Simultaneous Inference for Conditional Average Treatment Effects using Distributional Nearest Neighbor Estimation

Jakob R. Juergens
University of Wisconsin - Madison

Last edited: November 21, 2024

Abstract

This paper presents a computationally simple method of estimating heterogeneous treatment effects based on the Two-Scale Distributional Nearest Neighbor (TDNN) estimator of Demirkaya et al. (2024). As part of this analysis, I improve on conditions required for consistent variance estimation presented in the original paper and provide results for asymptotically valid pointwise inference in a nonparametric regression setup and extend the analysis to the estimation of conditional average treatment effects. Building on the framework of Ritzwoller and Syrgkanis (2024), I develop uniformly valid confidence bands for the TDNN estimator. I then show how to apply these to perform uniformly valid inference in both the nonparametric regression setup and the heterogeneous treatment effect setup. A main contribution is the development of a computationally simple method that leverages the theoretical results of the aforementioned papers.

Supplementary Material and R Package available at: Work in Progress

1 Introduction

LOREM IPSUM

After a short notation section, the remainder of this paper is organized as follows. LOREM IPSUM

1.1 Notation

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

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

For a data set $D_{[n]} = (Z_1, \ldots, Z_n)$ and a vector $\ell \in L_{n,s}$, denote by $D_{[n],-\ell}$ 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 $D_{n,-i}$. Similarly, given such a data set $D_{[n]}$ and index vector ℓ, denote by D_ℓ the data set only consisting of the observations in $D_{[n]}$ corresponding to 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 convergence in distribution, while $\rightarrow_p$ denotes convergence in probability and $\rightarrow_{a.s.}$ denotes almost sure convergence. The norm $\|\cdot\|_{\psi_1}$ denotes the ψ_1-Orlicz norm. Recall that random variables are sub-exponential if and only if they have a finite ψ_1-Orlicz norm.

TO-DO:

- $\lesssim$ needs an explanation

2 Setup

Throughout this paper, we will consider two distinct setups. The first is a pure nonparametric regression setup closely mirroring the structure of Demirkaya et al. (2024). This setup will be very useful to illustrate the inner workings of the estimator of interest and serve as a leading example for the theoretical results.

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 $Y \in \mathbb{R}$ 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 we impose the following conditions.

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

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

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

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

$$\begin{aligned} &\mathbf{D}_n = \{Z_i = (X_i, W_i, Y_i)\}_{i=1}^n \quad \text{from the model} \quad Y = \mathbb{1}(W=0)\mu_0(X) + \mathbb{1}(W=1)\mu_1(X) + \varepsilon, \\ &W_i \sim \text{Bern}(\pi(X_i)) \end{aligned} \tag{2.3}$$

where $Y \in \mathbb{R}$ 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 associated probability measure φ on $\mathcal{X}$ and ε is the unobservable model error on which we impose the following conditions.

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

Furthermore, $\mu_0 : \mathcal{X} \to \mathbb{R}$ and $\mu_1 : \mathcal{X} \to \mathbb{R}$ are the two unknown potential outcome functions and $\pi : \mathcal{X} \to [0,1]$ is a function describing the probability of treatment uptake, effectively corresponding to the propensity score. Let the distribution induced by this model be denoted by Q and thus $Z_i = (X_i, W_i, Y_i) \overset{iid}{\sim} Q$.

In this second setting, I will use the notation $\mathbf{D}^{(0)}$ and $\mathbf{D}^{(1)}$ to refer to the data subsets containing 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 manner.

Remark 1 (Potential Applications).
From an microeconometric perspective, these two setups cover a wide array of applications. While nonparametric regression is itself often advantageous to answer economic questions, the real strengths show when considering the second setup. LOREM IPSUM

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 (Assumption 1 and Assumption 2) the following conditions hold:

- *The feature space $\mathcal{X} = \operatorname{supp}(X)$ is a bounded, compact subset of $\mathbb{R}^k$*
- *The density $f(\cdot)$ is bounded away from 0 and ∞*
- *$f(\cdot)$ are $\mu(\cdot)$ are four times continuously differentiable with bounded second, third, and fourth-order partial derivatives in a neighborhood of x*

In the Heterogeneous Treatment Effect setting (Assumption 2), the following additional condition holds:

- *$\mu_0(\cdot)$ and $\mu_1(\cdot)$ are four times continuously differentiable with bounded second, third, and fourth-order partial derivatives in a neighborhood of x*

There is potential to relax these assumptions at the cost of requiring both less interpretable conditions and more technically sophisticated proofs. Additionally, we 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, have continuously varying variance. In other terms, $\sigma_\varepsilon^2 : \mathcal{X} \to \mathbb{R}_{>0}$ is a continuous function.

As $\mathcal{X}$ is a bounded and compact set, this implies that there exists a $\overline{\sigma}_\varepsilon^2 > 0$ such that for any $x \in \mathcal{X}$ we have $\sigma_\varepsilon^2(x) \leq \overline{\sigma}_\varepsilon^2$. Additionally, due to the assumptions on the regression functions, this ensures the existence of seconds moments of Y in both scenarios. Furthermore, to assure that there is a sufficient number of treated and untreated observations local to each point of interest asymptotically, we 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), we assume that there exist a constant $\mathfrak{p} \in (0, 1/2)$ such that

$$\forall x \in \mathcal{X} : \quad 0 < \mathfrak{p} \leq \pi(x) \leq 1 - \mathfrak{p} < 1. \tag{2.5}$$

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

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. LOREM IPSUM

3 Two-Scale Distributional Nearest Neighbor Estimator

While less economically enticing, I will introduce the TDNN estimator using the simple nonparametric regression setup first. I 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 commence by adapting it to tackle the problem of estimating heterogeneous treatment effects. As I will embed both estimation problems in the context of subsampled conditional moment regression to then build uniform inference procedures based on Ritzwoller and Syrgkanis (2024), the approach might at first seem unnatural. However, due to the constructions that follow in Section 5, this approach will be well worth the slightly cumbersome initial presentation.

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. I 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 approached 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 and Guyader (2010). Given a sample as described in Assumption 1 and a fixed feature vector x, we 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 we 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.

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

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 holds. Then, for any fixed $x \in \mathcal{X}$, we have that 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 setup, Demirkaya et al. (2024) develop a novel bias-correction method for the DNN estimator that leads to appealing finite-sample properties of the resulting Two-Scale Distributional Nearest Neighbor (TDNN) estimator. Their method is based on 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

$$W_1^*(s_1, s_2) = \frac{1}{1 - (s_1/s_2)^{-2/d}} \quad \text{and} \quad w_2^*(s_1, s_2) = 1 - W_1^*(s_1, s_2) \tag{3.12}$$

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 the 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 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 $x \in \operatorname{supp}(X) \subset \mathbb{R}^d$, 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

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

3.2 DNN and TDNN in Heterogeneous Treatment Effect Estimation

Motivated by the nonparametric regression setup, we set out to apply the underlying idea in the context of heterogeneous treatment effects. Similar to before, we 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 Double/Debiased Machine Learning (DML). 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, $\text{CATE}(x) = \mathbb{E}\left[Y_i\left(W_i = 1\right) - Y_i\left(W_i = 0\right) \mid X_i = x\right]$, we will employ a Neyman-orthogonal score function to curtail the influence of the nuisance parameters on our estimation.

$$\begin{aligned} M(x; \text{CATE}, \mu, p) = \mathbb{E}\left[m\left(Z_i; \text{CATE}, \mu, \pi\right) \mid X_i = x\right] = 0 \quad \text{where} \\ m\left(Z_i; \text{CATE}, \mu, \pi\right) = \mu_1\left(X_i\right) - \mu_0\left(X_i\right) + \beta\left(W_i, X_i\right)\left(Y_i - \mu_{W_i}\left(X_i\right)\right) - \text{CATE}\left(X_i\right) \end{aligned} \tag{3.16}$$

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

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

As a shorthand notation, we will furthermore use $m\left(Z_i; \mu, \pi\right) = m\left(Z_i; \text{CATE}, \mu, \pi\right) + \text{CATE}\left(X_i\right)$. 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.

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

However, due to the presence of infinite-dimensional nuisance parameters, it becomes attractive to proceed by using this weighted empirical moment equation embedded into the DML2 estimator of Chernozhukov et al. (2018). Applying these ideas to the context of estimating the CATE has been previously explored, for example by Semenova and 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. *(T)DNN-DML2 CATE-Estimator*
To estimate the Conditional Average Treatment Effect at a point of interest $x \in \mathcal{X}$, we 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.*
2. *For each $k \in [K]$, use the DNN estimator on the data set $\mathbf{D}_{I_k^C} \ldots$*
   (a) *to estimate the nuisance parameters μ_0 and μ_1:*
   $$\widehat{\mu}_{k,s}^{w}(x) = \widehat{\mu}_{w,s}\left(x; \mathbf{D}_{I_k^C}^{(w)}\right) \quad \textit{for } w = 0, 1 \tag{3.19}$$
   (b) *if π is unknown, i.e. we are not in a randomized experiment setting, additionally estimate π*
   $$\widehat{\pi}_{k,s}(x) = \widehat{\mu}_s\left(x; \mathbf{D}_{I_k^C}\right) \quad \textit{where the predicted variable is } W \tag{3.20}$$
3. *Construct the estimator $\widehat{\text{CATE}}(x)$ as the solution to the following equation.*
$$\begin{aligned} 0 &= \sum_{k=1}^{K} \sum_{i \in I_k} K(x, X_i) m\left(Z_i; \widehat{\text{CATE}}(x), \widehat{\mu}_{k,s}, \widehat{\pi}_{k,s}\right) \\ &= \sum_{i=1}^{n-s+1} \left[\frac{\binom{n-i}{s-1}}{\binom{n}{s}} \sum_{k=1}^{K} \mathbb{1}(i \in I_k)\, m\left(Z_{(i)}; \widehat{\text{CATE}}(x), \widehat{\mu}_{k,s}, \widehat{\pi}_{k,s}\right)\right] \\ &= \sum_{i=1}^{n-s+1} \left[\frac{\binom{n-i}{s-1}}{\binom{n}{s}} m\left(Z_{(i)}; \widehat{\text{CATE}}(x), \widehat{\mu}_{k(i),s}, \widehat{\pi}_{k(i),s}\right)\right] \end{aligned} \tag{3.21}$$

This description shows the case of the DNN estimator. The corresponding TDNN-based estimator is defined analogously, employing the TDNN estimator in the first-stage estimation procedure and using the corresponding weights of the TDNN-estimator in the second stage. Observe, that the weights $K(x, X_i)$ chosen in the second step are chosen according to the full sample. Plugging in for the score function in the equation that defines the estimator, we can observe the following.

$$\widehat{\text{CATE}}(x) = \sum_{i=1}^{n-s+1} \frac{\binom{n-i}{s-1}}{\binom{n}{s}} \left[\widehat{\mu}_{k(i),s}^{1}\left(X_{(i)}\right) - \widehat{\mu}_{k(i),s}^{0}\left(X_{(i)}\right) + \widehat{\beta}_{k(i),s}\left(W_{(i)}, X_{(i)}\right)\left(Y_{(i)} - \widehat{\mu}_{k(i),s}^{W_{(i)}}\left(X_{(i)}\right)\right)\right] \tag{3.22}$$

Thus, given first-stage estimates of the nuisance parameters, we 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 reweighting of the terms corresponding to different observations. Considering the first stage estimates, we can recognize that the estimation of μ^0 and μ^1 is effectively a nonparametric regression problem as previously described where we used the reduced data sets $\mathbf{D}^{(0)}$ and $\mathbf{D}^{(1)}$, respectively. In contrast, the estimation of π that is necessary in nearly all contexts but randomized experiments can be described further due to the binary outcome. For that purpose, let $Z_{(i|k)}$ denote the i'th closest observation in fold k akin to the construction shown in Equation 3.3 relative to x but with respect to the data in fold k.

$$\widehat{\pi}_{k,s}(x) = \sum_{i=1}^{n-m-s+1} \frac{\binom{n-m-i}{s-1}}{\binom{n-m}{s}} W_{(i|k)} \tag{3.23}$$

What these equations show is that the main computational cost associated with these methods comes from having to construct multiple orderings of the sample of interest. The essential strength of this approach: It is not necessary to solve any complex optimization problems to obtain the estimator. Furthermore, due to the prevalence of constructing orderings of data with respect to the euclidean norm, this is a well-studied problem with efficient algorithms available of the shelf.

- Approach: approximation of a complete U-statistic
- Idea: Dominance of Hájek Projection gives normality
- For each observation, the corresponding estimated nuisance parameters are asymptotically "close enough" compared to if we consider the Hájek projection of the fully constructed U-statistic

LOREM IPSUM

4 Pointwise Inference for the TDNN Estimator

To perform inference in the regression setup, Demirkaya et al. (2024) introduce variance estimators based on the Jackknife and Bootstrap. However, as they point out, their consistency results rely on a likely suboptimal rate condition for the subsampling scale. While Theorem 3.2 allows s_2 to be of the order $o(n)$, the variance estimators rely on the considerably stronger condition that $s_2 = o(n^{1/3})$. Establishing 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.

4.1 Jackknife Variance Estimators for Nonparametric Regression

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

$$\omega^2(x) = \mathrm{Var}_D\left(\widehat{\mu}_{s_1,s_2}(x;\mathbf{D}_n)\right) \tag{4.1}$$

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.

$$\widehat{\omega}^2_{JK}(x;\mathbf{D}_n) = \frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\mu}_{s_1,s_2}(x;\mathbf{D}_{n,-i}) - \widehat{\mu}_{s_1,s_2}(x;\mathbf{D}_n)\right)^2 \tag{4.2}$$

Theorem 4.1 (Closed Form Expression for the Jackknife-Variance Estimator)**.**
The Jackknife variance estimator for the DNN estimator has the following convenient closed-form representations.

$$\mathit{LOREMIPSUM} \tag{4.3}$$

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

$$\mathit{LOREMIPSUM} \tag{4.4}$$

As a generalization to the Jackknife, we can also consider the delete-d Jackknife that builds on the same working principle. Instead of removing one observation at a time, we remove d observations and average over all possible d-subsets removals. This leads to the following representation.

$$\widehat{\omega}^2_{JKD}(x;d,\mathbf{D}_n) = \frac{n-d}{d}\binom{n}{d}^{-1}\sum_{\ell\in L_{n,d}}\left(\widehat{\mu}_{s_1,s_2}(x;\mathbf{D}_{n,-\ell}) - \widehat{\mu}_{s_1,s_2}(x;\mathbf{D}_n)\right)^2 \tag{4.5}$$

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, due to the unwieldiness of the closed form, I refrain from deriving it.

In this section, we 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.

$$\widehat{\omega}_{PI}^2\left(x;\mathbf{D}_n\right)=\frac{s_2^2}{n^2}\sum_{i=1}^{n}\left[\left(\binom{n-1}{s-1}^{-1}\sum_{\ell\in L_{s_2-1}([n]\backslash\{i\})}h_{s_1,s_2}\left(x;D_{\ell\cup\{i\}}\right)\right)-\widehat{\mu}_{s_1,s_2}\left(x;\mathbf{D}_n\right)\right]^2 \tag{4.6}$$

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, we obtain the following result.

Theorem 4.2 (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

$$\frac{\widehat{\omega}_{PI}^2\left(x;\mathbf{D}_n\right)}{\omega^2\left(x;\mathbf{D}_n\right)}\longrightarrow_p 1. \tag{4.7}$$

In an analogous fashion to Theorems 5 and 6 from Demirkaya et al. (2024), we furthermore obtain the following consistency results for the presented variance estimators. 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), we then obtain the following results on the Jackknife and Bootstrap variance estimators respectively. As part of the proof of these results, we obtain general results on the consistency of Jackknife and Bootstrap variance estimators for infinite-order U-statistics beyond the TDNN estimator.

Theorem 4.3 (Jackknife Variance Estimator Consistency).
Let $0<\mathfrak{c}\leq s_1/s_2\leq 1-\mathfrak{c}<1$ and $s_2=o(n)$, then

$$\frac{\widehat{\omega}_{JK}^2\left(x;\mathbf{D}_n\right)}{\omega^2\left(x;\mathbf{D}_n\right)}\longrightarrow_p 1. \tag{4.8}$$

Theorem 4.4 (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

$$\frac{\widehat{\omega}_{JKD}^2\left(x;d,\mathbf{D}_n\right)}{\omega^2\left(x;\mathbf{D}_n\right)}\longrightarrow_p 1. \tag{4.9}$$

4.2 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, we need to introduce some additional notation, where, for simplicity, we 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.

$$|H_k| = \frac{|I_k|}{2} = m/2 \quad \textit{and} \quad H_k \subset I_k \tag{4.10}$$

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.

$$R^*_{n,s}\left(x; \mathbf{D}_{[n]}, \mathcal{I}\right) = \overline{\mathrm{CATE}}_{\mathcal{H}}(x) - \widehat{\mathrm{CATE}}(x) \tag{4.11}$$

Here, $\overline{\mathrm{CATE}}_{\mathcal{H}}(x)$ 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, we 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.

$$\begin{aligned} 0 &= \sum_{k=1}^{K} \sum_{i \in H_k} K(x, X_i \,|\, \mathcal{H}) m\left(Z_i; \overline{CATE}_{\mathcal{H}}(x), \widehat{\mu}_{k,s}, \widehat{\pi}_{k,s}\right) \\ &= \sum_{i=1}^{n/2-s+1} \left[\frac{\binom{n/2-i}{s-1}}{\binom{n/2}{s}} m\left(Z_{(i\,|\,\mathcal{H})}; \overline{\mathrm{CATE}}_{\mathcal{H}}(x), \widehat{\mu}_{k(i\,|\,\mathcal{H}),s}, \widehat{\pi}_{k(i\,|\,\mathcal{H}),s}\right) \right] \end{aligned} \tag{4.12}$$

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

$$\begin{aligned} \overline{\mathrm{CATE}}_{\mathcal{H}}(x) = \sum_{i=1}^{n/2-s+1} \frac{\binom{n/2-i}{s-1}}{\binom{n/2}{s}} &\Big[\widehat{\mu}^1_{k(i\,|\,\mathcal{H}),s}\left(X_{(i\,|\,\mathcal{H})}\right) - \widehat{\mu}^0_{k(i\,|\,\mathcal{H}),s}\left(X_{(i\,|\,\mathcal{H})}\right) \\ &+ \widehat{\beta}_{k(i\,|\,\mathcal{H}),s}\left(W_{(i\,|\,\mathcal{H})}, X_{(i\,|\,\mathcal{H})}\right)\left(Y_{(i\,|\,\mathcal{H})} - \mu_{W_{(i\,|\,\mathcal{H})}}\left(X_{(i\,|\,\mathcal{H})}\right)\right)\Big] \end{aligned} \tag{4.13}$$

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

$$R^*_{n,s}\left(x; \mathbf{D}_{[n]}, \mathcal{I}\right) = LOREMIPSUM \tag{4.14}$$

Theorem 4.5 (Consistent Variance Estimation for the (T)DNN-DML2 CATE Estimator).
LOREM IPSUM

4.3 Pointwise Inference with the TDNN Estimator

Theorem 4.6 (Pointwise Inference in Nonparametric Regression).
LOREM IPSUM

Theorem 4.7 (Pointwise Inference in Heterogeneous Treatment Effect Estimation).
LOREM IPSUM

5 Uniform Inference for the TDNN Estimator

Noteworthy properties of κ are its permutational symmetry in D_ℓ and that κ does not consider the response variable when assigning weights to the observations under consideration. The latter immediately implies a property that has been called "Honesty" by Wager and Athey (2018).

Definition 3 (Symmetry and Honesty - Adapted from Ritzwoller and Syrgkanis (2024))**.**

1. *The kernel $\kappa(\cdot,\cdot,D_\ell)$ is Honest in the sense that*

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

where $\perp\!\!\!\perp$ denotes conditional independence.

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

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

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

Definition 4 (Uniform Confidence Regions)**.**
A confidence region for the TDNN (or DNN) estimators that is uniformly valid at the rate $r_{n,d}$ is a family of random intervals

$$\widehat{\mathcal{C}}\left(x^{(p)}\right) := \left\{\widehat{C}(x_j^{(p)}) = \left[c_L(x_j^{(p)}), c_U(x_j^{(p)})\right] \,:\, j \in [p]\right\} \tag{5.1}$$

based on the observed data, such that

$$\sup_{P\in\mathbf{P}}\left|P\left(\mu(x^{(d)}) \in \widehat{\mathcal{C}}\left(x^{(d)}\right)\right)\right| \leq r_{n,d} \tag{5.2}$$

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

5.1 Low-Level

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

Theorem 5.1 (Ritzwoller and Syrgkanis (2024) - Theorem 4.1).
For any sequence of kernel orders $b = b_n$, where

$$\frac{1}{n} \frac{\nu_j^2}{\sigma_{b,j}^2} \to 0 \quad as \quad n \to \infty, \tag{5.3}$$

we have that

$$\sqrt{\frac{n}{\sigma_{b,j}^2 b^2}} \binom{n}{b}^{-1} \sum_{\mathbf{s} \in \mathbf{S}_{n,b}} u\left(x_j^{(p)}; D_{\mathbf{s}}\right) \rightsquigarrow \mathcal{N}(0,1), \quad as \quad n \to \infty. \tag{5.4}$$

Theorem 5.2 (Adapted from Ritzwoller and Syrgkanis (2024) - Theorem 4.2).
Define the terms

$$\bar{\psi}_{s_2}^2 = \max_{j \in [p]} \left\{ \nu_j^2 - s_2 \sigma_{s_2,j}^2 \right\} \quad and \quad \underline{\sigma}_{s_2}^2 = \min_{j \in [p]} \sigma_{s_2,j}^2. \tag{5.5}$$

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

$$\left\| h_{s_1,s_2}(x; D_\ell) \right\|_{\psi_1} \leq \phi \tag{5.6}$$

for each j in $[d]$, then

$$\sqrt{\frac{n}{s_2^2 \underline{\sigma}_{s_2}^2}} \left\| \widehat{\mu}_{s_1,s_2}(x^{(p)}; \mathbf{D}_n) - \mu(x^{(p)}) - \frac{s_2}{n} \sum_{i=1}^{n} h_{s_1,s_2}^{(1)}(x^{(p)}; \mathbf{z}_i) \right\|_\infty = \sqrt{\frac{n}{s_2^2 \underline{\sigma}_{s_2}^2}} \left\| \mathrm{HR}_{s_1,s_2}(x^{(p)}; \mathbf{D}_n) \right\|_\infty \lesssim \xi_{n,s_2}, \tag{5.7}$$

where

$$\xi_{n,s_2} = \left(\frac{C s_2 \log(pn)}{n} \right)^{s_2/2} \left(\left(\frac{n \bar{\psi}_{s_2}^2}{{s_2}^2 \underline{\sigma}_{s_2}^2} \right)^{1/2} + \left(\frac{\phi^2 s_2 \log^4(pn)}{\underline{\sigma}_{s_2}^2} \right)^{1/2} \right), \tag{5.8}$$

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

5.2 High-Level

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

Definition 5 (Half-Sample Bootstrap Root Approximation - Ritzwoller and Syrgkanis (2024)).
The Half-Sample Bootstrap Root Approximation of the sampling distribution of the root

$$R\left(x^{(p)};\mathbf{D}_n\right) := \widehat{\mu}\left(x^{(p)};\mathbf{D}_n\right) - \mu(x^{(p)}) \tag{5.9}$$

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

$$R^*\left(x^{(p)};\mathbf{D}_n\right) := \widehat{\mu}\left(x^{(p)};D_l\right) - \widehat{\mu}\left(x^{(p)};\mathbf{D}_n\right) \tag{5.10}$$

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

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

$$\widehat{\lambda}_j^2\left(x^{(p)};\mathbf{D}_n\right) = \mathrm{Var}\left(\sqrt{n}R^*(x_j^{(p)};\mathbf{D}_n)\,|\,\mathbf{D}_n\right) \quad \text{and} \quad \widehat{\Lambda}_n\left(x^{(p)};\mathbf{D}_n\right) = \mathrm{diag}\left(\left\{\widehat{\lambda}_j^2\left(x^{(p)};\mathbf{D}_n\right)\right\}_{j=1}^p\right) \tag{5.11}$$

$$\widehat{S}^*\left(x^{(p)};\mathbf{D}_n\right) := \sqrt{n}\left\|\left(\widehat{\Lambda}_n\left(x^{(p)};\mathbf{D}_n\right)\right)^{-1/2} R^*\left(x^{(p)};\mathbf{D}_n\right)\right\|_2 \tag{5.12}$$

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

Theorem 5.3 (Uniform Confidence Region - Ritzwoller and Syrgkanis (2024)).
Define the intervals

$$\widehat{\mathcal{C}}\left(x_j^{(p)};\mathbf{D}_n\right) := \widehat{\mu}\left(x_j^{(p)};\mathbf{D}_n\right) \pm n^{-1/2}\widehat{\lambda}_j\left(x^{(p)};\mathbf{D}_n\right)\mathrm{cv}\,(\alpha;\mathbf{D}_n) \tag{5.13}$$

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

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

Definition 6 (Shrinkage and Incrementality - Adapted from Ritzwoller and Syrgkanis (2024)).
We say that the kernel $\kappa\left(\cdot,\cdot,D_\ell\right)$ has a uniform shrinkage rate ϵ_b if

$$\sup_{P\in\mathbf{P}}\sup_{j\in[p]}\mathbb{E}\left[\max\left\{\left\|X_i - x_j^{(p)}\right\|_2 : \kappa\left(x_j^{(p)},X_i,D_\ell\right) > 0\right\}\right] \le \epsilon_b. \tag{5.14}$$

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

$$\inf_{P\in\mathbf{P}}\sup_{j\in[p]} Var\left(\mathbb{E}\left[\sum_{i\in\ell}\kappa\left(x_j^{(p)},X_i,D_\ell\right)m\left(Z_i;\mu\right)\mid l\in\ell, Z_l = Z\right]\right) \gtrsim b^{-1} \tag{5.15}$$

where Z is an independent random variable with distribution P.

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

$$\sup_{P \in \mathbf{P}} \sup_{j \in [p]} \mathbb{E}\left[\max\left\{\left\|X_i - x_j^{(p)}\right\|_2 : \mathrm{rk}(x_j^{(p)}; X_i, D_\ell) = 1\right\}\right] \leq \epsilon_b \tag{5.16}$$

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

$$\begin{aligned}
&\inf_{P \in \mathbf{P}} \sup_{j \in [p]} \mathrm{Var}\left(\mathbb{E}\left[\sum_{i \in \ell} \mathbb{1}\left(\mathrm{rk}(x_j^{(p)}; X_i, D_\ell) = 1\right)(Y_i - \mu(X_i)) \mid l \in \ell, Z_l = Z\right]\right) \\
&= \inf_{P \in \mathbf{P}} \sup_{j \in [p]} \mathrm{Var}\left(\sum_{i \in \ell} \mathbb{E}\left[\mathbb{1}\left(\mathrm{rk}(x_j^{(p)}; X_i, D_\ell) = 1\right)\varepsilon_i \mid l \in \ell, Z_l = Z\right]\right) \\
&= \inf_{P \in \mathbf{P}} \sup_{j \in [p]} \mathrm{Var}\left(\sum_{i=1}^{s} \mathbb{E}\left[\mathbb{1}\left(\mathrm{rk}(x_j^{(p)}; X_i, D_{1:s}) = 1\right)\varepsilon_i \mid l \in [s], Z_l = Z\right]\right) \\
&= \inf_{P \in \mathbf{P}} \sup_{j \in [p]} s^2 \cdot \mathrm{Var}\left(\mathbb{E}\left[\mathbb{1}\left(\mathrm{rk}(x_j^{(p)}; X_1, D_{1:s}) = 1\right)\varepsilon_1 \mid l \in [s], Z_l = Z\right]\right) \gtrsim b^{-1}
\end{aligned} \tag{5.17}$$

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

LOREM IPSUM

Assumption 8 (Boundedness - Adapted from Ritzwoller and Syrgkanis (2024)).
The absolute value of the function $m(\cdot;\mu)$ is bounded by the constant $(\theta+1)\phi$ almost surely.

$$|m(Z_i;\mu)| = |Y_i - \mu(X_i)| = |\varepsilon_i| \leq (\theta+1)\phi \quad a.s. \tag{5.18}$$

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

Assumption 9 (Sub-Exponential Norm Bound).

LOREM IPSUM

Assumption 10 (Moment Smoothness - Adapted from Ritzwoller and Syrgkanis (2024)).
Define the moments

$$M^{(1)}(x;\mu) = \mathbb{E}\left[m^{(1)}\left(Z_i;\mu\right) \mid X_i = x\right] \quad \textit{and} \quad M^{(2)}(x) = \mathbb{E}\left[m^{(2)}\left(Z_i\right) \mid X_i = x\right], \tag{5.19}$$

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

$$M^{(1)}(x;\mu) = -\mu(x) \quad \textit{and} \quad M^{(2)}(x) = \mu(x). \tag{5.20}$$

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

$$\forall x, x' \in \operatorname{supp}(X): \quad \sup_{P \in \mathbf{P}} \left|\mu(x) - \mu\left(x'\right)\right| \lesssim \left\|x - x'\right\|_2. \tag{5.21}$$

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

$$\inf_{P \in \mathbf{P}} \inf_{j \in [p]} \left|M^{(1)}\left(x_j^{(p)}\right)\right| = \inf_{P \in \mathbf{P}} \inf_{j \in [p]} \left|\mu\left(x_j^{(p)}\right)\right| \geq c \tag{5.22}$$

for some positive constant c.

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

5.3 Uniform Inference with the TDNN Estimator

Theorem 5.4 (Uniform Inference in Nonparametric Regression).
LOREM IPSUM

Theorem 5.5 (Uniform Inference in Heterogeneous Treatment Effect Estimation).
LOREM IPSUM

6 Simulations

Having developed theoretical results concerning uniform inference methods for the TDNN estimator, we 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, we consider a collection of setups. First, we focus on illustrating the bias correcting 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 obtainable by combining multiple subsampling scales.

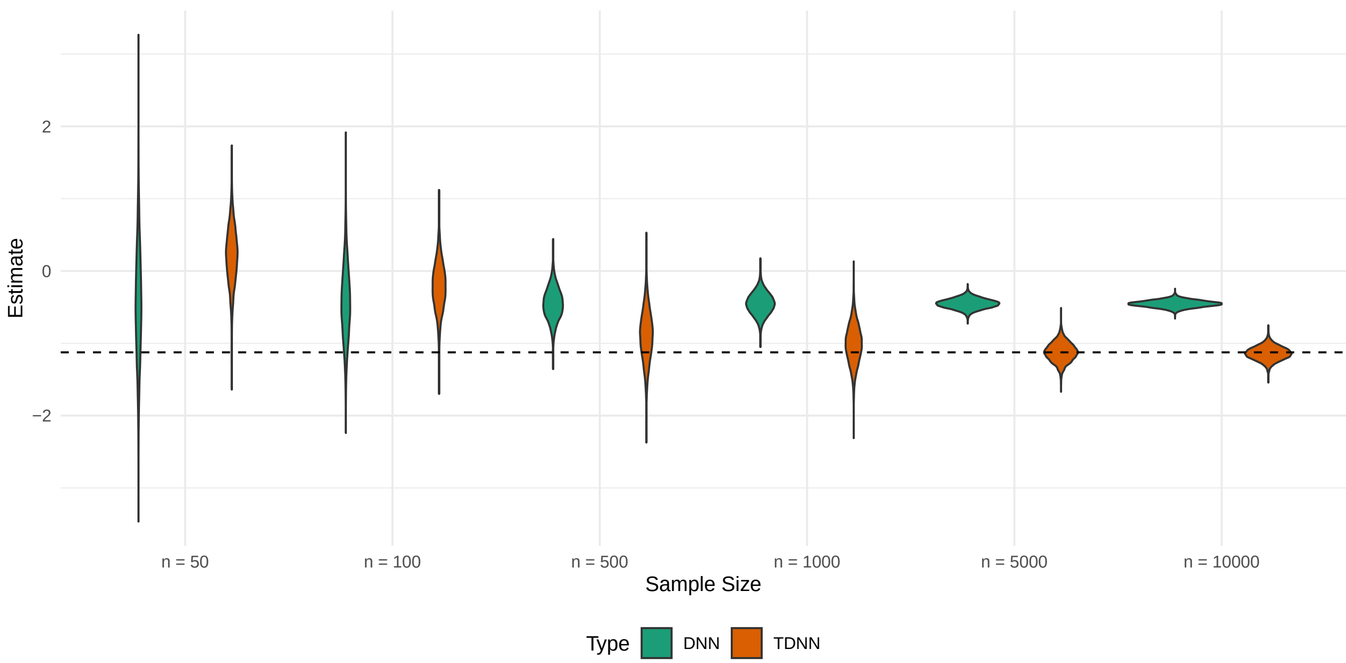

Figure 1: Comparison of the DNN ($s = 20$) and TDNN ($s_1 = 20, s_2 = 50$) Estimators for different sample sizes. The dashed line indicates the value of the unknown regression function at the point of interest. Simulation Setup replicates Setting 1 from Demirkaya et al. (2024) for 10000 Monte Carlo Replications.

As a second, potentially more illustrative example, we consider the estimation of a function of two arguments. Specifically, we 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 $\sigma_\varepsilon^2(x) = \frac{1}{100} |x_1 + x_2|^2$. The resulting surface is depicted in Figure 2.

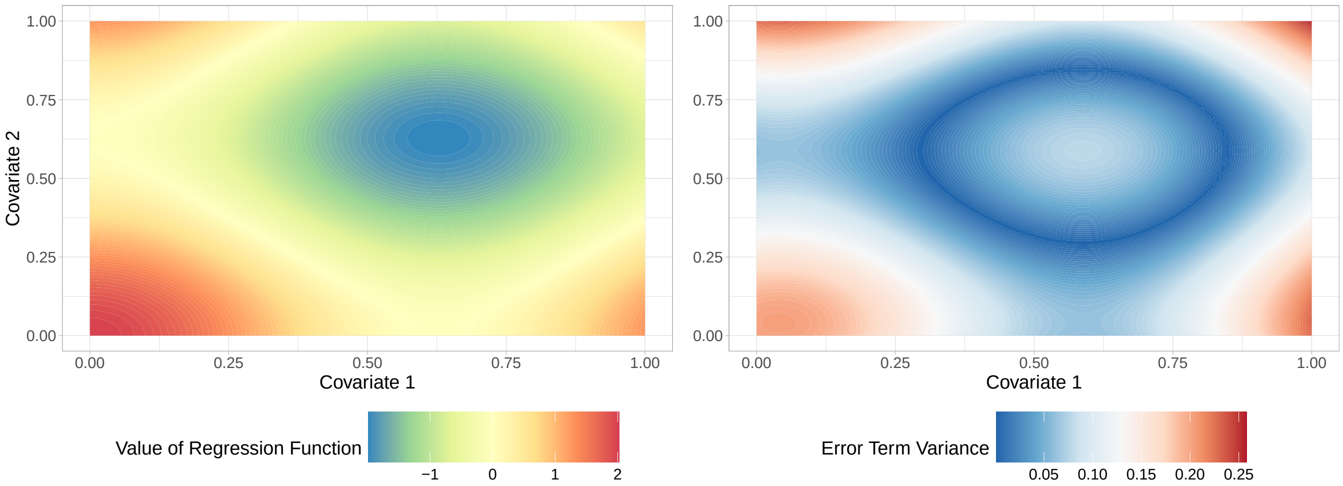

Figure 2: Value of the Regression Function (left) and Variance of the Error Term (right)

6.2 CATE-Estimation

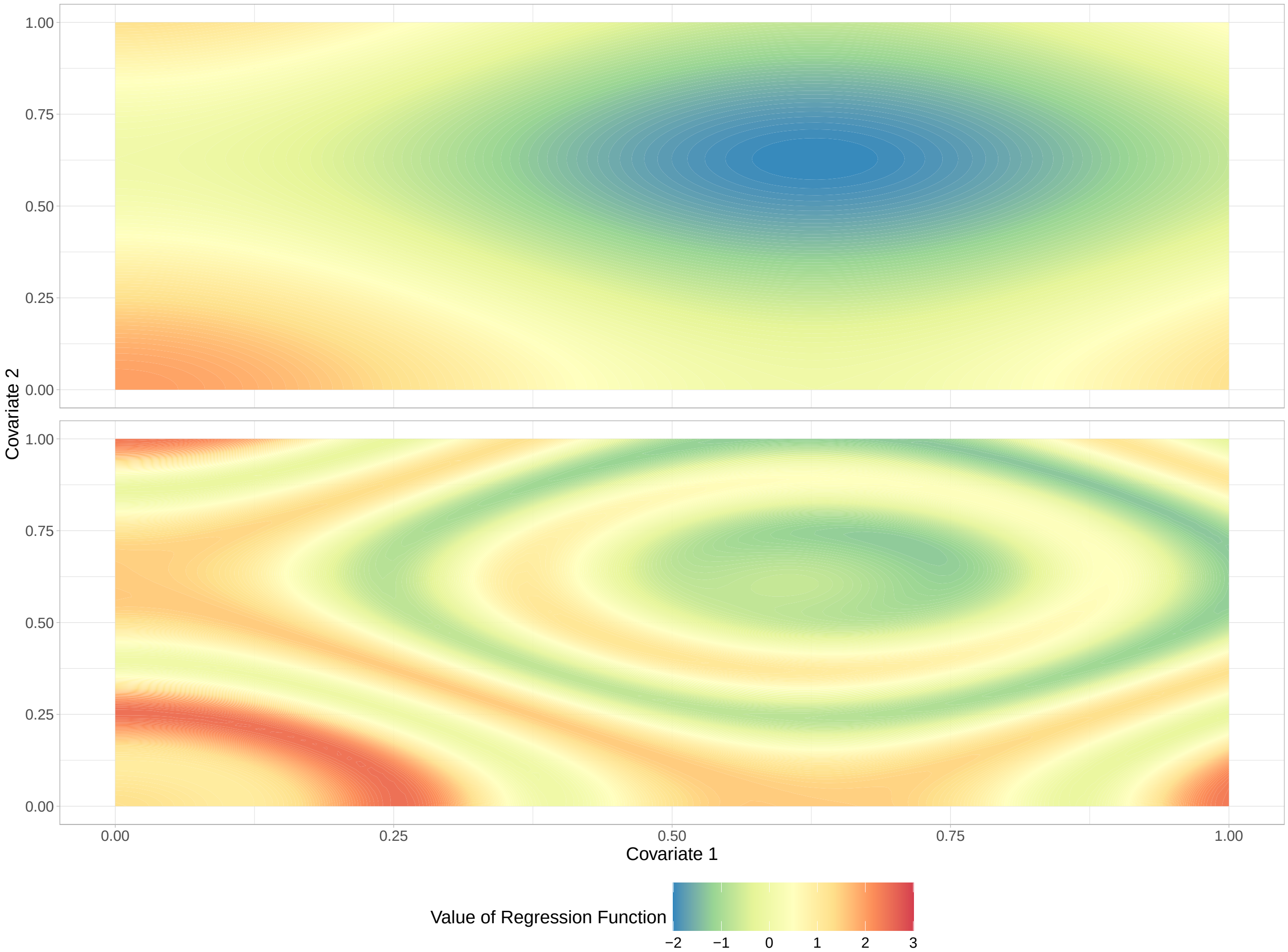

Figure 3: Value of the Regression Functions μ_0 (upper) and μ_1 (lower). Error term structure remains unchanged.

LOREM IPSUM

7 Application

LOREM IPSUM

8 Conclusion

LOREM IPSUM

References

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

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

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

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

Chernozhukov, Victor 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.

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

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

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

Peng, Wei, Lucas Mentch, and Leonard Stefanski (June 2021). *Bias, Consistency, and Alternative Perspectives of the Infinitesimal Jackknife*. arXiv:2106.05918 [math, stat]. DOI: 10.48550/arXiv.2106.05918.

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

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

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

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

A The (T)DNN Estimator as a Generalized U-Statistic

As the majority 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).

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

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 A.1 (Kernel of TDNN Estimator - Adapted from Lemma 8 of Demirkaya et al. (2024))**.**
The kernel of the TDNN estimator takes the following form.

$$\begin{aligned} h_{s_1,s_2}(x; D) &= w_1^* \left[\binom{s_2}{s_1}^{-1} \sum_{\ell \in L_{s_2,s_1}} h_{s_1}(x; D_\ell) \right] + w_2^* h_{s_2}(x; D) \\ &= w_1^* \tilde{\mu}_{s_1}(x; D) + w_2^* h_{s_2}(x; D) \end{aligned} \tag{A.2}$$

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

$$\psi_s^c(x; \mathbf{z}_1, \ldots, \mathbf{z}_c) = \mathbb{E}_D \left[h_s(x; D) \mid Z_1 = \mathbf{z}_1, \ldots, Z_c = \mathbf{z}_c \right] \tag{A.3}$$

$$h_s^{(1)}(x; \mathbf{z}_1) = \psi_s^1(x; \mathbf{z}_1) - \mu(x) \tag{A.4}$$

$$h_s^{(c)}(x; \mathbf{z}_1, \ldots, \mathbf{z}_c) = \psi_s^c(x; \mathbf{z}_1, \ldots, \mathbf{z}_c) - \sum_{j=1}^{c-1} \left(\sum_{\ell \in L_{n,j}} h_s^{(j)}(x; \mathbf{z}_\ell) \right) - \mu(x) \quad \text{for } c = 2, \ldots, s \tag{A.5}$$

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

$$\xi_s^c(x) = \mathrm{Var}_{1:c}\left(\psi_s^c(x; Z_1, \ldots, Z_c) \right) \tag{A.6}$$

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

$$\zeta_{s_1,s_2}^c(x) = \mathrm{Var}_{1:c}\left(\psi_{s_1,s_2}^c(x; Z_1, \ldots, Z_c) \right) \tag{A.7}$$

with an analogous definition of Z'.

A.1 Hoeffding-Decomposition

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

$$H_s^c\left(x;\mathbf{D}_n\right)=\binom{n}{c}^{-1}\sum_{\ell\in L_{n,c}}h_s^{(c)}(x;D_\ell)\quad\text{and}\quad H_{s_1,s_2}^c\left(x;\mathbf{D}_n\right)=\binom{n}{c}^{-1}\sum_{\ell\in L_{n,c}}h_{s_1,s_2}^{(c)}(x;D_\ell)\tag{A.8}$$

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

$$\tilde{\mu}_s\left(x;\mathbf{D}_n\right)=\mu(x)+\sum_{j=1}^{s}\binom{s}{j}H_s^j\left(x;\mathbf{D}_n\right)\quad\text{and}\quad\widehat{\mu}_{s_1,s_2}\left(x;\mathbf{D}_n\right)=\mu(x)+\sum_{j=1}^{s_2}\binom{s_2}{j}H_{s_1,s_2}^j\left(x;\mathbf{D}_n\right)\tag{A.9}$$

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.

$$\forall c=1,2,\ldots,j-1:\quad\mathbb{E}_D\left[h_{s_1,s_2}^{(j)}\left(x;D\right)\mid Z_1=\mathbf{z}_1,\ldots,Z_c=\mathbf{z}_c\right]=0\quad\text{and}\quad\mathbb{E}_D\left[h_{s_1,s_2}^{(j)}\left(x;D\right)\right]=0\tag{A.10}$$

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

$$\operatorname{Var}_D\left(\widehat{\mu}_{s_1,s_2}\left(x;D\right)\right)=\sum_{j=1}^{s_2}\binom{s_2}{j}^2\operatorname{Var}_D\left(H_{s_1,s_2}^j\left(x;D\right)\right)\tag{A.11}$$

$$\operatorname{Var}_D\left(H_{s_1,s_2}^j\left(x;D\right)\right)=\binom{n}{j}^{-1}\operatorname{Var}_D\left(h_{s_1,s_2}^{(j)}\left(x;D\right)\right)=:\binom{n}{j}^{-1}V_{s_1,s_2}^j\left(x\right)\tag{A.12}$$

Third, the following equivalent expression for the kernel variance.

$$\zeta_{s_1,s_2}^{s_2}\left(x\right)=\operatorname{Var}_D\left(h_{s_1,s_2}\left(x;D\right)\right)=\sum_{j=1}^{s_2}\binom{s_2}{j}V_{s_1,s_2}^j\tag{A.13}$$

B Useful Results

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

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

Lemma B.3.
As a consequence of Lemma B.2, we find the following limit results in the nonparametric regression setup.

$$\lim_{s \to \infty} \mathbb{E}_1 [Y_1 s\mathbb{E}_{2:s}[\kappa(x; Z_1, D)]] = \mu(x) \tag{B.2}$$

$$\lim_{s \to \infty} \mathbb{E}_1 \left[Y_1^2 s\mathbb{E}_{2:s}[\kappa(x; Z_1, D)]\right] = \mu^2(x) + \sigma_\varepsilon^2(x) \leq \mu^2(x) + \overline{\sigma}_\varepsilon^2 \tag{B.3}$$

Similarly, in the CATE estimation setup, we can make the following observations.

$$\lim_{s \to \infty} \mathbb{E}_1 [m(Z_1; \mu, \pi)\, s\mathbb{E}_{2:s}[\kappa(x; Z_1, D)]] = \mu_1(x) - \mu_0(x) \tag{B.4}$$

$$\begin{aligned} \lim_{s \to \infty} \mathbb{E}_1 \left[m^2(Z_1; \mu, \pi)\, s\mathbb{E}_{2:s}[\kappa(x; Z_1, D)]\right] &= (\mu_1(x) - \mu_0(x))^2 + \frac{\sigma_\varepsilon^2(x)}{\pi(x)(1 - \pi(x))} \\ &\leq (\mu_1(x) - \mu_0(x))^2 + \frac{\overline{\sigma}_\varepsilon^2}{\mathfrak{p}(1 - \mathfrak{p})} \end{aligned} \tag{B.5}$$

Proof of Lemma B.3. Starting with the first limit, we find the following.

$$\begin{aligned} \mathbb{E}_1 [Y_1 s\mathbb{E}_{2:s}[\kappa(x; Z_1, D)]] &= \mathbb{E}_1 [(\mu(X_1) + \varepsilon_1)\, s\mathbb{E}_{2:s}[\kappa(x; Z_1, D)]] \\ &= \mathbb{E}_1 [(\mu(X_1) + \mathbb{E}[\varepsilon_1 \mid X_1])\, s\mathbb{E}_{2:s}[\kappa(x; Z_1, D)]] \\ &= \mathbb{E}_1 [\mu(X_1)\, s\mathbb{E}_{2:s}[\kappa(x; Z_1, D)]] \overset{\text{(Lem B.2)}}{\longrightarrow} \mu(x) \quad \text{as} \quad s \to \infty \end{aligned} \tag{B.6}$$

Similarly, when considering the second limit, we can make the following observation.

$$
\begin{aligned}
\mathbb{E}_1\left[Y_1^2 s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] &= \mathbb{E}_1\left[\left(\mu\left(X_1\right)+\varepsilon_1\right)^2 s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \\
&= \mathbb{E}_1\left[\left(\mu^2\left(X_1\right)+2\mu\left(X_1\right)\varepsilon_1+\varepsilon_1^2\right) s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \\
&= \mathbb{E}_1\left[\left(\mu^2\left(X_1\right)+2\mu\left(X_1\right)\mathbb{E}\left[\varepsilon_1 \mid X_1\right]+\mathbb{E}\left[\varepsilon_1^2 \,\middle|\, X_1\right]\right) s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \\
&= \mathbb{E}_1\left[\left(\mu^2\left(X_1\right)+\sigma_\varepsilon^2(X_1)\right) s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \\
&\overset{\text{(Lem B.2)}}{\longrightarrow} \mu^2\left(x\right)+\sigma_\varepsilon^2(x) \quad \text{as} \quad s\to\infty
\end{aligned} \tag{B.7}
$$

In the CATE estimation setting, we can proceed analogously.

$$
\begin{aligned}
\mathbb{E}_1\left[m\left(Z_1;\mu,\pi\right)s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] &= \mathbb{E}_1\left[\left(\mu_1\left(X_1\right)-\mu_0\left(X_1\right)+\beta\left(W_1,X_1\right)\varepsilon_i\right)s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \\
&= \mathbb{E}_1\left[\left(\mu_1\left(X_1\right)-\mu_0\left(X_1\right)+\beta\left(W_1,X_1\right)\mathbb{E}\left[\varepsilon_i \mid \right]\right)s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \\
&= \mathbb{E}_1\left[\left(\mu_1\left(X_1\right)-\mu_0\left(X_1\right)\right)s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \\
&\overset{\text{(Lem B.2)}}{\longrightarrow} \mu_1\left(x\right)-\mu_0\left(x\right) \quad \text{as} \quad s\to\infty
\end{aligned} \tag{B.8}
$$

Similarly, we can find the following.

$$
\begin{aligned}
&\mathbb{E}_1\left[m^2\left(Z_i;\mu,\pi\right)s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] = \mathbb{E}_1\left[\left(\mu_1\left(X_1\right)-\mu_0\left(X_1\right)+\beta\left(W_1,X_1\right)\varepsilon_i\right)^2 s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \\
&\quad= \mathbb{E}_1\left[\left(\mu_1\left(X_i\right)-\mu_0\left(X_1\right)\right)^2 s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] + \underbrace{\mathbb{E}_1\left[\left(\mu_1\left(X_i\right)-\mu_0\left(X_1\right)\right)\beta\left(W_1,X_1\right)\mathbb{E}\left[\varepsilon_1 \mid X_1\right]s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right]}_{=0} \\
&\qquad+ \mathbb{E}_1\left[\left(\beta\left(W_1,X_1\right)\varepsilon_i\right)^2 s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \\
&\quad= \mathbb{E}_1\left[\left(\mu_1\left(X_1\right)-\mu_0\left(X_1\right)\right)^2 s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] + \mathbb{E}_1\left[\left(\frac{W_1}{\pi\left(X_1\right)}-\frac{1-W_1}{1-\pi\left(X_1\right)}\right)^2 \varepsilon_1^2 s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \\
&\quad= \underbrace{\mathbb{E}_1\left[\left(\mu_1\left(X_1\right)-\mu_0\left(X_1\right)\right)^2 s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right]}_{\longrightarrow(\mu_1(x)-\mu_0(x))^2 \quad \text{as} \quad s\to\infty} + \underbrace{\mathbb{E}_1\left[\mathbb{E}\left[\left(\frac{W_1}{\pi\left(X_1\right)}-\frac{1-W_1}{1-\pi\left(X_1\right)}\right)^2 \varepsilon_1^2 \,\middle|\, X_1\right] s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right]}_{(B)}
\end{aligned} \tag{B.9}
$$

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

$$
\begin{aligned}
(B) &= \mathbb{E}_1\left[\mathbb{E}\left[\left(\frac{W_1}{\pi\left(X_1\right)}-\frac{1-W_1}{1-\pi\left(X_1\right)}\right)^2 \varepsilon_1^2 \,\middle|\, X_1\right] s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right] \\
&= \mathbb{E}_1\left[\frac{\sigma_\varepsilon^2(X_1)\cdot s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]}{\pi^2\left(X_1\right)\left(1-\pi\left(X_1\right)\right)^2}\cdot\mathbb{E}\left[\left(W_1\left(1-\pi\left(X_1\right)\right)-\left(1-W_1\right)\pi\left(X_1\right)\right)^2 \,\middle|\, X_1\right]\right]
\end{aligned} \tag{B.10}
$$

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.

$$
\begin{aligned}
(B) &= \mathbb{E}_1\left[\frac{\sigma_\varepsilon^2(X_1)\cdot s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]}{\pi^2\left(X_1\right)\left(1-\pi\left(X_1\right)\right)^2}\cdot\mathbb{E}\left[W_1\left(1-\pi\left(X_1\right)\right)^2+\left(1-W_1\right)\pi^2\left(X_1\right) \,\middle|\, X_1\right]\right] \\
&= \mathbb{E}_1\left[\frac{\sigma_\varepsilon^2(X_1)\cdot s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]}{\pi^2\left(X_1\right)\left(1-\pi\left(X_1\right)\right)^2}\cdot\left(\pi(X_1)\left(1-\pi\left(X_1\right)\right)^2+(1-\pi(X_1))\pi^2\left(X_1\right)\right)\right] \\
&= \mathbb{E}_1\left[\frac{\sigma_\varepsilon^2(X_1)\cdot s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]}{\pi\left(X_1\right)\left(1-\pi\left(X_1\right)\right)}\right] \overset{\text{(Lem B.2)}}{\longrightarrow} \frac{\sigma_\varepsilon^2(x)}{\pi\left(x\right)\left(1-\pi\left(x\right)\right)} \quad \text{as} \quad s\to\infty
\end{aligned} \tag{B.11}
$$

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

$$\mathbb{E}_1\left[m^2\left(Z_i;\mu,\pi\right)s\mathbb{E}_{2:s}\left[\kappa(x;Z_1,D)\right]\right]\overset{\text{(Lem B.2)}}{\longrightarrow}\left(\mu_1\left(x\right)-\mu_0\left(x\right)\right)^2+\frac{\sigma_\varepsilon^2(x)}{\pi\left(x\right)\left(1-\pi\left(x\right)\right)}\quad\text{as}\quad s\to\infty \tag{B.12}$$

■

Lemma B.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 inequality holds.

$$\mathbb{E}_{D,D'}\left[Y_1Y'_{c+1}\,c(s-c)\,\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]\leq LOREMIPSUM \tag{B.13}$$

Similarly, in the CATE estimation setting (Setup 2), i.e. replacing observations drawn from P by observations drawn from Q, the following inequality holds.

$$\mathbb{E}_{D,D'}\left[m\left(Z_1;\mu,\pi\right)m\left(Z'_{c+1};\mu,\pi\right)\,c(s-c)\,\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]\leq LOREMIPSUM \tag{B.14}$$

Proof of Lemma B.4.
Consider first the following argument.

$$\mathbb{E}_{D,D'}\left[Y_1Y'_{c+1}\,c(s-c)\,\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)\right]=LOREMIPSUM \tag{B.15}$$

LOREM IPSUM ■

Lemma B.5 (Peng, Mentch, and Stefanski (2021) - Lemma 1)**.**
Suppose that $\sum X_i^2\xrightarrow{p}1,\sum\mathbb{E}\left[X_i^2\right]\to1$, and $\sum_{i=1}^n\mathbb{E}\left[Y_i^2\right]\to0$, then

$$\sum\left[X_i+Y_i\right]^2\xrightarrow{p}1\qquad\textit{and }\mathbb{E}\left[\sum\left(X_i+Y_i\right)^2\right]\to1. \tag{B.16}$$

Lemma B.6 (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 $D_{\ell,-i}=\{Z_l\,|\,l\in\ell\backslash\{i\}\}$.

C Proofs for Results in Section 3

D Proofs for Results in Section 4

D.1 Closed Form Representations

Proof of Theorem 4.1.
Recall the closed form representation of the DNN estimator as presented in Equation 3.8.

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

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.

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

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

LOREM IPSUM ■

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

$$\mathbb{E}_D \left[h_s \left(x; D\right)\right] = \mathbb{E}_1 \left[\mu \left(X_1\right) s \left(1 - \psi \left(B \left(x, \|X_1 - x\|\right)\right)\right)^{s-1}\right] \longrightarrow \mu(x) \quad as \quad s \to \infty \tag{D.3}$$

Proof of Lemma D.1. This result follows immediately from Lemma B.3. ■

Lemma D.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

$$\psi_s^1 \left(x; z_1\right) = \varepsilon_1 \mathbb{E}_D \left[\kappa \left(x; Z_1, D\right) \;\middle|\; X_1 = x_1\right] + \mathbb{E}_D \left[\sum_{i=1}^{s} \kappa \left(x; Z_i, D\right) \mu(X_i) \;\middle|\; X_1 = x_1\right] \tag{D.4}$$

Proof of Lemma D.2.

$$\begin{aligned}
\psi_s^1 \left(x; z_1\right) &= \mathbb{E}_D \left[h_s \left(x; D\right) \mid Z_1 = z_1\right] = \mathbb{E}_D \left[\sum_{i=1}^{s} \kappa \left(x; Z_i, D\right) Y_i \;\middle|\; Z_1 = z_1\right] \\
&= \mathbb{E}_D \left[\left(\mu(x_1) + \varepsilon_1\right) \kappa \left(x; Z_1, D\right) + \sum_{i=2}^{s} \kappa \left(x; Z_i, D\right) \mu(X_i) \;\middle|\; Z_1 = z_1\right] \\
&= \varepsilon_1 \mathbb{E}_D \left[\kappa \left(x; Z_1, D\right) \;\middle|\; X_1 = x_1\right] + \mathbb{E}_D \left[\sum_{i=1}^{s} \kappa \left(x; Z_i, D\right) \mu(X_i) \;\middle|\; X_1 = x_1\right]
\end{aligned} \tag{D.5}$$

■

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

$$\begin{aligned}\mathbb{E}_D\left[h_s(x;D)\right] &= \mathbb{E}_1\left[\left(\mu_1(X_1)-\mu_0(X_1)+\beta(W_1,X_1)\left(Y_1-\mu_{W_1}(X_1)\right)\right)s\left(1-\psi\left(B\left(x,\|X_1-x\|\right)\right)\right)^{s-1}\right]\\ &\longrightarrow \mathrm{CATE}(x) \quad \textit{as} \quad s\to\infty\end{aligned} \tag{D.6}$$

Proof of Lemma D.3. This result follows immediately from Lemma B.3. ∎

Lemma D.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

$$\psi_s^1(x;z_1)=\beta(W_1,X_1)\,\varepsilon_1\cdot\mathbb{E}\left[\kappa(x;Z_1,D)\mid X_1=x_1\right]+\mathbb{E}_D\left[\sum_{i=1}^{s}\kappa(x;Z_i,D)\left(\mu_1(X_i)-\mu_0(X_i)\right)\,\middle|\,Z_1=z_1\right] \tag{D.7}$$

Proof of Lemma D.4.

$$\begin{aligned}\psi_s^1(x;z_1) &= \mathbb{E}_D\left[h_s(x;D)\mid Z_1=z_1\right]\\ &= \mathbb{E}_D\left[\sum_{i=1}^{s}\kappa(x;Z_i,D)\left(\mu_1(X_i)-\mu_0(X_i)+\beta(W_i,X_i)\left(Y_i-\mu_{W_i}(X_i)\right)\right)\,\middle|\,Z_1=z_1\right]\\ &= \left(\mu_1(X_1)-\mu_0(X_1)+\beta(W_1,X_1)\,\varepsilon_1\right)\mathbb{E}\left[\kappa(x;Z_1,D)\mid X_1=x_1\right]\\ &\quad+\mathbb{E}_D\left[\sum_{i=2}^{s}\kappa(x;Z_i,D)\left(\mu_1(X_i)-\mu_0(X_i)\right)\,\middle|\,Z_1=z_1\right]\\ &= \beta(W_1,X_1)\,\varepsilon_1\cdot\mathbb{E}\left[\kappa(x;Z_1,D)\mid X_1=x_1\right]+\mathbb{E}_D\left[\sum_{i=1}^{s}\kappa(x;Z_i,D)\left(\mu_1(X_i)-\mu_0(X_i)\right)\,\middle|\,Z_1=z_1\right]\end{aligned} \tag{D.8}$$

∎

D.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 D.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

$$\Omega_s(x) = \mathbb{E}\left[h_s^2(x; Z_1, \ldots, Z_s)\right]. \tag{D.9}$$

Then,

$$\Omega_s(x) = \mathbb{E}_1\left[(\mu(X_1) + \varepsilon_1)^2 s\mathbb{E}_{2:s}[\kappa(x; Z_1, D)]\right] \lesssim \mu^2(x) + \overline{\sigma}_\varepsilon^2 + o(1) \quad \textit{as} \quad s \to \infty. \tag{D.10}$$

Proof of Lemma D.5.

$$\begin{aligned}\Omega_s(x) = \mathbb{E}\left[h_s^2(x; Z_1, \ldots, Z_s)\right] = \mathbb{E}_D\left[\left(\sum_{i=1}^{s}\kappa(x; Z_i, D)\, Y_i\right)^2\right] = \mathbb{E}_D\left[\sum_{i=1}^{s}\sum_{j=1}^{s}(\kappa(x; Z_i, D)\,\kappa(x; Z_j, D)\, Y_i Y_j)\right] \\ = \mathbb{E}_D\left[s\kappa(x; Z_1, D)\, Y_1^2\right] = \mathbb{E}_1\left[Y_1^2 s\mathbb{E}_{2:s}[\kappa(x; Z_1, D)]\right] \overset{\text{(Lem B.3)}}{\longrightarrow} \mu^2(x) + \sigma_\varepsilon^2(x) \quad \text{as} \quad s \to \infty\end{aligned} \tag{D.11}$$

Thus, we obtain the desired result. ■

Lemma D.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

$$\Omega_s^c(x) = \mathbb{E}\left[h_s(x; Z_1, \ldots, Z_c, Z_{c+1}, \ldots, Z_s) \cdot h_s\left(x; Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_s\right)\right]. \tag{D.12}$$

Then,

$$\Omega_s^c(x) \lesssim \frac{s^2 + cs - c^2}{s^2}\mu^2(x) + (c/s)\overline{\sigma}_\varepsilon^2 + o(1) \quad \textit{for} \quad s \quad \textit{sufficiently large} \tag{D.13}$$

and thus

$$\Omega_s^c(x) \lesssim \mu^2(x) + \overline{\sigma}_\varepsilon^2 + o(1) \quad \textit{as} \quad s \to \infty. \tag{D.14}$$

Proof of Lemma D.6.

$$
\begin{aligned}
\Omega_s^c(x) &= \mathbb{E}\left[h_s\left(x; Z_1, \ldots, Z_c, Z_{c+1}, \ldots, Z_s\right) \cdot h_s\left(x; Z_1, \ldots, Z_c, Z_{c+1}', \ldots, Z_s'\right)\right] \\
&= \mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{s} \kappa\left(x; Z_i, D\right) Y_i\right)\left(\sum_{j=1}^{c} \kappa\left(x; Z_j, D'\right) Y_j + \sum_{j=c+1}^{s} \kappa\left(x; Z_j', D'\right) Y_j'\right)\right] \\
&= \mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=1}^{c} \kappa\left(x; Z_i, D\right)\kappa\left(x; Z_j, D'\right) Y_i Y_j\right] + \mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=c+1}^{s} \kappa\left(x; Z_i, D\right)\kappa\left(x; Z_j', D'\right) Y_i Y_j'\right] \\
&\quad + \mathbb{E}_{D,D'}\left[\sum_{i=c+1}^{s}\sum_{j=1}^{c} \kappa\left(x; Z_i, D\right)\kappa\left(x; Z_j, D'\right) Y_i Y_j\right] + \mathbb{E}_{D,D'}\left[\sum_{i=c+1}^{s}\sum_{j=c+1}^{s} \kappa\left(x; Z_i, D\right)\kappa\left(x; Z_j', D'\right) Y_i Y_j'\right] \\
&= \underbrace{\mathbb{E}_{D,D'}\left[c\kappa\left(x; Z_1, D\right)\kappa\left(x; Z_1, D'\right) Y_1^2\right]}_{(A)} + \underbrace{\mathbb{E}_{D,D'}\left[c(s-c)\kappa\left(x; Z_1, D\right)\kappa\left(x; Z_{c+1}', D'\right) Y_1 Y_{c+1}'\right]}_{(B)} \\
&\quad + \underbrace{\mathbb{E}_{D,D'}\left[c(s-c)\kappa\left(x; Z_{c+1}, D\right)\kappa\left(x; Z_1, D'\right) Y_{c+1} Y_1\right]}_{(C)} \\
&\quad + \underbrace{\mathbb{E}_{D,D'}\left[(s-c)^2\kappa\left(x; Z_{c+1}, D\right)\kappa\left(x; Z_{c+1}', D'\right) Y_{c+1} Y_{c+1}'\right]}_{(D)}
\end{aligned}
\tag{D.15}
$$

Starting from this decomposition, we will analyze the terms one by one. First, by Lemma B.3, we find the following.

$$
\begin{aligned}
(A) &= \mathbb{E}_{D,D'}\left[c\kappa\left(x; Z_1, D\right)\kappa\left(x; Z_1, D'\right) Y_1^2\right] = (c/s)\mathbb{E}_1\left[Y_1^2 s\mathbb{E}_{2:s}\left[\kappa\left(x; Z_1, D\right)\kappa\left(x; Z_1, D'\right)\right]\right] \\
&\leq (c/s)\mathbb{E}_1\left[Y_1^2 s\mathbb{E}_{2:s}\left[\kappa\left(x; Z_1, D\right)\right]\right] \overset{\text{(Lem B.3)}}{\lesssim} (c/s)\left(\mu^2(x) + \sigma_\varepsilon(x)\right) + o(1)
\end{aligned}
\tag{D.16}
$$

Similarly, we can find that:

$$
\begin{aligned}
(B) &= \mathbb{E}_{D,D'}\left[c(s-c)\kappa\left(x; Z_1, D\right)\kappa\left(x; Z_{c+1}', D'\right) Y_1 Y_{c+1}'\right] \\
&= \mathbb{E}_{D,D'}\left[\mu(X_1)\mu(X_{c+1}')\, c(s-c)\,\kappa\left(x; Z_1, D\right)\kappa\left(x; Z_{c+1}', D'\right)\right] \\
&\leq \mathbb{E}_D\left[\left|\mu(X_1)\right| c\,\kappa\left(x; Z_1, D\right)\right]\mathbb{E}_{D'}\left[\left|\mu(X_{c+1}')\right|(s-c)\,\kappa\left(x; Z_{c+1}', D'\right)\right] \\
&= \frac{c(s-c)}{s^2} \leq \mathbb{E}_D\left[\left|\mu(X_1)\right| s\,\kappa\left(x; Z_1, D\right)\right]\mathbb{E}_{D'}\left[\left|\mu(X_{c+1}')\right| s\,\kappa\left(x; Z_{c+1}', D'\right)\right] \\
&= \frac{c(s-c)}{s^2}\left(\mathbb{E}_D\left[\left|\mu(X_1)\right| s\,\kappa\left(x; Z_1, D\right)\right]\right)^2 = \frac{c(s-c)}{s^2}\left(\mathbb{E}_1\left[\left|\mu(X_1)\right| s\left\{1 - \varphi\left(B\left(x, \|X_1 - x\|\right)\right)\right\}^{s-1}\right]\right)^2 \\
&\lesssim \frac{c(s-c)}{s^2}\mu^2(x) + o(1)
\end{aligned}
\tag{D.17}
$$

Following analogous steps, we find the same result for the third term.

$$
(C) = \mathbb{E}_{D,D'}\left[c(s-c)\kappa\left(x; Z_{c+1}, D\right)\kappa\left(x; Z_1, D'\right) Y_{c+1} Y_1\right] \lesssim \frac{c(s-c)}{s^2}\mu^2(x) + o(1) \tag{D.18}
$$

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

$$
\begin{aligned}
(D) &= \mathbb{E}_{D,D'}\left[(s-c)^2\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)Y_{c+1}Y'_{c+1}\right] \\
&= \mathbb{E}_{D,D'}\left[\mu(X_{c+1})\mu(X'_{c+1})\left(s-c\right)^2\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)\right] \\
&\leq \mathbb{E}_{D}\left[\left|\mu(X_{c+1})\right|\left(s-c\right)\kappa\left(x;Z_{c+1},D\right)\right]\mathbb{E}_{D'}\left[\left|\mu(X'_{c+1})\right|\left(s-c\right)\kappa\left(x;Z'_{c+1},D'\right)\right] \\
&= \frac{(s-c)^2}{s^2}\mathbb{E}_{D}\left[\left|\mu(X_{c+1})\right|s\,\kappa\left(x;Z_{c+1},D\right)\right]\mathbb{E}_{D'}\left[\left|\mu(X'_{c+1})\right|s\,\kappa\left(x;Z'_{c+1},D'\right)\right] \\
&= \frac{(s-c)^2}{s^2}\left(\mathbb{E}_{D}\left[\left|\mu(X_{c+1})\right|s\,\kappa\left(x;Z_{c+1},D\right)\right]\right)^2 \lesssim \frac{(s-c)^2}{s^2}\mu^2(x)+o(1)
\end{aligned}
\tag{D.19}
$$

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

Lemma D.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

$$\Upsilon_{s_1,s_2}(x) = \mathbb{E}\left[h_{s_1}(x; Z_1, \ldots, Z_{s_1}) \cdot h_{s_2}(x; Z_1, \ldots, Z_{s_1}, \ldots, Z_{s_2})\right]. \tag{D.20}$$

Then,

$$\Upsilon_{s_1,s_2}(x) \lesssim \mu^2(x) + \overline{\sigma}_\varepsilon^2 + o(1) \quad \textit{as} \quad s_1, s_2 \to \infty \quad \textit{with} \quad 0 < \mathfrak{c} \leq s_1/s_2 \leq 1 - \mathfrak{c} < 1. \tag{D.21}$$

Proof of Lemma D.7.

$$\begin{aligned}
\Upsilon_{s_1,s_2}(x) &= \mathbb{E}\left[h_{s_1}(x; Z_1, \ldots, Z_{s_1}) \cdot h_{s_2}(x; Z_1, \ldots, Z_{s_1}, \ldots, Z_{s_2})\right] \\
&= \mathbb{E}_D\left[\left(\sum_{i=1}^{s_1} \kappa(x; Z_i, D_{[s_1]}) Y_i\right)\left(\sum_{j=1}^{s_1} \kappa(x; Z_j, D) Y_j + \sum_{j=s_1+1}^{s_2} \kappa(x; Z_j, D) Y_j\right)\right] \\
&= \mathbb{E}_D\left[\sum_{i=1}^{s_1} \kappa(x; Z_i, D) Y_i^2\right] + \mathbb{E}_D\left[\sum_{i=1}^{s_1}\sum_{j=s_1+1}^{s_2} \kappa(x; Z_i, D_{[s_1]}) \kappa(x; Z_j, D) Y_i Y_j\right] \\
&= \mathbb{E}_D\left[Y_1^2\, s_1\, \kappa(x; Z_1, D)\right] + \mathbb{E}_D\left[Y_1 Y_{s_2}\, s_1(s_2 - s_1)\, \kappa(x; Z_1, D_{[s_1]}) \kappa(x; Z_{s_2}, D)\right] \\
&= \mathbb{E}_D\left[\left(\mu^2(X_1) + \sigma_\varepsilon^2(X_1)\right)\, s_1\, \kappa(x; Z_1, D)\right] + \mathbb{E}_D\left[\mu(X_1)\mu(X_{s_2})\, s_1(s_2 - s_1)\, \kappa(x; Z_1, D_{[s_1]}) \kappa(x; Z_{s_2}, D)\right] \\
&= \frac{s_1}{s_2}\mathbb{E}_D\left[\left(\mu^2(X_1) + \sigma_\varepsilon^2(X_1)\right)\, s_1\, \kappa(x; Z_1, D)\right] + \frac{s_2 - s_1}{s_2}\mathbb{E}_D\left[\mu(X_1)\mu(X_{s_2})\, s_1 s_2\, \kappa(x; Z_1, D_{[s_1]}) \kappa(x; Z_{s_2}, D)\right] \\
&\leq \frac{s_1}{s_2}\mathbb{E}_D\left[\left(\mu^2(X_1) + \sigma_\varepsilon^2(X_1)\right)\, s_2\, \kappa(x; Z_1, D)\right] \\
&\quad + \frac{s_2 - s_1}{s_2}\mathbb{E}_D\left[|\mu(X_1)|\, s_1\, \kappa(x; Z_1, D_{[s_1]})\right]\mathbb{E}_D\left[|\mu(X_{s_2})|\, s_2\, \kappa(x; Z_{s_2}, D)\right] \\
&\lesssim \mu^2(x) + \sigma_\varepsilon^2(x) + o(1) \leq \mu^2(x) + \overline{\sigma}_\varepsilon^2 + o(1).
\end{aligned} \tag{D.22}$$

■

Lemma D.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

$$\Upsilon^c_{s_1,s_2}(x) = \mathbb{E}\left[h_{s_1}\left(x; Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_{s_1}\right) \cdot h_{s_2}(x; Z_1, \ldots, Z_{s_2})\right]. \tag{D.23}$$

Then,

$$\begin{aligned} &\Upsilon^c_{s_1,s_2}(x) \lesssim \frac{cs_2 - c^2 + s_1 s_2}{s_1 s_2}\mu^2(x) + (c/s_1)\overline{\sigma}^2_\varepsilon + o(1) \\ &\textit{for} \quad s_1, s_2 \quad \textit{sufficiently large} \quad \textit{with} \quad 0 < \mathfrak{c} \leq s_1/s_2 \leq 1 - \mathfrak{c} < 1 \end{aligned} \tag{D.24}$$

and thus

$$\Upsilon^c_{s_1,s_2}(x) \lesssim \mu^2(x) + o(1) \quad \textit{as} \quad s_1, s_2 \to \infty \quad \textit{with} \quad 0 < \mathfrak{c} \leq s_1/s_2 \leq 1 - \mathfrak{c} < 1. \tag{D.25}$$

Proof of Lemma D.8.

$$\begin{aligned} \Upsilon^c_{s_1,s_2}(x) &= \mathbb{E}\left[h_{s_1}\left(x; Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_{s_1}\right) \cdot h_{s_2}(x; Z_1, \ldots, Z_{s_2})\right] \\ &= \mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{c} \kappa(x; Z_i, D')Y_i + \sum_{i=c+1}^{s_1} \kappa(x; Z'_i, D')Y'_i\right)\left(\sum_{j=1}^{c} \kappa(x; Z_j, D)Y_j + \sum_{j=c+1}^{s_2} \kappa(x; Z_j, D)Y_j\right)\right] \\ &= \underbrace{\mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=1}^{c} \kappa(x; Z_i, D')\kappa(x; Z_j, D)Y_iY_j\right]}_{(A)} + \underbrace{\mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{c} \kappa(x; Z_i, D')Y_i\right)\left(\sum_{j=c+1}^{s_2} \kappa(x; Z_j, D)Y_j\right)\right]}_{(B)} \\ &\quad + \underbrace{\mathbb{E}_{D,D'}\left[\sum_{i=c+1}^{s_1}\sum_{j=1}^{c} \kappa(x; Z'_i, D')\kappa(x; Z_j, D)Y'_iY_j\right]}_{(C)} + \underbrace{\mathbb{E}_{D,D'}\left[\left(\sum_{i=c+1}^{s_1} \kappa(x; Z'_i, D')Y'_i\right)\left(\sum_{j=c+1}^{s_2} \kappa(x; Z_j, D)Y_j\right)\right]}_{(D)} \end{aligned} \tag{D.26}$$

Again, we have four terms to analyze individually.

$$\begin{aligned} (A) &= \mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=1}^{c} \kappa(x; Z_i, D')\kappa(x; Z_j, D)Y_iY_j\right] \\ &= \mathbb{E}_{D,D'}\left[\sum_{i=1}^{c} Y_i^2 \kappa(x; Z_i, D')\kappa(x; Z_i, D)\right] \\ &= \mathbb{E}_{D,D'}\left[Y_1^2\, c\,\kappa(x; Z_1, D')\kappa(x; Z_1, D)\right] = \mathbb{E}_{D,D'}\left[\left(\mu^2(X_1) + \sigma^2_\varepsilon(X_1)\right)\, c\,\kappa(x; Z_1, D_{[c]})\kappa(x; Z_1, D'_{c+1:s_1})\right] \\ &= \mathbb{E}_D\left[\left(\mu^2(X_1) + \sigma^2_\varepsilon(X_1)\right)\, c\,\kappa(x; Z_1, D)\right] = \frac{c}{s_1}\mathbb{E}_D\left[\left(\mu^2(X_1) + \sigma^2_\varepsilon(X_1)\right)\, s_1\,\kappa(x; Z_1, D)\right] \\ &\lesssim (c/s_1)(\mu^2(x) + \sigma^2_\varepsilon(x)) + o(1) \leq (c/s_1)(\mu^2(x) + \overline{\sigma}^2_\varepsilon) + o(1) \end{aligned} \tag{D.27}$$

Considering the second term, we find the following.

$$
\begin{aligned}
(B) &= \mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{c}\kappa(x;Z_i,D')Y_i\right)\left(\sum_{j=c+1}^{s_2}\kappa(x;Z_j,D)Y_j\right)\right] = \mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=c+1}^{s_2}Y_iY_j\kappa(x;Z_i,D')\kappa(x;Z_j,D)\right] \\
&= \mathbb{E}_{D,D'}\left[c(s_2-c)\,Y_1Y_{s_1}\kappa(x;Z_1,D')\kappa(x;Z_{s_2},D)\right] = \frac{c(s_2-c)}{s_1s_2}\mathbb{E}_{D,D'}\left[Y_1Y_{s_2}\,s_1s_2\,\kappa(x;Z_1,D')\kappa(x;Z_{s_2},D)\right] \\
&\leq \frac{c(s_2-c)}{s_1s_2}\mathbb{E}_{D'}\left[|\mu(X_1)|\,s_1\,\kappa(x;Z_1,D')\right]\mathbb{E}_{D}\left[|\mu(X_{s_2})|\,s_2\,\kappa(x;Z_{s_2},D)\right] \\
&\lesssim \frac{c(s_2-c)}{s_1s_2}\mu^2(x)+o(1)
\end{aligned}
\tag{D.28}
$$

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

$$
\begin{aligned}
(C) &= \mathbb{E}_{D,D'}\left[\sum_{i=c+1}^{s_1}\sum_{j=1}^{c}\kappa(x;Z_i',D')\kappa(x;Z_j,D)Y_i'Y_j\right] = \mathbb{E}_{D,D'}\left[Y_{s_1}'Y_1\,(s_1-c)c\,\kappa(x;Z_{s_1}',D')\kappa(x;Z_1,D)\right] \\
&= \frac{(s_1-c)c}{s_1s_2}\mathbb{E}_{D,D'}\left[\mu(X_{s_1}')\mu(X_1)\,s_1s_2\,\kappa(x;Z_{s_1}',D')\kappa(x;Z_1,D)\right] \\
&\leq \frac{(s_1-c)c}{s_1s_2}\mathbb{E}_{D}\left[|\mu(X_{s_1}')|\,s_1\,\kappa(x;Z_{s_1}',D')\right]\mathbb{E}_{D}\left[|\mu(X_1)|\,s_2\,\kappa(x;Z_1,D)\right] \\
&\lesssim \frac{(s_1-c)c}{s_1s_2}\mu^2(x)+o(1)
\end{aligned}
\tag{D.29}
$$

Lastly, concerning the fourth term, observe the following.

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

■

Lemma D.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*

$$\zeta_{s_1,s_2}^{s_2}(x) \lesssim \mu^2(x) + \overline{\sigma}_\varepsilon + o(1) \quad \text{as} \quad s_1, s_2 \to \infty \quad \text{with} \quad 0 < \mathfrak{c} \leq s_1/s_2 \leq 1 - \mathfrak{c} < 1. \tag{D.31}$$

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

$$\begin{aligned}
\zeta_{s_1,s_2}^{s_2}(x) &= \operatorname{Var}\left(h_{s_1,s_2}(x; Z_1, \ldots, Z_{s_2})\right) = \operatorname{Var}_D\left(h_{s_1,s_2}(x; D)\right) \\
&\leq \mathbb{E}_D\left[h_{s_1,s_2}^2(x; D)\right] = \mathbb{E}_D\left[\left(w_1^* \tilde{\mu}_{s_1}(x; D) + w_2^* h_{s_2}(x; D)\right)^2\right] \\
&= \left(w_1^*\right)^2 \mathbb{E}_D\left[\tilde{\mu}_{s_1}^2(x; D)\right] + 2 w_1^* w_2^* \mathbb{E}_D\left[\tilde{\mu}_{s_1}(x; D) h_{s_2}(x; D)\right] + \left(w_2^*\right)^2 \Omega_{s_2}
\end{aligned} \tag{D.32}$$

Then, observe the following.

$$\begin{aligned}
\mathbb{E}_D\left[\tilde{\mu}_{s_1}^2(x; D)\right] &= \mathbb{E}_D\left[\left(\binom{s_2}{s_1}^{-1} \sum_{\ell \in L_{s_2,s_1}} h_{s_1}(x; D_\ell)\right)^2\right] = \binom{s_2}{s_1}^{-2} \mathbb{E}_D\left[\sum_{\iota,\iota' \in L_{s_2,s_1}} h_{s_1}(x; D_\iota) h_{s_1}(x; D_{\iota'})\right] \\
&= \binom{s_2}{s_1}^{-2} \sum_{c=0}^{s_1} \binom{s_2}{s_1}\binom{s_1}{c}\binom{s_2 - s_1}{s_1 - c} \Omega_{s_1}^c = \binom{s_2}{s_1}^{-1} \sum_{c=0}^{s_1} \binom{s_1}{c}\binom{s_2 - s_1}{s_1 - c} \Omega_{s_1}^c \\
&\lesssim \Omega_{s_1} \lesssim \mu(x)^2 + \sigma_\varepsilon^2 + o(1) \quad \text{as} \quad s \to \infty
\end{aligned} \tag{D.33}$$

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

$$\Omega_{s_2} \lesssim \mu(x)^2 + \sigma_\varepsilon^2 + o(1) \quad \text{as} \quad s \to \infty \tag{D.34}$$

Lastly, consider the following.

$$\begin{aligned}
\mathbb{E}_D\left[\tilde{\mu}_{s_1}(x; D) h_{s_2}(x; D)\right] &= \mathbb{E}_D\left[\binom{s_2}{s_1}^{-1} \sum_{\ell \in L_{s_2,s_1}} h_{s_1}(x; D_\ell) h_{s_2}(x; D)\right] \\
&= \mathbb{E}_D\left[h_{s_1}\left(x; D_{[s_1]}\right) h_{s_2}(x; D)\right] = \Upsilon_{s_1,s_2}(x)
\end{aligned} \tag{D.35}$$

Thus, we find the following.

$$\begin{aligned}
\zeta_{s_2,s_2}(x) &\lesssim \left(w_1^*\right)^2 \Omega_{s_1} + 2 w_1^* w_2^* \Upsilon_{s_1,s_2}(x) + \left(w_1^*\right)^2 \Omega_{s_2} \\
&\lesssim \left(w_1^* + w_2^*\right)^2 \left(\mu^2(x) + \sigma_\varepsilon\right) + o(1) = \mu^2(x) + \sigma_\varepsilon + o(1).
\end{aligned} \tag{D.36}$$

■

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

$$0 < \mathfrak{c} \leq s_1/s_2 \leq 1 - \mathfrak{c} < 1 \quad \textit{and} \quad s_2 = o(n), \tag{D.37}$$

it holds that

$$\zeta_{s_1,s_2}^{1}(x) \sim s_2^{-1}. \tag{D.38}$$

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

$$\Omega_s(x) = \mathbb{E}\left[h_s^2(x; Z_1, \ldots, Z_s)\right]. \tag{D.39}$$

Then,

$$\Omega_s(x) \lesssim (\mu_1(x) - \mu_0(x))^2 + \frac{\overline{\sigma}_\varepsilon^2}{\mathfrak{p}(1-\mathfrak{p})} + o(1) \tag{D.40}$$

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

$$\begin{aligned}
\Omega_s(x) &= \mathbb{E}\left[h_s^2(x; Z_1, \ldots, Z_s)\right] = \mathbb{E}_D\left[\left(\sum_{i=1}^{s} \kappa(x; Z_i, D)\, m(Z_i; \mu, \pi)\right)^2\right] \\
&= \mathbb{E}_D\left[\sum_{i=1}^{s}\sum_{j=1}^{s} \kappa(x; Z_i, D)\, \kappa(x; Z_j, D)\, m(Z_i; \mu, \pi)\, m(Z_j; \mu, \pi)\right] \\
&= \mathbb{E}_D\left[s\kappa(x; Z_1, D)\, m^2(Z_1; \mu, \pi)\right] = \mathbb{E}_1\left[m^2(Z_1; \mu, \pi)\, s\mathbb{E}_{2:s}\left[\kappa(x; Z_1, D)\right]\right] \\
&\xrightarrow{\text{(Lem B.3)}} (\mu_1(x) - \mu_0(x))^2 + \frac{\sigma_\varepsilon^2(x)}{\pi(x)(1-\pi(x))} \quad \text{as} \quad s \to \infty
\end{aligned} \tag{D.41}$$

This gives us the desired result.

$$\Omega_s(x) \lesssim (\mu_1(x) - \mu_0(x))^2 + \frac{\overline{\sigma}_\varepsilon^2}{\mathfrak{p}(1-\mathfrak{p})} + o(1) \tag{D.42}$$

■

Lemma D.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

$$\Omega_s^c(x) = \mathbb{E}\left[h_s\left(x; Z_1, \ldots, Z_c, Z_{c+1}, \ldots, Z_s\right) \cdot h_s\left(x; Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_s\right)\right]. \tag{D.43}$$

Then,

$$\Omega_s^c(x) \lesssim \left(\mu_1(x) - \mu_0(x)\right)^2 + \frac{\overline{\sigma}_\varepsilon^2}{\mathfrak{p}(1-\mathfrak{p})} + o(1). \tag{D.44}$$

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

$$\begin{aligned}
\Omega_s^c(x) &= \mathbb{E}\left[h_s\left(x; Z_1, \ldots, Z_c, Z_{c+1}, \ldots, Z_s\right) \cdot h_s\left(x; Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_s\right)\right] \\
&= \mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{s} \kappa\left(x; Z_i, D\right) m\left(Z_i; \mu, p\right)\right)\left(\sum_{j=1}^{c} \kappa\left(x; Z_j, D'\right) m\left(Z_j; \mu, p\right) + \sum_{j=c+1}^{s} \kappa\left(x; Z'_j, D'\right) m\left(Z'_j; \mu, p\right)\right)\right] \\
&= \underbrace{\mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{c} \kappa\left(x; Z_i, D\right) m\left(Z_i; \mu, p\right)\right)\left(\sum_{j=1}^{c} \kappa\left(x; Z_j, D'\right) m\left(Z_j; \mu, p\right)\right)\right]}_{(A)} \\
&\quad + \underbrace{\mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{c} \kappa\left(x; Z_i, D\right) m\left(Z_i; \mu, p\right)\right)\left(\sum_{j=c+1}^{s} \kappa\left(x; Z'_j, D'\right) m\left(Z'_j; \mu, p\right)\right)\right]}_{(B)} \\
&\quad + \underbrace{\mathbb{E}_{D,D'}\left[\left(\sum_{i=c+1}^{s} \kappa\left(x; Z_i, D\right) m\left(Z_i; \mu, p\right)\right)\left(\sum_{j=1}^{c} \kappa\left(x; Z_j, D'\right) m\left(Z_j; \mu, p\right)\right)\right]}_{(C)} \\
&\quad + \underbrace{\mathbb{E}_{D,D'}\left[\left(\sum_{i=c+1}^{s} \kappa\left(x; Z_i, D\right) m\left(Z_i; \mu, p\right)\right)\left(\sum_{j=c+1}^{s} \kappa\left(x; Z'_j, D'\right) m\left(Z'_j; \mu, p\right)\right)\right]}_{(D)}
\end{aligned} \tag{D.45}$$

Considering these terms one by one, we can make the following observations.

$$\begin{aligned}
(A) &= \mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{c} \kappa\left(x; Z_i, D\right) m\left(Z_i; \mu, p\right)\right)\left(\sum_{j=1}^{c} \kappa\left(x; Z_j, D'\right) m\left(Z_j; \mu, p\right)\right)\right] \\
&= \mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=1}^{c} \kappa\left(x; Z_i, D\right) \kappa\left(x; Z_j, D'\right) m\left(Z_i; \mu, p\right) m\left(Z_j; \mu, p\right)\right] \\
&= \mathbb{E}_1\left[m^2\left(Z_1; \mu, p\right) c\mathbb{E}_{2:s}\left[\kappa\left(x; Z_1, D\right)\kappa\left(x; Z_1, D'\right)\right]\right] \leq (c/s) \cdot \mathbb{E}_1\left[m^2\left(Z_1; \mu, p\right) s\mathbb{E}_{2:s}\left[\kappa\left(x; Z_1, D\right)\right]\right] \\
&\overset{\text{(Lem B.3)}}{\lesssim} (c/s)\left[\left(\mu_1(x) - \mu_0(x)\right)^2 + \frac{\sigma_\varepsilon^2(x)}{\pi(x)\left(1 - \pi(x)\right)}\right] + o(1)
\end{aligned} \tag{D.46}$$

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

$$\begin{aligned}
(B) &= \mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{c}\kappa\left(x;Z_i,D\right)m\left(Z_i;\mu,\pi\right)\right)\left(\sum_{j=c+1}^{s}\kappa\left(x;Z'_j,D'\right)m\left(Z'_j;\mu,\pi\right)\right)\right] \\
&= \mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=c+1}^{s}\kappa\left(x;Z_i,D\right)\kappa\left(x;Z'_j,D'\right)m\left(Z_i;\mu,\pi\right)m\left(Z'_j;\mu,\pi\right)\right] \\
&= \mathbb{E}_{D,D'}\left[c(s-c)\kappa\left(x;Z_1,D\right)\kappa\left(x;Z'_{c+1},D'\right)m\left(Z_1;\mu,\pi\right)m\left(Z'_{c+1};\mu,\pi\right)\right] \\
&\leq \mathbb{E}_{D}\left[c\kappa\left(x;Z_1,D\right)\left|m\left(Z_1;\mu,\pi\right)\right|\right]\mathbb{E}_{D'}\left[(s-c)\kappa\left(x;Z'_{c+1},D'\right)\left|m\left(Z'_{c+1};\mu,\pi\right)\right|\right] \\
&= \frac{c(s-c)}{s^2}\cdot\left(\mathbb{E}_1\left[\left|m\left(Z_1;\mu,\pi\right)\right|s\mathbb{E}_{2:s}\left[\kappa\left(x;Z_1,D\right)\right]\right]\right)^2 \\
&\lesssim \frac{c(s-c)}{s^2}\left(\mu_1\left(x\right)-\mu_0\left(x\right)\right)^2+o(1)
\end{aligned} \tag{D.47}$$

Applying the same principles to the third term we find a similar result.

$$\begin{aligned}
(C) &= \mathbb{E}_{D,D'}\left[\left(\sum_{i=c+1}^{s}\kappa\left(x;Z_i,D\right)m\left(Z_i;\mu,p\right)\right)\left(\sum_{j=1}^{c}\kappa\left(x;Z_j,D'\right)m\left(Z_j;\mu,p\right)\right)\right] \\
&\lesssim \frac{c(s-c)}{s^2}\left(\mu_1\left(x\right)-\mu_0\left(x\right)\right)^2+o(1)
\end{aligned} \tag{D.48}$$

$$\begin{aligned}
(D) &= \mathbb{E}_{D,D'}\left[\left(\sum_{i=c+1}^{s}\kappa\left(x;Z_i,D\right)m\left(Z_i;\mu,p\right)\right)\left(\sum_{j=c+1}^{s}\kappa\left(x;Z'_j,D'\right)m\left(Z'_j;\mu,p\right)\right)\right] \\
&= \mathbb{E}_{D,D'}\left[(s-c)^2\kappa\left(x;Z_{c+1},D\right)\kappa\left(x;Z'_{c+1},D'\right)m\left(Z_{c+1};\mu,p\right)m\left(Z'_{c+1};\mu,p\right)\right] \\
&\leq \frac{(s-c)^2}{s^2}\cdot\mathbb{E}_{D}\left[\left|m\left(Z_{c+1};\mu,p\right)\right|s\kappa\left(x;Z_{c+1},D\right)\right]\mathbb{E}_{D'}\left[\left|m\left(Z'_{c+1};\mu,p\right)\right|s\kappa\left(x;Z'_{c+1},D'\right)\right] \\
&= \frac{(s-c)^2}{s^2}\left(\mathbb{E}_1\left[\left|m\left(Z_1;\mu,p\right)\right|s\mathbb{E}_{2:s}\left[\kappa\left(x;Z_{c+1},D\right)\right]\right]\right)^2 \\
&\lesssim \frac{(s-c)^2}{s^2}\left(\mu_1\left(x\right)-\mu_0\left(x\right)\right)^2+o(1)
\end{aligned} \tag{D.49}$$

Thus, we find the desired result.

$$\begin{aligned}
\Omega_s^c\left(x\right) &= (A)+(B)+(C)+(D) \\
&\lesssim \frac{c}{s}\left[\left(\mu_1\left(x\right)-\mu_0\left(x\right)\right)^2+\frac{\overline{\sigma}_{\varepsilon}^2}{\mathfrak{p}\left(1-\mathfrak{p}\right)}\right]+2\frac{c(s-c)}{s^2}\left(\mu_1\left(x\right)-\mu_0\left(x\right)\right)^2+\frac{(s-c)^2}{s^2}\left(\mu_1\left(x\right)-\mu_0\left(x\right)\right)^2+o(1) \\
&= \left[\frac{cs+2c(s-c)+(s-c)^2}{s^2}\right]\left(\mu_1\left(x\right)-\mu_0\left(x\right)\right)^2+\frac{c}{s}\frac{\overline{\sigma}_{\varepsilon}^2}{\mathfrak{p}\left(1-\mathfrak{p}\right)}+o(1) \\
&\lesssim \left(\mu_1\left(x\right)-\mu_0\left(x\right)\right)^2+\frac{\overline{\sigma}_{\varepsilon}^2}{\mathfrak{p}\left(1-\mathfrak{p}\right)}+o(1)
\end{aligned} \tag{D.50}$$

■

Lemma D.13.
Let $D = \{Z_1, \ldots, Z_{s_2}\}$ be a vector of i.i.d. random variables drawn from Q for $s_2 > s_1$. Furthermore, let

$$\Upsilon_{s_1,s_2}(x) = \mathbb{E}\left[h_{s_1}(x; Z_1, \ldots, Z_{s_1}) \cdot h_{s_2}(x; Z_1, \ldots, Z_{s_1}, \ldots, Z_{s_2})\right]. \tag{D.51}$$

Then,

$$\Upsilon_{s_1,s_2}(x) \lesssim 2\left(\mu_1(x) - \mu_0(x)\right)^2 + \frac{\overline{\sigma}_\varepsilon^2}{\mathfrak{p}(1-\mathfrak{p})} + o(1) \quad \textit{as} \quad s_1, s_2 \to \infty \quad \textit{with} \quad 0 < \mathfrak{c} \leq s_1/s_2 \leq 1 - \mathfrak{c} < 1. \tag{D.52}$$

Proof of Lemma D.13. Consider first the following.

$$\begin{aligned}
\Upsilon_{s_1,s_2}(x) &= \mathbb{E}\left[h_{s_1}(x; Z_1, \ldots, Z_{s_1}) \cdot h_{s_2}(x; Z_1, \ldots, Z_{s_1}, \ldots, Z_{s_2})\right] \\
&= \mathbb{E}_D\left[\left(\sum_{i=1}^{s_1} \kappa(x; Z_i, D_{[s_1]}) m(Z_i; \mu, p)\right)\left(\sum_{j=1}^{s_1} \kappa(x; Z_j, D) m(Z_j; \mu, p) + \sum_{j=s_1+1}^{s_2} \kappa(x; Z_j, D) m(Z_j; \mu, p)\right)\right] \\
&= \underbrace{\mathbb{E}_D\left[\sum_{i=1}^{s_1}\sum_{j=1}^{s_1} \kappa(x; Z_i, D_{[s_1]}) \kappa(x; Z_j, D) m(Z_i; \mu, p)\, m(Z_j; \mu, p)\right]}_{(A)} \\
&\quad + \underbrace{\mathbb{E}_D\left[\sum_{i=1}^{s_1}\sum_{j=s_1+1}^{s_2} \kappa(x; Z_i, D_{[s_1]}) \kappa(x; Z_j, D) m(Z_i; \mu, p)\, m(Z_j; \mu, p)\right]}_{(B)}
\end{aligned} \tag{D.53}$$

Using this decomposition, we can make the following findings.

$$\begin{aligned}
(A) &= \mathbb{E}_D\left[\sum_{i=1}^{s_1}\sum_{j=1}^{s_1} \kappa(x; Z_i, D_{[s_1]}) \kappa(x; Z_j, D) m(Z_i; \mu, p)\, m(Z_j; \mu, p)\right] \\
&= \mathbb{E}_D\left[\sum_{i=1}^{s_1} \kappa(x; Z_i, D_{[s_1]}) m^2(Z_i; \mu, p)\right] = \mathbb{E}_D\left[m^2(Z_1; \mu, p)\, s_1 \kappa(x; Z_1, D_{[s_1]})\right] \\
&= \mathbb{E}_1\left[m^2(Z_1; \mu, p)\, s_1 \mathbb{E}_{2:s_2}\left[\kappa(x; Z_1, D_{[s_1]})\right]\right] = \mathbb{E}_1\left[m^2(Z_1; \mu, p)\, s_1 \mathbb{E}_{2:s_1}\left[\kappa(x; Z_1, D_{[s_1]})\right]\right] \\
&\overset{\text{(Lem B.3)}}{\lesssim} \left(\mu_1(x) - \mu_0(x)\right)^2 + \frac{\overline{\sigma}_\varepsilon^2}{\mathfrak{p}(1-\mathfrak{p})} + o(1)
\end{aligned} \tag{D.54}$$

$$\begin{aligned}
(B) &= \mathbb{E}_D\left[\sum_{i=1}^{s_1}\sum_{j=s_1+1}^{s_2} \kappa(x; Z_i, D_{[s_1]}) \kappa(x; Z_j, D) m(Z_i; \mu, p)\, m(Z_j; \mu, p)\right] \\
&= \mathbb{E}_D\left[s_1(s_2 - s_1) \kappa(x; Z_1, D_{[s_1]}) \kappa(x; Z_{s_2}, D) m(Z_1; \mu, p)\, m(Z_{s_2}; \mu, p)\right] \\
&\leq \frac{(s_2 - s_1)}{s_2} \mathbb{E}_D\left[\left|m(Z_1; \mu, p)\right| s_1 \kappa(x; Z_1, D_{[s_1]})\right] \mathbb{E}_D\left[\left|m(Z_{s_2}; \mu, p)\right| s_2 \kappa(x; Z_{s_2}, D)\right] \\
&\lesssim \frac{(s_2 - s_1)}{s_2} \left(\mu_1(x) - \mu_0(x)\right)^2 + o(1)
\end{aligned} \tag{D.55}$$

Thus, we obtain the desired result.

$$\Upsilon_{s_1,s_2}(x) = (A) + (B) \lesssim 2\left(\mu_1(x) - \mu_0(x)\right)^2 + \frac{\overline{\sigma}_\varepsilon^2}{\mathfrak{p}(1-\mathfrak{p})} + o(1) \tag{D.56}$$

■

Lemma D.14.
Let $D = \{Z_1, \ldots, Z_{s_2}\}$ be a vector of i.i.d. random variables drawn from Q 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

$$\Upsilon^c_{s_1,s_2}(x) = \mathbb{E}\left[h_{s_1}\left(x; Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_{s_1}\right) \cdot h_{s_2}\left(x; Z_1, \ldots, Z_{s_2}\right)\right]. \tag{D.57}$$

Then,

$$\begin{aligned} &\Upsilon^c_{s_1,s_2}(x) \lesssim 4\left(\mu_1(x) - \mu_0(x)\right)^2 + \frac{\sigma_\varepsilon^2(x)}{\mathfrak{p}(1-\mathfrak{p})} + o(1) \\ &\textit{for} \quad s_1, s_2 \quad \textit{sufficiently large} \quad \textit{with} \quad 0 < \mathfrak{c} \leq s_1/s_2 \leq 1 - \mathfrak{c} < 1. \end{aligned} \tag{D.58}$$

Proof of Lemma D.14.

$$\begin{aligned} \Upsilon^c_{s_1,s_2}(x) &= \mathbb{E}\left[h_{s_1}\left(x; Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_{s_1}\right) \cdot h_{s_2}\left(x; Z_1, \ldots, Z_{s_2}\right)\right] \\ &= \mathbb{E}_{D,D'}\left[\left(\sum_{i=1}^{c} \kappa(x; Z_i, D'_{[s_1]}) m\left(Z_i; \mu, p\right) + \sum_{i=c+1}^{s_1} \kappa(x; Z'_i, D'_{[s_1]}) m\left(Z'_i; \mu, p\right)\right)\left(\sum_{j=1}^{s_2} \kappa(x; Z_j, D) m\left(Z_j; \mu, p\right)\right)\right] \\ &= \underbrace{\mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=1}^{c} \kappa(x; Z_i, D'_{[s_1]})\kappa(x; Z_j, D) m\left(Z_i; \mu, p\right) m\left(Z_j; \mu, p\right)\right]}_{(A)} \\ &\quad + \underbrace{\mathbb{E}_{D,D'}\left[\sum_{i=1}^{c}\sum_{j=c+1}^{s_2} \kappa(x; Z_i, D'_{[s_1]})\kappa(x; Z_j, D) m\left(Z_i; \mu, p\right) m\left(Z_j; \mu, p\right)\right]}_{(B)} \\ &\quad + \underbrace{\mathbb{E}_{D,D'}\left[\sum_{i=c+1}^{s_1}\sum_{j=1}^{c} \kappa(x; Z'_i, D'_{[s_1]})\kappa(x; Z_j, D) m\left(Z'_i; \mu, p\right) m\left(Z_j; \mu, p\right)\right]}_{(C)} \\ &\quad + \underbrace{\mathbb{E}_{D,D'}\left[\sum_{i=c+1}^{s_1}\sum_{j=c+1}^{s_2} \kappa(x; Z'_i, D'_{[s_1]})\kappa(x; Z_j, D) m\left(Z'_i; \mu, p\right) m\left(Z_j; \mu, p\right)\right]}_{(D)} \end{aligned} \tag{D.59}$$

Now, considering the terms individually, we find the following.

$$\begin{aligned}(A) &= \mathbb{E}_{D,D'}\left[c\kappa(x;Z_1,D'_{[s_1]})\kappa(x;Z_1,D)m^2\left(Z_1;\mu,p\right)\right] = \frac{c}{s_2}\cdot\mathbb{E}_1\left[m^2\left(Z_1;\mu,p\right)s_2\mathbb{E}_{2:s_2}\left[\kappa(x;Z_1,D'_{[s_1]})\kappa(x;Z_1,D)\right]\right]\\ &\leq \frac{c}{s_2}\cdot\mathbb{E}_1\left[m^2\left(Z_1;\mu,p\right)s_2\mathbb{E}_{2:s_2}\left[\kappa(x;Z_1,D)\right]\right] \lesssim \frac{c}{s_2}\left(\left(\mu_1(x)-\mu_0(x)\right)^2+\frac{\sigma_\varepsilon^2(x)}{\mathfrak{p}(1-\mathfrak{p})}\right)+o(1)\end{aligned} \tag{D.60}$$

Similarly, we find the following.

$$\begin{aligned}(B) &= \mathbb{E}_{D,D'}\left[c(s_2-c)\kappa(x;Z_1,D'_{[s_1]})\kappa(x;Z_{c+1},D)m\left(Z_1;\mu,p\right)m\left(Z_{c+1};\mu,p\right)\right]\\ &= \frac{c(s_2-c)}{s_1s_2}\mathbb{E}_{D,D'}\left[m\left(Z_1;\mu,p\right)m\left(Z_{c+1};\mu,p\right)s_1s_2\kappa(x;Z_1,D'_{[s_1]})\kappa(x;Z_{c+1},D)\right]\\ &\leq \frac{c(s_2-c)}{s_1s_2}\mathbb{E}_{D,D'}\left[\left|m\left(Z_1;\mu,p\right)\right|s_1\kappa(x;Z_1,D'_{[s_1]})\right]\mathbb{E}_{D,D'}\left[\left|m\left(Z_{c+1};\mu,p\right)\right|s_2\kappa(x;Z_{c+1},D)\right]\\ &\lesssim \frac{c(s_2-c)}{s_1s_2}\left(\mu_1(x)-\mu_0(x)\right)^2+o(1)\end{aligned} \tag{D.61}$$

Applying the same argument to the third term, we find an analogous result.

$$\begin{aligned}(C) &= \mathbb{E}_{D,D'}\left[(s_1-c)c\kappa(x;Z'_{c+1},D'_{[s_1]})\kappa(x;Z_1,D)m\left(Z'_{c+1};\mu,p\right)m\left(Z_1;\mu,p\right)\right]\\ &\lesssim \frac{c(s_1-c)}{s_1s_2}\left(\mu_1(x)-\mu_0(x)\right)^2+o(1)\end{aligned} \tag{D.62}$$

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

$$\begin{aligned}(D) &= \mathbb{E}_{D,D'}\left[(s_1-c)(s_2-c)\kappa(x;Z'_{c+1},D'_{[s_1]})\kappa(x;Z_{c+1},D)m\left(Z'_{c+1};\mu,p\right)m\left(Z_{c+1};\mu,p\right)\right]\\ &= \frac{(s_1-c)(s_2-c)}{s_1s_2}\mathbb{E}_{D,D'}\left[m\left(Z'_{c+1};\mu,p\right)m\left(Z_{c+1};\mu,p\right)s_1s_2\kappa(x;Z'_{c+1},D'_{[s_1]})\kappa(x;Z_{c+1},D)\right]\\ &\lesssim \frac{(s_1-c)(s_2-s_1)}{s_1s_2}\left(\mu_1(x)-\mu_0(x)\right)^2+o(1)\end{aligned} \tag{D.63}$$

By combining these asymptotic bounds, we find the desired result.

$$\Upsilon^c_{s_1,s_2}(x) = (A)+(B)+(C)+(D) \lesssim 4\left(\mu_1(x)-\mu_0(x)\right)^2+\frac{\sigma_\varepsilon^2(x)}{\mathfrak{p}(1-\mathfrak{p})}+o(1) \tag{D.64}$$

■

D.6 Variance Estimator Consistency Theorems

Lemma D.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.

$$\frac{s}{n}\left(\frac{\zeta_s^s}{s\zeta_s^1}-1\right)\rightarrow 0 \tag{D.65}$$

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

$$\frac{n}{s^2}\frac{Var\left(\mathrm{U}_s\left(\mathbf{D}_{[n]}\right)\right)}{\zeta_s^1}\rightarrow 1. \tag{D.66}$$

Proof.

$$\begin{aligned}
1 \leq \frac{n}{s^2}\frac{\mathrm{Var}\left(\mathrm{U}_s\left(\mathbf{D}_{[n]}\right)\right)}{\zeta_s^1} &= \left(\frac{s^2}{n}\zeta_s^1\right)^{-1}\sum_{j=1}^{s}\binom{s}{j}^2\binom{n}{j}^{-1}V_s^j \\
&\leq 1+\left(\frac{s^2}{n}\zeta_s^1\right)^{-1}\frac{s^2}{n^2}\sum_{j=2}^{s}\binom{s}{j}V_s^j \\
&\leq 1+\frac{s}{n}\left(\frac{\zeta_s^s}{s\zeta_s^1}-1\right)\rightarrow 1.
\end{aligned} \tag{D.67}$$

■

Lemma D.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 D.15.*

Proof. Recall the results from Lemmas D.9 and D.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.

$$\frac{s_2}{n}\left(\frac{\zeta_{s_1,s_2}^{s_2}(x)}{s_2\zeta_{s_1,s_2}^{1}(x)}-1\right)\sim\frac{s_2}{n}\left(\mu^2(x)+\sigma_\varepsilon+o(1)-1\right)\sim\frac{s_2}{n}\rightarrow 0 \tag{D.68}$$

■

Proof of Theorem 4.2.
The desired result immediately follows from an application of Theorem 6 from Peng, Mentch, and Stefanski (2021). ■

Proof of Theorem 4.3.
Recall the definition of the Jackknife Variance estimator.

$$\widehat{\omega}^2_{JK}\left(x;\mathbf{D}_n\right) = \frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\mu}_{s_1,s_2}\left(x;\mathbf{D}_{n,-i}\right) - \widehat{\mu}_{s_1,s_2}\left(x;\mathbf{D}_n\right)\right)^2 \tag{D.69}$$

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

$$\begin{aligned}
\widehat{\omega}^2_{JK}\left(x;\mathbf{D}_n\right) &= \frac{n-1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{s_2}\binom{s_2}{j}H^j_{s_1,s_2}\left(\mathbf{D}_{n,-i}\right) - \sum_{j=1}^{s_2}\binom{s_2}{j}H^j_{s_1,s_2}\right)^2 \\
&= \frac{n-1}{n}\sum_{j=1}^{n}\left(\sum_{j=1}^{s_2}\binom{s_2}{j}\left(H^j_{s_1,s_2} - H^j_{s_1,s_2}\left(\mathbf{D}_{n,-i}\right)\right)\right)^2 \\
&= \frac{n-1}{n}\sum_{j=1}^{n}\left(\sum_{j=1}^{s_2}\binom{s_2}{j}\left(\binom{n}{j}^{-1}\sum_{\iota\in L_{n,j}}h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota) - \binom{n-1}{j}^{-1}\sum_{\ell\in L_j([n]\setminus\{i\})}h^{(j)}_{s_1,s_2}(\mathbf{D}_\ell)\right)\right)^2 \\
&= \frac{n-1}{n}\sum_{j=1}^{n}\left[\frac{s_2}{n}h^{(1)}_{s_1,s_2}(Z_i) + \sum_{j\neq i}\left(\frac{s_2}{n} - \frac{s_2}{n-1}\right)h^{(1)}_{s_1,s_2}(Z_j)\right. \\
&\quad\left. + \sum_{j=2}^{s_2}\binom{s_2}{j}\left(\binom{n}{j}^{-1}\sum_{\iota\in L_{n,j}}h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota) - \binom{n-1}{j}^{-1}\sum_{\ell\in L_j([n]\setminus\{i\})}h^{(j)}_{s_1,s_2}(\mathbf{D}_\ell)\right)\right]^2 \\
&= \frac{n-1}{n}\frac{s^2}{n^2}\sum_{j=1}^{n}\left[h^{(1)}_{s_1,s_2}(Z_i) - \frac{1}{n-1}\sum_{j\neq i}h^{(1)}_{s_1,s_2}(Z_j)\right. \\
&\quad\left. + \frac{n}{s}\sum_{j=2}^{s_2}\binom{s_2}{j}\left(\binom{n}{j}^{-1}\sum_{\iota\in L_{j-1}([n]\setminus\{i\})}h^{(j)}_{s_1,s_2}(\mathbf{D}_{\iota\cup\{i\}}) + \left[\binom{n}{j}^{-1} - \binom{n-1}{j}^{-1}\right]\sum_{\ell\in L_j([n]\setminus\{i\})}h^{(j)}_{s_1,s_2}(\mathbf{D}_\ell)\right)\right] \\
&=: \frac{n-1}{n}\frac{s^2}{n^2}\sum_{j=1}^{n}\left[h^{(1)}_{s_1,s_2}(Z_i) + T_i\right]^2
\end{aligned} \tag{D.70}$$

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.

$$\mathbb{E}\left[\left(h^{(1)}_{s_1,s_2}(Z_i)\right)^2\right] = V^1_{s_1,s_2} = \zeta^1_{s_1,s_2} \tag{D.71}$$

Furthermore, as a consequence of the independence of the observations and the uncorrelatedness of Hoeffding projections 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}{(n-es_2)^2}\zeta^{s_2}_{s_1,s_2} \\
&= \left(\frac{1}{n-1} + \frac{2es_2n}{(n-es_2)^2}\right)\zeta^1_{s_1,s_2} + 2e\left(\frac{1}{n} + \frac{n}{(n-es_2)^2}\right)\left(\zeta^{s_2}_{s_1,s_2} - s_2\zeta^1_{s_1,s_2}\right)
\end{aligned}
\tag{D.72}
$$

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

$$
\zeta^{s_2}_{s_1,s_2}(x) \lesssim \mu^2(x) + \sigma_\varepsilon + o(1) \quad \text{and} \quad \zeta^1_{s_1,s_2}(x) \sim s_2^{-1} \tag{D.73}
$$

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.

$$\frac{\mathbb{E}\left[T_i^2\right]}{V_{s_1,s_2}^1} \leq \frac{\left(\frac{1}{n-1} + \frac{2es_2n}{(n-es_2)^2}\right)\zeta_{s_1,s_2}^1 + 2e\left(\frac{1}{n} + \frac{n}{(n-es_2)^2}\right)\left(\zeta_{s_1,s_2}^{s_2} - s_2\zeta_{s_1,s_2}^1\right)}{\zeta_{s_1,s_2}^1}$$
$$= \frac{1}{n-1} + \frac{2es_2n}{(n-es_2)^2} + 2e\left(\frac{1}{n} + \frac{n}{(n-es_2)^2}\right)\left(\frac{\zeta_{s_1,s_2}^{s_2} - s_2\zeta_{s_1,s_2}^1}{\zeta_{s_1,s_2}^1}\right) \to 0 \tag{D.74}$$

Therefore, we can conclude that $h_s^{(1)}(Z_i)$ dominates T_i^2 in the expression of interest. Using Lemma B.5, we can thus conclude the following.

$$\frac{\frac{n}{s_2^2}\widehat{\omega}_{JK}^2(x;\mathbf{D}_n)}{V_{s_1,s_2}^1(x)} \to_p \frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n}\frac{\left(h_{s_1,s_2}^{(1)}(x;Z_i)\right)^2}{V_{s_1,s_2}^1(x)}$$
$$\to_p \frac{n-1}{n}\frac{\mathbb{E}\left[\left(h_{s_1,s_2}^{(1)}(x;Z_i)\right)^2\right]}{V_{s_1,s_2}^1(x)} \to 1 \tag{D.75}$$

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

Proof of Theorem 4.4.
Consider first the case absent additional randomization in the form of ω and recall the definition of the delete-d Jackknife Variance estimator.

$$\widehat{\omega}^2_{JKD}\left(x;d,\mathbf{D}_n\right)=\frac{n-d}{d}\binom{n}{d}^{-1}\sum_{\ell\in L_{n,d}}\left(\widehat{\mu}_{s_1,s_2}\left(x;\mathbf{D}_{n,-\ell}\right)-\widehat{\mu}_{s_1,s_2}\left(x;\mathbf{D}_n\right)\right)^2 \tag{D.76}$$

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

$$\begin{aligned}
\widehat{\omega}^2_{JKD}\left(x;d,\mathbf{D}_n\right)&=\frac{n-d}{d}\binom{n}{d}^{-1}\sum_{\ell\in L_{n,d}}\left(\sum_{j=1}^{s_2}\binom{s_2}{j}\left(H^j_{P_t}-H^j_{P_t}\left(\mathbf{D}_{n,-\ell}\right)\right)\right)^2\\
&=\frac{n-d}{d}\binom{n}{d}^{-1}\sum_{\ell\in L_{n,d}}\left(\sum_{j=1}^{s_2}\binom{s_2}{j}\left(\binom{n}{j}^{-1}\sum_{\iota\in L_{n,j}}h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota)-\binom{n-d}{j}^{-1}\sum_{\iota\in L_j([n]\setminus\ell)}h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota)\right)\right)^2\\
&=\frac{n-d}{d}\binom{n}{d}^{-1}\sum_{\ell\in L_{n,d}}\left[\frac{s_2}{n}\sum_{i\in\ell}h^{(1)}_{s_1,s_2}(Z_i)+\sum_{i\in[n]\setminus\ell}\left(\frac{s_2}{n}-\frac{s_2}{n-d}\right)h^{(1)}_{s_1,s_2}(Z_i)\right.\\
&\qquad\left.+\sum_{j=2}^{s_2}\binom{s_2}{j}\left(\binom{n}{j}^{-1}\sum_{\iota\in L_{n,j}}h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota)-\binom{n-d}{j}^{-1}\sum_{\iota\in L_j([n]\setminus\ell)}h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota)\right)\right]^2\\
&=\frac{n-d}{d}\binom{n}{d}^{-1}\left(\frac{s_2}{n}\right)^2\sum_{\ell\in L_{n,d}}\left[\sum_{i\in\ell}h^{(1)}_{s_1,s_2}(Z_i)-\frac{d}{n-d}\sum_{i\in[n]\setminus\ell}h^{(1)}_{s_1,s_2}(Z_i)\right.\\
&\qquad\left.+\frac{n}{s_2}\sum_{j=2}^{s_2}\binom{s_2}{j}\left(\binom{n}{j}^{-1}\sum_{\iota\in L_{n,j}}h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota)-\binom{n-d}{j}^{-1}\sum_{\iota\in L_j([n]\setminus\ell)}h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota)\right)\right]^2\\
&=:(n-d)\binom{n}{d}^{-1}\left(\frac{s_2}{n}\right)^2\sum_{\ell\in L_{n,d}}\left[\frac{1}{\sqrt{d}}\sum_{i\in\ell}h^{(1)}_{s_1,s_2}(Z_i)+T_\ell\right]^2
\end{aligned} \tag{D.77}$$

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 B.5. Luckily, since Lemma B.5 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 4.3. Thus, consider the following for an arbitrary fixed index-subset ℓ with cardinality d.

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{\sqrt{d}}\sum_{i\in\ell}h^{(1)}_{s_1,s_2}(Z_i)\right)^2\right]&=\frac{1}{d}\mathbb{E}\left[\sum_{i\in\ell}\sum_{j\in\ell}h^{(1)}_{s_1,s_2}(Z_i)h^{(1)}_{s_1,s_2}(Z_j)\right]=\frac{1}{d}\sum_{i\in\ell}\sum_{j\in\ell}\mathbb{E}\left[h^{(1)}_{s_1,s_2}(Z_i)h^{(1)}_{s_1,s_2}(Z_j)\right]\\
&=\frac{|\ell|}{d}\cdot\mathbb{E}\left[\left(h^{(1)}_{s_1,s_2}(Z_1)\right)^2\right]=\zeta_{P_t,1}
\end{aligned} \tag{D.78}$$

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.

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

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

$$
\begin{aligned}
T_\ell = &\frac{\sqrt{d}}{n-d} \sum_{i\in[n]\setminus\ell} h^{(1)}_{s_1,s_2}(Z_i) \\
&+ \frac{n}{s_2\sqrt{d}} \sum_{j=2}^{s_2} \binom{s_2}{j} \left(\binom{n}{j}^{-1} \left(\sum_{a=1}^{j} \sum_{\substack{\varkappa\in L_a(\ell)\\ \varrho\in L_{j-a}([n]\setminus\ell)}} h^{(j)}_{s_1,s_2}(D_{\varkappa\cup\varrho}) \right) + \left(\binom{n}{j}^{-1} - \binom{n-d}{j}^{-1} \right) \sum_{\iota\in L_j([n]\setminus\ell)} h^{(j)}_{s_1,s_2}(\mathbf{D}_\iota) \right)
\end{aligned}
\tag{D.80}
$$

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

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

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)}{\le} \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] \\
&\le \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] \\
&\le \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{D.82}
$$

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.

$$
\begin{aligned}
\mathbb{E}\left[(T_\ell)^2\right] &\lesssim \frac{d}{n-d}V^1_{s_1,s_2} + \frac{n}{s_2 d}\sum_{j=2}^{s_2}\left(\frac{e(s_2-1)}{n-1}\right)^{j-1} \frac{j(n-2)}{(n-d-s_2+1)(n-d-2s_t)}\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&= \frac{d}{n-d}V^1_{s_1,s_2} + \frac{2e\cdot n(n-2)}{(n-1)(n-d-s_2+1)(n-d-2s_t)d}\sum_{j=2}^{s_2}\frac{j}{s_2}\left(\frac{e(s_2-1)}{n-1}\right)^{j-1}\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&\lesssim \frac{d}{n-d}V^1_{s_1,s_2} + \frac{4e}{(n-d-s_2)s_2 d}\sum_{j=2}^{s_2} j\left(\frac{e(s_2-1)}{n-1}\right)^{j-1}\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&\leq \frac{d}{n-d}V^1_{s_1,s_2} + \frac{4e}{(n-d-s_2)s_2 d}\sum_{j=2}^{s_2}\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] + \frac{4e}{(n-d-s_2)s_2 d}\sum_{j=2}^{s_2}(j-1)\left(\frac{e(s_2-1)}{n-1}\right)^{j-1}\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] \\
&\leq \frac{d}{n-d}V^1_{s_1,s_2} + \frac{4e}{(n-d-s_2)s_2 d}\sum_{j=2}^{s_2}\left[\binom{s_2}{j}V^j_{s_1,s_2}\right] + \frac{4e\cdot\zeta^{s_2}_{s_1,s_2}}{(n-d-s_2)s_2 d}\sum_{j=1}^{\infty} j\left(\frac{e(s_2-1)}{n-1}\right)^{j} \\
&= \frac{d}{n-d}\zeta^1_{s_1,s_2} + \frac{4e}{(n-d-s_2)s_2 d}\left(\zeta^{s_2}_{s_1,s_2} - s_2\zeta^1_{s_1,s_2}\right) + \frac{4e\cdot\zeta^{s_2}_{s_1,s_2}}{(n-d-s_2)s_2 d}\cdot\frac{e(s_2-1)(n-1)}{(n-1-e(s_2-1))^2} \\
&= \left(\frac{d}{n-d} + \frac{e(s_2-1)(n-1)}{(n-1-e(s_2-1))^2}\right)\zeta^1_{s_1,s_2} + \frac{4e}{(n-d-s_2)s_2 d}\left(1 + \frac{e(s_2-1)(n-1)}{(n-1-e(s_2-1))^2}\right)\left(\zeta^{s_2}_{s_1,s_2} - s_2\zeta^1_{s_1,s_2}\right)
\end{aligned}
\tag{D.83}
$$

We continue as in the default Jackknife case.

$$
\begin{aligned}
\frac{\mathbb{E}\left[T_\ell^2\right]}{V^1_{s_1,s_2}} &\leq \frac{d}{n-d} + \frac{e(s_2-1)(n-1)}{(n-1-e(s_2-1))^2} + \frac{4e}{(n-d-s_2)s_2 d}\left(1 + \frac{e(s_2-1)(n-1)}{(n-1-e(s_2-1))^2}\right)\frac{\zeta^{s_2}_{s_1,s_2} - s_2\zeta^1_{s_1,s_2}}{\zeta^1_{s_1,s_2}} \\
&\to 0.
\end{aligned}
\tag{D.84}
$$

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. ■

D.7 Pointwise Inference Results

Proof of Theorem 4.6.

■

Proof of Theorem 4.7.

■

E Proofs for Results in Section 5