

1.1 Notation

Let $[n] = \{1, \dots, n\}$. Given a finite index set $\mathcal{I} \subset \mathbb{N}$, I introduce the following notational conventions.

$$L_s(\mathcal{I}) = \{(l_1, \dots, l_s) \in \mathcal{I}^s \,|\, \forall i \neq j : \; l_i \neq l_j\} \quad \text{and} \quad L_{n,s} = L_s\left([n]\right) \tag{1.1}$$

For a data set $\mathbf{D}_{[n]} = (Z_1, \dots, Z_n)$ and a vector $\ell \in L_{n,s}$, denote by $\mathbf{D}_{[n],-\ell}$ the data set where the observations corresponding to indices in ℓ have been removed. To simplify the notation in the case that a single observation (say the $i'th$ observation) is removed, I use the notation $\mathbf{D}_{n,-i}$. Similarly, given such a data set $\mathbf{D}_{[n]}$ and index vector ℓ,

denote by $\mathbf{D}_\ell$ the data set consisting only of the observations in $\mathbf{D}_{[n]}$ corresponding to the indices in ℓ. In an abuse of notation, when considering two index vectors ℓ and ι that do not share any entries, I denote by $\ell \cup \iota$ the concatenation of the two vectors, e.g., if $\ell = (8, 2, 5)$ and $\iota = (1, 6)$, then $\ell \cup \iota = (8, 2, 5, 1, 6)$.

In the following, $\rightsquigarrow$ denotes weak convergence, while $\to_p$ denotes convergence in probability, and $\to_{a.s.}$ denotes almost sure convergence. We will use the symbol $\lesssim$ to denote an inequality that holds for sufficiently large sample sizes n or kernel orders s. As we consider settings where these diverge together, the specific reference parameter will be clear from the context.

1.2 Related Literature

The related literature can be broadly categorized into three main strands: Nearest-Neighbor type estimators in nonparametric regression, variance estimation for (generalized) U-statistic type estimators including Random Forest, and estimation and inference for CATEs using double / debiased machine learning (DDML) methods. A great introduction to the Nearest-Neighbor method is given in Biau and Devroye (2015), illustrating the potential of the method for classification and regression tasks. Of particular interest in the context of this paper are the so-called "Weighted Nearest-Neighbor" methods for nonparametric regression. While this is a well-studied type of estimator in and of itself, I draw particular connections to bagged-nearest-neighbor type estimators. This class of estimators is built on the framework of "potential closest neighbors" as introduced by Lin and Jeon (2006). Relevant papers studying their properties are, among others, Biau, Cérou, and Guyader (2010), Biau and Devroye (2010), and Steele (2009). These papers also point out the close connections to RF and illustrate why studying the bagged nearest-neighbor method could potentially guide our analysis of RF. Recently, Demirkaya et al. (2024) developed a clever debiasing procedure for the bagged or, as they coin it, distributional nearest-neighbor estimator by combining multiple subsampling scales. The resulting TDNN estimator lies at the heart of this paper, and the results presented here should be seen in the context of the already established distributional approximations established in the paper.

U-statistics were introduced by Wassily Hoeffding in Hoeffding (1948) and have been a well-established tool in mathematical statistics for a long time. Thus, there is a significant body of literature that studies their properties, including outstanding introductions such as Lee (2019). Concerning variance estimation for U-statistics, two highly related papers are Arvesen (1969), exploring the theory of the Jackknife when applied to U-statistics, and Arcones and Gine (1992) which fulfills a similar role for the bootstrap. Building on the concept of U-statistics, Peng, Coleman, and Mentch (2022) introduced the notion of generalized U-statistics, unifying randomized, incomplete, and infinite-order U-statistics that have been previously established in the literature. While being a relatively novel development, there is a significant body of literature concerning infinite-order U-statistics, which share their structure with the TDNN estimator. As the purpose of variance estimation in the problem at hand is ultimately to employ distributional approximations, papers such as Chen and Kato (2019) and Song, Chen, and Kato (2019) are similarly of high relevance for potential applications. Due to the close connection to the random forest method introduced by Breiman (2001), there is also a relevant overlap with the literature on that topic. Thus, articles such as Wager, Hastie, and Efron (2014) and Wager and Athey (2018) are of special interest, especially since causal forests are considered the state-of-the-art technique for estimating CATEs.

In the context of estimation and inference regarding CATEs using DDML, Victor Chernozhukov, Chetverikov, et al. (2018) should be pointed out first. By combining cross-fitting with the use of Neyman-orthogonal moments, the authors built the foundation for many modern methods for estimation in the presence of high-dimensional nuisance parameters.

Several extensions to this highly influential idea have been proposed, some of which explicitly aim at estimating CATEs. An example is Semenova and Victor Chernozhukov (2021), who develops estimation and inference procedures for the best linear predictor of a class of causal functions that contain the CATE. Following a different approach, Victor Chernozhukov, Whitney K. Newey, and Syrgkanis (2024) introduce the concept of conditional influence functions and develop a Kernel-based method that similarly to the paper at hand aims at the nonparametric estimation of causal parameters such as the CATE. In a similar vein, Chernozhukov, W K Newey, and Singh (2022) is highly relevant, as it provides a very general analysis of DDML as a meta-algorithm, covering the estimation and inference for CATE.

2 Setup

Throughout this paper, I will consider two distinct setups focusing on separate but intertwined problems. As a running example to give immediate economic meaning to the statistical problems, I consider the problem of evaluating a job training program given only observational data in the style of LaLonde (1986).

Example 1 (Average Hourly Earnings and IT Training)**.**
Imagine obtaining a large data set containing hourly earnings Y of workers with age X_1 and X_2 years of education. A typical quantity of interest is the average hourly earnings given a specific combination of worker characteristics, for example the average earnings for a 20-year-old worker with 12 years of education (a recent high school graduate without further education).

To simplify the presentation, the running example ignores many widely discussed economic problems in this environment and instead focuses on a highly stylized problem with a limited set of characteristics. The first statistical setup is a pure nonparametric regression setup that closely mirrors the structure of Demirkaya et al. (2024) and will be useful to illustrate the inner workings of the estimator of interest.

Assumption 1 (Nonparametric Regression DGP)**.**
The observed data consists of an i.i.d. sample taking the following form.

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

where $Y \in \mathcal{Y} \subset \mathbb{R}$ is the response, $X \in \mathcal{X} \subset \mathbb{R}^k$ is a feature vector of fixed dimension k distributed according to a density function f with associated probability measure φ on $\mathcal{X}$, and $\mu(x)$ is the unknown mean regression function. ε is the unobservable model error on which I impose the following conditions.

$$\mathbb{E}\left[\varepsilon \,|\, X\right] = 0, \quad Var\left(\varepsilon \,|\, X = x\right) = \sigma_\varepsilon^2\left(x\right) \tag{2.2}$$

Let the distribution induced by this model be denoted by P and thus $Z_i = (X_i, Y_i) \overset{iid}{\sim} P$.

We will also consider a second statistical setting with more immediate econometric relevance: estimation of and inference on heterogeneous treatment effects in the potential outcomes framework. This statistical setup serves as a more immediately applicable version of the theoretical setup presented in Ritzwoller and Syrgkanis (2024) and brings their results closer to practitioners in the field of economics.

Example 1 (Average Hourly Earnings and IT Training - continued)**.**
Consider the introduction of a job training program that covers basic IT skills that potentially influence the average earnings of a worker with given characteristics. Think of providing said IT training to a 20-year-old high-school graduate without further education or a 56-year-old with a PhD. It seems unreasonable to expect the effect of basic IT training on hourly wages to be similar in these cases. Furthermore, it seems unlikely for these participants to choose to participate in the training with equal probability, adding an additional complicating factor.

Assumption 2 (Heterogeneous Treatment Effect DGP)**.**
The observed data consists of an i.i.d. sample taking the following form.

$$\begin{aligned} &\mathbf{D}_n = \{Z_i = (X_i, W_i, Y_i)\}_{i=1}^n \quad \textit{from the model} \quad Y = \mathbb{1}(W=0)\mu_0^0(X) + \mathbb{1}(W=1)\mu_0^1(X) + \varepsilon, \\ &W_i \sim \text{Bern}\left(\pi_0\left(X_i\right)\right) \end{aligned} \tag{2.3}$$

where $Y \in \mathcal{Y} \subset \mathbb{R}$ is the response and $W \in \{0,1\}$ is an observed treatment indicator. $X \in \mathcal{X} \subset \mathbb{R}^k$ is a vector of covariates of fixed dimension k distributed according to a density function f with an associated probability measure φ on $\mathcal{X}$ and ε is the unobservable model error on which I impose the following conditions.

$$\varepsilon \perp\!\!\!\perp W \,|\, X, \quad \mathbb{E}\left[\varepsilon \,|\, X\right] = 0, \quad Var\left(\varepsilon \,|\, X = x\right) = \sigma_\varepsilon^2\left(x\right) \tag{2.4}$$

Furthermore, $\mu_0^0 : \mathcal{X} \to \mathbb{R}$ and $\mu_0^1 : \mathcal{X} \to \mathbb{R}$ are the two unknown potential outcome functions and $\pi_0 : \mathcal{X} \to [0,1]$ is a function describing the probability of treatment uptake, effectively corresponding to the propensity score.

- *We denote the vector of functional nuisance parameters by $\eta_0 = \left(\mu_0^0, \mu_0^1, \pi_0\right)'$.*
- *Let the distribution induced by this model be denoted by Q and thus $Z_i = (X_i, W_i, Y_i) \overset{iid}{\sim} Q$ supported on $\mathcal{Z} \subset \mathcal{X} \times \{0,1\} \times \mathcal{Y}$.*

In this second setting, I will use the notation $\mathbf{D}^{(0)}$ and $\mathbf{D}^{(1)}$ to refer to the data subsets that contain only observations with $W = 0$ and $W = 1$, respectively. Clearly, this model can be interpreted in the context of the potential outcomes framework in the usual way.

Throughout this paper, I will additionally rely on a number of assumptions that are more technical in nature.

Assumption 3 (Technical Assumptions)**.**
In both settings (Setup 1 and Setup 2) the following conditions hold as applicable to the settings, respectively.

- *The feature space* $\mathcal{X} = \text{supp}(X)$ *is a bounded, compact subset of* $\mathbb{R}^k$
- *The density* $f(\cdot)$ *is bounded away from 0 and* ∞*.*

$$\forall x \in \mathcal{X}: \quad 0 < \underline{\mathfrak{f}} \leq f(x) \leq \bar{\mathfrak{f}} < \infty \tag{2.5}$$

- $f(\cdot)$, $\mu(\cdot)$, $\mu_0^0(\cdot)$, $\mu_0^1(\cdot)$, *and* $\pi_0(\cdot)$ *are four times continuously differentiable with bounded second, third, and fourth-order partial derivatives. Specifically, in mathematical terms:*

$$\forall w \in \{0,1\} \quad \forall x \in \mathcal{X} \quad \forall (i,j,k,l) \in [d]^4: \tag{2.6}$$

$$\begin{aligned}
-\infty \;<\; \underline{\mathfrak{f}}' \;&\leq \partial_{i,j} f(x), \quad \partial_{i,j,k} f(x), \quad \partial_{i,j,k,l} f(x) \quad \leq \; \bar{\mathfrak{f}}' \;< \infty \\
-\infty \;<\; \underline{\mathfrak{m}}' \;&\leq \partial_{i,j} \mu(x), \quad \partial_{i,j,k} \mu(x), \quad \partial_{i,j,k,l} \mu(x) \quad \leq \bar{\mathfrak{m}}' \;< \infty \\
-\infty \;<\; \underline{\mathfrak{m}}' \;&\leq \partial_{i,j} \mu_0^w(x), \partial_{i,j,k} \mu_0^w(x), \partial_{i,j,k,l} \mu_0^w(x) \leq \bar{\mathfrak{m}}' \;< \infty \\
-\infty \;<\; \underline{\mathfrak{p}}' \;&\leq \partial_{i,j} \pi_0(x), \; \partial_{i,j,k} \pi_0(x), \; \partial_{i,j,k,l} \pi_0(x) \; \leq \; \bar{\mathfrak{p}}' \;< \infty
\end{aligned}$$

- $\mu(\cdot), \mu_0^0(\cdot), \mu_0^1(\cdot) \in L^2(\mathcal{X})$ *are square-integrable functions on* $\mathcal{X}$*.*

There is considerable potential to relax these assumptions at the cost of requiring both less interpretable conditions and more technically sophisticated proofs. For example, the bounded derivatives condition can be relaxed to hold only in a neighborhood of x while requiring a weaker, more complex condition on the behavior of the derivatives beyond that neighborhood. Furthermore, it is necessary to point out again that this setup considers a highly stylized example, as applying these conditions to the IT training program quickly runs into problems.

Example 1 (Average Hourly Earnings and IT Training - continued)**.**
Consider two types of workers: those with 12 years of education, i.e., a high school diploma and those who drop out slightly before finishing high school and therefore have slightly less than 12 years of education. Given the continuous differentiability assumption on the regression function, I do not allow for a discontinuity in average earnings between high school graduates and dropouts. This seems unreasonable as employers would ceteris paribus prefer the worker with a high school diploma in most circumstances. Similarly, I would expect discontinuities in the distribution of educational attainment, for example right after the standard time necessary to achieve common educational milestones. For the sake of a simplified exposition, I will ignore problems such as this going forward.

Additionally, I require a rather standard assumption in localized regression approaches, namely that the variance changes continuously.

Assumption 4 (Error Distribution Assumptions)**.**
The error terms ε defined in Setup 1 and Setup 2, respectively, fulfill the following conditions.

- *ε has continuously varying variance. In other words, $\sigma_\varepsilon^2 : \mathcal{X} \to \mathbb{R}_{>0}$ is a continuous function.*
- *$\sigma_\varepsilon^2 \in L^2(\mathcal{X})$ is a square-integrable function on $\mathcal{X}$*

As $\mathcal{X}$ is a compact and bounded set, this implies that there exists a $\overline{\sigma}_\varepsilon^2 > 0$ such that for any $x \in \mathcal{X}$ we have $\sigma_\varepsilon^2(x) \leq \overline{\sigma}_\varepsilon^2$. Readers of Demirkaya et al. (2024) will recognize that this setup, in contrast to the original paper, allows for heteroskedasticity of the error terms. This comes at basically no cost as the original proofs can be used nearly unchanged to prove the corresponding theorems on distributional approximations. Additionally, due to the assumptions on the regression functions, this ensures the existence of seconds moments of Y in both scenarios. Furthermore, to ensure that there are a sufficient number of treated and untreated observations local to each point of interest asymptotically, I require the following condition on the treatment assignment and uptake mechanism.

Assumption 5 (Non-Trivial Treatment Overlap)**.**
In the Heterogeneous Treatment Effect Setup (Assumption 2), I assume that there exists a constant $\mathfrak{p} \in (0, 1/2)$ such that

$$\forall x \in \mathcal{X}: \quad 0 < \mathfrak{p} \leq \pi_0(x) \leq 1 - \mathfrak{p} < 1. \tag{2.7}$$

This assumption seems rather strong when considering the full universe of potential treatment recipients. In reality, we can constrain this assumption of overlap to neighborhoods of points of interest x. As long as there is sufficient overlap in those neighborhoods the ideas of our identification strategy continue to hold locally.

Example 1 (Average Hourly Earnings and IT Training - continued)**.**
In the IT training example, this condition requires that for each combination of characteristics, there are workers who choose to participate in the IT training and those who do not. More precisely, I assume that as I consider larger and larger data sets, there are both types of workers with characteristics that are very similar to those we are currently interested in. Intuitively this ensures that we have a suitable basis of workers to achieve meaningful comparisons.

Assumption 6 (Stable Unit Treatment Value Assumption (SUTVA))**.**
For any n, let $\mathfrak{W}_n : \mathcal{X}^n \to \{0,1\}^n$ and $\mathfrak{W}'_n : \mathcal{X}^n \to \{0,1\}^n$ be two functions characterizing treatment assignment among a group of n potential observations. Fixing a collection of potential observations corresponding to a collection of feature vectors $\mathbf{X} \in \mathcal{X}^n$ for the potential observations and $i \in [n]$, we impose that given $[\mathfrak{W}_n(\mathbf{X})]_i = [\mathfrak{W}'_n(\mathbf{X})]_i$, the following holds.

$$\begin{aligned} Y_i =& \mathbb{1}\left([\mathfrak{W}_n(\mathbf{X})]_i = 0\right)\mu_0^0(\mathbf{X}_i) + \mathbb{1}\left([\mathfrak{W}_n(\mathbf{X})]_i = 1\right)\mu_0^1(\mathbf{X}_i) + \epsilon_i \\ =& \mathbb{1}\left([\mathfrak{W}'_n(\mathbf{X})]_i = 0\right)\mu_0^0(\mathbf{X}_i) + \mathbb{1}\left([\mathfrak{W}'_n(\mathbf{X})]_i = 1\right)\mu_0^1(\mathbf{X}_i) + \epsilon_i = Y'_i \end{aligned} \tag{2.8}$$

Technically, since we are assuming i.i.d. observations in the characterization of the CATE setup, this is already implied. However, due to the importance of the SUTVA assumption in the treatment estimation literature, it seems appropriate to explicitly point out that it is implicitly assumed that the assumption holds.

3 Distributional Nearest Neighbor Estimators

While less economically attractive, I will introduce the TDNN estimator using the simple nonparametric regression setup first. We will do this by first considering the simpler (one-scale) distributional nearest-neighbor estimator, which naturally extends to its two-scale variant as shown in Demirkaya et al. (2024). Then, having established the method, I will begin by adapting it to tackle the problem of estimating conditional average treatment effects.

3.1 DNN and TDNN in Nonparametric Regression

We can rephrase the nonparametric regression problem in terms of estimating specific conditional moments. In the case at hand, this means that our problem can be phrased in the following way.

$$M(x;\mu) = \mathbb{E}\left[m(Z_i;\mu) \,|\, X_i = x\right] = 0 \quad \text{where} \quad m(Z_i;\mu) = Y_i - \mu(X_i). \tag{3.1}$$

Due to the absence of nuisance parameters, conditions such as local Neyman-orthogonality vacuously hold. We point this out to highlight a contrast that we will encounter when studying the treatment effect setting. In the simpler non-parametric regression setting, we can approach the problem by solving the corresponding empirical conditional moment equation.

$$M_n(x;\mu,\mathbf{D}_n) = \sum_{i=1}^{n} K(x,X_i)m(Z_i;\mu) = 0 \tag{3.2}$$

In this equation, $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a data-dependent Kernel function measuring the "distance" between the point of interest and an observation. Notationally, this makes the local and data-dependent approach of this procedure explicit. One estimator that fulfills the purpose of estimating μ nonparametrically is the Distributional Nearest Neighbor (DNN) estimator. With a name coined by Demirkaya et al. (2024), the DNN estimator is based on important work by Steele (2009) and Biau, Cérou, and Guyader (2010). Given a sample as described in Assumption 1 and a fixed feature vector x, I first order the sample based on the distance to the point of interest.

$$||X_{(1)} - x||_2 \leq ||X_{(2)} - x||_2 \leq \ldots \leq ||X_{(n)} - x||_2 \tag{3.3}$$

Here draws are broken according to the natural indices of the observations in a deterministic way to simplify the derivations going forward. While the distance induced by the euclidean norm is a useful tool for developing an intuition for the method, the idea is not inherently connected to it. In fact, any distance induced by a norm that captures the geometry of the feature space in a suitable way can be used to construct an analogous weighting scheme. The generated ordering implies an associated ordering on the response variables and I denote by $Y_{(i)}$ the response corresponding to $X_{(i)}$. Let $\text{rk}(x;X_i,D)$ denote the *rank* that is assigned to observation i in a sample D relative to a point of interest x, setting $\text{rk}(x;X_i,D) = \infty$ if $Z_i \notin D$. Similarly, let $Y_{(1)}(x;D)$ indicate the response value of the closest neighbor in set D. This enables us to define a data-driven kernel function κ following the notation of Ritzwoller and Syrgkanis (2024).

$$\kappa(x;Z_i,D,\xi) = \mathbb{1}\left(\text{rk}(x;X_i,D) = 1\right) \tag{3.4}$$

Here, ξ is an additional source of randomness in the construction of the base learner that comes into play when analyzing, for example, random forests as proposed by Breiman (2001) using the CART-algorithm described in Breiman et al. (2017). As the DNN estimator does not incorporate such additional randomness, the term is omitted in further considerations. In future research, additional randomness such as, for example, column subsampling could be

considered, in turn making the addition of ξ necessary again. Using κ, it is straightforward to find an expression for the distance function K in Equation 3.2 corresponding to the DNN estimator.

$$K(x, X_i) = \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} \mathbb{1}(i \in \ell) \frac{\kappa(x; Z_i, D_\ell)}{s!} = \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} \frac{\mathbb{1}\left(\mathrm{rk}(x; Z_i, D_\ell) = 1\right)}{s!} \tag{3.5}$$

Inserting into Equation 3.2, this gives us the following empirical conditional moment equation.

$$M_n(x; \mu, \mathbf{D}_n) = \sum_{i=1}^{n} \left(\binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} \frac{\mathbb{1}\left(\mathrm{rk}(x; Z_i, D_\ell) = 1\right)}{s!} \right) (Y_i - \mu(X_i)) = 0 \tag{3.6}$$

Solving this empirical conditional moment equation then yields the DNN estimator $\tilde{\mu}_s(x)$ with subsampling scale s. Defining the kernel function, $h_s(x; D_\ell) := (s!)^{-1} Y_{(1)}(x; D_\ell)$, it is given by the following U-statistic.

$$\tilde{\mu}_s(x; \mathbf{D}_n) = \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} h_s(x; D_\ell) \tag{3.7}$$

Steele (2009) shows that the DNN estimator has a simple closed form representation based on the original ordered sample.

$$\tilde{\mu}_s(x; \mathbf{D}_n) = \binom{n}{s}^{-1} \sum_{i=1}^{n-s+1} \binom{n-i}{s-1} Y_{(i)} \tag{3.8}$$

This representation will allow me to derive computationally simple representations for the practical use of the procedures presented in this paper. This is in contrast to most U-statistic based methods that inherently rely on evaluating the kernel on individual subsets, incurring a potentially prohibitive computational cost. Furthermore, this representation motivates an asymptotic approximation of the weights assigned to each observation that starkly reduces the potentially computationally intensive computation of large binomial coefficients. For this purpose let $\alpha_s = s/n$ leading to the following approximation of the DNN estimator using asymptotic weights.

$$\tilde{\mu}_s(x; \mathbf{D}_n) \approx \sum_{i=1}^{n-s+1} \alpha_s \left(1 - \alpha_s\right)^{i-1} Y_{(i)} \tag{3.9}$$

It is worthwhile to point out that the role of s in the implicit bias-variance trade-off of the DNN estimator runs counter to the role of k in the usual k-NN regression. Where a larger k is usually associated with a lower variance at the cost of a higher bias, a larger s does the opposite. This is due to the fact that a higher s reduces the number of observations that can occur as the closest observation in any given s-subset. As a special example that illustrates the relationship, consider the DNN estimator choosing $s = n$ recovering the simple 1-NN regression estimator. As part of their paper, Demirkaya et al. (2024) develop an explicit expression for the first-order bias term of the DNN estimator and the following distributional approximation result.

Theorem 3.1 (Demirkaya et al. (2024) - Theorem 2)**.**
Assume that we observe data as described in Assumption 1 and that Assumption 3 is valid. Then, for any fixed $x \in \mathcal{X}$*, we have for some positive sequence* ω_n *of order* $\sqrt{s/n}$

$$\frac{\tilde{\mu}_s(x;\mathbf{D}_n) - \mu(x) - B(s) - R(s)}{\omega_n} \rightsquigarrow \mathcal{N}(0,1) \tag{3.10}$$

as $n, s \to \infty$ *with* $s = o(n)$*. Here,* $B(s)$ *and* $R(s)$ *are defined as the following bias terms.*

$$B(s) = \Gamma(2/k+1)\frac{f(x)\operatorname{tr}(\mu''(x)) + 2\mu'(x)^T f'(x)}{2dV_d^{2/k} f(x)^{1+2/k}} s^{-2/k} \quad \text{and} \quad R(s) = \begin{cases} O\left(s^{-3}\right), & k = 1 \\ O\left(s^{-4/k}\right), & k \geq 2 \end{cases} \tag{3.11}$$

where...

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

Starting from this set-up, Demirkaya et al. (2024) develop a novel bias correction method for the DNN estimator that leads to appealing finite-sample properties of the resulting Two-Scale Distributional Nearest Neighbor (TDNN) estimator. Their method is based on the explicit formula for the first-order bias term of the DNN estimator, which in turn allows them to eliminate it through a clever combination of two DNN estimators. Choosing two subsampling scales $1 \leq s_1 < s_2 \leq n$ and two corresponding weights

$$w_1^*(s_1, s_2) = \frac{1}{1 - (s_1/s_2)^{-2/k}} \quad \text{and} \quad w_2^*(s_1, s_2) = 1 - w_1^*(s_1, s_2) \tag{3.12}$$

they define the corresponding TDNN estimator as follows.

$$\widehat{\mu}_{s_1,s_2}(x;\mathbf{D}_n) = w_1^*(s_1, s_2)\tilde{\mu}_{s_1}(x;\mathbf{D}_n) + w_2^*(s_1, s_2)\tilde{\mu}_{s_2}(x;\mathbf{D}_n) \tag{3.13}$$

This leads to the elimination of the first-order bias term shown in Theorem 3.1 leading to desirable finite-sample properties. Furthermore, the authors show that this construction improves the quality of the normal approximation.

Assumption 7 (Bounded Ratio of Kernel-Orders)**.**
There is a constant $\mathfrak{c} \in (0, 1/2)$ *such that the ratio of kernel orders is bounded in the following way.*

$$\forall n: \quad 0 < \mathfrak{c} \leq s_1/s_2 \leq 1 - \mathfrak{c} < 1. \tag{3.14}$$

We make this assumption to avoid edge cases, where asymptotically the TDNN estimator converges to one of the DNN estimators that make it up. As this edge case is irrelevant in practice and as it would be simpler to employ the corresponding DNN estimator in the first place, this is not a practically substantial restriction.

Theorem 3.2 (Demirkaya et al. (2024) - Theorem 3)**.**
Assume that we observe data as described in Assumption 1 and that Assumption 3 holds. Furthermore, let $s_1, s_2 \to \infty$ *with* $s_1 = o(n)$ *and* $s_2 = o(n)$ *be such that Assumption 7 holds for some* $\mathfrak{c} \in (0, 1/2)$*. Then, for any fixed* $x \in \text{supp}(X) \subset \mathbb{R}^d$*, it holds that for some positive sequence* σ_n *of order* $(s_2/n)^{1/2}$*,*

$$\sigma_n^{-1} \left(\widehat{\mu}_{s_1,s_2} (x; \mathbf{D}_n) - \mu(x) - \Lambda \right) \rightsquigarrow \mathcal{N}(0, 1) \tag{3.15}$$

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

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

3.2 DNN in Heterogeneous Treatment Effect Estimation

Motivated by the nonparametric regression setup, I set out to apply the underlying idea in the context of heterogeneous treatment effects. Similarly to before, I start by specifying a moment corresponding to our object of interest, taking into account the additional factors that come into play. Due to the presence of a high-dimensional nuisance parameter in the form of the function q, it is natural to apply the concepts of DDML. This approach closely follows the leading example of Ritzwoller and Syrgkanis (2024). The main goal at this stage is to construct a highly practical method based on their ideas that leverages the computational simplicity of the distributional nearest-neighbor framework.

While considering the problem of point-estimation of a conditional average treatment effect given a feature vector x, $\theta_0(x) = \mathbb{E}\left[Y_i (W_i = 1) - Y_i (W_i = 0) \,|\, X_i = x \right]$, I will employ a Neyman-orthogonal score function to curtail the influence of the nuisance parameters on our estimation.

$$\begin{aligned} &M(x; \theta_0, \eta) = \mathbb{E}\left[m(Z_i; \theta, \eta) \,|\, X_i = x \right] = 0 \quad \text{where} \\ &m(Z_i; \theta, \eta) = \mu^1(X_i) - \mu^0(X_i) + \beta(W_i, X_i) \left(Y_i - \mu^{W_i}(X_i) \right) - \theta(X_i) \end{aligned} \tag{3.16}$$

Here, I make use of the following notation, that is common in the potential outcomes framework, and the well-known Horvitz-Thompson weight.

$$\text{for } w = 1, 2: \quad \mu_w(x) = \mathbb{E}\left[Y_i \,|\, W_i = w,\ X_i = x \right] \quad \text{and} \quad \beta(w, x) = \frac{w}{\pi(x)} - \frac{1 - w}{1 - \pi(x)} \tag{3.17}$$

As a shorthand notation, I will furthermore use $m(Z_i; \eta) = m(Z_i; \theta_0, \eta) + \theta_0(X_i)$. This notation will mainly be used to shorten the presentation of proofs in the appendix. Proceeding in an analogous fashion to the nonparametric regression setup leads us to the following empirical moment equation, where $\widehat{\mu}$ and $\widehat{\pi}$ are first-stage estimators and K is the data-driven kernel function defined in Equation 3.5.

$$M_n(x; \widehat{\eta}) = \sum_{i=1}^{n} K(x, X_i) m(Z_i; \widehat{\eta}) = 0 \tag{3.18}$$

However, due to the presence of infinite-dimensional nuisance parameters, it becomes attractive to proceed using this weighted empirical moment equation embedded in the DML2 estimator of Victor Chernozhukov, Chetverikov, et al. (2018). Applying these ideas to the context of estimating the CATE has been previously explored, for example by

Semenova and Victor Chernozhukov (2021) For the sake of simplicity, I will assume that $m = n/K$, i.e. the desired number of observations in each fold, is an integer going forward.

Definition 1. *DNN-DML2 CATE-Estimator*
To estimate the Conditional Average Treatment Effect at a point of interest $x \in \mathcal{X}$, proceed as follows.

1. *Take a K-fold random partition $\mathcal{I} = (I_k)_{k=1}^K$ of the observation indices $[n]$ such that the size of each fold I_k is $m = n/K$. For each $k \in [K]$, define $I_k^C = [n] \backslash I_k$. Furthermore, for the observation being assigned rank $i \in [n]$, denote by $k_{(i)}$ the fold that the observation appears in. In contrast, for observation j (**NOT** the observation being assigned rank j) denote the corresponding fold by k_j.*

2. *For each $k \in [K]$, use a first-stage estimator for the functional nuisance parameters on the data set $\mathbf{D}_{I_k^C}$...*

   (a) to estimate the nuisance parameters μ_0^0 and μ_0^1:

$$\widehat{\mu}_k^w(x) = \widehat{\mu}\left(x; \mathbf{D}_{I_k^C}^{(w)}\right) \quad \text{for } w = 0, 1 \tag{3.19}$$

   (b) if π_0 is unknown, i.e. we are not in a randomized experiment setting, additionally estimate π_0

$$\widehat{\pi}_k(x) = \widehat{\mu}\left(x; \mathbf{D}_{I_k^C}\right) \quad \text{where the predicted variable is } W \tag{3.20}$$

   For each fold k, denote the vector of estimates by $\widehat{\eta}_k = \left(\widehat{\mu}_k^0, \widehat{\mu}_k^1, \widehat{\pi}_k\right)'$. For each observation i, define the combined estimator as follows $\widehat{\eta}(Z_i) = \widehat{\eta}_{k_i}(Z_i)$.

3. *Construct the estimator $\widehat{\theta}(x)$ as the solution to the following equation.*

$$0 = \sum_{k=1}^{K} \sum_{i \in I_k} K(x, X_i) m\left(Z_i; \widehat{\theta}(x), \widehat{\eta}_k\right) = \sum_{i=1}^{n-s+1} \frac{\binom{n-i}{s-1}}{\binom{n}{s}} m\left(Z_{(i)}; \widehat{\theta}(x), \widehat{\eta}_{k_{(i)}}\right) \tag{3.21}$$

Observe that the weights $K(x, X_i)$ chosen in the second step are chosen according to the whole sample - not according to the chosen folds. Using a lower choice of subsampling scale for this estimation step can help avoid extreme values in the Neyman-orthogonal score function due to estimated propensity scores close to zero or one. This is because a lower subsampling scale averages over a larger number of observations and thus can contribute to better smoothing properties for the propensity score. Using the score function in the equation that defines the estimator, we can observe the following.

$$\begin{aligned} \widehat{\theta}(x) &= \sum_{i=1}^{n-s+1} \frac{\binom{n-i}{s-1}}{\binom{n}{s}} \left[\widehat{\mu}_{k_{(i)}}^1\left(X_{(i)}\right) - \widehat{\mu}_{k_{(i)}}^0\left(X_{(i)}\right) + \widehat{\beta}_{k_{(i)}}\left(W_{(i)}, X_{(i)}\right)\left(Y_{(i)} - \widehat{\mu}_{k_{(i)}}^{W_{(i)}}\left(X_{(i)}\right)\right)\right] \\ &= \sum_{i=1}^{n-s+1} \frac{\binom{n-i}{s-1}}{\binom{n}{s}} m(Z_{(i)}, \widehat{\eta}_{k_{(i)}}) \end{aligned} \tag{3.22}$$

Thus, given first-stage estimates of nuisance parameters, I have a closed-form representation of the CATE estimator for a given partition of $[n]$. Furthermore, given these first-stage estimates, the evaluation of the CATE-estimator at a different point of interest is merely a re-weighting of the terms corresponding to different observations.

As is typical in the literature on doubly-robust inference, we need to make assumptions on the convergence rate of

the first-stage estimators to allow for the desired type of inference. Specifically, I take the following slightly modified assumptions from Assumption 3.2 of Victor Chernozhukov, Chetverikov, et al. (2018).

Assumption 8 (DDML-Rate Conditions)**.**
LOREM IPSUM

1. *Given a random subset I of $[n]$ of size $m = n/k$, the nuisance parameter estimator $\widehat{\eta}(\mathbf{D}_{I^C})$ belongs to the realization set $\mathcal{T}_n$ with probability $1 - \Delta_n$, where $\mathcal{T}_n$ contains η_0 and is constrained by the following conditions. Denote the event that $\widehat{\eta}(\mathbf{D}_{I^C}) \in \mathcal{T}_n$ by $\mathcal{E}_n$.*

2. *The following moment condition holds:*

$$m_n := \sup_{\eta \in \mathcal{T}_n} \left(\mathbb{E}_Z \left[|m(Z; \theta_0, \eta)|^q\right]\right)^{1/q} \leq c_1 \tag{3.23}$$

3. *The following conditions on the statistical rates r_n, r'_n, and λ'_n hold:*

$$r'_n := \sup_{\eta \in \mathcal{T}_n} \left(\mathbb{E}_Z \left[|m(Z; \theta_0, \eta) - m(Z; \theta_0, \eta_0)|^2\right]\right)^{1/2} \leq \delta_n, \tag{3.24}$$

$$\lambda'_n := \sup_{r \in (0,1), \eta \in \mathcal{T}_n} \left|\partial_r^2 \mathbb{E}_Z \left[m(Z; \theta_0, \eta_0 + r(\eta - \eta_0))\right]\right| \leq \delta_n / \sqrt{n}. \tag{3.25}$$

4. *The variance of the score is non-zero.*

$$LOREMIPSUM \tag{3.26}$$

In contrast to the paper that these assumptions are taken from, there is an absence of conditions on what is named ψ^a in the original paper. This is due to the fact that in the case of the CATE, this term is -1 and thus independent of the nuisance parameters. Thus, specific constraints on the effects of the first-stage estimation problem are unnecessary. In a potential generalization of the method proposed in this paper to other doubly-robust inference problems, this aspect should be addressable with relatively minimal additional technical arguments along the lines in Victor Chernozhukov, Chetverikov, et al. (2018).

4 DNN Estimators as Generalized U-Statistics

As most of the theoretical results in Demirkaya et al. (2024) rely on representations as a U-statistic, it is helpful to introduce additional concepts and notation at this stage. Recalling Equation 3.7, the DNN and TDNN estimators can be expressed in the following U-statistic form and are thus a type of generalized complete U-statistic as introduced by Peng, Coleman, and Mentch (2022).

$$\tilde{\mu}_s(x;\mathbf{D}_n) = \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s}} h_s(x;D_\ell) \quad \text{and} \quad \widehat{\mu}_{s_1,s_2}(x;\mathbf{D}_n) = \binom{n}{s}^{-1} \sum_{\ell \in L_{n,s_2}} h_{s_1,s_2}(x;D_\ell) \tag{4.1}$$

It is worth pointing out that in contrast to the DNN estimator, the kernel for the TDNN estimator is of order $s_2 > s_1$. The authors derive an explicit formula for the kernel that shows the connection between the DNN and TDNN estimators. This connection will prove useful going forward.

Lemma 4.1 (Kernel of TDNN Estimator - Adapted from Lemma 8 of Demirkaya et al. (2024))**.** *The kernel of the TDNN estimator takes the following form.*

$$\begin{aligned} h_{s_1,s_2}(x;D) &= w_1^* \left[\binom{s_2}{s_1}^{-1} \sum_{\ell \in L_{s_2,s_1}} h_{s_1}(x;D_\ell) \right] + w_2^* h_{s_2}(x;D) \\ &= w_1^* \tilde{\mu}_{s_1}(x;D) + w_2^* h_{s_2}(x;D) \end{aligned} \tag{4.2}$$

Borrowing the notational conventions from Lee (2019), I additionally introduce the following notation.

$$\psi_s^c(x;\mathbf{z}_1,\ldots,\mathbf{z}_c) = \mathbb{E}_D\left[h_s(x;D) \mid Z_1 = \mathbf{z}_1,\ldots,Z_c = \mathbf{z}_c\right] \tag{4.3}$$

$$h_s^{(1)}(x;\mathbf{z}_1) = \psi_s^1(x;\mathbf{z}_1) - \mu(x) \tag{4.4}$$

$$h_s^{(c)}(x;\mathbf{z}_1,\ldots,\mathbf{z}_c) = \psi_s^c(x;\mathbf{z}_1,\ldots,\mathbf{z}_c) - \sum_{j=1}^{c-1} \left(\sum_{\ell \in L_{n,j}} h_s^{(j)}(x;\mathbf{z}_\ell) \right) - \mu(x) \quad \text{for } c = 2,\ldots,s \tag{4.5}$$

In contrast to the notational inspiration, the subsampling size s is made explicit. Since we are dealing with an infinite-order U-statistic, s will be diverging with n. Completely analogous, define the corresponding objects for the TDNN estimator. For the DNN estimator and any $1 \leq c \leq s$, define

$$\xi_s^c(x) = \mathrm{Var}_{1:c}\left(\psi_s^c(x;Z_1,\ldots,Z_c)\right) \tag{4.6}$$

where $Z'_{c+1},\ldots,Z'_n$ are i.i.d. from P and independent of $Z_1,\ldots,Z_n$ and thus $\xi_s^s(x) = \mathrm{Var}\left(h_s(x;Z_1,\ldots,Z_s)\right)$. Similarly, for the TDNN estimator and any $1 \leq c \leq s_2$, let

$$\zeta_{s_1,s_2}^c(x) = \mathrm{Var}_{1:c}\left(\psi_{s_1,s_2}^c(x;Z_1,\ldots,Z_c)\right) \tag{4.7}$$

with an analogous definition of Z'. As a byproduct (or main purpose depending on the perspective) these terms can be used to derive the Hoeffding decomposition of the TDNN estimator.

$$H_s^c(x;\mathbf{D}_n) = \binom{n}{c}^{-1} \sum_{\ell \in L_{n,c}} h_s^{(c)}(x;D_\ell) \quad \text{and} \quad H_{s_1,s_2}^c(x;\mathbf{D}_n) = \binom{n}{c}^{-1} \sum_{\ell \in L_{n,c}} h_{s_1,s_2}^{(c)}(x;D_\ell) \tag{4.8}$$

These projection terms can then be used to construct the following Hoeffding decompositions.

$$\tilde{\mu}_s\left(x;\mathbf{D}_n\right) = \mu(x) + \sum_{j=1}^{s}\binom{s}{j} H_s^j\left(x;\mathbf{D}_n\right) \quad \text{and} \quad \widehat{\mu}_{s_1,s_2}\left(x;\mathbf{D}_n\right) = \mu(x) + \sum_{j=1}^{s_2}\binom{s_2}{j} H_{s_1,s_2}^j\left(x;\mathbf{D}_n\right) \tag{4.9}$$

Standard results for U-statistics (see, for example, Lee (2019)) now give us a number of useful results. First, an immediate result on the expectations of the Hoeffding-projection kernels.

$$\forall c = 1,2,\ldots,j-1: \quad \mathbb{E}_D\left[h_{s_1,s_2}^{(j)}\left(x;D\right) \mid Z_1 = \mathbf{z}_1,\ldots,Z_c = \mathbf{z}_c\right] = 0 \quad \text{and} \quad \mathbb{E}_D\left[h_{s_1,s_2}^{(j)}\left(x;D\right)\right] = 0 \tag{4.10}$$

Second, I obtain a useful variance decomposition in terms of the Hoeffding-projection variances.

$$\operatorname{Var}_D\left(\widehat{\mu}_{s_1,s_2}\left(x;D\right)\right) = \sum_{j=1}^{s_2}\binom{s_2}{j}^2 \operatorname{Var}_D\left(H_{s_1,s_2}^j\left(x;D\right)\right) \tag{4.11}$$

$$\operatorname{Var}_D\left(H_{s_1,s_2}^j\left(x;D\right)\right) = \binom{n}{j}^{-1} \operatorname{Var}_D\left(h_{s_1,s_2}^{(j)}\left(x;D\right)\right) =: \binom{n}{j}^{-1} V_{s_1,s_2}^j\left(x\right) \tag{4.12}$$

Third, the following equivalent expression for the kernel variance.

$$\zeta_{s_1,s_2}^{s_2}\left(x\right) = \operatorname{Var}_D\left(h_{s_1,s_2}\left(x;D\right)\right) = \sum_{j=1}^{s_2}\binom{s_2}{j} V_{s_1,s_2}^j \tag{4.13}$$

4.1 CATE-Estimators as Generalized U-Statistics

Given estimates of the functional nuisance parameters, the proposed CATE estimators can be analyzed as generalized U-statistics in the same way as in the nonparametric regression context. As most of the theoretical results on these estimators will similarly rely on Hoeffding projection arguments, I will introduce analogous notation in this more general scenario. First, observe that the DNN-DML2 CATE estimator can be rewritten as follows to explicitly show its construction as a generalized U-statistic.

$$\begin{aligned}
\widehat{\theta}\left(x;\mathbf{D}\right) &= \sum_{i=1}^{n-s+1} \frac{\binom{n-i}{s-1}}{\binom{n}{s}} \left[\widehat{\mu}_{k_{(i)}}^1\left(X_{(i)}\right) - \widehat{\mu}_{k_{(i)}}^0\left(X_{(i)}\right) + \widehat{\beta}_{k_{(i)}}\left(W_{(i)},X_{(i)}\right)\left(Y_{(i)} - \widehat{\mu}_{k_{(i)}}^{W_{(i)}}\left(X_{(i)}\right)\right)\right] \\
&= \binom{n}{s}^{-1} \sum_{\ell\in L_{n,s}} \sum_{i=1}^{n} \frac{\mathbb{1}\left(\operatorname{rk}(x;Z_i,D_\ell)=1\right)}{s!} \left[\widehat{\mu}_{k_i}^1\left(X_i\right) - \widehat{\mu}_{k_i}^0\left(X_i\right) + \widehat{\beta}_{k_i}\left(W_i,X_i\right)\left(Y_i - \widehat{\mu}_{k_i}^{W_i}\left(X_i\right)\right)\right] \\
&= \binom{n}{s}^{-1} \sum_{\ell\in L_{n,s}} \underbrace{\sum_{i=1}^{n} \frac{\mathbb{1}\left(\kappa(x;Z_i,D_\ell)=1\right)}{s!} m(Z_i,\widehat{\eta}_{k_i})}_{\chi_s(x;\mathbf{D}_\ell,\widehat{\eta})} \\
&= \binom{n}{s}^{-1} \sum_{\ell\in L_{n,s}} \chi_s(x;\mathbf{D}_\ell,\widehat{\eta})
\end{aligned} \tag{4.14}$$

In the same fashion as in the previous case, I can now define the Hoeffding decomposition for the estimator. Note that in this step, the functional nuisance parameter estimates $\widehat{\mu}$ and $\widehat{\pi}$ are considered as depending on the data that

we form the expectation over and are thus itself random and part of the expectation.

$$\vartheta_s^c\left(x; \mathbf{z}_1, \ldots, \mathbf{z}_c, \widehat{\eta}\right) = \mathbb{E}_D\left[\chi_s(x; \mathbf{D}, \widehat{\eta}) \,|\, Z_1 = \mathbf{z}_1, \ldots, Z_c = \mathbf{z}_c\right] \tag{4.15}$$

$$\chi_s^{(1)}\left(x; \mathbf{z}_1, \widehat{\eta}\right) = \vartheta_s^1\left(x; \mathbf{z}_1, \widehat{\eta}\right) - \mathbb{E}_D\left[\chi_s(x; \mathbf{D}, \widehat{\eta})\right] \tag{4.16}$$

As before, I define the higher-order projection terms, i.e. for $c = 2, \ldots, s$, in the following way.

$$\chi_s^{(c)}\left(x; \mathbf{z}_1, \ldots, \mathbf{z}_c, \widehat{\eta}\right) = \vartheta_s^c(x; \mathbf{z}_1, \ldots, \mathbf{z}_c, \widehat{\eta}) - \sum_{j=1}^{c-1} \left(\sum_{\ell \in L_{n,j}} \chi_s^{(j)}(x; \mathbf{z}_\ell, \widehat{\eta}) \right) - \mathbb{E}_D\left[\chi_s(x; \mathbf{D}, \widehat{\eta})\right] \tag{4.17}$$

In anticipation of arguments involving empirical process theory, I define as follows.

$$\chi_{s,0}(x; \mathbf{D}_{[s]}) = \chi_s(x; \mathbf{D}_{[s]}, \eta_0) \tag{4.18}$$

$$\vartheta_{s,0}^c\left(x; \mathbf{z}_1, \ldots, \mathbf{z}_c\right) = \vartheta_s^c\left(x; \mathbf{z}_1, \ldots, \mathbf{z}_c, \eta_0\right) \tag{4.19}$$

$$\chi_{s,0}^{(1)}\left(x; \mathbf{z}_1\right) = \chi_s^{(1)}\left(x; \mathbf{z}_1, \eta_0\right) = \vartheta_{s,0}^1\left(x; \mathbf{z}_1\right) - \mathbb{E}_D\left[\chi_{s,0}\left(x; \mathbf{D}\right)\right] \tag{4.20}$$

$$\begin{aligned} \chi_{s,0}^{(c)}\left(x; \mathbf{z}_1, \ldots, \mathbf{z}_c\right) &= \chi_s^{(c)}\left(x; \mathbf{z}_1, \ldots, \mathbf{z}_c, \eta_0\right) \\ &= \vartheta_{s,0}^c(x; \mathbf{z}_1, \ldots, \mathbf{z}_c) - \sum_{j=1}^{c-1} \left(\sum_{\ell \in L_{n,j}} \chi_{s,0}^{(j)}(x; \mathbf{z}_\ell) \right) - \mathbb{E}_D\left[\chi_{s,0}^{(c)}\left(x; Z_1, \ldots, Z_c\right)\right] \end{aligned} \tag{4.21}$$

These definitions allow us to use the following decomposition, where now I acknowledge the randomness in the first-stage estimation.

$$\chi_s^{(c)}\left(x; \mathbf{D}_\ell, \widehat{\eta}\right) = \chi_{s,0}^{(c)}\left(x; \mathbf{D}_\ell\right) + \underbrace{\chi_s^{(c)}\left(x; \mathbf{D}_\ell, \widehat{\eta}\right) - \chi_{s,0}^{(c)}\left(x; \mathbf{D}_\ell\right)}_{R_c(x; \mathbf{D}_\ell)} \tag{4.22}$$

Considering this decomposition will be of great usefulness when analyzing the asymptotic properties of the CATE estimators. Let $\widehat{\mu}_k$ and $\widehat{\pi}_k$ denote the nuisance parameter estimates calculated on the complement of fold k. Using the full expression for the CATE estimator, the following is obtained.

$$\begin{aligned} \widehat{\theta}\left(x; \mathbf{D}\right) = \underbrace{\mathbb{E}_D\left[\widehat{\theta}\left(x; \mathbf{D}\right)\right]}_{\text{Centering-Term}} + \underbrace{\frac{s}{n} \sum_{i=1}^{n} \chi_{s,0}^{(1)}\left(x; Z_i\right)}_{\text{Oracle-Hájek-Projection}} + \underbrace{\frac{s}{k} \sum_{l=1}^{k} \frac{1}{m} \sum_{i \in \mathcal{I}_k} \left(\underbrace{\chi_s^{(1)}\left(x; Z_i, \widehat{\eta}_k\right) - \chi_{s,0}^{(1)}\left(x; Z_i\right)}_{R_{1,k}(x; Z_i)} \right)}_{\text{Oracle-Hájek-Projection Error}} \\ + \underbrace{\sum_{j=2}^{s} \binom{s}{j} \binom{n}{j}^{-1} \sum_{\ell \in L_{n,j}} \chi_{s,0}^{(j)}\left(x; \mathbf{D}_\ell\right)}_{\text{Oracle-Hájek-Residual}} + \underbrace{\sum_{j=2}^{s} \binom{s}{j} \binom{n}{j}^{-1} \sum_{\ell \in L_{n,j}} R_j\left(x; \mathbf{D}_\ell\right)}_{\text{Higher-Order Error Terms}} \end{aligned} \tag{4.23}$$

I call these terms Oracle-Hájek projection and Oracle-Hájek residual because of the additional randomness due to the first-stage estimation of the functional nuisance parameters that is reflected in the R_j terms. An integral part of showing the asymptotic properties of the estimator is to show that a number of terms in this representation are asymptotically negligible as long as the first-stage estimator converges fast enough.

5 Pointwise Inference for DNN Estimators

To perform inference in the regression setup, Demirkaya et al. (2024) introduce variance estimators based on Jackknife and Bootstrap. However, as they point out, their consistency results rely on a likely suboptimal rate condition for the subsampling scale. Although Theorem 3.2 allows s_2 to be of the order $o(n)$, the variance estimation results rely on the considerably stronger condition $s_2 = o(n^{1/3})$. Establishing consistent variance estimation under weaker assumptions on the subsampling rates could broaden the scope of the TDNN estimator for inferential purposes considerably. Furthermore, it can contribute to a better balance between variance and bias, as the choice of the kernel orders is crucial when considering the finite sample properties of the estimator. In this paper, I will focus specifically on variance estimators based on the Jackknife and show consistency results under $s = o(n)$. This is motivated by the closed-form representation of the estimators in question leading to computationally simple formulas for the exact Jackknife variance estimators.

5.1 Jackknife Variance Estimators for Nonparametric Regression

Define the following variance we need to estimate to perform pointwise inference at a point of interest x.

$$\omega^2(x) = \mathrm{Var}_D\left(\widehat{\mu}_{s_1,s_2}(x;\mathbf{D}_n)\right) \tag{5.1}$$

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

$$\widehat{\omega}^2_{JK}(x;\mathbf{D}_n) = \frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\mu}_{s_1,s_2}(x;\mathbf{D}_{n,-i}) - \widehat{\mu}_{s_1,s_2}(x;\mathbf{D}_n)\right)^2 \tag{5.2}$$

Theorem 5.1 (Closed Form Expression for the Jackknife-Variance Estimator)**.**
The Jackknife variance estimator for the DNN estimator has the following convenient closed-form representations.

$$LOREMIPSUM \tag{5.3}$$

Similarly, the Jackknife variance estimator for the TDNN estimator admits the following representation.

$$LOREMIPSUM \tag{5.4}$$

As a generalization to the Jackknife, we can also consider the delete-d Jackknife that builds on the same working principle but averages over all possible d-subset removals. This leads to the following representation.

$$\widehat{\omega}^2_{JKD}(x;d,\mathbf{D}_n) = \frac{n-d}{d}\binom{n}{d}^{-1}\sum_{\ell\in L_{n,d}}\left(\widehat{\mu}_{s_1,s_2}(x;\mathbf{D}_{n,-\ell}) - \widehat{\mu}_{s_1,s_2}(x;\mathbf{D}_n)\right)^2 \tag{5.5}$$

Similar to the Jackknife, it is possible to derive a closed form representation for the delete-d Jackknife. The derivation would proceed along the exact same lines as in the Jackknife case. However, because of the unwieldiness of the closed form, I refrain from deriving it.

In this section, I will loosen that restrictive condition to make use of the attractive performance of U-statistics with

large subsampling rates in the context of inference. The PIJK variance estimator applied to the TDNN estimator is as follows.

$$\widehat{\omega}_{PI}^2\left(x;\mathbf{D}_n\right)=\frac{s_2^2}{n^2}\sum_{i=1}^{n}\left[\left(\binom{n-1}{s-1}^{-1}\sum_{\ell\in L_{s_2-1}([n]\setminus\{i\})}h_{s_1,s_2}\left(x;D_{\ell\cup\{i\}}\right)\right)-\widehat{\mu}_{s_1,s_2}\left(x;\mathbf{D}_n\right)\right]^2 \tag{5.6}$$

LOREM IPSUM

Analyzing the kernel of the TDNN estimators, it can be shown that the conditions of Theorem 6 of Peng, Mentch, and Stefanski (2021) apply under the regime $s_2=o(n)$. Thus, I obtain the following result.

Theorem 5.2 (Pseudo-Infinitesimal Jackknife Variance Estimator Consistency)**.**
Let $0<\mathfrak{c}\leq s_1/s_2\leq 1-\mathfrak{c}<1$ *and* $s_2=o(n)$*, then*

$$\frac{\widehat{\omega}_{PI}^2\left(x;\mathbf{D}_n\right)}{\omega^2\left(x;\mathbf{D}_n\right)}\longrightarrow_p 1. \tag{5.7}$$

In an analogous fashion to Theorems 5 and 6 of Demirkaya et al. (2024), I further obtain the following consistency results for the variance estimators presented. As they point out, proving these results goes beyond the techniques presented in Arvesen (1969), instead relying on results for infinite-order U-statistics. Following the ideas from Peng, Mentch, and Stefanski (2021), I then obtain the following results on the Jackknife and Bootstrap variance estimators, respectively. As part of the proof of these results, I obtain general results on the consistency of Jackknife and Bootstrap variance estimators for infinite-order U-statistics beyond the TDNN estimator.

Theorem 5.3 (Jackknife Variance Estimator Consistency)**.**
Let $0<\mathfrak{c}\leq s_1/s_2\leq 1-\mathfrak{c}<1$ *and* $s_2=o(n)$*, then*

$$\frac{\widehat{\omega}_{JK}^2\left(x;\mathbf{D}_n\right)}{\omega^2\left(x;\mathbf{D}_n\right)}\longrightarrow_p 1. \tag{5.8}$$

Theorem 5.4 (delete-d Jackknife Variance Estimator Consistency)**.**
Let $0<\mathfrak{c}\leq s_1/s_2\leq 1-\mathfrak{c}<1$*,* $s_2=o(n)$*, and* $d=o(n)$*, then*

$$\frac{\widehat{\omega}_{JKD}^2\left(x;d,\mathbf{D}_n\right)}{\omega^2\left(x;\mathbf{D}_n\right)}\longrightarrow_p 1. \tag{5.9}$$

5.2 Asymptotic Normality of the (T)DNN-DML2 CATE Estimator

As a first intermediate result in the analysis of the asymptotic properties of the CATE estimator, we can find the following concerning an oracale variant of the proposed estimator defined as the solution to the following problem.

$$0 = \sum_{i=1}^{n} K(x, X_i) m\left(Z_i; \widehat{\theta}_0(x), \eta_0\right) = \sum_{i=1}^{n-s+1} \frac{\binom{n-i}{s-1}}{\binom{n}{s}} m\left(Z_{(i)}; \widehat{\theta}_0(x), \eta_0\right) \tag{5.10}$$

This is a rather basic result that can be seen as a corollary to the nonparametric regression case.

Theorem 5.5 (PRELIMINARY! THIS IS NOT YET FORMALLY SHOWN)**.**
Assume that we observe data as described in Assumption 2 and that Assumption 3 is valid. Then, for any fixed $x \in \mathcal{X}$, we have for some positive sequence ω_n of order $\sqrt{s/n}$

$$\frac{\widehat{\theta}_0(x; \mathbf{D}_n) - \theta_0(x)}{\omega_n} \rightsquigarrow \mathcal{N}(0, 1) \tag{5.11}$$

as $n, s \to \infty$ with $s = o(n)$.

LOREM IPSUM

5.3 Variance Estimation for the (T)DNN-DML2 CATE Estimator

Ideas:

- Ignoring the occurrence of left-out observation in nuisance parameter estimation and do basic Jackknife - does this lead to bias?
- Leave Fold-Out Bootstrap with slowly diverging number of folds ($k \to \infty$, $m = o(n)$) - Effectively a variant of delete-d bootstrap
- Leave out two folds in the estimator's first step. Then use each previously left out fold for Jackknife construction to eliminate contamination from nuisance parameters
- Modify approach presented in Ritzwoller and Syrgkanis (2024) Appendix F.4 - modified half-sample k-fold cross-split bootstrap root

A fitting variance estimator given the context of this paper in the literature can be obtained by modifying a construction presented in Ritzwoller and Syrgkanis (2024). Specifically, the procedure is based on a variation of the approach presented in Appendix F.4 of the aforementioned paper and makes use of a carefully constructed bootstrap-root. Thus, I need to introduce some additional notation, where, for simplicity, I assume that m, i.e. the number of observations in each I_k, is even.

Definition 2 (Crossfitting Half-Sample)**.**
Given a K-fold partition $\mathcal{I} = (I_k)_{k=1}^K$ *of* $[n]$*, a corresponding half-sample of* $\mathcal{I}$ *is a collection of subsets* $\mathcal{H} = (H_k)_{k=1}^K$ *such that for all* $k \in [K]$*, the following holds.*

$$|H_k| = \frac{|I_k|}{2} = m/2 \quad \textit{and} \quad H_k \subset I_k \tag{5.12}$$

The set of all such half-samples of $\mathcal{I}$ *is denoted by* $\mathfrak{H}(\mathcal{I})$*.*

This bootstrap root will take the following structure.

$$R_{n,s}^* \left(x; \mathbf{D}_{[n]}, \mathcal{I}\right) = \overline{\theta}_{\mathcal{H}}(x) - \widehat{\theta}(x) \tag{5.13}$$

Here, $\overline{\theta}_{\mathcal{H}}(x)$ is the solution to the following equation, where $\mathcal{I}$ is a fixed partition of $[n]$ and $\mathcal{H}$ is a fixed half-sample corresponding to $\mathcal{I}$. In analogy to the previously established notation, I let $K(x, X_i \,|\, \mathcal{H})$ denote the kernel as previously established but with respect to the chosen half-sample, and $Z_{(i\,|\,\mathcal{H})}$ denote the i'th closest observation to the point of interest $\mathbf{x}$ that is contained in $\mathcal{H}$. Furthermore, $k(i \,|\, \mathcal{H})$ denotes the fold $k \in [K]$ that the i'th closest observation in $\mathcal{H}$ is contained in.

$$\begin{aligned} 0 &= \sum_{k=1}^{K} \sum_{i \in H_k} K(x, X_i \,|\, \mathcal{H}) m\left(Z_i; \overline{CATE}_{\mathcal{H}}(x), \widehat{\mu}_{k,s}, \widehat{\pi}_{k,s}\right) \\ &= \sum_{i=1}^{n/2-s+1} \left[\frac{\binom{n/2-i}{s-1}}{\binom{n/2}{s}} m\left(Z_{(i\,|\,\mathcal{H})}; \overline{\theta}_{\mathcal{H}}(x), \widehat{\mu}_{k(i\,|\,\mathcal{H}),s}, \widehat{\pi}_{k(i\,|\,\mathcal{H}),s}\right) \right] \end{aligned} \tag{5.14}$$

Plugging in for the moment under consideration once more, I find the following.

$$\begin{aligned} \overline{\theta}_{\mathcal{H}}(x) = \sum_{i=1}^{n/2-s+1} \frac{\binom{n/2-i}{s-1}}{\binom{n/2}{s}} \Big[&\widehat{\mu}^1_{k(i\,|\,\mathcal{H}),s}\left(X_{(i\,|\,\mathcal{H})}\right) - \widehat{\mu}^0_{k(i\,|\,\mathcal{H}),s}\left(X_{(i\,|\,\mathcal{H})}\right) \\ &+ \widehat{\beta}_{k(i\,|\,\mathcal{H}),s}\left(W_{(i\,|\,\mathcal{H})}, X_{(i\,|\,\mathcal{H})}\right)\left(Y_{(i\,|\,\mathcal{H})} - \mu_{W_{(i\,|\,\mathcal{H})}}\left(X_{(i\,|\,\mathcal{H})}\right)\right) \Big] \end{aligned} \tag{5.15}$$

Recognizing the similarity to $\widehat{\theta}(x)$, I can further simplify in the following way.

$$R_{n,s}^* \left(x; \mathbf{D}_{[n]}, \mathcal{I}\right) = \textit{LOREMIPSUM} \tag{5.16}$$

Theorem 5.6 (Consistent Variance Estimation for the (T)DNN-DML2 CATE Estimator)**.**
LOREM IPSUM

5.4 Pointwise Inference with the TDNN Estimator

Theorem 5.7 (Pointwise Inference in Nonparametric Regression)**.**
LOREM IPSUM

Theorem 5.8 (Pointwise Inference in Heterogeneous Treatment Effect Estimation)**.**
LOREM IPSUM

6 Simulations

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

6.1 Nonparametric Regression

To investigate the practicality of the nonparametric regression estimators presented in this paper, I consider a collection of setups. First, I focus on illustrating the bias correction properties of the TDNN estimator by replicating some of the findings of Demirkaya et al. (2024). One such promising example is shown in Figure 1 highlighting the potential improvements available by combining multiple subsampling scales.

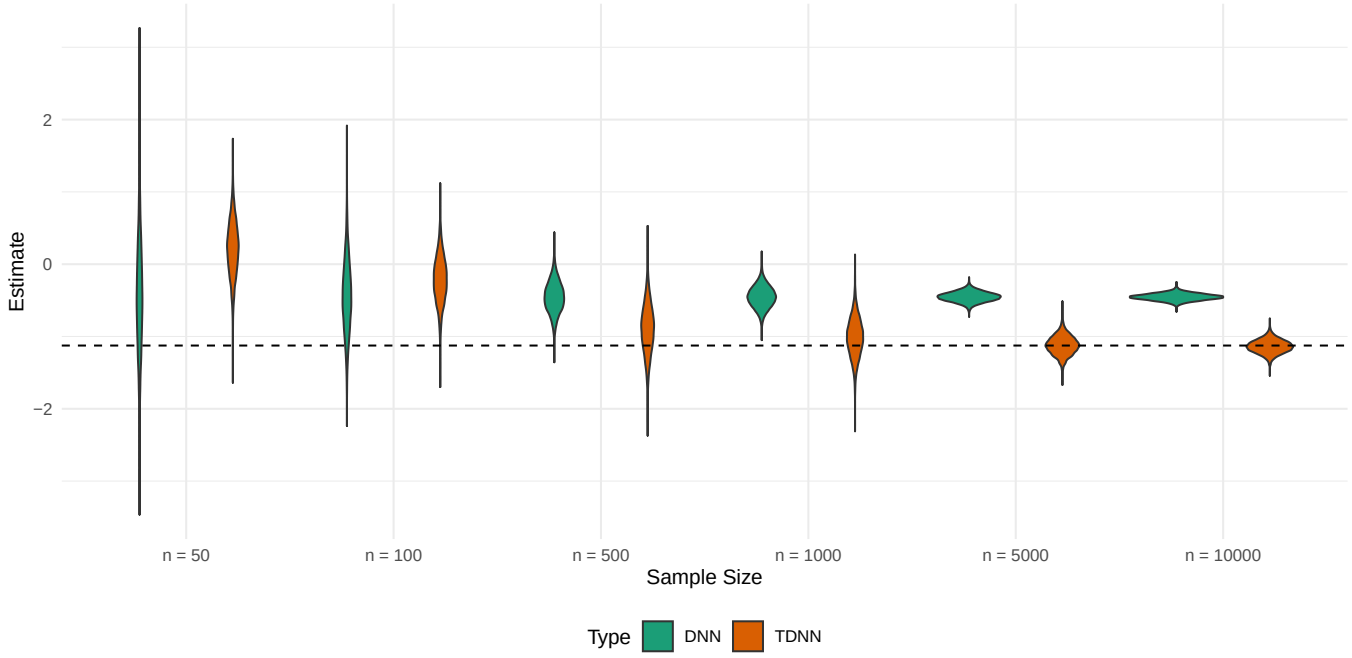

Figure 1: Comparison of the DNN ($s = 20$) and TDNN ($s_1 = 20, s_2 = 50$) Estimators for different sample sizes. The dashed line indicates the value of the unknown regression function at the point of interest. Simulation Setup replicates Setting 1 from Demirkaya et al. (2024) for 10000 Monte Carlo Replications.

As can be seen in Figure 1, a suitable choice of subsampling scales can effectively reduce the bias of the TDNN estimator compared to the DNN estimator. This reinforces the idea that the TDNN estimator can be a useful tool in practice that has the potential to improve on well-established nearest neighbor methods.

As a second, potentially more illustrative example, I consider the estimation of a function of two arguments. Specifically, I consider the function $\mu(x) = 5 \cdot (\cos(x_1) + \cos(x_2))$ on $[0,1]^2$ with heteroskedastic error terms whose variance is determined by $\sigma_\varepsilon^2(x) = \frac{1}{16}\left(x_1^2 + x_2^2\right)^2$. The resulting surface is shown in Figure 2.

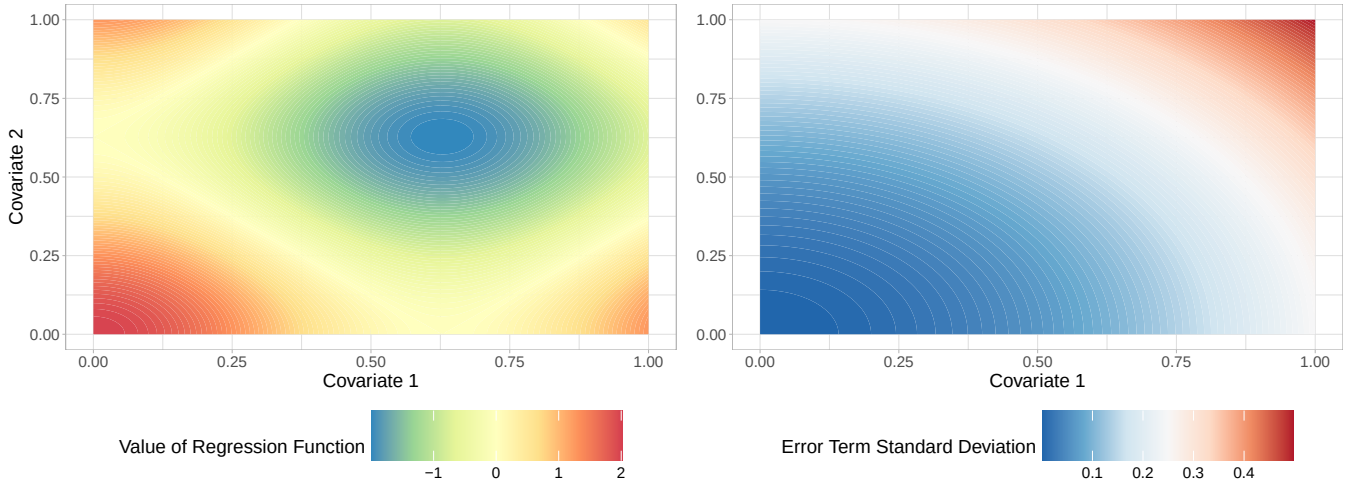

Figure 2: Value of the Regression Function (left) and Variance of the Error Term (right)

To analyze the behavior of the estimator in this setting, I run a number of Monte Carlo simulations each consisting of 10000 simulation runs. Although the theoretical analysis was of purely asymptotic nature, these simulation results can provide a modicum of guidance when it comes to choices such as the kernel orders employed in the estimation procedure. Each run consists of 10000 observations that are uniformly distributed on $[0,1]^2$, and I find the following concerning the bias and variance of the estimators given different kernel orders s_1 and s_2.

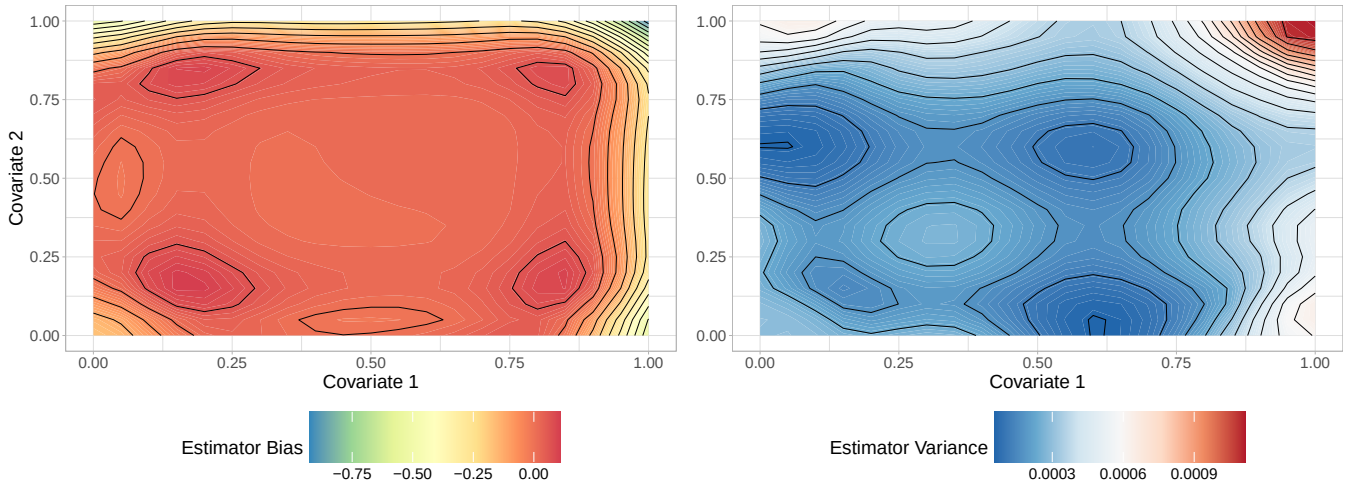

Figure 3: Approximate Bias (left) and Variance (right) of the TDNN Estimator with $s_1 = 10$ and $s_2 = 25$

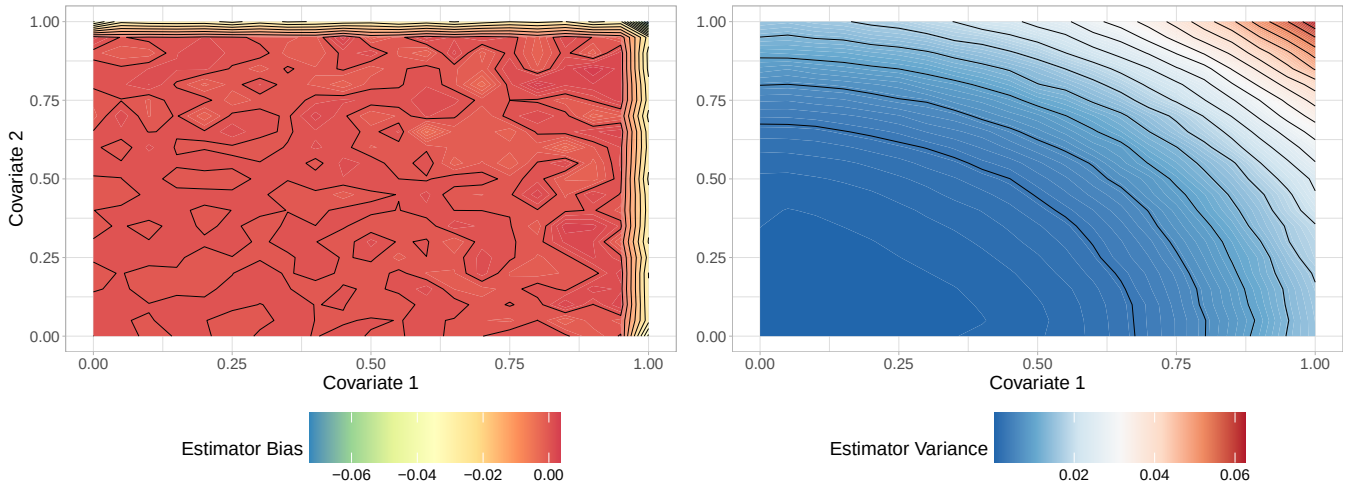

Figure 4: Approximate Bias (left) and Variance (right) of the TDNN Estimator with $s_1 = 1000$ and $s_2 = 2500$

6.2 CATE-Estimation

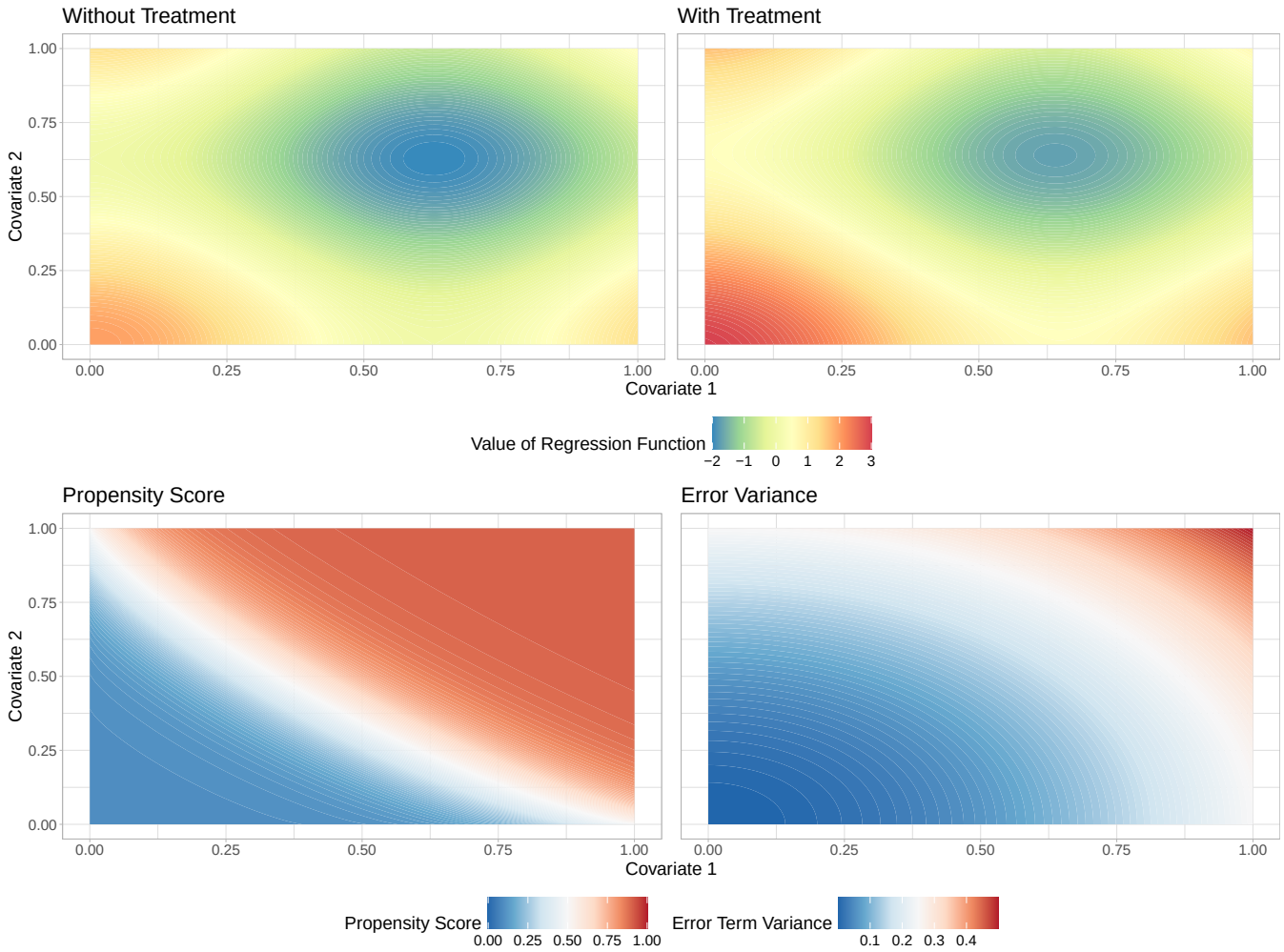

Figure 5: Value of the Regression Functions μ_0 (upper) and μ_1 (lower). Error term structure remains unchanged.

LOREM IPSUM

7 Application

LOREM IPSUM

8 Conclusion and Outlook

LOREM IPSUM

8.1 Outlook

The promising results of this paper lay out a few avenues for continued research. A first idea to potentially improve on the performance of the implemented weighting scheme is to use weights that correspond to a column-subsampled version of the (T)DNN weighting scheme. In the context of RF this is a commonly employed technique that helps to de-correlate the individual trees that are used to construct the RF. This decorrelation has been shown to be an integral part of the desirable performance that RF delivers in practice via simulation studies. In the context of this paper, a similar effect can be expected to take place while having the potential to still be expressible in a relatively simple, principled closed-form weighting scheme.

LOREM IPSUM

References

Arcones, Miguel A. and Evarist Gine (June 1992). "On the Bootstrap of U and V Statistics". In: *The Annals of Statistics* 20.2. DOI: 10.1214/aos/1176348650.

Arvesen, James N. (Dec. 1969). "Jackknifing U-Statistics". en. In: *The Annals of Mathematical Statistics* 40.6, pp. 2076–2100. DOI: 10.1214/aoms/1177697287.

Biau, Gérard, Frédéric Cérou, and Arnaud Guyader (2010). "On the Rate of Convergence of the Bagged Nearest Neighbor Estimate". en. In: *Journal of Machine Learning Research* 11.22, pp. 687–712.

Biau, Gérard and Luc Devroye (Nov. 2010). "On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification". In: *Journal of Multivariate Analysis* 101.10, pp. 2499–2518. DOI: 10.1016/j.jmva.2010.06.019.

– (2015). *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-25388-6.

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

Breiman, Leo et al. (Oct. 2017). *Classification and Regression Trees*. New York: Chapman and Hall/CRC. DOI: 10.1201/9781315139470.

Chen, Xiaohui and Kengo Kato (Dec. 2019). "Randomized incomplete U-statistics in high dimensions". In: *The Annals of Statistics* 47.6, pp. 3127–3156. DOI: 10.1214/18-AOS1773.

Chernozhukov, V, W K Newey, and R Singh (June 2022). "A simple and general debiased machine learning theorem with finite-sample guarantees". In: *Biometrika* 110.1, pp. 257–264. DOI: 10.1093/biomet/asac033.

Chernozhukov, Victor, Denis Chetverikov, et al. (Feb. 2018). "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21.1, pp. C1–C68. DOI: 10.1111/ectj.12097.

Chernozhukov, Victor, Whitney K. Newey, and Vasilis Syrgkanis (Dec. 2024). *Conditional Influence Functions*. DOI: 10.48550/arXiv.2412.18080.

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

Hoeffding, Wassily (Sept. 1948). "A Class of Statistics with Asymptotically Normal Distribution". In: *The Annals of Mathematical Statistics* 19.3, pp. 293–325. DOI: 10.1214/aoms/1177730196.

LaLonde, Robert J. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". In: *The American Economic Review* 76.4, pp. 604–620.

Lee, A J. (Mar. 2019). *U-Statistics*. en. 0th ed. Routledge. DOI: 10.1201/9780203734520.

Lin, Yi and Yongho Jeon (June 2006). "Random Forests and Adaptive Nearest Neighbors". en. In: *Journal of the American Statistical Association* 101.474, pp. 578–590. DOI: 10.1198/016214505000001230.

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

Peng, Wei, Lucas Mentch, and Leonard Stefanski (2021). *Bias, Consistency, and Alternative Perspectives of the Infinitesimal Jackknife*. DOI: 10.48550/ARXIV.2106.05918.

Ritzwoller, David M. and Vasilis Syrgkanis (Sept. 2024). *Simultaneous Inference for Local Structural Parameters with Random Forests*. DOI: 10.48550/arXiv.2405.07860.

Semenova, Vira and Victor Chernozhukov (May 2021). "Debiased machine learning of conditional average treatment effects and other causal functions". In: *The Econometrics Journal* 24.2, pp. 264–289. DOI: 10.1093/ectj/utaa027.

Song, Yanglei, Xiaohui Chen, and Kengo Kato (Jan. 2019). "Approximating high-dimensional infinite-order U-statistics: Statistical and computational guarantees". In: *Electronic Journal of Statistics* 13.2, pp. 4794–4848. DOI: 10.1214/19-EJS1643.

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

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

Wager, Stefan, Trevor Hastie, and Bradley Efron (2014). "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife". In: *Journal of Machine Learning Research* 15.48, pp. 1625–1651.

A Useful Results

A.1 Properties of the κ Function

Lemma A.1 (Demirkaya et al. (2024) - Lemma 12)**.**
Let $D = \{Z_1, \ldots, Z_s\}$ an i.i.d. sample drawn from P. The indicator functions $\kappa(x; Z_i, D)$ satisfy the following properties.

1. *For any $i \neq j$, we have $\kappa(x; Z_i, D)\,\kappa(x; Z_j, D) = 0$ with probability one;*
2. *$\sum_{i=1}^{s} \kappa(x; Z_i, D) = 1$;*
3. *$\forall i \in [s]: \quad \mathbb{E}_{1:s}[\kappa(x; Z_i, D)] = s^{-1}$*
4. *$\mathbb{E}_{2:s}[\kappa(x; Z_1, D)] = \{1 - \varphi(B(x, \|X_1 - x\|))\}^{s-1}$*

Here $\mathbb{E}_{i:s}$ denotes the expectation with respect to $\{Z_i, Z_{i+1}, \ldots, Z_s\}$. Furthermore, φ denotes the probability measure on $\mathbb{R}^d$ induced by the random vector X.

Lemma A.2 (Demirkaya et al. (2024) - Lemma 13)**.**
For any L^1 function f that is continuous at x, it holds that

$$\lim_{s\to\infty} \mathbb{E}_1[f(X_1)\, s\mathbb{E}_{2:s}[\kappa(x; Z_1, D)]] = f(x). \tag{A.1}$$

As an additional tool, we will make use of the following analogous results concerning products of two kernel functions with nonzero expectation. These results will then be used to construct an analogon of Lemma A.2 for the corresponding cases. This will serve very similar purposes in the analysis of (conditional) covariance terms as the previous results serve for (conditional) expectations.

Lemma A.3.
Fix sample size n, subsampling scale s, and c such that $0 < c \leq s \leq n$. Let $D = \{Z_1, Z_2, \ldots, Z_c, Z_{c+1}, \ldots Z_s\}$ be an i.i.d. data set drawn from P as described in Setup 1. Let $D' = \{Z_1, Z_2, \ldots, Z_c, Z'_{c+1}, \ldots Z'_s\}$ be a second data set that shares the first c observations with D. The remaining $s - c$ observations of D', i.e. $\{Z'_{c+1}, \ldots Z'_s\}$, are i.i.d. draws from P that are independent of D.

Then the following three statements hold.

$$\forall i \in [c]: \quad \mathbb{E}_{D,D'}[\kappa(x; Z_i, D)\,\kappa(x; Z_i, D')] = (2s - c)^{-1} = \omega(e^{-s}) \tag{A.2}$$

$$\begin{aligned}\forall i \in [c]\ \forall j \in \{c+1, \ldots, s\}: \quad \mathbb{E}_{D,D'}\left[\kappa(x; Z_i, D)\,\kappa\left(x; Z'_j, D'\right)\right] &= \frac{1}{(2s-c)(2s-c-1)} \sum_{i=0}^{s-c-1} \binom{s-c-1}{i} \binom{2s-c-2}{i}^{-1} \\ &= \omega(e^{-s})\end{aligned} \tag{A.3}$$

$$\begin{aligned}\forall i, j \in \{c+1, \ldots, s\}: \quad \mathbb{E}_{D,D'}\left[\kappa(x; Z_i, D)\,\kappa\left(x; Z'_j, D'\right)\right] &= \frac{2}{(2s-c)(2s-c-1)} \sum_{i=0}^{s-c-1} \binom{s-c-1}{i} \binom{2s-c-1}{s-1+i}^{-1} \\ &= \omega(e^{-s})\end{aligned} \tag{A.4}$$

Proof of Lemma A.3.
Without loss of generality, we will consider the cases of $i = 1$ and $j = c + 1$ for the first two equations.

$$\begin{aligned}\mathbb{E}_{D,D'}\left[\kappa\left(x; Z_1, D\right)\kappa\left(x; Z_1, D'\right)\right] &= \mathbb{E}_{D,D'}\left[\kappa\left(x; Z_1, D_{1:c}\right)\kappa\left(x; Z_1, D_{(c+1):s}\right)\kappa\left(x; Z_1, D'_{(c+1):s}\right)\right] \\ &= \mathbb{E}\left[\kappa\left(x; Z_1, D_{[2s-c]}\right)\right] = (2s-c)^{-1}\end{aligned} \tag{A.5}$$

Considering the second case, we find the following.

$$\begin{aligned}\mathbb{E}_{D,D'}\left[\kappa\left(x; Z_1, D\right)\kappa\left(x; Z'_{c+1}, D'\right)\right] &= \frac{1}{(2s-c)!}\sum_{i=0}^{s-c-1}\binom{s-c-1}{i} i!\left((s-1)+(s-c-1-i)\right)! \\ = \frac{1}{(2s-c)!}\sum_{i=0}^{s-c-1}\binom{s-c-1}{i} i!\,(2s-c-2-i)! &= \frac{(2s-c-2)!}{(2s-c)!}\sum_{i=0}^{s-c-1}\binom{s-c-1}{i}\binom{2s-c-2}{i}^{-1}\end{aligned} \tag{A.6}$$

While unintuitive at first, the terms in this expression have intuitive meaning when we consider this as a combinatoric problem. Consider lining up the observations in order of their distance to the point of interest and counting the cases for which the expression in the expectation is equal to one. First, there are $(2s-c)!$ possible orderings of the observations with probability one, leading to the denominator. Next, notice that only those orderings where $\|X'_{c+1} - x\| \leq \|X_1 - x\|$ and $\|X_1 - x\| \leq \|X_i - x\|$ for any $i = 2, \ldots, c$ can possibly lead to a non-zero realization of the kernel term. Furthermore, out of the $(s-c-1)$ observations in $D'_{(c+2):s}$, it is possible for $i = 0, \ldots, s-c-1$ observations to lie at a distance to the point of interest that is smaller than $\|X_1 - x\|$ but larger than $\|X'_{c+1} - x\|$ in any permutation. The sum adjusts for those possible configurations. Next, we can make the following observation concerning the expression we just derived.

$$\begin{aligned}\mathbb{E}_{D,D'}\left[\kappa\left(x; Z_1, D\right)\kappa\left(x; Z'_{c+1}, D'\right)\right] &\geq \frac{1}{(2s-c)(2s-c-1)}\sum_{i=0}^{s-c-1}\frac{(s-c-i)^i}{i!}\frac{i!}{(2s-c-2)^i} \\ &= \frac{1}{(2s-c)(2s-c-1)}\left(1+\sum_{i=0}^{s-c-1}\left(\frac{s-c-i}{2s-c-2}\right)^i\right) \\ &\geq \frac{1}{(2s-c)^2}\end{aligned} \tag{A.7}$$

We can now observe the following using the small Omega Bachmann-Landau notation.

$$\lim_{s\to\infty}\frac{(2s-c)^{-2}}{e^{-s}} = \infty \quad \Longrightarrow \quad \mathbb{E}_{D,D'}\left[\kappa\left(x; Z_1, D\right)\kappa\left(x; Z'_{c+1}, D'\right)\right] = \omega(e^{-x}) \tag{A.8}$$

Considering the third case, without loss of generality, we consider the case of $i = j = c + 1$. We find the following.

$$\begin{aligned}
&\mathbb{E}_{D,D'}\left[\kappa\left(x; Z_{c+1}, D\right)\kappa\left(x; Z'_{c+1}, D'\right)\right] \\
&= \mathbb{E}_{D,D'}\left[\kappa\left(x; Z_{c+1}, D_{1:c}\right)\kappa\left(x; Z'_{c+1}, D_{1:c}\right)\kappa\left(x; Z_{c+1}, D_{(c+1):s}\right)\kappa\left(x; Z'_{c+1}, D'_{(c+1):s}\right)\right] \\
&= \frac{2}{(2s-c)!}\sum_{i=0}^{s-c-1}\binom{s-c-1}{i}(s-1+i)!(s-c-1-i)! \\
&= \frac{2(2s-c-2)!}{(2s-c)!}\sum_{i=0}^{s-c-1}\binom{s-c-1}{i}\binom{2s-c-2}{s-1+i}^{-1} \\
&= \frac{2}{(2s-c)(2s-c-1)}\sum_{i=0}^{s-c-1}\binom{s-c-1}{i}\binom{2s-c-2}{s-1+i}^{-1}
\end{aligned} \tag{A.9}$$

The third case follows from a similar combinatorial logic as the second. We consider without loss of generality the case that $\|X'_{c+1} - x\| \leq \|X_{c+1} - x\|$ and adjust for this fact by multiplying the whole expression by two. Notice now that any number $i = 0, \ldots, s-c-1$ of observations in $D'_{(c+2):s}$ can be farther away from x than X_{c+1} or at a distance that is between $\|X'_{c+1} - x\|$ and $\|X_{c+1} - x\|$. The summation adjusts for all possible permutations that fulfill this criterion. Furthermore, we can make the following observation.

$$\begin{aligned}
\mathbb{E}_{D,D'}\left[\kappa\left(x; Z_{c+1}, D\right)\kappa\left(x; Z'_{c+1}, D'\right)\right] &\geq \frac{1}{(2s-c)^2}\sum_{i=0}^{s-c-1}\binom{s-c-1}{i}\binom{2s-c-2}{s-1+i}^{-1} \\
&= \frac{1}{(2s-c)^2}\sum_{i=0}^{s-c-1}\left(\frac{(s-c-1)!}{i!(s-c-1-i)!}\cdot\frac{(s-1+i)!(s-c-1-i)!}{(2s-c-2)!}\right) \\
&= \frac{1}{(2s-c)^2}\cdot\frac{(s-c-1)!}{(2s-c-2)!}\sum_{i=0}^{s-c-1}\frac{(s-1+i)!}{i!} \\
&= \frac{1}{(2s-c)^2}\cdot\frac{(s-c-1)!}{(2s-c-2)!}\left(\frac{(2s-c-2)!}{(s-c-1)!} + \sum_{i=0}^{s-c-2}\frac{(s-1+i)!}{i!}\right) \\
&\geq \frac{1}{(2s-c)^2}
\end{aligned} \tag{A.10}$$

We can now observe the following using the small Omega Bachmann-Landau notation.

$$\lim_{s\to\infty}\frac{(2s-c)^{-2}}{e^{-s}} = \infty \quad \Longrightarrow \quad \mathbb{E}_{D,D'}\left[\kappa\left(x; Z_{c+1}, D\right)\kappa\left(x; Z'_{c+1}, D'\right)\right] = \omega(e^{-x}) \tag{A.11}$$

■

Lemma A.4.

Fix sample size n, subsampling scale s, and c such that $0 < c \leq s \leq n$. Let $D = \{Z_1, Z_2, \ldots, Z_c, Z_{c+1}, \ldots Z_s\}$ be an i.i.d. data set drawn from P as described in Setup 1. Let $D' = \left\{Z_1, Z_2, \ldots, Z_c, Z'_{c+1}, \ldots Z'_s\right\}$ be a second data set that shares the first c observations with D. The remaining $s - c$ observations of D', i.e. $\left\{Z'_{c+1}, \ldots Z'_s\right\}$, are i.i.d. draws from P that are independent of D.

Then, the following statements hold.

$$\forall i \in [c]: \quad \mathbb{E}\left[\kappa\left(x; Z_i, D\right)\kappa\left(x; Z_i, D'\right) \mid X_i\right] = \left\{1 - \varphi\left(B\left(x, \|X_i - x\|\right)\right)\right\}^{2s-c-1} \tag{A.12}$$

$$\begin{aligned}\forall i \in [c]\ \forall j \in \{c+1, \ldots, s\}: \quad & \mathbb{E}\left[\kappa\left(x; Z_i, D\right)\kappa\left(x; Z'_j, D'\right) \mid X_i, X'_j\right] \\ & = \mathbb{1}\left(\|X'_j - x\| \leq \|X_i - x\|\right) \cdot \left\{1 - \varphi\left(B\left(x, \|X_i - x\|\right)\right)\right\}^{s-1} \cdot \left\{1 - \varphi\left(B\left(x, \left\|X'_j - x\right\|\right)\right)\right\}^{s-c-1}\end{aligned} \tag{A.13}$$

$$\begin{aligned}\forall i, j \in \{c+1, \ldots, s\}: \quad & \mathbb{E}\left[\kappa\left(x; Z_i, D\right)\kappa\left(x; Z'_j, D'\right) \mid X_i, X'_j\right] \\ & = \left\{1 - \varphi\left(B\left(x, \min\left(\|X_i - x\|, \left\|X'_j - x\right\|\right)\right)\right)\right\}^{s-c-1} \cdot \left\{1 - \varphi\left(B\left(x, \max\left(\|X_i - x\|, \left\|X'_j - x\right\|\right)\right)\right)\right\}^{s-1}\end{aligned} \tag{A.14}$$

Proof of Lemma A.4.

Without loss of generality, we will consider the cases of $i = 1$ for the first equation.

$$\begin{aligned}\mathbb{E}\left[\kappa\left(x; Z_1, D\right)\kappa\left(x; Z_1, D'\right) \mid X_1\right] &= \mathbb{E}\left[\kappa\left(x; Z_1, D_{[c]}\right)\kappa\left(x; Z_1, D_{(c+1):s}\right)\kappa\left(x; Z_1, D'_{(c+1):s}\right) \,\middle|\, X_1\right] \\ &= \mathbb{E}\left[\kappa\left(x; Z_1, D_{[c]}\right) \mid X_1\right] \cdot \mathbb{E}\left[\kappa\left(x; Z_1, D_{(c+1):s}\right) \mid X_1\right] \cdot \mathbb{E}\left[\kappa\left(x; Z_1, D'_{(c+1):s}\right) \,\middle|\, X_1\right] \\ &= \left\{1 - \varphi\left(B\left(x, \|X_i - x\|\right)\right)\right\}^{c-1} \cdot \left\{1 - \varphi\left(B\left(x, \|X_i - x\|\right)\right)\right\}^{s-c} \cdot \left\{1 - \varphi\left(B\left(x, \|X_i - x\|\right)\right)\right\}^{s-c} \\ &= \left\{1 - \varphi\left(B\left(x, \|X_i - x\|\right)\right)\right\}^{2s-c-1}\end{aligned} \tag{A.15}$$

Without loss of generality, we will consider the cases of $i = 1$ and $j = c + 1$ for the second equation.

$$\begin{aligned}&\mathbb{E}\left[\kappa\left(x; Z_1, D\right)\kappa\left(x; Z'_{c+1}, D'\right) \mid X_1, X'_{c+1}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\kappa\left(x; Z_1, D_{1:c}\right)\kappa\left(x; Z_1, D_{(c+1):s}\right)\kappa\left(x; Z'_{c+1}, D_{1:c}\right)\kappa\left(x; Z'_{c+1}, D'_{(c+1):s}\right) \,\middle|\, X_1, \ldots, X_c, X'_{c+1}\right] \,\middle|\, X_1, X'_{c+1}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\kappa\left(x; Z_1, D_{(c+1):s}\right) \cdot \kappa\left(x; Z'_{c+1}, D'_{(c+1):s}\right) \,\middle|\, X_1, \ldots, X_c, X'_{c+1}\right] \cdot \kappa\left(x; Z_1, D_{1:c}\right) \cdot \kappa\left(x; Z'_{c+1}, D_{1:c}\right) \,\middle|\, X_1, X'_{c+1}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\kappa\left(x; Z_1, D_{(c+1):s}\right) \cdot \kappa\left(x; Z'_{c+1}, D'_{(c+1):s}\right) \,\middle|\, X_1, X'_{c+1}\right] \cdot \kappa\left(x; Z_1, D_{1:c}\right) \cdot \kappa\left(x; Z'_{c+1}, D_{1:c}\right) \,\middle|\, X_1, X'_{c+1}\right] \\ &= \mathbb{E}\left[\kappa\left(x; Z_1, D_{(c+1):s}\right) \cdot \kappa\left(x; Z'_{c+1}, D'_{(c+1):s}\right) \,\middle|\, X_1, X'_{c+1}\right] \cdot \mathbb{E}\left[\kappa\left(x; Z_1, D_{1:c}\right) \cdot \kappa\left(x; Z'_{c+1}, D_{1:c}\right) \mid X_1, X'_{c+1}\right] \\ &= \mathbb{E}\left[\kappa\left(x; Z_1, D_{(c+1):s}\right) \mid X_1\right] \cdot \mathbb{E}\left[\kappa\left(x; Z'_{c+1}, D'_{(c+1):s}\right) \,\middle|\, X'_{c+1}\right] \cdot \mathbb{1}\left(\|X'_{c+1} - x\| \leq \|X_1 - x\|\right) \cdot \mathbb{E}\left[\kappa\left(x; Z_1, D_{1:c}\right) \mid X_1\right] \\ &= \mathbb{1}\left(\|X'_{c+1} - x\| \leq \|X_1 - x\|\right) \cdot \mathbb{E}\left[\kappa\left(x; Z_1, D\right) \mid X_1\right] \cdot \mathbb{E}\left[\kappa\left(x; Z'_{c+1}, D'_{(c+1):s}\right) \,\middle|\, X'_{c+1}\right] \\ &= \mathbb{1}\left(\|X'_{c+1} - x\| \leq \|X_1 - x\|\right) \cdot \left\{1 - \varphi\left(B\left(x, \|X_1 - x\|\right)\right)\right\}^{s-1} \cdot \left\{1 - \varphi\left(B\left(x, \left\|X'_{c+1} - x\right\|\right)\right)\right\}^{s-c-1}\end{aligned} \tag{A.16}$$

For the third case, without loss of generality, we consider the case of $i = j = c + 1$.

$$\begin{aligned}
&\mathbb{E}\left[\kappa\left(x; Z_{c+1}, D\right)\kappa\left(x; Z'_{c+1}, D'\right) \mid X_{c+1}, X'_{c+1}\right] \\
&\quad= \mathbb{E}\left[\mathbb{E}\left[\kappa\left(x; Z_{c+1}, D\right)\kappa\left(x; Z'_{c+1}, D'\right) \mid X_1, \ldots, X_c, X_{c+1}, X'_{c+1}\right] \mid X_{c+1}, X'_{c+1}\right] \\
&\quad= \mathbb{E}\left[\mathbb{E}\left[\kappa\left(x; Z_{c+1}, D_{1:(c+1)}\right)\kappa\left(x; Z'_{c+1}, D'_{1:(c+1)}\right)\right.\right. \\
&\qquad\left.\left.\kappa\left(x; Z_{c+1}, D_{(c+1):s}\right)\kappa\left(x; Z'_{c+1}, D'_{(c+1):s}\right) \mid X_1, \ldots, X_c, X_{c+1}, X'_{c+1}\right] \mid X_{c+1}, X'_{c+1}\right] \\
&\quad= \mathbb{E}\left[\mathbb{E}\left[\kappa\left(x; Z_{c+1}, D_{1:(c+1)}\right)\kappa\left(x; Z'_{c+1}, D'_{1:(c+1)}\right) \mid X_1, \ldots, X_c, X_{c+1}, X'_{c+1}\right]\right. \\
&\qquad\left.\kappa\left(x; Z_{c+1}, D_{(c+1):s}\right)\kappa\left(x; Z'_{c+1}, D'_{(c+1):s}\right) \mid X_{c+1}, X'_{c+1}\right] \\
&\quad= \mathbb{E}\left[\kappa\left(x; Z_{c+1}, D_{1:(c+1)}\right)\kappa\left(x; Z'_{c+1}, D'_{1:(c+1)}\right) \mid X_{c+1}, X'_{c+1}\right] \\
&\qquad\cdot \mathbb{E}\left[\kappa\left(x; Z_{c+1}, D_{(c+1):s}\right) \mid X_{c+1}\right] \cdot \mathbb{E}\left[\kappa\left(x; Z'_{c+1}, D'_{(c+1):s}\right) \mid X'_{c+1}\right]
\end{aligned} \tag{A.17}$$

Without loss of generality, consider the case that $\|X_{c+1} - x\| \leq \|X'_{c+1} - x\|$.

$$\mathbb{E}\left[\kappa\left(x; Z_{c+1}, D_{1:(c+1)}\right)\kappa\left(x; Z'_{c+1}, D'_{1:(c+1)}\right) \mid X_{c+1}, X'_{c+1}\right] = \mathbb{E}\left[\kappa\left(x; Z'_{c+1}, D'_{1:(c+1)}\right) \mid X'_{c+1}\right] \tag{A.18}$$

Furthermore, observe the following.

$$\mathbb{E}\left[\kappa\left(x; Z'_{c+1}, D'_{1:(c+1)}\right) \mid X'_{c+1}\right] \cdot \mathbb{E}\left[\kappa\left(x; Z'_{c+1}, D'_{(c+1):s}\right) \mid X'_{c+1}\right] = \mathbb{E}\left[\kappa\left(x; Z'_{c+1}, D'\right) \mid X'_{c+1}\right] \tag{A.19}$$

Thus, we can find the following.

$$\begin{aligned}
&\mathbb{E}\left[\kappa\left(x; Z_1, D\right)\kappa\left(x; Z'_{c+1}, D'\right) \mid X_1, X'_{c+1}\right] \\
&\quad= \mathbb{1}\left(\|X'_{c+1} - x\| \leq \|X_{c+1} - x\|\right) \cdot \left\{1 - \varphi\left(B\left(x, \|X_{c+1} - x\|\right)\right)\right\}^{s-1} \cdot \left\{1 - \varphi\left(B\left(x, \left\|X'_{c+1} - x\right\|\right)\right)\right\}^{s-c-2} \\
&\qquad+ \mathbb{1}\left(\|X'_{c+1} - x\| > \|X_{c+1} - x\|\right) \cdot \left\{1 - \varphi\left(B\left(x, \|X_{c+1} - x\|\right)\right)\right\}^{s-c-1} \cdot \left\{1 - \varphi\left(B\left(x, \left\|X'_{c+1} - x\right\|\right)\right)\right\}^{s-1} \\
&\quad= \left\{1 - \varphi\left(B\left(x, \min\left(\|X_{c+1} - x\|, \left\|X'_{c+1} - x\right\|\right)\right)\right)\right\}^{s-c-1} \cdot \left\{1 - \varphi\left(B\left(x, \max\left(\|X_{c+1} - x\|, \left\|X'_{c+1} - x\right\|\right)\right)\right)\right\}^{s-1}
\end{aligned} \tag{A.20}$$

■

Lemma A.5.

Fix sample size n, subsampling scale s, and c such that $0 < c \leq s \leq n$. Let $D = \{Z_1, Z_2, \ldots, Z_c, Z_{c+1}, \ldots Z_s\}$ be an i.i.d. data set drawn from P as described in Setup 1. Let $D' = \left\{Z_1, Z_2, \ldots, Z_c, Z'_{c+1}, \ldots Z'_s\right\}$ be a second data set that shares the first c observations with D. The remaining $s - c$ observations of D', i.e. $\left\{Z'_{c+1}, \ldots Z'_s\right\}$, are i.i.d. draws from P that are independent of D.

For any $L^2(\mathcal{X})$ function f that is continuous at x, it holds that

$$\lim_{s\to\infty} \underbrace{\mathbb{E}_1\left[f^2(X_1)(2s-c)\mathbb{E}_{2:s}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z_1,D'\right)\right]\right]}_{(A)} = f^2(x) \tag{A.21}$$

$$\lim_{s\to\infty} \underbrace{\mathbb{E}_{1,(c+1)'}\left[f(X_1)f(X'_{c+1})\cdot\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_1,Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\right]}_{(B)} = f^2(x) \tag{A.22}$$

$$\lim_{s\to\infty} \underbrace{\mathbb{E}_{c+1}\left[f(X_{c+1})f(X'_{c+1})\cdot\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_{c+1},Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\right]}_{(C)} = f^2(x) \tag{A.23}$$

Proof of Lemma A.5.

We will largely argue along the same lines as the original proof in Demirkaya et al. (2024). Thus, consider first the following inequalities.

$$\begin{aligned}\left|(A)-f^2(x)\right| &= \left|\mathbb{E}_1\left[f^2(X_1)(2s-c)\mathbb{E}_{2:s}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z_1,D'\right)\right]\right]-f^2(x)\right| \\ &\leq \mathbb{E}_1\left[\left|f^2(X_1)-f^2(x)\right|(2s-c)\mathbb{E}_{2:s}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z_1,D'\right)\right]\right]\end{aligned} \tag{A.24}$$

$$\begin{aligned}\left|(B)-f^2(x)\right| &= \left|\mathbb{E}_{1,(c+1)'}\left[f(X_1)f(X'_{c+1})\cdot\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_1,Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\right]-f^2(x)\right| \\ &\leq \mathbb{E}_{1,(c+1)'}\left[\left|f(X_1)f(X'_{c+1})-f^2(x)\right|\cdot\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_1,Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\right]\end{aligned} \tag{A.25}$$

$$\begin{aligned}\left|(C)-f^2(x)\right| &= \left|\mathbb{E}_{c+1}\left[f(X_{c+1})f(X'_{c+1})\cdot\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_{c+1},Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\right]-f^2(x)\right| \\ &\leq \mathbb{E}_{c+1}\left[\left|f(X_{c+1})f(X'_{c+1})-f^2(x)\right|\cdot\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_{c+1},Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\right]\end{aligned} \tag{A.26}$$

Now, fix an arbitrary $\epsilon > 0$. By continuity of f at x, there exists a $\delta > 0$, such that the following holds.

$$\forall X, X' \in B(x,\delta): \quad \left|f(X)\cdot f(X')-f^2(x)\right| < \epsilon \tag{A.27}$$

We can consider decompositions of these terms in analogy to Demirkaya et al. (2024), i.e. by considering cases with

observations lying within this sphere or outside of it, and observe the following.

$$\begin{aligned}
&\mathbb{E}_1\left[\left|f^2(X_1)-f^2(x)\right|(2s-c)\mathbb{E}_{2:s}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z_1,D'\right)\mathbb{1}\left(X_1\in B(x,\delta)\right)\right]\right]\\
&\quad\leq\epsilon\cdot\mathbb{E}_1\left[(2s-c)\mathbb{E}_{2:s}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z_1,D'\right)\mathbb{1}\left(X_1\in B(x,\delta)\right)\right]\right]\\
&\quad\leq\epsilon\cdot\mathbb{E}_1\left[(2s-c)\mathbb{E}_{2:s}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z_1,D'\right)\right]\right]=\epsilon
\end{aligned}\tag{A.28}$$

$$\begin{aligned}
&\mathbb{E}_{1,(c+1)'}\left[\left|f(X_1)f(X'_{c+1})-f^2(x)\right|\cdot\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_1,Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\mathbb{1}\left(X_1,X'_{c+1}\in B(x,\delta)\right)\right]\\
&\quad\leq\epsilon\cdot\mathbb{E}_{1,(c+1)'}\left[\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_1,Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\mathbb{1}\left(X_1,X'_{c+1}\in B(x,\delta)\right)\right]\\
&\quad\leq\epsilon\cdot\mathbb{E}_{1,(c+1)'}\left[\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_1,Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\right]=\epsilon
\end{aligned}\tag{A.29}$$

$$\begin{aligned}
&\mathbb{E}_{c+1}\left[\left|f(X_{c+1})f(X'_{c+1})-f^2(x)\right|\cdot\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_{c+1},Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\mathbb{1}\left(X_{c+1},X'_{c+1}\in B(x,\delta)\right)\right]\\
&\quad\leq\epsilon\cdot\mathbb{E}_{c+1}\left[\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_{c+1},Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\mathbb{1}\left(X_{c+1},X'_{c+1}\in B(x,\delta)\right)\right]\\
&\quad\leq\epsilon\cdot\mathbb{E}_{c+1}\left[\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_{c+1},Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\right]=\epsilon
\end{aligned}\tag{A.30}$$

Considering next the parts of the expectation that are not covered by the previous cases, we can find the following. As in the original proof, we use the fact that if X or X' do not lie within $B(x,\delta)$, then the following holds

$$B(x,\delta)\subseteq B(x,\max\left(\|X-x\|,\|X'-x\|\right)).\tag{A.31}$$

This allows us to find the following.

$$\begin{aligned}
&\mathbb{E}_1\left[\left|f^2(X_1)-f^2(x)\right|(2s-c)\mathbb{E}_{2:s}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z_1,D'\right)\left(1-\mathbb{1}\left(X_1\in B(x,\delta)\right)\right)\right]\right]\\
&\quad\leq\mathbb{E}_1\left[\left|f^2(X_1)-f^2(x)\right|(2s-c)\left\{1-\varphi\left(B\left(x,\delta\right)\right)\right\}^{2s-c-1}\left(1-\mathbb{1}\left(X_1\in B(x,\delta)\right)\right)\right]\\
&\quad\leq(2s-c)1\cdot\left\{1-\varphi\left(B\left(x,\delta\right)\right)\right\}^{2s-c-1}\cdot\mathbb{E}_1\left[\left|f^2(X_1)-f^2(x)\right|\right]
\end{aligned}\tag{A.32}$$

In the second case, first recall the form of the conditional expectation from Lemma A.4.

$$\begin{aligned}
&\mathbb{E}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid X_1,X'_{c+1}\right]\\
&\quad=\mathbb{1}\left(\|X'_{c+1}-x\|\leq\|X_1-x\|\right)\cdot\left\{1-\varphi\left(B\left(x,\|X_1-x\|\right)\right)\right\}^{s-1}\cdot\left\{1-\varphi\left(B\left(x,\left\|X'_{c+1}-x\right\|\right)\right)\right\}^{s-c-1}
\end{aligned}\tag{A.33}$$

The indicator variable in this expression is only non-zero if $\max\left(\|X_1-x\|,\|X'_{c+1}-x\|\right)=\|X_1-x\|$. Thus, in light of the conditioning, we can observe the following.

$$B(x,\delta)\subseteq B(x,\|X_1-x\|)\tag{A.34}$$

Thus, we can make the following observation.

$$\begin{aligned}
&\mathbb{E}_{1,(c+1)'}\left[\left|f(X_1)f(X'_{c+1})-f^2(x)\right|\cdot\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_1,Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\left(1-\mathbb{1}\left(X_1,X'_{c+1}\in B(x,\delta)\right)\right)\right]\\
&\quad\overset{\text{Lem A.3}}{\leq}(2s-c)^2\cdot\mathbb{E}_{1,(c+1)'}\Big[\left|f(X_1)f(X'_{c+1})-f^2(x)\right|\\
&\qquad\cdot\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_1,Z'_{c+1}\right]\left(1-\mathbb{1}\left(X_1,X'_{c+1}\in B(x,\delta)\right)\right)\Big]\\
&\quad\overset{\text{Lem A.4}}{=}(2s-c)^2\cdot\mathbb{E}_{1,(c+1)'}\Big[\left|f(X_1)f(X'_{c+1})-f^2(x)\right|\cdot\mathbb{1}\left(\|X'_{c+1}-x\|\leq\|X_1-x\|\right)\\
&\qquad\cdot\left\{1-\varphi\left(B\left(x,\|X_1-x\|\right)\right)\right\}^{s-1}\cdot\left\{1-\varphi\left(B\left(x,\left\|X'_{c+1}-x\right\|\right)\right)\right\}^{s-c-1}\cdot\left(1-\mathbb{1}\left(X_1,X'_{c+1}\in B(x,\delta)\right)\right)\Big]\\
&\quad\leq(2s-c)^2\cdot\left\{1-\varphi\left(B\left(x,\delta\right)\right)\right\}^{s-1}\cdot\mathbb{E}_{1,(c+1)'}\Big[\left|f(X_1)f(X'_{c+1})-f^2(x)\right|\cdot\mathbb{1}\left(\delta<\|X'_{c+1}-x\|\leq\|X_1-x\|\right)\Big]\\
&\quad\leq(2s-c)^2\cdot\left\{1-\varphi\left(B\left(x,\delta\right)\right)\right\}^{s-1}\cdot\mathbb{E}_{1,(c+1)'}\left[\left|f(X_1)f(X'_{c+1})-f^2(x)\right|\right]
\end{aligned}\tag{A.35}$$

Similarly, considering the third case, we observe the following.

$$\begin{aligned}
&\mathbb{E}_{c+1}\left[\left|f(X_{c+1})f(X'_{c+1})-f^2(x)\right|\cdot\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_{c+1},Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\left(1-\mathbb{1}\left(X_{c+1},X'_{c+1}\in B(x,\delta)\right)\right)\right]\\
&\quad\overset{\text{Lem A.3}}{\leq}(2s-c)^2\cdot\mathbb{E}_{c+1}\Big[\left|f(X_{c+1})f(X'_{c+1})-f^2(x)\right|\\
&\qquad\cdot\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_{c+1},Z'_{c+1}\right]\left(1-\mathbb{1}\left(X_{c+1},X'_{c+1}\in B(x,\delta)\right)\right)\Big]\\
&\quad\overset{\text{Lem A.4}}{=}(2s-c)^2\cdot\mathbb{E}_{c+1}\Big[\left|f(X_{c+1})f(X'_{c+1})-f^2(x)\right|\cdot\left\{1-\varphi\left(B\left(x,\min\left(\|X_{c+1}-x\|,\left\|X'_{c+1}-x\right\|\right)\right)\right)\right\}^{s-c-1}\\
&\qquad\cdot\left\{1-\varphi\left(B\left(x,\max\left(\|X_{c+1}-x\|,\left\|X'_{c+1}-x\right\|\right)\right)\right)\right\}^{s-1}\cdot\left(1-\mathbb{1}\left(X_{c+1},X'_{c+1}\in B(x,\delta)\right)\right)\Big]\\
&\quad\leq(2s-c)^2\cdot\mathbb{E}_{c+1}\Big[\left|f(X_{c+1})f(X'_{c+1})-f^2(x)\right|\cdot\left\{1-\varphi\left(B\left(x,\min\left(\|X_{c+1}-x\|,\left\|X'_{c+1}-x\right\|\right)\right)\right)\right\}^{s-c-1}\\
&\qquad\cdot\left\{1-\varphi\left(B\left(x,\delta\right)\right)\right\}^{s-1}\cdot\left(1-\mathbb{1}\left(X_{c+1},X'_{c+1}\in B(x,\delta)\right)\right)\Big]\\
&\quad\leq(2s-c)^2\cdot\left\{1-\varphi\left(B\left(x,\delta\right)\right)\right\}^{s-1}\cdot\mathbb{E}_{c+1}\Big[\left|f(X_{c+1})f(X'_{c+1})-f^2(x)\right|\Big]
\end{aligned}\tag{A.36}$$

Concerning the resulting terms in these three expressions, we can then make the following observations.

$$\mathbb{E}\Big[\left|f^2(X)-f^2(x)\right|\Big]\leq\mathbb{E}\Big[f^2(X)\Big]+f^2(x)=\|f\|_{L_2}^2+f^2(x)\tag{A.37}$$

$$\mathbb{E}\Big[\left|f(X)f(X')-f^2(x)\right|\Big]\leq\mathbb{E}\Big[\left|f(X)f(X')\right|\Big]+f^2(x)\leq\mathbb{E}\Big[\left|f(X)\right|\cdot\left|f(X')\right|\Big]+f^2(x)=\|f\|_{L^1}^2+f^2(x)\tag{A.38}$$

As f is an $L^2(\mathcal{X})$ function on a bounded domain, observe that $\|f\|_{L^1}$ is finite. Thus, we can find the following.

$$\mathbb{E}_1\left[\left|f^2(X_1)-f^2(x)\right|(2s-c)\mathbb{E}_{2:s}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z_1,D'\right)\left(1-\mathbb{1}\left(X_1\in B(x,\delta)\right)\right)\right]\right]\longrightarrow 0\quad\text{as}\quad s\to\infty\tag{A.39}$$

$$\begin{aligned}
&\mathbb{E}_{1,(c+1)'}\left[\left|f(X_1)f(X'_{c+1})-f^2(x)\right|\cdot\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_1,Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\left(1-\mathbb{1}\left(X_1,X'_{c+1}\in B(x,\delta)\right)\right)\right]\\
&\quad\longrightarrow 0\quad\text{as}\quad s\to\infty
\end{aligned}\tag{A.40}$$

$$\mathbb{E}_{c+1}\left[\left|f(X_{c+1})f(X'_{c+1}) - f^2(x)\right| \cdot \frac{\mathbb{E}_{D,D'}\left[\kappa\left(x; Z_{c+1}, D\right)\kappa\left(x; Z'_{c+1}, D'\right) \mid Z_{c+1}, Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x; Z_{c+1}, D\right)\kappa\left(x; Z'_{c+1}, D'\right)\right]}\left(1 - \mathbb{1}\left(X_{c+1}, X'_{c+1} \in B(x,\delta)\right)\right)\right]$$
$$\longrightarrow 0 \quad \text{as} \quad s \to \infty \tag{A.41}$$

Combining these findings, for large enough s we can bound the terms $\|(A) - f^2(x)\|$, $\|(B) - f^2(x)\|$, and $\|(C) - f^2(x)\|$ by 2ϵ, respectively. As ϵ was arbitrary this concludes the proof. ■

Lemma A.6.

Fix sample size n, subsampling scale s, and c such that $0 < c \leq s \leq n$. Let $D = \{Z_1, Z_2, \ldots, Z_c, Z_{c+1}, \ldots Z_s\}$ be an i.i.d. data set drawn from P as described in Setup 1. Let $D' = \left\{Z_1, Z_2, \ldots, Z_c, Z'_{c+1}, \ldots Z'_s\right\}$ be a second data set that shares the first c observations with D. The remaining $s - c$ observations of D', i.e. $\left\{Z'_{c+1}, \ldots Z'_s\right\}$, are i.i.d. draws from P that are independent of D.

Then the following inequalities hold.

$$\forall i \in [c]\ \forall j \in \{c+1, \ldots, s\}: \quad \mathbb{E}_{D,D'}\left[\kappa\left(x; Z_i, D\right)\kappa\left(x; Z'_j, D'\right)\right] \leq \frac{s}{(2s-c)(2s-c-1)(c+1)} \tag{A.42}$$

$$\forall i, j \in \{c+1, \ldots, s\}: \quad \mathbb{E}_{D,D'}\left[\kappa\left(x; Z_i, D\right)\kappa\left(x; Z'_j, D'\right)\right] \leq \frac{2(s-c)}{(2s-c)^2(2s-c-1)} \tag{A.43}$$

Proof of Lemma A.6.

Recall the results of Lemma A.3 and make the following observations.

$$\begin{aligned}
\mathbb{E}_{D,D'}\left[\kappa\left(x; Z_i, D\right)\kappa\left(x; Z'_j, D'\right)\right] &= \frac{1}{(2s-c)(2s-c-1)} \sum_{i=0}^{s-c-1} \binom{s-c-1}{i}\binom{2s-c-2}{i}^{-1} \\
&\leq \frac{1}{(2s-c)(2s-c-1)} \sum_{i=0}^{s-c-1} \frac{(s-c-1)^i}{i!} \cdot \frac{i!}{(2s-c-1-i)^i} = \frac{1}{(2s-c)(2s-c-1)} \sum_{i=0}^{s-c-1} \left(\frac{s-c-1}{2s-c-1-i}\right)^i \\
&\leq \frac{1}{(2s-c)(2s-c-1)} \sum_{i=0}^{s-c-1} \left(\frac{s-c-1}{2s-c-1-i}\right)^i \leq \frac{1}{(2s-c)(2s-c-1)} \sum_{i=0}^{s-c-1} \left(\frac{s-c-1}{s}\right)^i \\
&\leq \frac{1}{(2s-c)(2s-c-1)} \sum_{i=0}^{\infty} \left(\frac{s-c-1}{s}\right)^i = \frac{s}{(2s-c)(2s-c-1)(c+1)}
\end{aligned} \tag{A.44}$$

Similarly, for the second case, we can make the following observation.

$$
\begin{aligned}
\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_i,D\right)\kappa\left(x;Z_j',D'\right)\right] &= \frac{2}{(2s-c)(2s-c-1)}\sum_{i=0}^{s-c-1}\binom{s-c-1}{i}\binom{2s-c-1}{s-1+i}^{-1} \\
&= \frac{2}{(2s-c)(2s-c-1)}\sum_{i=0}^{s-c-1}\binom{s-c-1}{i}\binom{2s-c-1}{s-c-i}^{-1} \\
&= \frac{2}{(2s-c)(2s-c-1)}\sum_{i=0}^{s-c-1}\frac{(s-c-1)!}{(s-c-1-i)!i!}\frac{(s-1+i)!(s-c-i)!}{(2s-c-1)!} \\
&= \frac{2}{(2s-c)(2s-c-1)}\cdot\frac{(s-c-1)!(s-1)!}{(2s-c-1)!}\cdot\sum_{i=0}^{s-c-1}(s-c-i)\binom{s-1+i}{i} \\
&\leq \frac{2}{(2s-c)(2s-c-1)}\cdot\frac{(s-c)!(s-1)!}{(2s-c-1)!}\cdot\sum_{i=0}^{s-c-1}\binom{s-1+i}{i} \\
&= \frac{2}{(2s-c)(2s-c-1)}\cdot\binom{2s-c}{s-c}^{-1}\cdot\binom{2s-c-1}{s-c-1} = \frac{2(s-c)}{(2s-c)^2(2s-c-1)}
\end{aligned}
\tag{A.45}
$$

■

Lemma A.7 (Peng, Mentch, and Stefanski (2021) - Lemma 1)**.**
Suppose that $\sum X_i^2 \xrightarrow{p} 1, \sum \mathbb{E}\left[X_i^2\right] \to 1$*, and* $\sum_{i=1}^n \mathbb{E}\left[Y_i^2\right] \to 0$*, then*

$$
\sum\left[X_i+Y_i\right]^2 \xrightarrow{p} 1 \quad \textit{and} \quad \mathbb{E}\left[\sum\left(X_i+Y_i\right)^2\right] \to 1. \tag{A.46}
$$

Lemma A.8 (Honesty of the DNN/TDNN Estimators)**.**
The DNN and TDNN estimator kernels $\kappa\left(\cdot,\cdot,D_\ell\right)$ *are Honest in the sense of Wager and Athey (2018).*

$$
\kappa\left(x,X_i,D_\ell\right) \perp\!\!\!\perp Y_i \mid X_i, D_{\ell,-i},
$$

where $\perp\!\!\!\perp$ *denotes conditional independence and* $D_{\ell,-i} = \{Z_l \mid l \in \ell \backslash \{i\}\}$.

A.2 Consequences of DDML Rate-Assumptions

Recall first the conditions imposed in Assumption 8.

$$r_n' := \sup_{\eta \in \mathcal{T}_n} \left(\mathbb{E}_Z \left[\left| m\left(Z; \theta_0, \eta\right) - m\left(Z; \theta_0, \eta_0\right) \right|^2 \right] \right)^{1/2} \leq \delta_n \tag{A.47}$$

$$\lambda_n' := \sup_{r \in (0,1), \eta \in \mathcal{T}_n} \left| \partial_r^2 \mathbb{E}_Z \left[m\left(Z; \theta_0, \eta_0 + r\left(\eta - \eta_0\right)\right) \right] \right| \leq \delta_n / \sqrt{n} \tag{A.48}$$

We are interested in how these conditions translate into statements on the Oracle-Hoeffding decomposition errors. Thus, recall the form of the error terms as introduced in Equation 4.22.

$$R_c\left(x; \mathbf{D}_\ell\right) = \chi_s^{(c)}\left(x; \mathbf{D}_\ell, \widehat{\eta}\right) - \chi_{s,0}^{(c)}\left(x; \mathbf{D}_\ell\right) \tag{A.49}$$

More specifically, we are interested in the part absent the centering term as these mostly cancel out across the error terms of different order.

$$\tilde{R}_c\left(x; \mathbf{D}_\ell\right) = R_c\left(x; \mathbf{D}_\ell\right) - (-1)^c \cdot \left(\mathbb{E}_D \left[\chi_s(x; \mathbf{D}_{[s]}, \widehat{\eta}) - \chi_{s,0}\left(x; \mathbf{D}_{[s]}\right) \right] \right) \tag{A.50}$$

Lemma A.9.
LOREM IPSUM

Proof of Lemma A.9.
Consider the following argument.

$$\begin{aligned}
&\sup_{\eta \in \mathcal{T}_n} \left(\mathbb{E}_Z \left[\left| \chi_s^{(1)}\left(x; Z, \eta\right) - \chi_{s,0}^{(1)}\left(x; Z\right) + \left(\mathbb{E}_D \left[\chi_s(x; \mathbf{D}_{[s]}, \widehat{\eta}) - \chi_{s,0}\left(x; \mathbf{D}_{[s]}\right) \right] \right) \right|^2 \right] \right)^{1/2} \\
&= \sup_{\eta \in \mathcal{T}_n} \left(\mathbb{E}_Z \left[\left| \vartheta_s^1\left(x; Z, \eta\right) - \vartheta_{s,0}^1\left(x; Z\right) \right|^2 \right] \right)^{1/2} \\
&= \sup_{\eta \in \mathcal{T}_n} \Bigg(\mathbb{E}_Z \Bigg[\Bigg| \frac{m\left(Z, \theta_0, \eta\right) - m\left(Z, \theta_0, \eta_0\right)}{s!} \cdot \mathbb{E}_D \left[\kappa(x; Z_1, \mathbf{D}_{[s]}) \,\middle|\, Z_1 = Z \right] \\
&+ \frac{s-1}{s!} \cdot \mathbb{E}_D \left[\kappa(x; Z_2, \mathbf{D}_{[s]}) \cdot \left(m\left(Z_2, \theta_0, \eta\right) - m\left(Z_2, \theta_0, \eta_0\right) \right) \,\middle|\, Z_1 = Z \right] \Bigg|^2 \Bigg] \Bigg)^{1/2} \\
&\overset{\triangle\text{-ineq.}}{\leq} \sup_{\eta \in \mathcal{T}_n} \Bigg\{ \frac{1}{s!} \left(\mathbb{E}_Z \left[\left| \left(m\left(Z, \theta_0, \eta\right) - m\left(Z, \theta_0, \eta_0\right) \right) \cdot \mathbb{E}_D \left[\kappa(x; Z_1, \mathbf{D}_{[s]}) \,\middle|\, Z_1 = Z \right] \right|^2 \right] \right)^{1/2} \\
&\quad + \frac{s-1}{s!} \left(\mathbb{E}_Z \left[\left| \mathbb{E}_D \left[\kappa(x; Z_2, \mathbf{D}_{[s]}) \cdot \left(m\left(Z_2, \theta_0, \eta\right) - m\left(Z_2, \theta_0, \eta_0\right) \right) \,\middle|\, Z_1 = Z \right] \right|^2 \right] \right)^{1/2} \Bigg\}
\end{aligned} \tag{A.51}$$

Continuing from this step, we can further observe the following.

$$
\begin{aligned}
&\sup_{\eta\in\mathcal{T}_n}\left(\mathbb{E}_Z\left[\left|\chi_s^{(1)}\left(x;Z,\eta\right)-\chi_{s,0}^{(1)}\left(x;Z\right)+\left(\mathbb{E}_D\left[\chi_s(x;\mathbf{D}_{[s]},\widehat{\eta})-\chi_{s,0}\left(x;\mathbf{D}_{[s]}\right)\right]\right)\right|^2\right]\right)^{1/2}\\
&\quad\leq\frac{1}{s!}\sup_{\eta\in\mathcal{T}_n}\left(\mathbb{E}_Z\left[\left(m\left(Z,\theta_0,\eta\right)-m\left(Z,\theta_0,\eta_0\right)\right)^2\cdot\mathbb{E}_D\left[\kappa(x;Z_1,\mathbf{D}_{[s]})\,\middle|\,Z_1=Z\right]^2\right]\right)^{1/2}\\
&\qquad+\frac{s-1}{s!}\sup_{\eta\in\mathcal{T}_n}\left(\mathbb{E}_Z\left[\mathbb{E}_D\left[\kappa(x;Z_2,\mathbf{D}_{[s]})\cdot\left(m\left(Z_2,\theta_0,\eta\right)-m\left(Z_2,\theta_0,\eta_0\right)\right)\,\middle|\,Z_1=Z\right]^2\right]\right)^{1/2}\\
&\quad\leq\frac{1}{s!}\sup_{\eta\in\mathcal{T}_n}\left(\mathbb{E}_Z\left[\left(m\left(Z,\theta_0,\eta\right)-m\left(Z,\theta_0,\eta_0\right)\right)^2\right]\right)^{1/2}+\frac{s-1}{s!}\sup_{\eta\in\mathcal{T}_n}\left(\mathbb{E}_Z\left[\mathbb{E}_D\left[m\left(Z_2,\theta_0,\eta\right)-m\left(Z_2,\theta_0,\eta_0\right)\,\middle|\,Z_1=Z\right]^2\right]\right)^{1/2}\\
&\quad\leq\frac{\delta_n}{(s-1)!}
\end{aligned}
\tag{A.52}
$$

LOREM IPSUM ■

Lemma A.10.
LOREM IPSUM

Proof of Lemma A.10.

$$
\begin{aligned}
&\sup_{r\in(0,1),\eta\in\mathcal{T}_n}\left|\partial_r^2\mathbb{E}_Z\left[\vartheta_s^1\left(x;Z,\eta_0+r\left(\eta-\eta_0\right)\right)\right]\right|=\sup_{r\in(0,1),\eta\in\mathcal{T}_n}\left|\partial_r^2\mathbb{E}_Z\left[\mathbb{E}_D\left[\chi_s(x;\mathbf{D}_{[s]},\eta_0+r\left(\eta-\eta_0\right))\,\middle|\,Z_1=Z\right]\right]\right|\\
&\quad=\sup_{r\in(0,1),\eta\in\mathcal{T}_n}\left|\partial_r^2\mathbb{E}_D\left[\chi_s(x;\mathbf{D}_{[s]},\eta_0+r\left(\eta-\eta_0\right))\right]\right|\\
&\quad=\frac{1}{s!}\cdot\sup_{r\in(0,1),\eta\in\mathcal{T}_n}\left|\partial_r^2\mathbb{E}_D\left[m(Z,\theta_0,\eta_0+r\left(\eta-\eta_0\right))\cdot s\kappa(x;Z,\mathbf{D}_{[s]})\right]\right|\\
&\quad=\frac{1}{s!}\cdot\sup_{r\in(0,1),\eta\in\mathcal{T}_n}\left|\partial_r^2\mathbb{E}_1\left[m(Z_1,\theta_0,\eta_0+r\left(\eta-\eta_0\right))\cdot s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,\mathbf{D}_{[s]})\right]\right]\right|\\
&\quad=LOREMIPSUM
\end{aligned}
\tag{A.53}
$$

effectively, we need a decaying second derivative in a shrinking neighborhood of x. This might be easier to do with a slightly weaker but more explicit rate condition that I would need to show for the CATE moment. I might be able to do something with DCT here! LOREM IPSUM ■

B Proofs for Results in Section 5

B.1 Closed Form Representations

Proof of Theorem 5.1.
Recall the closed form representation of the DNN estimator as presented in Equation 3.8 and its asymptotic approximation in Equation 3.9.

$$\tilde{\mu}_s(x;\mathbf{D}_n) = \binom{n}{s}^{-1} \sum_{i=1}^{n-s+1} \binom{n-i}{s-1} Y_{(i)} \approx \sum_{i=1}^{n-s+1} \alpha_s (1-\alpha_s)^{i-1} Y_{(i)} \tag{B.1}$$

Plugging into the Jackknife variance estimator for the DNN estimator now gives us the following where we assume that n is sufficiently large for $n-s+1$ to be larger than s.

$$\begin{aligned} \widehat{\omega}^2_{\mathrm{JK}} &= \frac{n-1}{n} \sum_{i=1}^{n} \left(\tilde{\mu}_s(x;\mathbf{D}_{n,-i}) - \tilde{\mu}_s(x;\mathbf{D}_n)\right)^2 \\ &= \end{aligned} \tag{B.2}$$

Even more simple, we can use the approximate weights to find the following representation. For this purpose recall that $\alpha_s = s/n$ and define $\tilde{\alpha}_s = s/(n-1)$. Thus, $\tilde{\alpha}_s = \frac{n}{n-1}\alpha_s$.

$$\begin{aligned} \widehat{\omega}^2_{\mathrm{JK}} &= \frac{n-1}{n} \sum_{i=1}^{n} \left(\tilde{\mu}_s(x;\mathbf{D}_{n,-i}) - \tilde{\mu}_s(x;\mathbf{D}_n)\right)^2 \\ &\approx \frac{n-1}{n} \Bigg[\sum_{i=1}^{n-s+1} \Bigg(\sum_{j=1}^{i-1} \left(\tilde{\alpha}_s (1-\tilde{\alpha}_s)^{j-1} - \alpha_s (1-\alpha_s)^{j-1} \right) Y_{(j)} \\ &\qquad + \sum_{j=i+1}^{n-s+2} \left(\tilde{\alpha}_s (1-\tilde{\alpha}_s)^{j-1} - \alpha_s (1-\alpha_s)^{j} \right) Y_{(j)} - \alpha_s (1-\alpha_s)^{i-1} Y_{(i)} \Bigg)^2 \\ &\qquad + \sum_{i=n-s+2}^{n} \left(\sum_{j=1}^{n-s+1} \left(\tilde{\alpha}_s (1-\tilde{\alpha}_s)^{j-1} - \alpha_s (1-\alpha_s)^{j-1} \right) Y_{(j)} \right)^2 \Bigg] \\ &= \alpha_s^2 \cdot \frac{n-1}{n} \Bigg[\sum_{i=1}^{n-s+1} \Bigg(\sum_{j=1}^{i-1} \left(\frac{n}{n-1} \left(\frac{n-1-s}{n-1} \right)^{j-1} - \left(\frac{n-s}{n} \right)^{j-1} \right) Y_{(j)} \\ &\qquad + \sum_{j=i+1}^{n-s+2} \left(\frac{n}{n-1} \left(\frac{n-1-s}{n-1} \right)^{j-1} - \left(\frac{n-s}{n} \right)^{j} \right) Y_{(j)} - \left(\frac{n-s}{n} \right)^{i-1} Y_{(i)} \Bigg)^2 \\ &\qquad + \sum_{i=n-s+2}^{n} \left(\sum_{j=1}^{n-s+1} \left(\frac{n}{n-1} \left(\frac{n-1-s}{n-1} \right)^{j-1} - \left(\frac{n-s}{n} \right)^{j-1} \right) Y_{(j)} \right)^2 \Bigg] \end{aligned} \tag{B.3}$$

The closed form of the Jackknife variance estimator for the TDNN estimator follows from the same approach.

LOREM IPSUM ■

B.2 NPR - Kernel (Conditional) Expectations

As part of deriving consistency results for the variance estimators under consideration, we need to do a careful analysis of the Kernel of the DNN and TDNN estimators. In this section of the appendix we will thus derive the expectations of the kernel and its corresponding Hájek projection. First, we start with the nonparametric regression setup.

Lemma B.1 (NPR - DNN Kernel Expectation)**.**
Let x denote a point of interest. Then

$$\mathbb{E}_D\left[h_s\left(x;D\right)\right] = \mathbb{E}_1\left[Y_1 s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \longrightarrow \mu\left(x\right) \quad as \quad s\to\infty \tag{B.4}$$

Proof of Lemma B.1. This result follows immediately from Lemma A.2 and the following observation.

$$\begin{aligned}\mathbb{E}_1\left[Y_1 s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] &= \mathbb{E}_1\left[\left(\mu\left(X_1\right)+\varepsilon_1\right)s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right]\\ &= \mathbb{E}_1\left[\left(\mu\left(X_1\right)+\mathbb{E}\left[\varepsilon_1 \mid X_1\right]\right)s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right]\\ &= \mathbb{E}_1\left[\mu\left(X_1\right)s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \overset{\text{(Lem A.2)}}{\longrightarrow} \mu\left(x\right) \quad \text{as} \quad s\to\infty\end{aligned} \tag{B.5}$$

■

Lemma B.2 (NPR - DNN Hajék Kernel Expectation)**.**
Let $z_1=(x_1,y_1)$ denote a specific realization of Z and x denote a point of interest. Then

$$\psi_s^1\left(x;z_1\right) = \varepsilon_1\mathbb{E}_D\left[\kappa\left(x;Z_1,D\right)\,\middle|\,X_1=x_1\right] + \mathbb{E}_D\left[\sum_{i=2}^{s}\kappa\left(x;Z_i,D\right)\mu(X_i)\,\middle|\,X_1=x_1\right] \tag{B.6}$$

Proof of Lemma B.2.

$$\begin{aligned}\psi_s^1\left(x;z_1\right) &= \mathbb{E}_D\left[h_s\left(x;D\right)\mid Z_1=z_1\right] = \mathbb{E}_D\left[\sum_{i=1}^{s}\kappa\left(x;Z_i,D\right)Y_i\,\middle|\,Z_1=z_1\right]\\ &= \mathbb{E}_D\left[\left(\mu(x_1)+\varepsilon_1\right)\kappa\left(x;Z_1,D\right)+\sum_{i=2}^{s}\kappa\left(x;Z_i,D\right)\mu(X_i)\,\middle|\,Z_1=z_1\right]\\ &= \varepsilon_1\mathbb{E}_D\left[\kappa\left(x;Z_1,D\right)\,\middle|\,X_1=x_1\right]+\mathbb{E}_D\left[\sum_{i=2}^{s}\kappa\left(x;Z_i,D\right)\mu(X_i)\,\middle|\,X_1=x_1\right]\end{aligned} \tag{B.7}$$

■

B.3 CATE - Kernel (Conditional) Expectations

Next, we address the CATE estimation setup, where we first consider the scenario where the nuisance parameters are assumed to be known a priori. In a second step, we will show that asymptotically, the estimation of nuisance parameters as described in Definition 1, does not alter the asymptotic analysis of the estimator. For clarity, we point out that in contexts relating to the estimation of the conditional average treatment effect, the kernel or score function h_s could hypothetically signify the first or second stage kernel. As the first stage is effectively covered by the nonparametric regression setup, we will take h_s in these contexts to mean the kernel weighted Neyman-orthogonal score associated with the CATE.

Lemma B.3 (CATE - DNN Kernel Expectation)**.**
Let x denote a point of interest. Then

$$\begin{aligned}\mathbb{E}_D\left[h_s\left(x;D\right)\right] &= \mathbb{E}_1\left[m(Z_1,\eta_0)s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \\ &\longrightarrow \theta_0(x) \quad as \quad s\to\infty\end{aligned} \tag{B.8}$$

Proof of Lemma B.3. This result follows immediately from Lemma A.2 and the following observation.

$$\begin{aligned}\mathbb{E}_1\left[m\left(Z_1;\eta_0\right)s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] &= \mathbb{E}_1\left[\left(\mu_0^1\left(X_1\right)-\mu_0^0\left(X_1\right)+\beta\left(W_1,X_1\right)\varepsilon_i\right)s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \\ &= \mathbb{E}_1\left[\left(\mu_0^1\left(X_1\right)-\mu_0^0\left(X_1\right)+\beta\left(W_1,X_1\right)\mathbb{E}\left[\varepsilon_1\mid X_1\right]\right)s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \\ &= \mathbb{E}_1\left[\left(\mu_0^1\left(X_1\right)-\mu_0^0\left(X_1\right)\right)s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \\ &\overset{\text{(Lem A.2)}}{\longrightarrow} \mu_0^1\left(x\right)-\mu_0^0\left(x\right)=\theta_0(x) \quad \text{as} \quad s\to\infty\end{aligned} \tag{B.9}$$

■

Lemma B.4 (CATE - DNN Hajék Kernel Expectation)**.**
Let $z_1=(x_1,W_1,y_1)$ denote a specific realization of Z and x denote a point of interest. Then

$$\psi_s^1\left(x;z_1\right)=\beta\left(W_1,X_1\right)\varepsilon_1\cdot\mathbb{E}\left[\kappa\left(x;Z_1,D\right)\mid X_1=x_1\right]+\mathbb{E}_D\left[\sum_{i=2}^{s}\kappa\left(x;Z_i,D\right)\left(\mu_0^1\left(X_i\right)-\mu_0^0\left(X_i\right)\right)\,\middle|\,Z_1=z_1\right] \tag{B.10}$$

Proof of Lemma B.4.

$$\begin{aligned}\psi_s^1\left(x;z_1\right) &= \mathbb{E}_D\left[\chi_{s,0}\left(x;D\right)\mid Z_1=z_1\right] \\ &= \mathbb{E}_D\left[\sum_{i=1}^{s}\kappa\left(x;Z_i,D\right)m(Z_i,\eta_0)\,\middle|\,Z_1=z_1\right] \\ &= \left(\mu_0^1\left(X_1\right)-\mu_0^0\left(X_1\right)+\beta\left(W_1,X_1\right)\varepsilon_1\right)\mathbb{E}\left[\kappa\left(x;Z_1,D\right)\mid X_1=x_1\right] \\ &\quad+\mathbb{E}_D\left[\sum_{i=2}^{s}\kappa\left(x;Z_i,D\right)\left(\mu_0^1\left(X_i\right)-\mu_0^0\left(X_i\right)\right)\,\middle|\,Z_1=z_1\right] \\ &= \beta\left(W_1,X_1\right)\varepsilon_1\cdot\mathbb{E}\left[\kappa\left(x;Z_1,D\right)\mid X_1=x_1\right]+\mathbb{E}_D\left[\sum_{i=2}^{s}\kappa\left(x;Z_i,D\right)\left(\mu_0^1\left(X_i\right)-\mu_0^0\left(X_i\right)\right)\,\middle|\,Z_1=z_1\right]\end{aligned} \tag{B.11}$$

■

B.4 NPR - Kernel Variances & Covariances

Similar to the previous section of proofs, we will continue by analyzing the variances and covariances of the kernels under consideration. These results will play an important role in the derivation of consistency properties for the variance estimators. Similar to the previous part, we will first consider the nonparametric regression setup and then proceed to the conditional average treatment effect setup.

Lemma B.5 (Adapted from Demirkaya et al. (2024))**.**
Let $D = \{Z_1, \ldots, Z_s\}$ be a vector of i.i.d. random variables drawn from P. Furthermore, let

$$\Omega_s(x) = \mathbb{E}\left[h_s^2(x; Z_1, \ldots, Z_s)\right]. \tag{B.12}$$

Then,

$$\Omega_s(x) = \mathbb{E}_1\left[(\mu(X_1) + \varepsilon_1)^2 s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right] \lesssim \mu^2(x) + \overline{\sigma}_\varepsilon^2 + o(1) \quad as \quad s \to \infty. \tag{B.13}$$

Proof of Lemma B.5.
This result follows immediately from Lemma A.2 and the following observation.

$$\begin{aligned}
\Omega_s(x) &= \mathbb{E}\left[h_s^2(x; Z_1, \ldots, Z_s)\right] = \mathbb{E}_D\left[\left(\sum_{i=1}^s \kappa(x; Z_i, D) Y_i\right)^2\right] = \mathbb{E}_D\left[\sum_{i=1}^s \sum_{j=1}^s \left(\kappa(x; Z_i, D)\kappa(x; Z_j, D) Y_i Y_j\right)\right] \\
&= \mathbb{E}_D\left[s\kappa(x; Z_1, D) Y_1^2\right] = \mathbb{E}_1\left[Y_1^2 s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right] = \mathbb{E}_1\left[(\mu(X_1) + \varepsilon_1)^2 s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right] \\
&= \mathbb{E}_1\left[\left(\mu^2(X_1) + 2\mu(X_1)\varepsilon_1 + \varepsilon_1^2\right) s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right] \\
&= \mathbb{E}_1\left[\left(\mu^2(X_1) + 2\mu(X_1)\mathbb{E}\left[\varepsilon_1 \mid X_1\right] + \mathbb{E}\left[\varepsilon_1^2 \mid X_1\right]\right) s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right] \\
&= \mathbb{E}_1\left[\left(\mu^2(X_1) + \sigma_\varepsilon^2(X_1)\right) s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right] \overset{\text{(Lem A.2)}}{\longrightarrow} \mu^2(x) + \sigma_\varepsilon^2(x) \quad \text{as} \quad s \to \infty
\end{aligned} \tag{B.14}$$

Furthermore, we have the following inequality.

$$\mu^2(x) + \sigma_\varepsilon^2(x) \leq \mu^2(x) + \overline{\sigma}_\varepsilon^2 \tag{B.15}$$

Thus, we obtain the desired result. ■

Lemma B.6.
Let $D = \{Z_1, \ldots, Z_s\}$ be a vector of i.i.d. random variables drawn from P. Let $D' = \{Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_s\}$ where $Z'_{c+1}, \ldots, Z'_s$ are i.i.d. draws from P that are independent of D. Furthermore, let

$$\Omega_s^c(x) = \mathbb{E}\left[h_s(x; Z_1, \ldots, Z_c, Z_{c+1}, \ldots, Z_s) \cdot h_s\left(x; Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_s\right)\right]. \tag{B.16}$$

Then,

$$\Omega_s^c(x) \lesssim \mu^2(x) + \overline{\sigma}_\varepsilon^2 + o(1) \quad as \quad s \to \infty. \tag{B.17}$$

Proof of Lemma B.6.

$$
\begin{aligned}
\Omega_s^c(x) &= \mathbb{E}\left[h_s(x; Z_1, \ldots, Z_c, Z_{c+1}, \ldots, Z_s) \cdot h_s\left(x; Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_s\right)\right] \\
&= \mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{s} \kappa(x; Z_i, D) Y_i\right)\left(\sum_{j=1}^{c} \kappa(x; Z_j, D') Y_j + \sum_{j=c+1}^{s} \kappa\left(x; Z'_j, D'\right) Y'_j\right)\right] \\
&= \mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=1}^{c} \kappa(x; Z_i, D)\kappa(x; Z_j, D') Y_i Y_j\right] + \mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=c+1}^{s} \kappa(x; Z_i, D)\kappa\left(x; Z'_j, D'\right) Y_i Y'_j\right] \\
&\quad + \mathbb{E}_{D,D'}\left[\sum_{i=c+1}^{s}\sum_{j=1}^{c} \kappa(x; Z_i, D)\kappa(x; Z_j, D') Y_i Y_j\right] + \mathbb{E}_{D,D'}\left[\sum_{i=c+1}^{s}\sum_{j=c+1}^{s} \kappa(x; Z_i, D)\kappa\left(x; Z'_j, D'\right) Y_i Y'_j\right] \\
&= \underbrace{\mathbb{E}_{D,D'}\left[c\kappa(x; Z_1, D)\kappa(x; Z_1, D') Y_1^2\right]}_{(A)} + 2 \cdot \underbrace{\mathbb{E}_{D,D'}\left[c(s-c)\kappa(x; Z_1, D)\kappa\left(x; Z'_{c+1}, D'\right) Y_1 Y'_{c+1}\right]}_{(B)} \\
&\quad + \underbrace{\mathbb{E}_{D,D'}\left[(s-c)^2\kappa(x; Z_{c+1}, D)\kappa\left(x; Z'_{c+1}, D'\right) Y_{c+1} Y'_{c+1}\right]}_{(C)}
\end{aligned}
\tag{B.18}
$$

Starting from this decomposition, we will analyze the terms one by one using Lemma A.5.

$$
\begin{aligned}
(A) &= \mathbb{E}_{D,D'}\left[c\kappa(x; Z_1, D)\kappa(x; Z_1, D') Y_1^2\right] = \frac{c}{2s-c} \cdot \mathbb{E}_1\left[\left(\mu(X_1) + \varepsilon_1\right)^2 \cdot (2s-c) \cdot \mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\kappa(x; Z_1, D')\right]\right] \\
&= \frac{c}{2s-c} \cdot \mathbb{E}_1\left[\left(\mu^2(X_1) + \sigma^2(X_1)\right) \cdot (2s-c) \cdot \mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\kappa(x; Z_1, D')\right]\right] \\
&\overset{\text{(Lem A.5)}}{\lesssim} \frac{c}{2s-c}\left(\mu^2(x) + \sigma_\varepsilon(x)\right) + o(1)
\end{aligned}
\tag{B.19}
$$

Similarly, we can find the following.

$$
\begin{aligned}
(B) &= \mathbb{E}_{D,D'}\left[c(s-c)\kappa(x; Z_1, D)\kappa\left(x; Z'_{c+1}, D'\right) Y_1 Y'_{c+1}\right] \\
&\overset{\text{Lem A.6}}{\leq} \frac{c(s-c)s}{(2s-c)(2s-c-1)(c+1)} \\
&\qquad \cdot \mathbb{E}_{1,(c+1)'}\left[\left(\mu(X_1) + \varepsilon_1\right) \cdot \left(\mu(X'_{c+1}) + \varepsilon'_{c+1}\right) \cdot \frac{\mathbb{E}_{D,D'}\left[\kappa(x; Z_1, D)\kappa\left(x; Z'_{c+1}, D'\right) \mid Z_1, Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa(x; Z_1, D)\kappa\left(x; Z'_{c+1}, D'\right)\right]}\right] \\
&\leq \frac{(s-c)s}{(2s-c)(2s-c-1)} \cdot \mathbb{E}_{1,(c+1)'}\left[\mu(X_1) \cdot \mu(X'_{c+1}) \cdot \frac{\mathbb{E}_{D,D'}\left[\kappa(x; Z_1, D)\kappa\left(x; Z'_{c+1}, D'\right) \mid Z_1, Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa(x; Z_1, D)\kappa\left(x; Z'_{c+1}, D'\right)\right]}\right] \\
&\overset{\text{Lem A.5}}{\lesssim} \frac{(s-c)s}{(2s-c)(2s-c-1)} \cdot \mu^2(x) + o(1)
\end{aligned}
\tag{B.20}
$$

The third term can be asymptotically bounded in the following way.

$$
\begin{aligned}
(C) &= \mathbb{E}_{D,D'}\left[(s-c)^2\kappa(x; Z_{c+1}, D)\kappa\left(x; Z'_{c+1}, D'\right) Y_{c+1} Y'_{c+1}\right] \\
&\overset{\text{Lem A.6}}{\leq} \frac{2(s-c)^3}{(2s-c)^2(2s-c-1)} \cdot \mathbb{E}_{c+1}\left[\mu(X_{c+1}) \cdot \mu(X'_{c+1}) \cdot \frac{\mathbb{E}_{D,D'}\left[\kappa(x; Z_{c+1}, D)\kappa\left(x; Z'_{c+1}, D'\right) \mid Z_{c+1}, Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa(x; Z_{c+1}, D)\kappa\left(x; Z'_{c+1}, D'\right)\right]}\right] \\
&\overset{\text{Lem A.5}}{\lesssim} \frac{2(s-c)}{(2s-c-1)} \cdot \mu^2(x) + o(1)
\end{aligned}
\tag{B.21}
$$

The result of Lemma B.6 follows immediately by summing up the asymptotic bounds for the individual terms. ■

Lemma B.7.
Let $D = \{Z_1, \ldots, Z_{s_2}\}$ be a vector of i.i.d. random variables drawn from P for $s_2 > s_1$. Furthermore, let

$$\Upsilon_{s_1,s_2}(x) = \mathbb{E}\left[h_{s_1}(x; Z_1, \ldots, Z_{s_1}) \cdot h_{s_2}(x; Z_1, \ldots, Z_{s_1}, \ldots, Z_{s_2})\right]. \tag{B.22}$$

Then,

$$\Upsilon_{s_1,s_2}(x) \lesssim \mu^2(x) + \overline{\sigma}_\varepsilon^2 + o(1) \quad as \quad s_1, s_2 \to \infty \quad with \quad 0 < \mathfrak{c} \leq s_1/s_2 \leq 1 - \mathfrak{c} < 1. \tag{B.23}$$

Proof of Lemma B.7.

$$\begin{aligned}
\Upsilon_{s_1,s_2}(x) &= \mathbb{E}\left[h_{s_1}(x; Z_1, \ldots, Z_{s_1}) \cdot h_{s_2}(x; Z_1, \ldots, Z_{s_1}, \ldots, Z_{s_2})\right] \\
&= \mathbb{E}_D\left[\left(\sum_{i=1}^{s_1} \kappa(x; Z_i, D_{[s_1]}) Y_i\right)\left(\sum_{j=1}^{s_1} \kappa(x; Z_j, D) Y_j + \sum_{j=s_1+1}^{s_2} \kappa(x; Z_j, D) Y_j\right)\right] \\
&= \mathbb{E}_D\left[\sum_{i=1}^{s_1} \kappa(x; Z_i, D) Y_i^2\right] + \mathbb{E}_D\left[\sum_{i=1}^{s_1}\sum_{j=s_1+1}^{s_2} \kappa(x; Z_i, D_{[s_1]})\kappa(x; Z_j, D) Y_i Y_j\right] \\
&= \mathbb{E}_D\left[Y_1^2\, s_1\, \kappa(x; Z_1, D)\right] + \mathbb{E}_D\left[Y_1 Y_{s_2}\, s_1(s_2 - s_1)\, \kappa(x; Z_1, D_{[s_1]})\kappa(x; Z_{s_2}, D)\right] \\
&= \mathbb{E}_D\left[\left(\mu^2(X_1) + \sigma_\varepsilon^2(X_1)\right)\, s_1\, \kappa(x; Z_1, D)\right] + \mathbb{E}_D\left[\mu(X_1)\mu(X_{s_2})\, s_1(s_2 - s_1)\, \kappa(x; Z_1, D_{[s_1]})\kappa(x; Z_{s_2}, D)\right] \\
&= \frac{s_1}{s_2}\mathbb{E}_D\left[\left(\mu^2(X_1) + \sigma_\varepsilon^2(X_1)\right)\, s_1\, \kappa(x; Z_1, D)\right] + \frac{s_2 - s_1}{s_2}\mathbb{E}_D\left[\mu(X_1)\mu(X_{s_2})\, s_1 s_2\, \kappa(x; Z_1, D_{[s_1]})\kappa(x; Z_{s_2}, D)\right] \\
&\leq \frac{s_1}{s_2}\mathbb{E}_D\left[\left(\mu^2(X_1) + \sigma_\varepsilon^2(X_1)\right)\, s_2\, \kappa(x; Z_1, D)\right] \\
&\quad + \frac{s_2 - s_1}{s_2}\mathbb{E}_D\left[|\mu(X_1)|\, s_1\, \kappa(x; Z_1, D_{[s_1]})\right]\mathbb{E}_D\left[|\mu(X_{s_2})|\, s_2\, \kappa(x; Z_{s_2}, D)\right] \\
&\lesssim \mu^2(x) + \sigma_\varepsilon^2(x) + o(1) \leq \mu^2(x) + \overline{\sigma}_\varepsilon^2 + o(1).
\end{aligned} \tag{B.24}$$

■

Lemma B.8.

Let $D = \{Z_1, \ldots, Z_{s_2}\}$ be a vector of i.i.d. random variables drawn from P for $s_2 > s_1$. Let $D' = \{Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_{s_1}\}$ where $Z'_{c+1}, \ldots, Z'_{s_1}$ are i.i.d. draws from P that are independent of D. Furthermore, let

$$\Upsilon^c_{s_1,s_2}(x) = \mathbb{E}\left[h_{s_1}\left(x; Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_{s_1}\right) \cdot h_{s_2}\left(x; Z_1, \ldots, Z_{s_2}\right)\right]. \tag{B.25}$$

Then,

$$\begin{aligned} &\Upsilon^c_{s_1,s_2}(x) \lesssim \frac{cs_2 - c^2 + s_1 s_2}{s_1 s_2}\mu^2(x) + (c/s_1)\overline{\sigma}^2_\varepsilon + o(1) \\ &\text{for} \quad s_1, s_2 \quad \text{sufficiently large} \quad \text{with} \quad 0 < \mathfrak{c} \le s_1/s_2 \le 1 - \mathfrak{c} < 1 \end{aligned} \tag{B.26}$$

and thus

$$\Upsilon^c_{s_1,s_2}(x) \lesssim \mu^2(x) + o(1) \quad \text{as} \quad s_1, s_2 \to \infty \quad \text{with} \quad 0 < \mathfrak{c} \le s_1/s_2 \le 1 - \mathfrak{c} < 1. \tag{B.27}$$

Proof of Lemma B.8.

$$\begin{aligned} \Upsilon^c_{s_1,s_2}(x) &= \mathbb{E}\left[h_{s_1}\left(x; Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_{s_1}\right) \cdot h_{s_2}\left(x; Z_1, \ldots, Z_{s_2}\right)\right] \\ &= \mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{c}\kappa(x; Z_i, D')Y_i + \sum_{i=c+1}^{s_1}\kappa(x; Z'_i, D')Y'_i\right)\left(\sum_{j=1}^{c}\kappa(x; Z_j, D)Y_j + \sum_{j=c+1}^{s_2}\kappa(x; Z_j, D)Y_j\right)\right] \\ &= \underbrace{\mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=1}^{c}\kappa(x; Z_i, D')\kappa(x; Z_j, D)Y_iY_j\right]}_{(A)} + \underbrace{\mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{c}\kappa(x; Z_i, D')Y_i\right)\left(\sum_{j=c+1}^{s_2}\kappa(x; Z_j, D)Y_j\right)\right]}_{(B)} \\ &\quad + \underbrace{\mathbb{E}_{D,D'}\left[\sum_{i=c+1}^{s_1}\sum_{j=1}^{c}\kappa(x; Z'_i, D')\kappa(x; Z_j, D)Y'_iY_j\right]}_{(C)} + \underbrace{\mathbb{E}_{D,D'}\left[\left(\sum_{i=c+1}^{s_1}\kappa(x; Z'_i, D')Y'_i\right)\left(\sum_{j=c+1}^{s_2}\kappa(x; Z_j, D)Y_j\right)\right]}_{(D)} \end{aligned} \tag{B.28}$$

Again, we have four terms to analyze individually.

$$\begin{aligned} (A) &= \mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=1}^{c}\kappa(x; Z_i, D')\kappa(x; Z_j, D)Y_iY_j\right] \\ &= \mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}Y_i^2\kappa(x; Z_i, D')\kappa(x; Z_i, D)\right] \\ &= \mathbb{E}_{D,D'}\left[Y_1^2\, c\,\kappa(x; Z_1, D')\kappa(x; Z_1, D)\right] = \mathbb{E}_{D,D'}\left[\left(\mu^2(X_1) + \sigma^2_\varepsilon(X_1)\right)\, c\,\kappa(x; Z_1, D_{[c]})\kappa(x; Z_1, D'_{c+1:s_1})\right] \\ &= \mathbb{E}_D\left[\left(\mu^2(X_1) + \sigma^2_\varepsilon(X_1)\right)\, c\,\kappa(x; Z_1, D)\right] = \frac{c}{s_1}\mathbb{E}_D\left[\left(\mu^2(X_1) + \sigma^2_\varepsilon(X_1)\right)\, s_1\,\kappa(x; Z_1, D)\right] \\ &\lesssim (c/s_1)(\mu^2(x) + \sigma^2_\varepsilon(x)) + o(1) \le (c/s_1)(\mu^2(x) + \overline{\sigma}^2_\varepsilon) + o(1) \end{aligned} \tag{B.29}$$

Considering the second term, we find the following.

$$
\begin{aligned}
(B) &= \mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{c}\kappa(x;Z_i,D')Y_i\right)\left(\sum_{j=c+1}^{s_2}\kappa(x;Z_j,D)Y_j\right)\right] = \mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=c+1}^{s_2}Y_iY_j\kappa(x;Z_i,D')\kappa(x;Z_j,D)\right] \\
&= \mathbb{E}_{D,D'}\left[c(s_2-c)\,Y_1Y_{s_1}\kappa(x;Z_1,D')\kappa(x;Z_{s_2},D)\right] = \frac{c(s_2-c)}{s_1s_2}\mathbb{E}_{D,D'}\left[Y_1Y_{s_2}\,s_1s_2\,\kappa(x;Z_1,D')\kappa(x;Z_{s_2},D)\right] \\
&\le \frac{c(s_2-c)}{s_1s_2}\mathbb{E}_{D'}\left[|\mu(X_1)|\,s_1\,\kappa(x;Z_1,D')\right]\mathbb{E}_{D}\left[|\mu(X_{s_2})|\,s_2\,\kappa(x;Z_{s_2},D)\right] \\
&\lesssim \frac{c(s_2-c)}{s_1s_2}\mu^2(x)+o(1)
\end{aligned}
\tag{B.30}
$$

Similarly, by simplifying the third term, we find the following.

$$
\begin{aligned}
(C) &= \mathbb{E}_{D,D'}\left[\sum_{i=c+1}^{s_1}\sum_{j=1}^{c}\kappa(x;Z_i',D')\kappa(x;Z_j,D)Y_i'Y_j\right] = \mathbb{E}_{D,D'}\left[Y_{s_1}'Y_1\,(s_1-c)c\,\kappa(x;Z_{s_1}',D')\kappa(x;Z_1,D)\right] \\
&= \frac{(s_1-c)c}{s_1s_2}\mathbb{E}_{D,D'}\left[\mu(X_{s_1}')\mu(X_1)\,s_1s_2\,\kappa(x;Z_{s_1}',D')\kappa(x;Z_1,D)\right] \\
&\le \frac{(s_1-c)c}{s_1s_2}\mathbb{E}_{D}\left[|\mu(X_{s_1}')|\,s_1\,\kappa(x;Z_{s_1}',D')\right]\mathbb{E}_{D}\left[|\mu(X_1)|\,s_2\,\kappa(x;Z_1,D)\right] \\
&\lesssim \frac{(s_1-c)c}{s_1s_2}\mu^2(x)+o(1)
\end{aligned}
\tag{B.31}
$$

Lastly, concerning the fourth term, observe the following.

$$
\begin{aligned}
(D) &= \mathbb{E}_{D,D'}\left[\left(\sum_{i=c+1}^{s_1}\kappa(x;Z_i',D')Y_i'\right)\left(\sum_{j=c+1}^{s_2}\kappa(x;Z_j,D)Y_j\right)\right] = \mathbb{E}_{D,D'}\left[\sum_{i=c+1}^{s_1}\sum_{j=c+1}^{s_2}\kappa(x;Z_i',D')\kappa(x;Z_j,D)Y_i'Y_j\right] \\
&= \mathbb{E}_{D,D'}\left[\mu(X_{s_1}')\mu(X_{s_2})\,(s_1-c)(s_2-c)\,\kappa(x;Z_{s_1}',D')\kappa(x;Z_{s_2},D)\right] \\
&= \frac{(s_1-c)(s_2-c)}{s_1s_2}\mathbb{E}_{D,D'}\left[\mu(X_{s_1}')\mu(X_{s_2})\,s_1s_2\,\kappa(x;Z_{s_1}',D')\kappa(x;Z_{s_2},D)\right] \\
&\le \frac{(s_1-c)(s_2-c)}{s_1s_2}\mathbb{E}_{D'}\left[|\mu(X_{s_1}')|\,s_1\,\kappa(x;Z_{s_1}',D')\right]\mathbb{E}_{D}\left[|\mu(X_{s_2})|\,s_2\,\kappa(x;Z_{s_2},D)\right] \\
&\lesssim \frac{(s_1-c)(s_2-c)}{s_1s_2}\mu^2(x)+o(1)
\end{aligned}
\tag{B.32}
$$

■

Lemma B.9 (Kernel Variance of the TDNN Kernel)**.** *For the kernel of the TDNN estimator with subsampling scales s_1 and s_2, it holds that*

$$\zeta_{s_1,s_2}^{s_2}(x) \lesssim \mu^2(x) + \overline{\sigma}_\varepsilon + o(1) \quad as \quad s_1, s_2 \to \infty \quad with \quad 0 < \mathfrak{c} \leq s_1/s_2 \leq 1 - \mathfrak{c} < 1. \tag{B.33}$$

Proof of Lemma B.9. Consider first the following decomposition.

$$\begin{aligned}
\zeta_{s_1,s_2}^{s_2}(x) &= \mathrm{Var}\left(h_{s_1,s_2}(x; Z_1, \ldots, Z_{s_2})\right) = \mathrm{Var}_D\left(h_{s_1,s_2}(x; D)\right) \\
&\leq \mathbb{E}_D\left[h_{s_1,s_2}^2(x; D)\right] = \mathbb{E}_D\left[\left(w_1^* \tilde{\mu}_{s_1}(x; D) + w_2^* h_{s_2}(x; D)\right)^2\right] \\
&= \left(w_1^*\right)^2 \mathbb{E}_D\left[\tilde{\mu}_{s_1}^2(x; D)\right] + 2 w_1^* w_2^* \mathbb{E}_D\left[\tilde{\mu}_{s_1}(x; D)\, h_{s_2}(x; D)\right] + \left(w_2^*\right)^2 \Omega_{s_2}
\end{aligned} \tag{B.34}$$

Then, observe the following.

$$\begin{aligned}
\mathbb{E}_D\left[\tilde{\mu}_{s_1}^2(x; D)\right] &= \mathbb{E}_D\left[\left(\binom{s_2}{s_1}^{-1} \sum_{\ell \in L_{s_2,s_1}} h_{s_1}(x; D_\ell)\right)^2\right] = \binom{s_2}{s_1}^{-2} \mathbb{E}_D\left[\sum_{\iota,\iota' \in L_{s_2,s_1}} h_{s_1}(x; D_\iota)\, h_{s_1}(x; D_{\iota'})\right] \\
&= \binom{s_2}{s_1}^{-2} \sum_{c=0}^{s_1} \binom{s_2}{s_1} \binom{s_1}{c} \binom{s_2 - s_1}{s_1 - c} \Omega_{s_1}^c = \binom{s_2}{s_1}^{-1} \sum_{c=0}^{s_1} \binom{s_1}{c} \binom{s_2 - s_1}{s_1 - c} \Omega_{s_1}^c \\
&\lesssim \Omega_{s_1} \lesssim \mu(x)^2 + \sigma_\varepsilon^2 + o(1) \quad \text{as} \quad s \to \infty
\end{aligned} \tag{B.35}$$

Recall that by Lemma B.5, we have the following.

$$\Omega_{s_2} \lesssim \mu(x)^2 + \sigma_\varepsilon^2 + o(1) \quad \text{as} \quad s \to \infty \tag{B.36}$$

Lastly, consider the following.

$$\begin{aligned}
\mathbb{E}_D\left[\tilde{\mu}_{s_1}(x; D)\, h_{s_2}(x; D)\right] &= \mathbb{E}_D\left[\binom{s_2}{s_1}^{-1} \sum_{\ell \in L_{s_2,s_1}} h_{s_1}(x; D_\ell)\, h_{s_2}(x; D)\right] \\
&= \mathbb{E}_D\left[h_{s_1}\left(x; D_{[s_1]}\right) h_{s_2}(x; D)\right] = \Upsilon_{s_1,s_2}(x)
\end{aligned} \tag{B.37}$$

Thus, we find the following.

$$\begin{aligned}
\zeta_{s_2,s_2}(x) &\lesssim \left(w_1^*\right)^2 \Omega_{s_1} + 2 w_1^* w_2^* \Upsilon_{s_1,s_2}(x) + \left(w_1^*\right)^2 \Omega_{s_2} \\
&\lesssim \left(w_1^* + w_2^*\right)^2 \left(\mu^2(x) + \sigma_\varepsilon\right) + o(1) = \mu^2(x) + \sigma_\varepsilon + o(1).
\end{aligned} \tag{B.38}$$

■

Lemma B.10 (Lemma 10 - Demirkaya et al. (2024))**.** *For the kernel of the TDNN estimator with subsampling scales s_1 and s_2 satisfying*

$$0 < \mathfrak{c} \leq s_1/s_2 \leq 1 - \mathfrak{c} < 1 \quad \textit{and} \quad s_2 = o(n), \tag{B.39}$$

it holds that

$$\zeta^1_{s_1,s_2}(x) \sim s_2^{-1}. \tag{B.40}$$

B.5 CATE - Kernel Variances & Covariances

Next, we will continue by showing analogous properties in the CATE setting. Similar to before, we will start under the assumption that the functional nuisance parameters are known a priori, to then show that the estimation of said parameters does not impact the asymptotic behavior of the estimator.

Lemma B.11.
Let $D = \{Z_1, \dots, Z_s\}$ be a vector of i.i.d. random variables generated by the setup shown in Assumption 2. Furthermore, let

$$\Omega_s(x) = \mathbb{E}\left[\chi^2_{s,0}(x; Z_1, \dots, Z_s)\right]. \tag{B.41}$$

Then,

$$\Omega_s(x) \lesssim \left(\mu_0^1(x) - \mu_0^0(x)\right)^2 + \frac{\overline{\sigma}^2_\varepsilon}{\mathfrak{p}(1-\mathfrak{p})} + o(1) \tag{B.42}$$

Proof of Lemma B.11.
First, notice that we can decompose the quantity of interest in the following way.

$$\begin{aligned}
\Omega_s(x) &= \mathbb{E}\left[\chi^2_{s,0}(x; Z_1, \dots, Z_s)\right] = \mathbb{E}_D\left[\left(\sum_{i=1}^s \kappa(x; Z_i, D)\, m(Z_i; \eta_0)\right)^2\right] \\
&= \mathbb{E}_D\left[\sum_{i=1}^s\sum_{j=1}^s \kappa(x; Z_i, D)\,\kappa(x; Z_j, D)\, m(Z_i; \eta_0)\, m(Z_j; \eta_0)\right] = \mathbb{E}_D\left[s\kappa(x; Z_1, D)\, m^2(Z_1; \eta_0)\right] \\
&= \mathbb{E}_1\left[m^2(Z_1; \eta_0)\, s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right] \\
&= \mathbb{E}_1\left[\left(\mu_0^1(X_1) - \mu_0^0(X_1) + \beta(W_1, X_1)\,\varepsilon_1\right)^2 s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right] \\
&= \mathbb{E}_1\left[\left(\mu_0^1(X_i) - \mu_0^0(X_1)\right)^2 s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right] + \mathbb{E}_1\left[\left(\beta(W_1, X_1)\,\varepsilon_1\right)^2 s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right] \\
&= \mathbb{E}_1\left[\left(\mu_0^1(X_1) - \mu_0^0(X_1)\right)^2 s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right] + \mathbb{E}_1\left[\left(\frac{W_1}{\pi_0(X_1)} - \frac{1-W_1}{1-\pi_0(X_1)}\right)^2 \varepsilon_1^2 s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right] \\
&= \underbrace{\mathbb{E}_1\left[\left(\mu_0^1(X_1) - \mu_0^0(X_1)\right)^2 s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right]}_{\xrightarrow{\text{Lem A.2}} \left(\mu_0^1(x) - \mu_0^0(x)\right)^2 \quad \text{as} \quad s\to\infty} + \underbrace{\mathbb{E}_1\left[\mathbb{E}\left[\left(\frac{W_1}{\pi_0(X_1)} - \frac{1-W_1}{1-\pi_0(X_1)}\right)^2 \varepsilon_1^2 \,\middle|\, X_1\right] s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right]}_{(B)}
\end{aligned} \tag{B.43}$$

Continuing with the second term, marked by (B), we find the following.

$$\begin{aligned}
(B) &= \mathbb{E}_1\left[\mathbb{E}\left[\left(\frac{W_1}{\pi_0(X_1)} - \frac{1-W_1}{1-\pi_0(X_1)}\right)^2 \varepsilon_1^2 \,\middle|\, X_1\right] s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right] \\
&= \mathbb{E}_1\left[\frac{\sigma^2_\varepsilon(X_1)\cdot s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]}{\pi_0^2(X_1)\left(1-\pi_0(X_1)\right)^2} \cdot \mathbb{E}\left[\left(W_1\left(1-\pi_0(X_1)\right) - (1-W_1)\,\pi_0(X_1)\right)^2 \,\middle|\, X_1\right]\right]
\end{aligned} \tag{B.44}$$

Observe that $W_1(1-W_1)=0$, $W_1^2=W_1$, and $(1-W_1)^2=1-W_1$, which allows us to use the following simplification.

$$\begin{aligned}(B) &= \mathbb{E}_1\left[\frac{\sigma_\varepsilon^2(X_1)\cdot s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]}{\pi_0^2\left(X_1\right)\left(1-\pi_0\left(X_1\right)\right)^2}\cdot\mathbb{E}\left[W_1\left(1-\pi_0\left(X_1\right)\right)^2+\left(1-W_1\right)\pi_0^2\left(X_1\right)\Big|\,X_1\right]\right]\\ &= \mathbb{E}_1\left[\frac{\sigma_\varepsilon^2(X_1)\cdot s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]}{\pi_0^2\left(X_1\right)\left(1-\pi_0\left(X_1\right)\right)^2}\cdot\pi_0(X_1)\left(1-\pi_0\left(X_1\right)\right)\cdot\left(1-\pi_0\left(X_1\right)+\pi_0\left(X_1\right)\right)\right]\\ &= \mathbb{E}_1\left[\frac{\sigma_\varepsilon^2(X_1)}{\pi_0\left(X_1\right)\left(1-\pi_0\left(X_1\right)\right)}\cdot s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right]\overset{\text{(Lem A.2)}}{\longrightarrow}\frac{\sigma_\varepsilon^2(x)}{\pi_0\left(x\right)\left(1-\pi_0\left(x\right)\right)}\quad\text{as}\quad s\to\infty\end{aligned}\tag{B.45}$$

Recombining the terms of interest, we find the desired limit bound.

$$\mathbb{E}_1\left[m^2\left(Z_i;\eta_0\right)s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right]\overset{\text{(Lem A.2)}}{\longrightarrow}\left(\mu_0^1\left(x\right)-\mu_0^0\left(x\right)\right)^2+\frac{\sigma_\varepsilon^2(x)}{\pi_0\left(x\right)\left(1-\pi_0\left(x\right)\right)}\quad\text{as}\quad s\to\infty\tag{B.46}$$

This gives us the desired result.

$$\Omega_s\left(x\right)\lesssim\left(\mu_0^1\left(x\right)-\mu_0^0\left(x\right)\right)^2+\frac{\overline{\sigma}_\varepsilon^2}{\mathfrak{p}\left(1-\mathfrak{p}\right)}+o(1)\tag{B.47}$$

■

Lemma B.12.

Let $D=\{Z_1,\ldots,Z_s\}$ be a vector of i.i.d. random variables drawn from as described in Setup 2.
Let $D'=\{Z_1,\ldots,Z_c,Z'_{c+1},\ldots,Z'_s\}$ where $Z'_{c+1},\ldots,Z'_s$ are i.i.d. draws from the model that are independent of D. Furthermore, let

$$\Omega_s^c\left(x\right)=\mathbb{E}\left[\chi_{s,0}\left(x;Z_1,\ldots,Z_c,Z_{c+1},\ldots,Z_s\right)\cdot\chi_{s,0}\left(x;Z_1,\ldots,Z_c,Z'_{c+1},\ldots,Z'_s\right)\right].\tag{B.48}$$

Then,

$$\Omega_s^c\left(x\right)\lesssim C\left[\left(\mu_0^1\left(x\right)-\mu_0^0\left(x\right)\right)^2+\frac{\overline{\sigma}_\varepsilon^2}{\mathfrak{p}\left(1-\mathfrak{p}\right)}\right]+o(1).\tag{B.49}$$

Proof of Lemma B.12. First, we decompose the term of interest in a similar fashion to before.

$$\begin{aligned}
\Omega_s^c(x) &= \mathbb{E}\left[\chi_{s,0}\left(x; Z_1, \ldots, Z_c, Z_{c+1}, \ldots, Z_s\right) \cdot \chi_{s,0}\left(x; Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_s\right)\right] \\
&= \mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{s} \kappa\left(x; Z_i, D\right) m\left(Z_i; \eta_0\right)\right)\left(\sum_{j=1}^{c} \kappa\left(x; Z_j, D'\right) m\left(Z_j; \eta_0\right) + \sum_{j=c+1}^{s} \kappa\left(x; Z'_j, D'\right) m\left(Z'_j; \eta_0\right)\right)\right] \\
&= \underbrace{\mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{c} \kappa\left(x; Z_i, D\right) m\left(Z_i; \eta_0\right)\right)\left(\sum_{j=1}^{c} \kappa\left(x; Z_j, D'\right) m\left(Z_j; \eta_0\right)\right)\right]}_{(A)} \\
&\quad + \underbrace{2\,\mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{c} \kappa\left(x; Z_i, D\right) m\left(Z_i; \eta_0\right)\right)\left(\sum_{j=c+1}^{s} \kappa\left(x; Z'_j, D'\right) m\left(Z'_j; \eta_0\right)\right)\right]}_{(B)} \\
&\quad + \underbrace{\mathbb{E}_{D,D'}\left[\left(\sum_{i=c+1}^{s} \kappa\left(x; Z_i, D\right) m\left(Z_i; \eta_0\right)\right)\left(\sum_{j=c+1}^{s} \kappa\left(x; Z'_j, D'\right) m\left(Z'_j; \eta_0\right)\right)\right]}_{(C)}
\end{aligned} \tag{B.50}$$

Considering these terms one by one, we can make the following observations. Here, we rely on the same argument structure as in the proof of Lemma B.6 and observations from the proof of Lemma B.11.

$$\begin{aligned}
(A) &= \mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{c} \kappa\left(x; Z_i, D\right) m\left(Z_i; \eta_0\right)\right)\left(\sum_{j=1}^{c} \kappa\left(x; Z_j, D'\right) m\left(Z_j; \eta_0\right)\right)\right] \\
&= \mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=1}^{c} \kappa\left(x; Z_i, D\right) \kappa\left(x; Z_j, D'\right) m\left(Z_i; \eta_0\right) m\left(Z_j; \eta_0\right)\right] \\
&= \frac{c}{2s-c} \cdot \mathbb{E}_1\left[m^2\left(Z_1; \eta_0\right) \cdot (2s-c) \cdot \mathbb{E}_{2:s}\left[\kappa\left(x; Z_1, D\right) \kappa\left(x; Z_1, D'\right)\right]\right] \\
&\overset{\text{Lem A.5}}{\lesssim} \frac{c}{2s-c} \cdot \left(\left(\mu_0^1(x) - \mu_0^0(x)\right)^2 + \frac{\sigma_\varepsilon^2(x)}{\pi_0(x)\left(1-\pi_0(x)\right)}\right) + o(1) \\
&\leq \frac{c}{2s-c} \cdot \left(\left(\mu_0^1(x) - \mu_0^0(x)\right)^2 + \frac{\overline{\sigma}_\varepsilon^2}{\mathfrak{p}\left(1-\mathfrak{p}\right)}\right) + o(1)
\end{aligned} \tag{B.51}$$

Similarly, for the second term, we can make the following observation.

$$
\begin{aligned}
(B) &= \mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{c}\kappa\left(x;Z_i,D\right)m\left(Z_i;\eta_0\right)\right)\left(\sum_{j=c+1}^{s}\kappa\left(x;Z'_j,D'\right)m\left(Z'_j;\eta_0\right)\right)\right] \\
&= \mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=c+1}^{s}\kappa\left(x;Z_i,D\right)\kappa\left(x;Z'_j,D'\right)m\left(Z_i;\eta_0\right)m\left(Z'_j;\eta_0\right)\right] \\
&= \mathbb{E}_{D,D'}\left[c(s-c)\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)m\left(Z_1;\eta_0\right)m\left(Z'_{c+1};\eta_0\right)\right] \\
&\overset{\text{Lem A.6}}{\leq} \frac{c(s-c)s}{(2s-c)(2s-c-1)(c+1)}\cdot\mathbb{E}_{1,(c+1)'}\left[m\left(Z_1;\eta_0\right)m\left(Z'_{c+1};\eta_0\right)\cdot\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_1,Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\right] \\
&\overset{\text{Lem A.5}}{\lesssim} \frac{c(s-c)s}{(2s-c)(2s-c-1)(c+1)}\cdot\left(\mu_0^1\left(x\right)-\mu_0^0\left(x\right)\right)^2+o(1)
\end{aligned}
\tag{B.52}
$$

Finally, for the third term, we can make the following observation.

$$
\begin{aligned}
(C) &= \mathbb{E}_{D,D'}\left[\left(\sum_{i=c+1}^{s}\kappa\left(x;Z_i,D\right)m\left(Z_i;\eta_0\right)\right)\left(\sum_{j=c+1}^{s}\kappa\left(x;Z'_j,D'\right)m\left(Z'_j;\eta_0\right)\right)\right] \\
&= \mathbb{E}_{D,D'}\left[(s-c)^2\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)m\left(Z_{c+1};\eta_0\right)m\left(Z'_{c+1};\eta_0\right)\right] \\
&\overset{\text{Lem A.6}}{\leq} \frac{2(s-c)^3}{(2s-c)^2(2s-c-1)}\cdot\mathbb{E}_{c+1}\left[m\left(Z_{c+1};\eta_0\right)m\left(Z'_{c+1};\eta_0\right)\cdot\frac{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\mid Z_{c+1},Z'_{c+1}\right]}{\mathbb{E}_{D,D'}\left[\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]}\right] \\
&\overset{\text{Lem A.5}}{\lesssim} \frac{2(s-c)}{(2s-c-1)}\cdot\left(\mu_0^1\left(x\right)-\mu_0^0\left(x\right)\right)^2+o(1)
\end{aligned}
\tag{B.53}
$$

Thus, we find the desired result.

$$
\begin{aligned}
\Omega_s^c\left(x\right) &= (A)+2\cdot(B)+(C) \\
&\lesssim \left(\frac{c}{2s-c}+\frac{4(s-c)}{2s-c-1}+\frac{2(s-c)}{2s-c-1}\right)\cdot\left(\mu_0^1\left(x\right)-\mu_0^0\left(x\right)\right)^2+\frac{c}{2s-c}\cdot\frac{\overline{\sigma}_\varepsilon^2}{\mathfrak{p}\left(1-\mathfrak{p}\right)}+o(1) \\
&\leq \frac{6s-5c}{2s-c-1}\cdot\left(\mu_0^1\left(x\right)-\mu_0^0\left(x\right)\right)^2+\frac{c}{2s-c}\cdot\frac{\overline{\sigma}_\varepsilon^2}{\mathfrak{p}\left(1-\mathfrak{p}\right)}+o(1)
\end{aligned}
\tag{B.54}
$$

■

B.6 Asymptotic Normality Results

B.7 NPR-Estimators - Asymptotic Normality

B.8 CATE-Estimators - Asymptotic Normality

Recall the decomposition of the DNN-DML2 estimator introduced in equation 4.23.

$$\begin{aligned}\widehat{\theta}\left(x;\mathbf{D}\right) = &\underbrace{\mathbb{E}_D\left[\widehat{\theta}\left(x;\mathbf{D}\right)\right]}_{\text{Centering-Term}} + \underbrace{\frac{s}{n}\sum_{i=1}^{n}\chi_{s,0}^{(1)}\left(x;Z_i\right)}_{\text{Oracle-Hájek-Projection}} + \underbrace{\frac{s}{k}\sum_{l=1}^{k}\frac{1}{m}\sum_{i\in\mathcal{I}_k}\left(\underbrace{\chi_s^{(1)}\left(x;Z_i,\widehat{\eta}_k\right)-\chi_{s,0}^{(1)}\left(x;Z_i\right)}_{R_{1,k}(x;Z_i)}\right)}_{\text{Oracle-Hájek-Projection Error}} \\ &\underbrace{+\sum_{j=2}^{s}\binom{s}{j}\binom{n}{j}^{-1}\sum_{\ell\in L_{n,j}}\chi_{s,0}^{(j)}\left(x;\mathbf{D}_\ell\right)}_{\text{Oracle-Hájek-Residual}} + \underbrace{\sum_{j=2}^{s}\binom{s}{j}\binom{n}{j}^{-1}\sum_{\ell\in L_{n,j}}R_j\left(x;\mathbf{D}_\ell\right)}_{\text{Higher-Order Error Terms}}\end{aligned} \tag{B.55}$$

where we have the following definition from Equation 4.22.

$$\chi_s^{(c)}\left(x;\mathbf{D}_\ell,\widehat{\eta}\right) = \chi_{s,0}^{(c)}\left(x;\mathbf{D}_\ell\right) + \underbrace{\chi_s^{(c)}\left(x;\mathbf{D}_\ell,\widehat{\eta}\right)-\chi_{s,0}^{(c)}\left(x;\mathbf{D}_\ell\right)}_{R_c(x;\mathbf{D}_\ell)} \tag{B.56}$$

Recall that we are ultimately even more interested in the approximation errors of the following form.

$$\tilde{R}_c\left(x;\mathbf{D}_\ell\right) = R_c\left(x;\mathbf{D}_\ell\right) - (-1)^c\cdot\left(\mathbb{E}_D\left[\chi_s(x;\mathbf{D}_{[s]},\widehat{\eta})-\chi_{s,0}\left(x;\mathbf{D}_{[s]}\right)\right]\right) \tag{B.57}$$

Lemma B.13 (Behavior of Oracle-Hájek-Projection Error)**.**

Proof of Lemma B.13.
Consider first the average Oracle-error within a given fold k and observe the following.

$$\bar{R}_{1,k}\left(x\right) = \frac{1}{m}\sum_{l\in I_k}\tilde{R}_{1,k}\left(x,Z_l\right) = \frac{1}{m}\sum_{l\in I_k}\left(\vartheta_s^1\left(x;Z_l,\widehat{\eta}_k\right)-\vartheta_{s,0}^1\left(x;Z_l\right)\right) \tag{B.58}$$

Define the following empirical process notation, where f is any Q-integrable function on $\mathcal{Z}$.

$$\mathbb{G}_{m,k}\left[f(Z)\right] = \sqrt{\frac{1}{m}}\sum_{i\in I_k}\Big(f(Z_i)-\mathbb{E}_Z\left[f(Z)\right]\Big) \tag{B.59}$$

Here, in analogy to step 3 in the proof of Theorem 3.1 in Victor Chernozhukov, Chetverikov, et al. (2018), we can now observe the following which follows from the triangle inequality.

$$\left|\bar{R}_{1,k}\left(x\right)\right| \leq \sqrt{\frac{1}{m}}\cdot\left(\mathcal{I}_{3,k}^{(1)}+\mathcal{I}_{4,k}^{(1)}\right) \tag{B.60}$$

where

$$\mathcal{I}_{3,k}^{(1)} := \left|\mathbb{G}_{m,k}\left[\vartheta_s^1\left(x;Z,\widehat{\eta}_k\right)\right]-\mathbb{G}_{m,k}\left[\vartheta_{s,0}^1\left(x;Z\right)\right]\right| \tag{B.61}$$

$$\mathcal{I}_{4,k}^{(1)} := \sqrt{m} \cdot \left| \mathbb{E}_Z \left[\vartheta_s^1 \left(x; Z, \widehat{\eta}_k \right) \middle| \mathbf{D}_{I_k^C} \right] - \mathbb{E}_Z \left[\vartheta_{s,0}^1 \left(x; Z \right) \right] \right| \tag{B.62}$$

Notice that conditional on $\mathbf{D}_{I_k^C}$ the first-stage estimate $\widehat{\eta}_k$ is non-stochastic. This allows us to make the following observation given the event $\mathcal{E}_n$ obtains.

$$\textit{LOREMIPSUM} \tag{B.63}$$

LOREM IPSUM ■

Lemma B.14 (Behavior of Higher-Order Error Terms)**.**

Proof of Lemma B.14.

■

B.9 Variance Estimator Consistency Theorems

Lemma B.15 (Asymptotic Dominance of Hájek Projection)**.**
Let $U_s\left(\mathbf{D}_{[n]}\right)$ be a non-randomized complete generalized U-statistic with kernel h_s. Let the kernel variance terms ζ_s^s and ζ_s^1 be defined in analogy to Section 3. Assume that the following condition holds.

$$\frac{s}{n}\left(\frac{\zeta_s^s}{s\zeta_s^1}-1\right)\to 0 \tag{B.64}$$

Then, asymptotically, the Hájek projection term dominates the variance of the U-statistic in the following sense.

$$\frac{n}{s^2}\frac{Var\left(\mathrm{U}_s\left(\mathbf{D}_{[n]}\right)\right)}{\zeta_s^1}\to 1. \tag{B.65}$$

Proof.

$$\begin{aligned}
1\leq \frac{n}{s^2}\frac{\mathrm{Var}\left(\mathrm{U}_s\left(\mathbf{D}_{[n]}\right)\right)}{\zeta_s^1} &= \left(\frac{s^2}{n}\zeta_s^1\right)^{-1}\sum_{j=1}^{s}\binom{s}{j}^2\binom{n}{j}^{-1}V_s^j \\
&\leq 1+\left(\frac{s^2}{n}\zeta_s^1\right)^{-1}\frac{s^2}{n^2}\sum_{j=2}^{s}\binom{s}{j}V_s^j \\
&\leq 1+\frac{s}{n}\left(\frac{\zeta_s^s}{s\zeta_s^1}-1\right)\to 1.
\end{aligned} \tag{B.66}$$

■

Lemma B.16 (Hájek Dominance for TDNN Estimator)**.**
*Let $0<\mathfrak{c}\leq s_1/s_2\leq 1-\mathfrak{c}<1$ and $s_2=o(n)$, then under Assumptions **??**, **??** and **??**, then the TDNN estimator fulfills the asymptotic Hájek dominance condition shown in Lemma B.15.*

Proof. Recall the results from Lemmas B.9 and B.10.

$$\zeta_{s_1,s_2}^{s_2}(x)\lesssim \mu^2(x)+\sigma_\varepsilon+o(1)\quad\text{and}\quad \zeta_{s_1,s_2}^{1}(x)\sim s_2^{-1}$$

Using these results, we can find the following.

$$\frac{s_2}{n}\left(\frac{\zeta_{s_1,s_2}^{s_2}(x)}{s_2\zeta_{s_1,s_2}^{1}(x)}-1\right)\sim\frac{s_2}{n}\left(\mu^2(x)+\sigma_\varepsilon+o(1)-1\right)\sim\frac{s_2}{n}\to 0 \tag{B.67}$$

■

Proof of Theorem 5.2.
The desired result immediately follows from an application of Theorem 6 from Peng, Mentch, and Stefanski (2021). ■

Proof of Theorem 5.3.
Recall the definition of the Jackknife Variance estimator.

$$\widehat{\omega}^2_{JK}\left(x;\mathbf{D}_n\right) = \frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\mu}_{s_1,s_2}\left(x;\mathbf{D}_{n,-i}\right) - \widehat{\mu}_{s_1,s_2}\left(x;\mathbf{D}_n\right)\right)^2 \tag{B.68}$$

Using the Hoeffding-decomposition of the original U-statistic, we can reformulate this expression in the following way.

$$\begin{aligned}
\widehat{\omega}^2_{JK}\left(x;\mathbf{D}_n\right) &= \frac{n-1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{s_2}\binom{s_2}{j}H^j_{s_1,s_2}\left(\mathbf{D}_{n,-i}\right) - \sum_{j=1}^{s_2}\binom{s_2}{j}H^j_{s_1,s_2}\right)^2 \\
&= \frac{n-1}{n}\sum_{j=1}^{n}\left(\sum_{j=1}^{s_2}\binom{s_2}{j}\left(H^j_{s_1,s_2} - H^j_{s_1,s_2}\left(\mathbf{D}_{n,-i}\right)\right)\right)^2 \\
&= \frac{n-1}{n}\sum_{j=1}^{n}\left(\sum_{j=1}^{s_2}\binom{s_2}{j}\left(\binom{n}{j}^{-1}\sum_{\iota\in L_{n,j}}h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota) - \binom{n-1}{j}^{-1}\sum_{\ell\in L_j([n]\setminus\{i\})}h^{(j)}_{s_1,s_2}(\mathbf{D}_\ell)\right)\right)^2 \\
&= \frac{n-1}{n}\sum_{j=1}^{n}\left[\frac{s_2}{n}h^{(1)}_{s_1,s_2}(Z_i) + \sum_{j\neq i}\left(\frac{s_2}{n} - \frac{s_2}{n-1}\right)h^{(1)}_{s_1,s_2}(Z_j)\right. \\
&\quad \left. + \sum_{j=2}^{s_2}\binom{s_2}{j}\left(\binom{n}{j}^{-1}\sum_{\iota\in L_{n,j}}h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota) - \binom{n-1}{j}^{-1}\sum_{\ell\in L_j([n]\setminus\{i\})}h^{(j)}_{s_1,s_2}(\mathbf{D}_\ell)\right)\right]^2 \\
&= \frac{n-1}{n}\frac{s^2}{n^2}\sum_{j=1}^{n}\left[h^{(1)}_{s_1,s_2}(Z_i) - \frac{1}{n-1}\sum_{j\neq i}h^{(1)}_{s_1,s_2}(Z_j)\right. \\
&\quad \left. + \frac{n}{s}\sum_{j=2}^{s_2}\binom{s_2}{j}\left(\binom{n}{j}^{-1}\sum_{\iota\in L_{j-1}([n]\setminus\{i\})}h^{(j)}_{s_1,s_2}(\mathbf{D}_{\iota\cup\{i\}}) + \left[\binom{n}{j}^{-1} - \binom{n-1}{j}^{-1}\right]\sum_{\ell\in L_j([n]\setminus\{i\})}h^{(j)}_{s_1,s_2}(\mathbf{D}_\ell)\right)\right] \\
&=: \frac{n-1}{n}\frac{s^2}{n^2}\sum_{j=1}^{n}\left[h^{(1)}_{s_1,s_2}(Z_i) + T_i\right]^2
\end{aligned} \tag{B.69}$$

Observe that due to the independence of the observations and the uncorrelatedness of Hoeffding projections of differing orders, $h^{(1)}_{s_1,s_2}(Z_i)$ and T_i are uncorrelated and both have mean zero. Now, continuing to follow the line of argument in Peng, Mentch, and Stefanski (2021), observe the following.

$$\mathbb{E}\left[\left(h^{(1)}_{s_1,s_2}(Z_i)\right)^2\right] = V^1_{s_1,s_2} = \zeta^1_{s_1,s_2} \tag{B.70}$$

Furthermore, as a consequence of the independence of the observations and the uncorrelatedness of Hoeffding projec-

tions of differing order, we find that

$$
\begin{aligned}
\mathbb{E}\left[T_i^2\right] &= \frac{1}{n-1}V^1_{s_1,s_2} + \frac{n^2}{s_2^2}\sum_{j=2}^{s_2}\binom{s_2}{j}^2\left\{\binom{n}{j}^{-2}\binom{n-1}{j-1}V^j_{s_1,s_2} + \left[\binom{n}{j}^{-1} - \binom{n-1}{j}^{-1}\right]^2\binom{n-1}{j}V^j_{s_1,s_2}\right\} \\
&= \frac{1}{n-1}V^1_{s_1,s_2} + \frac{n^2}{s_2^2}\sum_{j=2}^{s_2}\binom{s_2}{j}^2\left\{\binom{n}{j}^{-2}\frac{j}{n-j}\binom{n-1}{j}V^j_{s_1,s_2} + \binom{n}{j}^{-2}\left[1-\binom{n}{j}\binom{n-1}{j}^{-1}\right]^2\binom{n-1}{j}V^j_{s_1,s_2}\right\} \\
&= \frac{1}{n-1}V^1_{s_1,s_2} + \frac{n^2}{s_2^2}\sum_{j=2}^{s_2}\binom{s_2}{j}^2\binom{n}{j}^{-2}\left(\frac{j}{n-j} + \left(1-\frac{n}{n-j}\right)^2\right)\binom{n-1}{j}V^j_{s_1,s_2} \\
&= \frac{1}{n-1}V^1_{s_1,s_2} + \frac{n^2}{s_2^2}\sum_{j=2}^{s_2}\binom{s_2}{j}\binom{n}{j}^{-2}\binom{n-1}{j}\cdot\left(\frac{j}{n-j} + \frac{j^2}{(n-j)^2}\right)\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&= \frac{1}{n-1}V^1_{s_1,s_2} + \frac{n^2}{s_2^2}\sum_{j=2}^{s_2}\binom{s_2}{j}\binom{n}{j}^{-1}\frac{n-j}{n}\cdot\frac{j}{n}\left(\frac{n}{n-j} + \frac{jn}{(n-j)^2}\right)\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&= \frac{1}{n-1}V^1_{s_1,s_2} + \sum_{j=2}^{s_2}\frac{j}{s_2}\binom{s_2-1}{j-1}\binom{n-1}{j-1}^{-1}\frac{n-j}{n}\left(\frac{n}{n-j} + \frac{j}{n}\right)\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&\leq \frac{1}{n-1}V^1_{s_1,s_2} + \sum_{j=2}^{s_2}\frac{j}{s_2}\left(e\frac{s_2-1}{n-1}\right)^{j-1}\frac{n-j}{n}\left(\frac{n}{n-j} + \frac{j}{n}\right)\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&\lesssim \frac{1}{n-1}V^1_{s_1,s_2} + 2\sum_{j=2}^{s_2}\frac{j}{s_2}\left(e\frac{s_2-1}{n-1}\right)^{j-1}\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&\leq \frac{1}{n-1}V^1_{s_1,s_2} + 2e\sum_{j=2}^{s_2}\frac{1}{s_2}\frac{s_2-1}{n-1}\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] + 2\sum_{j=2}^{s_2}\frac{j-1}{s_2}\left(e\frac{s_2-1}{n-1}\right)^{j-1}\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&\leq \frac{1}{n-1}V^1_{s_1,s_2} + \frac{2e}{n-1}\sum_{j=2}^{s_2}\frac{s_2-1}{s_2}\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] + 2\sum_{j=2}^{s_2}\frac{j-1}{s_2}\left(e\frac{s_2-1}{n-1}\right)^{j-1}\zeta^{s_2}_{s_1,s_2} \\
&= \frac{1}{n-1}V^1_{s_1,s_2} + \frac{2e}{n}\sum_{j=2}^{s_2}\frac{n(s_2-1)}{(n-1)s_2}\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] + 2\zeta^{s_2}_{s_1,s_2}\sum_{j=2}^{s_2}\frac{j-1}{s_2}\left(e\frac{s_2-1}{n-1}\right)^{j-1} \\
&\leq \frac{1}{n-1}V^1_{s_1,s_2} + \frac{2e}{n}\sum_{j=2}^{s_2}\binom{s_2}{j}V^j_{s_1,s_2} + \frac{2\zeta^{s_2}_{s_1,s_2}}{s_2}\sum_{j=1}^{\infty}j\left(e\frac{s_2-1}{n-1}\right)^j \\
&\leq \frac{1}{n-1}V^1_{s_1,s_2} + \frac{2e}{n}\sum_{j=2}^{s_2}\binom{s_2}{j}V^j_{s_1,s_2} + \frac{2\zeta^{s_2}_{s_1,s_2}}{s_2}\sum_{j=1}^{\infty}j\left(e\frac{s_2}{n}\right)^j \\
&= \frac{1}{n-1}\zeta^1_{s_1,s_2} + \frac{2e}{n}\left(\zeta^{s_2}_{s_1,s_2} - s_2\zeta^1_{s_1,s_2}\right) + \frac{2en}{\left(n-es_2\right)^2}\zeta^{s_2}_{s_1,s_2} \\
&= \left(\frac{1}{n-1} + \frac{2es_2n}{\left(n-es_2\right)^2}\right)\zeta^1_{s_1,s_2} + 2e\left(\frac{1}{n} + \frac{n}{\left(n-es_2\right)^2}\right)\left(\zeta^{s_2}_{s_1,s_2} - s_2\zeta^1_{s_1,s_2}\right)
\end{aligned}
\tag{B.71}
$$

Recall the results of Lemmas B.9 and B.10.

$$
\zeta^{s_2}_{s_1,s_2}(x) \lesssim \mu^2(x) + \sigma_\varepsilon + o(1) \quad \text{and} \quad \zeta^1_{s_1,s_2}(x) \sim s_2^{-1} \tag{B.72}
$$

This immediately implies that $\frac{s_2}{n}\left(\frac{\zeta^{s_2}_{s_1,s_2}}{s_2\zeta^1_{s_1,s_2}} - 1\right) \to 0$. Using this result and the previous asymptotic upper bound, we

can find the following.

$$\begin{aligned}\frac{\mathbb{E}\left[T_i^2\right]}{V_{s_1,s_2}^1} &\leq \frac{\left(\frac{1}{n-1}+\frac{2es_2n}{(n-es_2)^2}\right)\zeta_{s_1,s_2}^1 + 2e\left(\frac{1}{n}+\frac{n}{(n-es_2)^2}\right)\left(\zeta_{s_1,s_2}^{s_2}-s_2\zeta_{s_1,s_2}^1\right)}{\zeta_{s_1,s_2}^1} \\ &= \frac{1}{n-1}+\frac{2es_2n}{(n-es_2)^2}+2e\left(\frac{1}{n}+\frac{n}{(n-es_2)^2}\right)\left(\frac{\zeta_{s_1,s_2}^{s_2}-s_2\zeta_{s_1,s_2}^1}{\zeta_{s_1,s_2}^1}\right)\to 0\end{aligned} \tag{B.73}$$

Therefore, we can conclude that $h_s^{(1)}\left(Z_i\right)$ dominates T_i^2 in the expression of interest. Using Lemma A.7, we can thus conclude the following.

$$\begin{aligned}\frac{\frac{n}{s_2^2}\widehat{\omega}_{JK}^2\left(x;\mathbf{D}_n\right)}{V_{s_1,s_2}^1\left(x\right)} &\to_p \frac{n-1}{n}\frac{1}{n}\sum_{i=1}^n \frac{\left(h_{s_1,s_2}^{(1)}(x;Z_i)\right)^2}{V_{s_1,s_2}^1\left(x\right)} \\ &\to_p \frac{n-1}{n}\frac{\mathbb{E}\left[\left(h_{s_1,s_2}^{(1)}(x;Z_i)\right)^2\right]}{V_{s_1,s_2}^1\left(x\right)}\to 1\end{aligned} \tag{B.74}$$

The desired rate-consistency then immediately follows from an application of Lemma B.15. ■

Proof of Theorem 5.4.
Consider first the case absent additional randomization in the form of ω and recall the definition of the delete-d Jackknife Variance estimator.

$$\widehat{\omega}^2_{JKD}\left(x; d, \mathbf{D}_n\right) = \frac{n-d}{d}\binom{n}{d}^{-1} \sum_{\ell \in L_{n,d}} \left(\widehat{\mu}_{s_1,s_2}\left(x; \mathbf{D}_{n,-\ell}\right) - \widehat{\mu}_{s_1,s_2}\left(x; \mathbf{D}_n\right)\right)^2 \tag{B.75}$$

Now, as in the proof for the conventional Jackknife variance estimator, we make use of the Hoeffding-decomposition in the following way.

$$\begin{aligned}
\widehat{\omega}^2_{JKD}\left(x; d, \mathbf{D}_n\right) &= \frac{n-d}{d}\binom{n}{d}^{-1} \sum_{\ell \in L_{n,d}} \left(\sum_{j=1}^{s_2} \binom{s_2}{j} \left(H^j_{P_t} - H^j_{P_t}\left(\mathbf{D}_{n,-\ell}\right)\right)\right)^2 \\
&= \frac{n-d}{d}\binom{n}{d}^{-1} \sum_{\ell \in L_{n,d}} \left(\sum_{j=1}^{s_2} \binom{s_2}{j} \left(\binom{n}{j}^{-1} \sum_{\iota \in L_{n,j}} h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota) - \binom{n-d}{j}^{-1} \sum_{\iota \in L_j([n]\setminus\ell)} h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota)\right)\right)^2 \\
&= \frac{n-d}{d}\binom{n}{d}^{-1} \sum_{\ell \in L_{n,d}} \left[\frac{s_2}{n}\sum_{i\in\ell} h^{(1)}_{s_1,s_2}(Z_i) + \sum_{i\in[n]\setminus\ell}\left(\frac{s_2}{n} - \frac{s_2}{n-d}\right) h^{(1)}_{s_1,s_2}(Z_i)\right. \\
&\quad \left. + \sum_{j=2}^{s_2} \binom{s_2}{j} \left(\binom{n}{j}^{-1} \sum_{\iota \in L_{n,j}} h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota) - \binom{n-d}{j}^{-1} \sum_{\iota \in L_j([n]\setminus\ell)} h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota)\right)\right]^2 \\
&= \frac{n-d}{d}\binom{n}{d}^{-1} \left(\frac{s_2}{n}\right)^2 \sum_{\ell \in L_{n,d}} \left[\sum_{i\in\ell} h^{(1)}_{s_1,s_2}(Z_i) - \frac{d}{n-d}\sum_{i\in[n]\setminus\ell} h^{(1)}_{s_1,s_2}(Z_i)\right. \\
&\quad \left. + \frac{n}{s_2}\sum_{j=2}^{s_2} \binom{s_2}{j} \left(\binom{n}{j}^{-1} \sum_{\iota \in L_{n,j}} h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota) - \binom{n-d}{j}^{-1} \sum_{\iota \in L_j([n]\setminus\ell)} h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota)\right)\right]^2 \\
&=: (n-d)\binom{n}{d}^{-1} \left(\frac{s_2}{n}\right)^2 \sum_{\ell \in L_{n,d}} \left[\frac{1}{\sqrt{d}}\sum_{i\in\ell} h^{(1)}_{s_1,s_2}(Z_i) + T_\ell\right]^2
\end{aligned} \tag{B.76}$$

We want to proceed in an analogous way to the proof of the pure Jackknife result. Thus, we want to show that $\sum_{i\in\ell} h^{(1)}_{s_1,s_2}(Z_i)$ dominates T_ℓ in the sense of Lemma A.7. Luckily, since Lemma A.7 does not depend on any particular independence assumptions of summands etc. this is a relatively straightforward adaptation of the strategy shown in the proof of Theorem 5.3. Thus, consider the following for an arbitrary fixed index-subset ℓ with cardinality d.

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{\sqrt{d}}\sum_{i\in\ell} h^{(1)}_{s_1,s_2}(Z_i)\right)^2\right] &= \frac{1}{d}\mathbb{E}\left[\sum_{i\in\ell}\sum_{j\in\ell} h^{(1)}_{s_1,s_2}(Z_i) h^{(1)}_{s_1,s_2}(Z_j)\right] = \frac{1}{d}\sum_{i\in\ell}\sum_{j\in\ell}\mathbb{E}\left[h^{(1)}_{s_1,s_2}(Z_i) h^{(1)}_{s_1,s_2}(Z_j)\right] \\
&= \frac{|\ell|}{d}\cdot\mathbb{E}\left[\left(h^{(1)}_{s_1,s_2}(Z_1)\right)^2\right] = \zeta_{P_t,1}
\end{aligned} \tag{B.77}$$

For the error term we introduce a case distinction. Case one corresponds to parameter choices where $s_2 \geq d$ and thus takes the following form.

$$\begin{aligned}
T_\ell = \frac{\sqrt{d}}{n-d} \sum_{i \in [n] \setminus \ell} h^{(1)}_{s_1,s_2}(Z_i) \\
+ \frac{n}{s_2\sqrt{d}} \Bigg\{ \sum_{j=2}^{d} \binom{s_2}{j} \Bigg(\binom{n}{j}^{-1} \Bigg(\sum_{a=1}^{j} \sum_{\substack{\varkappa \in L_a(\ell) \\ \varrho \in L_{j-a}([n]\setminus\ell)}} h^{(j)}_{s_1,s_2}(D_{\varkappa \cup \varrho}) \Bigg) + \Bigg(\binom{n}{j}^{-1} - \binom{n-d}{j}^{-1} \Bigg) \sum_{\iota \in L_j([n]\setminus\ell)} h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota) \Bigg) \\
+ \sum_{j=d+1}^{s_2} \binom{s_2}{j} \Bigg(\binom{n}{j}^{-1} \Bigg(\sum_{a=1}^{d} \sum_{\substack{\varkappa \in L_a(\ell) \\ \varrho \in L_{j-a}([n]\setminus\ell)}} h^{(j)}_{s_1,s_2}(D_{\varkappa \cup \varrho}) \Bigg) + \Bigg(\binom{n}{j}^{-1} - \binom{n-d}{j}^{-1} \Bigg) \sum_{\iota \in L_j([n]\setminus\ell)} h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota) \Bigg) \Bigg\}
\end{aligned} \tag{B.78}$$

Case two covers setups of the form $s_2 < d$ and thus takes the following form.

$$\begin{aligned}
T_\ell = \frac{\sqrt{d}}{n-d} \sum_{i \in [n] \setminus \ell} h^{(1)}_{s_1,s_2}(Z_i) \\
+ \frac{n}{s_2\sqrt{d}} \sum_{j=2}^{s_2} \binom{s_2}{j} \Bigg(\binom{n}{j}^{-1} \Bigg(\sum_{a=1}^{j} \sum_{\substack{\varkappa \in L_a(\ell) \\ \varrho \in L_{j-a}([n]\setminus\ell)}} h^{(j)}_{s_1,s_2}(D_{\varkappa \cup \varrho}) \Bigg) + \Bigg(\binom{n}{j}^{-1} - \binom{n-d}{j}^{-1} \Bigg) \sum_{\iota \in L_j([n]\setminus\ell)} h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota) \Bigg)
\end{aligned} \tag{B.79}$$

Having separated these two cases, we continue by investigating the expectation of their respective squares. Beginning with case one, we find the following.

$$\begin{aligned}
\mathbb{E}\left[(T_\ell)^2\right] &= \frac{d}{n-d} V^1_{s_1,s_2} \\
&\quad + \frac{n^2}{s_2^2 d} \sum_{j=2}^{d} \binom{s_2}{j}^2 \Bigg(\binom{n}{j}^{-2} \sum_{a=1}^{j} \left[\binom{d}{a} \binom{n-d}{j-a} \right] + \left[\binom{n}{j}^{-1} - \binom{n-d}{j}^{-1} \right]^2 \binom{n-d}{j} \Bigg) V^j_{s_1,s_2} \\
&\quad + \frac{n^2}{s_2^2 d} \sum_{j=d+1}^{s_2} \binom{s_2}{j}^2 \Bigg(\binom{n}{j}^{-2} \sum_{a=1}^{d} \left[\binom{d}{a} \binom{n-d}{j-a} \right] + \left[\binom{n}{j}^{-1} - \binom{n-d}{j}^{-1} \right]^2 \binom{n-d}{j} \Bigg) V^j_{s_1,s_2} \\
&\overset{(\star)}{=} \frac{d}{n-d} V^1_{s_1,s_2} \\
&\quad + \frac{n^2}{s_2^2 d} \sum_{j=2}^{d} \binom{s_2}{j}^2 \binom{n}{j}^{-2} \Bigg(\binom{n}{j} - \binom{n-d}{j} + \left[1 - \binom{n}{j} \binom{n-d}{j}^{-1} \right]^2 \binom{n-d}{j} \Bigg) V^j_{s_1,s_2} \\
&\quad + \frac{n}{s_2 d} \sum_{j=d+1}^{s_2} \frac{\binom{s_2-1}{j-1}}{\binom{n-1}{j-1}} \frac{\binom{n-d}{j}}{\binom{n}{j}} \Bigg(\sum_{a=1}^{d} \frac{\binom{d}{a}\binom{n-d}{j-a}}{\binom{n-d}{j}} + \left[1 - \binom{n}{j} \binom{n-d}{j}^{-1} \right]^2 \Bigg) \binom{s_2}{j} V^j_{s_1,s_2}
\end{aligned} \tag{B.80}$$

The equality marked by $(\star)$ holds by the Chu-Vandermonde identity - specifically with respect to the equivalent expression for the sum in the second term.

Continuing the analysis, we find the following.

$$
\begin{aligned}
\mathbb{E}\left[(T_\ell)^2\right] &= \frac{d}{n-d}V^1_{s_1,s_2} \\
&\quad + \frac{n}{s_2 d}\sum_{j=2}^{d}\frac{\binom{s_2-1}{j-1}\binom{n-d}{j}}{\binom{n-1}{j-1}\binom{n}{j}}\left(\binom{n}{j}\binom{n-d}{j}^{-1} - 1 + \left[1-\binom{n}{j}\binom{n-d}{j}^{-1}\right]^2\right)\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&\quad + \frac{n}{s_2 d}\sum_{j=d+1}^{s_2}\frac{\binom{s_2-1}{j-1}\binom{n-d}{j}}{\binom{n-1}{j-1}\binom{n}{j}}\left(\frac{\binom{n}{j}}{\binom{n-d}{j}}\sum_{a=1}^{d}\frac{\binom{d}{a}\binom{n-d}{j-a}}{\binom{n}{j}} + \left[1-\binom{n}{j}\binom{n-d}{j}^{-1}\right]^2\right)\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&= \frac{d}{n-d}V^1_{s_1,s_2} \\
&\quad + \frac{n}{s_2 d}\sum_{j=2}^{d}\frac{\binom{s_2-1}{j-1}\binom{n-d}{j}}{\binom{n-1}{j-1}\binom{n}{j}}\left(\binom{n}{j}^2\binom{n-d}{j}^{-2} - \binom{n}{j}\binom{n-d}{j}^{-1}\right)\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&\quad + \frac{n}{s_2 d}\sum_{j=d+1}^{s_2}\frac{\binom{s_2-1}{j-1}\binom{n-d}{j}}{\binom{n-1}{j-1}\binom{n}{j}}\left(\frac{\binom{n}{j}}{\binom{n-d}{j}\binom{n}{d}}\sum_{a=1}^{d}\binom{j}{a}\binom{n-j}{d-a} + \left[1-\binom{n}{j}\binom{n-d}{j}^{-1}\right]^2\right)\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&\overset{(\star\star)}{=} \frac{d}{n-d}V^1_{s_1,s_2} \\
&\quad + \frac{n}{s_2 d}\sum_{j=2}^{d}\frac{\binom{s_2-1}{j-1}}{\binom{n-1}{j-1}}\left(\binom{n}{j}\binom{n-d}{j}^{-1} - 1\right)\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&\quad + \frac{n}{s_2 d}\sum_{j=d+1}^{s_2}\frac{\binom{s_2-1}{j-1}\binom{n-d}{j}}{\binom{n-1}{j-1}\binom{n}{j}}\left(\frac{\binom{n}{j}}{\binom{n-d}{j}}\left[1-\binom{n-j}{d}\binom{n}{d}^{-1}\right] + \left[1-\binom{n}{j}\binom{n-d}{j}^{-1}\right]^2\right)\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&= \frac{d}{n-d}V^1_{s_1,s_2} + \frac{n}{s_2}\sum_{j=2}^{d}\frac{\binom{s_2-1}{j-1}}{\binom{n-1}{j-1}}\left(\binom{n}{j}\binom{n-d}{j}^{-1} - 1\right)\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&\quad + \frac{n}{s_2 d}\sum_{j=d+1}^{s_2}\frac{\binom{s_2-1}{j-1}\binom{n-d}{j}}{\binom{n-1}{j-1}\binom{n}{j}}\left(\frac{\binom{n}{j}}{\binom{n-d}{j}} - 1 + \left[1-\binom{n}{j}\binom{n-d}{j}^{-1}\right]^2\right)\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&= \frac{d}{n-d}V^1_{s_1,s_2} + \frac{n}{s_2 d}\sum_{j=2}^{s_2}\frac{\binom{s_2-1}{j-1}}{\binom{n-1}{j-1}}\left(\binom{n}{j}\binom{n-d}{j}^{-1} - 1\right)\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&= \frac{d}{n-d}V^1_{s_1,s_2} + \frac{n}{s_2 d}\sum_{j=2}^{s_2}\frac{\binom{s_2-1}{j-1}}{\binom{n-1}{j-1}}\left(\prod_{i=0}^{d-1}\left(1+\frac{j}{n-i-j}\right) - 1\right)\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&\overset{(\star\star\star)}{\leq} \frac{d}{n-d}V^1_{s_1,s_2} + \frac{n}{s_2 d}\sum_{j=2}^{s_2}\frac{\binom{s_2-1}{j-1}}{\binom{n-1}{j-1}}\frac{\sum_{i=0}^{d-1}\frac{j}{n-i-j}}{1-\sum_{i=0}^{d-1}\frac{j}{n-i-j}}\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&\leq \frac{d}{n-d}V^1_{s_1,s_2} + \frac{n}{s_2 d}\sum_{j=2}^{s_2}\frac{\binom{s_2-1}{j-1}}{\binom{n-1}{j-1}}\frac{j(n-j)}{(n-d-j+1)(n-d-2j)}\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&\leq \frac{d}{n-d}V^1_{s_1,s_2} + \frac{n}{s_2 d}\sum_{j=2}^{s_2}\left(\frac{e(s_2-1)}{n-1}\right)^{j-1}\frac{j(n-2)}{(n-d-s_2+1)(n-d-2s_t)}\left[\binom{s_2}{j}V^j_{s_1,s_2}\right]
\end{aligned}
\tag{B.81}
$$

The equality marked by $(\star\star)$ holds by the Chu-Vandermonde identity applied to the third summand, whereas the inequality marked by the equality marked by $(\star\star\star)$ follows from a Weierstrass-Product type inequality. Furthermore, this derivation shows that we do not really need to distinguish between the two described cases for the error term.

Proceeding this way allows us to continue our analysis similar to the proof for the simple leave-one-out Jackknife.

$$
\begin{aligned}
\mathbb{E}\left[(T_\ell)^2\right] &\lesssim \frac{d}{n-d} V^1_{s_1,s_2} + \frac{n}{s_2 d} \sum_{j=2}^{s_2} \left(\frac{e(s_2-1)}{n-1}\right)^{j-1} \frac{j(n-2)}{(n-d-s_2+1)(n-d-2s_t)} \left[\binom{s_2}{j} V^j_{s_1,s_2}\right] \\
&= \frac{d}{n-d} V^1_{s_1,s_2} + \frac{2e \cdot n(n-2)}{(n-1)(n-d-s_2+1)(n-d-2s_t)d} \sum_{j=2}^{s_2} \frac{j}{s_2} \left(\frac{e(s_2-1)}{n-1}\right)^{j-1} \left[\binom{s_2}{j} V^j_{s_1,s_2}\right] \\
&\lesssim \frac{d}{n-d} V^1_{s_1,s_2} + \frac{4e}{(n-d-s_2)s_2 d} \sum_{j=2}^{s_2} j \left(\frac{e(s_2-1)}{n-1}\right)^{j-1} \left[\binom{s_2}{j} V^j_{s_1,s_2}\right] \\
&\leq \frac{d}{n-d} V^1_{s_1,s_2} + \frac{4e}{(n-d-s_2)s_2 d} \sum_{j=2}^{s_2} \left[\binom{s_2}{j} V^j_{s_1,s_2}\right] + \frac{4e}{(n-d-s_2)s_2 d} \sum_{j=2}^{s_2} (j-1) \left(\frac{e(s_2-1)}{n-1}\right)^{j-1} \left[\binom{s_2}{j} V^j_{s_1,s_2}\right] \\
&\leq \frac{d}{n-d} V^1_{s_1,s_2} + \frac{4e}{(n-d-s_2)s_2 d} \sum_{j=2}^{s_2} \left[\binom{s_2}{j} V^j_{s_1,s_2}\right] + \frac{4e \cdot \zeta^{s_2}_{s_1,s_2}}{(n-d-s_2)s_2 d} \sum_{j=1}^{\infty} j \left(\frac{e(s_2-1)}{n-1}\right)^{j} \\
&= \frac{d}{n-d} \zeta^1_{s_1,s_2} + \frac{4e}{(n-d-s_2)s_2 d} \left(\zeta^{s_2}_{s_1,s_2} - s_2 \zeta^1_{s_1,s_2}\right) + \frac{4e \cdot \zeta^{s_2}_{s_1,s_2}}{(n-d-s_2)s_2 d} \cdot \frac{e(s_2-1)(n-1)}{\left(n-1-e(s_2-1)\right)^2} \\
&= \left(\frac{d}{n-d} + \frac{e(s_2-1)(n-1)}{\left(n-1-e(s_2-1)\right)^2}\right) \zeta^1_{s_1,s_2} + \frac{4e}{(n-d-s_2)s_2 d} \left(1 + \frac{e(s_2-1)(n-1)}{\left(n-1-e(s_2-1)\right)^2}\right) \left(\zeta^{s_2}_{s_1,s_2} - s_2 \zeta^1_{s_1,s_2}\right)
\end{aligned} \tag{B.82}
$$

We continue as in the default Jackknife case.

$$
\begin{aligned}
\frac{\mathbb{E}\left[T^2_\ell\right]}{V^1_{s_1,s_2}} &\leq \frac{d}{n-d} + \frac{e(s_2-1)(n-1)}{\left(n-1-e(s_2-1)\right)^2} + \frac{4e}{(n-d-s_2)s_2 d} \left(1 + \frac{e(s_2-1)(n-1)}{\left(n-1-e(s_2-1)\right)^2}\right) \frac{\zeta^{s_2}_{s_1,s_2} - s_2 \zeta^1_{s_1,s_2}}{\zeta^1_{s_1,s_2}} \\
&\to 0.
\end{aligned} \tag{B.83}
$$

Now, following the exact same logic as in the proof for the consistency of the Jackknife variance estimator, we obtain consistency of the delete-d Jackknife variance estimator. ■

B.10 Pointwise Inference Results

Proof of Theorem 5.7.

■

Proof of Theorem 5.8.

■