Causal-DRF: Conditional Kernel Treatment Effect Estimation Using Distributional Random Forest

Jeffrey Näf
University of Geneva

Junhyung Park
ETH Zürich

Herbert Susmann
NYU Grossman School of Medicine

Abstract

The conditional average treatment effect (CATE) is a commonly targeted statistical parameter for measuring the effect of a treatment conditional on covariates. However, the CATE will fail to capture effects of treatments beyond differences in conditional expectations. Inspired by causal forests for CATE estimation, we develop a forest-based method to estimate the conditional kernel treatment effect (CKTE), based on the recently introduced Distributional Random Forest (DRF) algorithm. Adapting the splitting criterion of DRF, we show how one forest fit can be used to obtain a consistent and asymptotically normal estimator of the CKTE, as well as an approximation of its sampling distribution. This allows to study the difference in distribution between control and treatment group and thus yields a more comprehensive understanding of the treatment effect.

1 INTRODUCTION

The definition, identification and estimation of treatment effects have gained increased attention in statistics and machine learning. Consider random variables $(Y^1, Y^0, W, \mathbf{X})$, where Y^1, Y^0 correspond to potential outcomes of treatment and control, respectively (Rubin, 2005), W is the treatment assignment, and $\mathbf{X}$ additional covariates. The so-called "fundamental problem of causal inference" (Holland, 1986) is that only one of the potential outcomes is observed; that is, one only observes $Y = WY^1 + (1-W)Y^0$. For example, treatment W could be thought of as indicating the prescription of a drug treatment ($W = 1$), Y a measure of health such as blood pressure, and $\mathbf{X}$ consisting of individual baseline covariates. A popular choice to investigate the effectiveness of the treatment given $\mathbf{X} = \mathbf{x}$ is the conditional average treatment effect (CATE), defined as $\mathbb{E}[Y^1 - Y^0 \mid \mathbf{X} = \mathbf{x}]$. Under the assumption of no unmeasured confounding, the CATE is identifiable in terms of the observed data $(Y, W, \mathbf{X})$. The CATE is an attractive target, as its conditional structure contains information about treatment heterogeneity across population subgroups, as opposed to the average treatment effect (ATE), which averages treatment effects over the entire population. However, the CATE is limited in that a treatment might affect the outcome beyond just its conditional expectation. For example, giving a drug to several new patients with the same covariates $\mathbf{x}$ could decrease blood pressure on average, but also lead to an undesirable increase in variance. Therefore, it is natural to consider more general measures of differences between treatment and control distributions to obtain a comprehensive picture of the treatment effect. Previous work on causal estimands that go beyond identifying differences in expectations focused, for example, on estimating cumulative distribution functions of the potential outcomes (Chernozhukov et al., 2013, 2020), or on estimating the difference in treatment and control groups of specified distributional parameters, such as the quantile treatment effect (Hohberg et al., 2020; Park et al., 2021). Recently, (Kallus and Oprescu, 2023) provided a unified doubly-robust approach to estimate a variety of distributional parameters to compare treatment and control groups.

In this paper, we study the conditional kernel treatment effect (CKTE). The CKTE can be seen as an analog to the CATE in a reproducing kernel Hilbert space (RKHS). For a reproducing kernel k with associated Hilbert space $\mathcal{H}$ and a potential multivariate outcome $\mathbf{Y}^1, \mathbf{Y}^0$, the CKTE is defined as

$$\tau_k(\mathbf{x}) = \mathbb{E}[k(\mathbf{Y}^1, \cdot) \mid \mathbf{X} = \mathbf{x}] - \mathbb{E}[k(\mathbf{Y}^0, \cdot) \mid \mathbf{X} = \mathbf{x}], \tag{1}$$

see e.g., (Park et al., 2021). If the outcome is univariate and we choose the trivial kernel $k(y_1, y_2) = y_1 y_2$, then $\tau_k(\mathbf{x})$ reduces to the traditional CATE; thus, the CKTE

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

can be seen as a generalization of the CATE. With other kernel choices, the CKTE provides a more comprehensive distributional comparison of the control and treatment groups. For instance, via the conditional witness function $\mathbf{y} \mapsto \tau_k(\mathbf{x})(\mathbf{y})$ and with *characteristic* kernels, one can read off the sign of the differences in the local conditional densities around $\mathbf{y}$. That is, $\tau_k(\mathbf{x})(\mathbf{y}) > 0$ if and only if the difference between the conditional density of the treatment group and that of the control group in the neighborhood of $\mathbf{y}$ is greater than zero, and likewise if $\tau_k(\mathbf{x})(\mathbf{y}) \leq 0$ (Park et al., 2021). The magnitude indicates how pronounced this local discrepancy is, and the kernel bandwidth controls the size of the neighborhood. Moreover, for characteristic kernels, the hypothesis $H_0\colon \|\tau_k(\mathbf{x})\|_{\mathcal{H}} = 0$ is equivalent to testing the hypothesis $H_0\colon \mathbb{P}^0_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$. Thus, for characteristic kernels, our pointwise test at a fixed covariate value $\mathbf{x}$ is a no distributional effect test: it assesses whether the conditional treatment and control distributions coincide. By itself, this test does not indicate the direction of the effect, nor which regions of the outcome space are more likely under treatment versus control. That information is instead provided by the conditional witness function $\mathbf{y} \mapsto \tau_k(\mathbf{y})$, whose sign indicates whether outcomes near $\mathbf{y}$ are locally more likely under treatment or control.

Estimating the CKTE involves estimating the conditional mean embedding (CME) $\mathbb{E}[k(\mathbf{Y}^j, \cdot) \mid \mathbf{X} = \mathbf{x}]$, for the groups $j = 0, 1$. This was done in (Park et al., 2021) using classical kernel regression for each group. More recently, (Näf et al., 2023) used two distributional random forests (DRFs) for the task. Building on random forests (Breiman, 2001), DRFs provide forest-based estimates of the kernel mean embedding, conditional on $\mathbf{X}$. Using DRF instead of the usual kernel estimators is attractive, because DRF inherits the favorable empirical performance of forest-based methods. In particular, DRF tends to deliver accurate estimates, even when $\mathbf{X}$ is high dimensional, and does so without the need for parameter tuning. For instance, the CKTE estimator in (Park et al., 2021) requires to choose two kernels and two regularization parameters, while (Näf et al., 2023) simply use the default choices of tuning parameters and kernel advocated in (Ćevid et al., 2022).

We build on the DRF-based approach of Näf et al. (2023) for estimating the CKTE. Specifically, inspired by causal forests (Wager and Athey, 2018), we modify the DRF splitting criterion so that the CKTE is estimated directly from a single forest fit shared across both groups. We refer to our new method as "Causal-DRF". Adapting the results derived in (Näf et al., 2023), we show how an asymptotically valid test of $H_0\colon \|\tau_k(\mathbf{x})\|_{\mathcal{H}} = 0$, as well as uniformly valid confidence bands around $\tau_k(\mathbf{x})(\mathbf{y})$, can be constructed. This is notoriously difficult for kernel approaches, as laid out, for instance, in (Martinez Taboada et al., 2023); because kernel-based tests involve degenerate U-statistics, approximating the (asymptotic) distribution, even if known, is computationally intensive. This is solved in (Martinez Taboada et al., 2023) for the kernel treatment effect using data splitting, allowing for an elegant asymptotic normality result for their test statistic. In contrast, in our methodology an additional resampling step in the forest construction provides a natural approximation of the sampling distribution of our test statistic for fixed $\mathbf{x}$ without increasing the complexity of the algorithm. This allows for the construction of an asymptotically valid test of $H_0\colon \mathbb{P}^0_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$ as well as uniformly valid confidence bands for the witness function from one fit of the forest. Our Causal-DRF thus brings similar advantages to the estimation of the CKTE as the widely known causal-forest of (Wager and Athey, 2018) brought to the CATE estimation.

We illustrate the capabilities of Causal-DRF through simulations and by applying the method to the 1991 Survey of Income and Program Participation (SIPP) pension data (previously analyzed in (Benjamin, 2003; Chernozhukov and Hansen, 2004), and (Kallus and Oprescu, 2023), among others). This dataset contains three wealth measures that form the variable of interest ($\mathbf{Y}$), 401(k) eligibility as treatment assignment (W), and a range of covariates such as age and income ($\mathbf{X}$). With Causal-DRF we can directly obtain the CKTE jointly for the three measures of wealth, using a single forest fit. This is in contrast to (Chernozhukov and Hansen, 2004) who use 3 independent quantile regressions and to (Kallus and Oprescu, 2023) who combine the wealth variables of the dataset into one. This allows us to analyze the effect of 401(k) eligibility on the three wealth measures using the witness function and to test the equality of the conditional distribution for both groups for single test points.

1.1 Related Work

Several recent papers consider the (unconditional) kernel treatment effect, such as (Muandet et al., 2021) and (Fawkes et al., 2022). In particular, (Martinez Taboada et al., 2023) develop an efficient test for $H_0 : \mathbb{P}_{\mathbf{Y}^0} = \mathbb{P}_{\mathbf{Y}^1}$ using the kernel treatment effect. However, just as with the ATE, the kernel treatment effect obscures the heterogeneity of the treatment across different individuals. As such, (Park et al., 2021) and (Näf et al., 2023), which consider the *conditional* kernel treatment effect by separately estimating the conditional mean embeddings on both the treatment and control groups, are of special interest. The closest prior work to our current proposal is the one taken in (Näf et al., 2023) in which two separate DRFs are fitted to obtain the conditional

kernel treatment effect with uncertainty quantification and our theoretical analysis is built directly on their work. In fact, on a technical level, we begin the theoretical analysis in Appendix D.2 by fixing a problematic argument in their proof of a crucial result. With this fix, our proofs continue similarly as in (Näf et al., 2023), although considerably more care is needed in case of Causal-DRF.

The main methodological difference from Näf et al. (2023) is that Causal-DRF uses a single forest with a CKTE-targeted splitting criterion shared across both groups, rather than fitting two separate forests. In case of the CATE it is well-known that separate estimation of the mean function can be problematic. For instance, different estimation errors in the two mean estimates may lead to spurious effects, especially when one of the two groups has fewer observations available than the other. We refer to (Wager and Taylor, 2023, page 47) for a more detailed explanation of these two problems in the context of the CATE estimation. The estimation of the CKTE will likely display similar behavior. This is particularly apparent in the case of kernel estimates as in (Park et al., 2021), as the two kernel mean embedding estimates are likely to have different regularization parameters, leading to a regularization bias in the CKTE. Although there is no explicit regularization with DRF, similar problems likely arise. Similarly, covariate shifts in the two groups might lead to further bias if the propensity score varies strongly with $\mathbf{X}$. In the simulation studies we consider, using two separate DRFs can have detrimental coverage for smaller sample sizes, even in a relatively simple example, while Causal-DRF successfully maintains coverage even for $n = 250$ samples. Nonetheless, we note that this shared-partition construction is not uniformly superior to separate estimation; rather, the two approaches impose different inductive biases, as in the corresponding CATE literature (Curth and van der Schaar, 2021). When the treatment and control groups share learnable structure, the shared-partition design can be advantageous, whereas separate fits may be preferable when group-specific structure dominates. In our simulation settings, Causal-DRF tends to yield improved finite-sample coverage and competitive or better estimation accuracy, while also requiring only one forest fit.

1.2 Notation and RKHS Background

Here we collect the notation used throughout the main paper. We observe an i.i.d. sample $\mathbf{Z}_i = (\mathbf{Y}_i, W_i, \mathbf{X}_i)$, $i = 1, \dots, n$, where $\mathbf{Y}_i$ is an outcome in $\mathbb{R}^d$, $W_i \in \{0, 1\}$ indicates treatment status and $\mathbf{X}_i$ are covariates taking values in $\mathbb{R}^p$. We denote the conditional distributions in each group as $\mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ and $\mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$, respectively. We collect the data as $\mathcal{Z}_n = \{\mathbf{Z}_1, \dots, \mathbf{Z}_n\}$.

Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ be the RKHS induced by the positive definite, bounded, and continuous kernel $k \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (see Hsing and Eubank 2015, Chapter 2.7 for an introduction to RKHS theory). Continuity of k ensures that $\mathcal{H}$ is separable (Hsing and Eubank, 2015, Theorem 2.7.5). For a random element ξ taking values in the (separable) Hilbert space $\mathcal{H}$ with $\mathbb{E}[\|\xi\|_{\mathcal{H}}] < \infty$, we define its expected value in $\mathcal{H}$ by $\mathbb{E}[\xi] = \int_\Omega \xi d\mathbb{P} \in \mathcal{H}$, where the integral is to be understood in a Bochner sense (Hsing and Eubank, 2015, Chapter 3). Because $\mathcal{H}$ is separable, this integral is well defined. When $\mathbb{E}[\|\xi\|^2_{\mathcal{H}}] < \infty$, let the variance of $\xi \in \mathcal{H}$ be written as $\mathrm{Var}(\xi) = \mathbb{E}[\|\xi\|^2_{\mathcal{H}}] - \|\mathbb{E}[\xi]\|^2_{\mathcal{H}}$. For a sequence of random elements ξ_n in $\mathcal{H}$, we denote by $\xi_n \xrightarrow{D} \xi$ convergence in distribution, i.e., for all bounded and continuous functions $F \colon \mathcal{H} \to \mathbb{R}$, we have $\mathbb{E}[F(\xi_n)] \to \mathbb{E}[F(\xi)]$ as $n \to \infty$. If, for all $f \in \mathcal{H}$, we have $\langle \xi, f \rangle \sim N(0, \sigma_f^2)$ for some $\sigma_f > 0$, we write $\xi \sim N(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ a self-adjoint Hilbert-Schmidt (HS) operator satisfying $\langle \mathbf{\Sigma} f, f \rangle = \sigma_f^2$. In this case, we also write $\xi_n \xrightarrow{D} N(0, \mathbf{\Sigma})$, if $\xi_n \xrightarrow{D} \xi$.

2 MOTIVATIONAL EXAMPLES

We consider four toy examples of potential causal effects for univariate outcomes ($d = 1$), covariates of dimension $p = 5$, and sample size $n = 1000$. In all examples, we take $X_1, \dots, X_p$ to be drawn i.i.d. from the uniform distribution on $[2, 3]$, and $\mathbb{P}(W = 1 \mid \mathbf{X}) = \mathbb{P}(W = 1) = 0.5$. In the first example, there is no effect such that $\mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}, \mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ are both Gaussian with mean 0 and variance 1 ($N(0, 1)$). For the second example, there is a mean effect, where $\mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} = N(0, 1)$, while $\mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} = N(x_1, 1)$. In the third example, there is no mean effect; however, the variance decreases when changing from the control to the treatment group: $\mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} = N(0, 1)$ and $\mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} = N(0, 1/x_1^2)$. Finally, we consider an example where both the variance and the mean shift when changing from control to treatment group $\mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} = N(0, 1)$ and $\mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} = N(x_1, x_1^2)$. In all examples, we choose the test point to be $\mathbf{x} = (2.5, \dots, 2.5)^\top$. In addition, we assume in this example that an increase in Y is desired due to the treatment.

Figures 1 and 2 show both the true underlying conditional densities as well as our estimation method with confidence bands. As mentioned above, the estimated witness functions follow the difference in densities between the treatment group in red and the control group in blue. When the difference in densities is negative at $\mathbf{y}$, the value of the estimated witness function tends to be negative, and vice-versa. In particular, negative val-

ues indicate that it is less likely to observe an outcome from the treatment population at $\mathbf{y}$. Thus the witness function maps out the behavior of the treatment effect for fixed $\mathbf{x}$. As the additional confidence bands are simultaneously valid for all y, any crossing of zero of the lower or upper confidence bands indicates a rejection of $H_0 \colon \mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$. In the left graph in Figure 1 the upper and lower confidence bands are far from zero, indicating that the method recognizes that there is no effect. For all other examples, Causal-DRF indicates correctly that there is a significant effect, as in each example the confidence bands have regions that do not intersect zero. Even in challenging cases such as Example 3, where there is only a variance shift, Figure 2 shows a clear rejection of H_0.

These examples indicate how the CKTE, estimated through Causal-DRF, can give a clearer picture of conditional treatment effects that go beyond differences in expectation. Moreover, the confidence bands offer a convenient tool to assess whether the effect is statistically significant. We note that the analysis for the last two cases is quite different if only the CATE is considered. First, the CATE will miss the decrease in variance in the third example. Second, while the last example may appear to be a success when only expectations are considered, only examining CATE will lead to missing differing variances in the treatment and control groups.

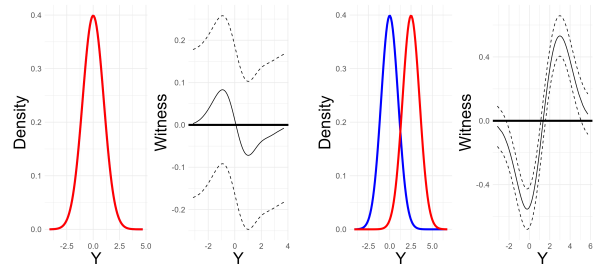

Figure 1: Density and Causal-DRF estimates for Example 1 (left, no effect) and Example 2 (right, mean shift). Blue: Density for the control group $W = 0$, Red: Density for the treatment group $W = 1$.

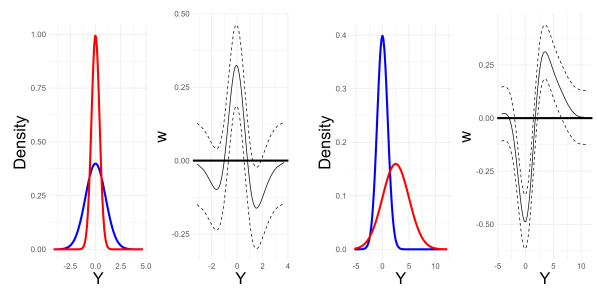

Figure 2: Density and Causal-DRF estimates for Example 3 (left, variance effect) and Example 4 (right, mean shift + variance effect). Blue: Density for the control group $W = 0$, Red: Density for the treatment group $W = 1$.

3 CAUSAL-DISTRIBUTIONAL RANDOM FORESTS

We start by reviewing the core DRF algorithm; for a full treatment, see Ćevid et al. (2022) and Näf et al. (2023). The DRF estimate of the CME can be viewed as a weighted average over the training points. These weights $\hat{w}_i(\mathbf{x})$ are found by fitting a random forest type algorithm within the RKHS $(\mathcal{H}, k)$. Specifically, a forest of N trees is constructed. Each tree recursively splits the observations into sets $\mathcal{I}_L = \{i : X_{ij} < c\}$ and $\mathcal{I}_R = \{i : X_{ij} \geq c\}$, with a splitting value $c \in \mathbb{R}$ and a splitting feature j chosen by a splitting criterion that is allowed to depend on $\mathbf{Y}$. In the original DRF algorithm, the Maximum Mean Discrepancy (MMD) statistic (Gretton et al., 2007) is used as a splitting criterion. This choice of splitting criterion can be interpreted as an analogue of the traditional CART splitting criterion (Breiman, 2001), but rather applied in the RKHS. Combining this idea with principles of forest construction developed by (Wager and Athey, 2018; Athey et al., 2019), we develop a novel splitting criterion in the RKHS that can be used to estimate $\tau_k(\mathbf{x})$. That is, if for a given node, the estimate in the left child is given as

$$\hat{\tau}_L = \frac{1}{|\mathcal{I}_L, W_i = 1|} \sum_{i \in \mathcal{I}_L, W_i = 1} k(\mathbf{Y}_i, \cdot) - \frac{1}{|\mathcal{I}_L, W_i = 0|} \sum_{i \in \mathcal{I}_L, W_i = 0} k(\mathbf{Y}_i, \cdot),$$

and similarly with $\hat{\tau}_R$, then we split such that a weighted version of

$$\begin{aligned} &\|\hat{\tau}_L - \hat{\tau}_R\|^2_{\mathcal{H}} \\ &= \Big\| \sum_{i \in \mathcal{I}_L} \left(\frac{W_i}{|\mathcal{I}_L, W_i = 1|} - \frac{1 - W_i}{|\mathcal{I}_L, W_i = 0|} \right) k(\mathbf{Y}_i, \cdot) \\ &- \sum_{i \in \mathcal{I}_R} \left(\frac{W_i}{|\mathcal{I}_R, W_i = 1|} - \frac{1 - W_i}{|\mathcal{I}_R, W_i = 0|} \right) k(\mathbf{Y}_i, \cdot) \Big\|^2_{\mathcal{H}} \end{aligned} \tag{2}$$

is maximal. This is a weighted version of the MMD statistic. Thus, using the same random approximation as in (Ćevid et al., 2022) for a translation-invariant kernel (see **(K3)** in Appendix D) this can be efficiently approximated using $S \in \mathbb{N}$ random functionals. For a set of candidate variables, $\mathcal{I}_l$, $\mathcal{I}_R$ are chosen to maximize this criterion. Details are given in Appendix C.

For each fitted tree $l = 1, \ldots, N$, let $\mathcal{L}(\mathbf{x})$ be the leaf

node of $\mathbf{x}$. The estimate of $\tau_k(\mathbf{x})$ is then an average over the N trees. This approach has the same computational complexity of the original DRF algorithm, specifically $\mathcal{O}(S \times N \times p \times n \log n)$; see (Ćevid et al., 2022). Collecting all $k(\mathbf{Y}_j, \cdot)$, one can rewrite $\hat{\tau}_{n,k}(\mathbf{x})$ conveniently as

$$\hat{\tau}_{n,k}(\mathbf{x}) = \sum_{i=1}^{n} \hat{w}_i(\mathbf{x}) k(\mathbf{Y}_i, \cdot), \tag{3}$$

with

$$\hat{w}_i(\mathbf{x}) = \frac{1}{N} \sum_{l=1}^{N} \Big(\frac{\mathbf{1}\{\mathbf{X}_i \in \mathcal{L}_l(\mathbf{x})\} W_i}{|\{j : \mathbf{X}_j \in \mathcal{L}_l(\mathbf{x}), W_j = 1\}|} - \frac{\mathbf{1}\{\mathbf{X}_i \in \mathcal{L}_l(\mathbf{x})\}(1 - W_i)}{|\{j : \mathbf{X}_j \in \mathcal{L}_l(\mathbf{x}), W_j = 0\}|} \Big).$$

Finally, to construct tests and confidence bands, we consider the following subsampling approach similar to that of (Näf et al., 2023). For a subset $\mathcal{S} \subset \{1, \ldots, n\}$, denote by $\hat{\tau}^{\mathcal{S}}_{n,k}(\mathbf{x})$ the version of $\hat{\tau}_{n,k}(\mathbf{x})$ that uses trees built with data from $\mathcal{S}$. We draw $\mathcal{S}$ by sampling n i.i.d. random variables $U_i \sim \text{Bernoulli}(1/2)$ and set $\mathcal{S} = \{i \colon U_i = 1\}$. Thus, we fit B subforests with L trees, such that $N = B \times L$. The resulting estimates $\hat{\tau}^{\mathcal{S}_b}_{n,k}(\mathbf{x})$, $b = 1, \ldots, B$, correspond to an approximation of the sampling distribution of $\hat{\tau}_{n,k}(\mathbf{x})$; we formalize this notion in the following section. Note that $|\mathcal{S}|/n \to 1/2$ almost surely and this approach is motivated from the half-sampling approach of Athey et al. (2019), though our independent sampling allows for a somewhat simpler theoretical treatment. The point estimate of the CME is then calculated by averaging the point predictions over all the half-samples, $\hat{\tau}_{n,k}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}^{\mathcal{S}_b}_{n,k}(\mathbf{x})$ which corresponds to (3). Algorithms 1 and 2 in Appendix D provide pseudocode for this procedure. This approach again has the same complexity as one fit of DRF (Ćevid et al., 2022), except that $N = B \times L$ should be chosen quite large.

The Causal-DRF estimator can be conveniently written in matrix form. Consider the matrix $\mathbf{K} = (k(\mathbf{Y}_i, \mathbf{Y}_j))_{i=1,\ldots n, j=1,\ldots n}$. Denote by $\hat{\mathbf{w}} \in \mathbb{R}^n$ the vector that concatenates the weights in (3) and similarly let $\hat{\mathbf{w}}^{\mathcal{S}_b} \in \mathbb{R}^n$ be the weights of the bth subforest. Let $\mathbf{k} = (k(\mathbf{Y}_1, \cdot), \ldots, k(\mathbf{Y}_n, \cdot))^\top$, and denote by $\mathbf{k}(\mathbf{y}) = (k(\mathbf{Y}_1, \mathbf{y}), \ldots, k(\mathbf{Y}_n, \mathbf{y}))^\top$ for $\mathbf{y} \in \mathbb{R}^d$. Then, $\hat{\tau}_{n,k}(\mathbf{x}) = \hat{\mathbf{w}}^\top \mathbf{k}$, $\hat{\tau}^{\mathcal{S}_b}_{n,k}(\mathbf{x}) = (\hat{\mathbf{w}}^{\mathcal{S}_b})^\top \mathbf{k}$. Following the software implementation, we will take k to be the Gaussian kernel given in (26) in Appendix D. Note that with this choice, it holds that $\sup_{\mathbf{y},\mathbf{y}'} k(\mathbf{y}, \mathbf{y}') \leq 1$. For the Gaussian kernel a bandwidth σ needs to be chosen, which may influence the performance of the algorithm. Here, we follow (Ćevid et al., 2022; Näf et al., 2023) and choose σ using the median heuristic of (Gretton et al., 2012). This appears to work quite well, though we note the (empirically) optimal choice of σ for (Causal-)DRF remains an open question, as does the influence of σ on the empirical performance.

To calculate tests and confidence bands for $\hat{\tau}_{n,k}(\mathbf{x})$, we first compute $q_{n,\alpha}$, the $1 - \alpha$ quantile of the B many draws from

$$\|\hat{\tau}^{\mathcal{S}_b}_{n,k}(\mathbf{x}) - \hat{\tau}_{n,k}(\mathbf{x})\|^2_{\mathcal{H}} = (\hat{\mathbf{w}}^{\mathcal{S}_b} - \hat{\mathbf{w}})^\top \mathbf{K} (\hat{\mathbf{w}}^{\mathcal{S}_b} - \hat{\mathbf{w}}). \tag{4}$$

Then a test for $H_0 \colon \mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}} = \mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}}$ can be constructed as

$$\phi(\mathcal{Z}_n) = \mathbf{1}\left\{ \|\hat{\tau}_{n,k}(\mathbf{x})\|^2_{\mathcal{H}} > q_{n,\alpha} \right\}, \tag{5}$$

while a simultaneous $1 - \alpha$ confidence band around $\hat{\tau}_{n,k}(\mathbf{x})$ can be constructed as:

$$\mathcal{B}(\mathbf{y}) = [\hat{\tau}_{n,k}(\mathbf{x})(\mathbf{y}) - \sqrt{q_{n,\alpha}}, \hat{\tau}_{n,k}(\mathbf{x})(\mathbf{y}) + \sqrt{q_{n,\alpha}}]. \tag{6}$$

The test in (5) should be interpreted as a test of equality of the conditional outcome distributions at the covariate value $\mathbf{x}$. In particular, rejection indicates that treatment has a distributional effect at $\mathbf{x}$, but the test statistic alone does not reveal the direction of that effect or where the two distributions differ. Such directional and region-specific information is instead conveyed by the witness function and its confidence band in (6): regions where the band lies strictly above zero correspond to outcomes locally more likely under treatment, while regions where it lies strictly below zero correspond to outcomes locally more likely under control. We note that the confidence band in (6) is uniform in $\mathbf{y}$ but only valid for a fixed $\mathbf{x}$. For several test points $\mathbf{x}$, one has to account for multiple testing issues.

In the next section we prove consistency and asymptotic normality of $\hat{\tau}_{n,k}(\mathbf{x})$, that ϕ is an asymptotically valid test and that $\mathcal{B}(\mathbf{y})$ provides asymptotically uniform coverage over $\mathbf{y} \in \mathbb{R}^d$.

4 THEORETICAL DEVELOPMENT

In this section we establish theoretical guarantees for the Causal-DRF algorithm, including its consistency and asymptotic normality. Notably, we are able to establish similar guarantees as (Näf et al., 2023), who rely on fitting separate forests for the treatment and control groups.

For convenience, denote the kernel embeddings as $\mu^{(j)}(\mathbf{x}) = \mathbb{E}[k(\mathbf{Y}^j, \cdot) \mid \mathbf{X} = \mathbf{x}]$, for $j \in \{0, 1\}$ and $\mu(\mathbf{x}) = \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}]$. To begin, we assume the forest satisfies Assumptions (**F1**)–(**F6**) and (**F1***) in Appendix D.1. These assumptions are a combination and refinement of the assumptions taken in (Wager and Athey, 2018) and (Näf et al., 2023). In addition, we place the

following assumptions on the data-generating process; **(D2)**-**(D5)** hold for $j \in \{0, 1\}$.

(**D1**) $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. on $[0, 1]^p$, and admit a density bounded away from 0 and infinity.

(**D2**) The mappings $\mathbf{x} \mapsto \mu^{(j)}(\mathbf{x})$ are Lipschitz .

(**D3**) $\mathrm{Var}(k(\mathbf{Y}^j, \cdot) \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}[\|k(\mathbf{Y}^j, \cdot)\|^2_{\mathcal{H}} \mid \mathbf{X} = \mathbf{x}] - \|\mathbb{E}[k(\mathbf{Y}^j, \cdot) \mid \mathbf{X} = \mathbf{x}]\|^2_{\mathcal{H}} > 0$.

(**D4**) For all $f \in \mathcal{H} \setminus \{0\}$, $\mathrm{Var}(\langle k(\mathbf{Y}^j, \cdot), f\rangle \mid \mathbf{X} = \mathbf{x}) = \mathrm{Var}(f(\mathbf{Y}^j) \mid \mathbf{X} = \mathbf{x}) > 0$.

(**D5**) For all $f \in \mathcal{H} \setminus \{0\}$, $\mathbf{x} \mapsto \mathbb{E}[\left|f(\mathbf{Y}^j)\right|^2 \mid \mathbf{X} = \mathbf{x}]$ is Lipschitz .

As noted in (Näf et al., 2023, page 10), **(D3)** is automatically satisfied for the Gaussian kernel under the assumption that the conditional distributions $\mathbf{Y}^j \mid \mathbf{X} = \mathbf{x}$ for $j \in \{0, 1\}$ are not point measures. We also adopt the following traditional causal assumptions.

(**C1**) Positivity (overlap): there exists $\varepsilon > 0$ such that for all $\mathbf{x}$,

$$\varepsilon < \mathbb{P}(W = 1 \mid \mathbf{X} = \mathbf{x}) < 1 - \varepsilon \tag{7}$$

(**C2**) Unconfoundedness (no unmeasured confounding): $(\mathbf{Y}^0, \mathbf{Y}^1) \perp\!\!\!\perp W \mid \mathbf{X}$

Finally, we place assumptions (**K1**)–(**K4**) (Appendix D) on the kernel function k. All of these assumptions are met by the Gaussian kernel (26).

Next, we consider the setting where the number of trees goes to infinity ($N \to \infty$). The forest estimator $\hat{\tau}_{n,k}(\mathbf{x})$ is then the average over all possible $\binom{n}{s_n}$ many trees:

$$\hat{\tau}_{n,k}(\mathbf{x}) = \frac{1}{\binom{n}{s_n}} \sum_{i_1 < i_2 < \cdots < i_{s_n}} \mathbb{E}_{\mathcal{E}}\left[\Gamma(\mathbf{x}; \mathcal{E}, \{\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}\})\right], \tag{8}$$

where $\mathcal{E}$ encodes potential auxiliary randomness from the trees (such as the dimensions considered in the splitting criterion), and

$$\mathbb{E}_{\mathcal{E}}\left[\Gamma(\mathbf{x}; \mathcal{E}, \{\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}\})\right] = \frac{1}{|\{j : \mathbf{X}_j \in \mathcal{L}(\mathbf{x}), W_j = 1\}|} \sum_{\mathbf{X}_j \in \mathcal{L}(\mathbf{x}), W_j = 1} k(\mathbf{Y}_j, \cdot) - \frac{1}{|\{j : \mathbf{X}_j \in \mathcal{L}(\mathbf{x}), W_j = 0\}|} \sum_{\mathbf{X}_j \in \mathcal{L}(\mathbf{x}), W_j = 0} k(\mathbf{Y}_j, \cdot). \tag{9}$$

This simplification is also used in (Wager and Athey, 2018; Athey et al., 2019; Ćevid et al., 2022), and (Näf et al., 2023). Similarly, $\hat{\tau}^{\mathcal{S}}_{n,k}(\mathbf{x})$ is also an infinite forest predictor but using only the samples in $\mathcal{S}$. Given these assumptions, we have that:

Theorem 4.1. *Assume conditions **(F1)**–**(F6)**, **(F1*)**, **(D1)**–**(D5)**, **(K1)**–**(K4)** and **(C1)**–**(C2)** hold. Then, there exists $\sigma_n \to 0$, such that, for some self-adjoint HS operator $\mathbf{\Sigma}_{\mathbf{x}}$, we have*

$$\xi_n(\mathbf{x}) = \frac{1}{\sigma_n}\left(\hat{\tau}_{n,k}(\mathbf{x}) - \tau_k(\mathbf{x})\right) \xrightarrow{D} N(0, \mathbf{\Sigma}_{\mathbf{x}}). \tag{10}$$

We note that to prove consistency, more specifically,

$$\|\hat{\tau}_{n,k}(\mathbf{x}) - \tau_k(\mathbf{x})\|_{\mathcal{H}} = \mathcal{O}_p\left(n^{-\gamma}\right) \tag{11}$$

for any $\gamma \leq \frac{1}{2}\min\left(1 - \beta, \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p} \cdot \beta\right)$, it is enough to assume (**F1**)–(**F5**), (**D1**) and (**D2**), (**K1**)–(**K4**) and (**C1**)–(**C2**). Here $\alpha \leq 0.2$ is the constant taken from Assumption **(F4)**, while β is the subsampling rate defining s_n in **(F5)**, which is chosen as $s_n = n^{\beta}$. Moreover, using similar arguments as in (Bénard et al., 2024, Proposition 3), the assumption of a bounded kernel **(K1)** also implies

$$\mathbb{E}[\|\hat{\tau}_{n,k}(\mathbf{X}) - \tau_k(\mathbf{X})\|_{\mathcal{H}}] \to 0, \tag{12}$$

where expectation is taken over the training data as well as $\mathbf{X}$.

Remark 4.2. Theorem 4.1 is similar to Theorem 6 in Näf et al. (2023), but the new tree construction includes several additional technical challenges. In particular, the key to proving Theorem 4.1 is to control the ratio of variances of the first order approximation (Theorem E.5), which is now more difficult with W_i involved. We refer to Remark E.4 and to Theorem E.5 for more details.

Unfortunately, the convergence in (10) is of limited use in practice, as one would also need to estimate self-adjoint HS operator. Instead, we follow a resampling approach as motivated by Athey et al. (2019); Näf et al. (2023) to approximate the sampling distribution of $\xi_n(\mathbf{x})$. Consider the infinite Causal-DRF $\hat{\tau}^{\mathcal{S}}_{n,k}(\mathbf{x})$, constructed using only the samples in $\mathcal{S}$, where $\mathcal{S}$ was sampled as described in Section 3. Under such a sampling scheme, randomness in $\hat{\tau}^{\mathcal{S}}_{n,k}(\mathbf{x})$ can be decomposed into randomness in the data and from random assignment to $\mathcal{S}$. Thus, conditional on the data, we show that

$$\xi^{\mathcal{S}}_n(\mathbf{x}) = \frac{1}{\sigma_n}\left(\hat{\tau}^{\mathcal{S}}_{n,k}(\mathbf{x}) - \hat{\tau}_{n,k}(\mathbf{x})\right) \tag{13}$$

converges to the same (Gaussian) variable as $\xi_n(\mathbf{x})$ (10). With this result in hand, we can randomly sample $\mathcal{S}$ and

use the empirical distribution of $\xi_n^{\mathcal{S}}(\mathbf{x})$ to approximate the asymptotic distribution of $\xi_n(\mathbf{x})$. Formalizing this result requires extending standard bootstrap arguments (Kosorok, 2003, 2008). Specifically, Theorem 4.3 shows that

$$\xi_n^{\mathcal{S}}(\mathbf{x}) \xrightarrow[W]{D} N(0, \mathbf{\Sigma_x}), \tag{14}$$

where $\xrightarrow[W]{D}$ denotes "conditional convergence in distribution", characterized by the condition

$$\sup_{h \in \mathrm{BL}_1(\mathcal{H})} \left| \mathbb{E}[h(\xi_n^{\mathcal{S}}(\mathbf{x})) \mid \mathcal{Z}_n] - \mathbb{E}[h(\xi(\mathbf{x}))] \right| \xrightarrow{p} 0, \tag{15}$$

where $\mathrm{BL}_1(\mathcal{H})$ denotes the space of all bounded Lipschitz functions from $\mathcal{H}$ to $\mathbb{R}$ with Lipschitz constant bounded by 1 (i.e. $h \in \mathrm{BL}_1(\mathcal{H})$ satisfies $\sup_{f \in \mathcal{H}} |h(f)| \leq 1$ and $|h(f_1) - h(f_2)| \leq \|f_1 - f_2\|_{\mathcal{H}}$ for all $f_1, f_2 \in \mathcal{H}$). Condition (15) implies that $\xi_n^{\mathcal{S}}(\mathbf{x})$ converges to $\xi(\mathbf{x})$ in distribution in probability, conditionally on the data $\mathcal{Z}_n$. For details, see for example (González-Rodríguez and Colubi, 2017); (Kosorok, 2008, Chapter 10).

Theorem 4.3. *Assume conditions **(F1)**–**(F6)**, **(F1*)**, **(D1)**–**(D5)**, **(K1)**–**(K4)** and **(C1)**–**(C2)** hold. Then,* (14) *holds.*

Theorem 4.3 allows us to approximate the distribution of $\xi_n(\mathbf{x})$ with the distribution of $\xi_n^{\mathcal{S}}(\mathbf{x})$. Fortunately, the distribution of $\xi_n^{\mathcal{S}}(\mathbf{x})$ can be approximated using the results of a single forest fit, as in (Athey et al., 2019) and (Näf et al., 2023). To do so, we construct the forest by (1) sampling B subsets $\mathcal{S}_1, \ldots, \mathcal{S}_B$ of $\{1, \ldots, n\}$, (2) fitting a Causal-DRF with L trees and calculating the prediction $\hat{\tau}_{n,k}^{\mathcal{S}_b}(\mathbf{x})$ within each half-sample $b = 1, \ldots, B$, and (3) forming the final prediction $\hat{\tau}_{n,k}^{\mathcal{S}_b}(\mathbf{x})$ by averaging the predictions of all the half-samples: $(\hat{\tau}_n^{\mathcal{S}_b}(\mathbf{x}))_{b=1}^B$. Combining the continuous mapping theorem with the above results, it follows immediately that

$$\frac{1}{\sigma_n^2} \left\| \hat{\tau}_{n,k}^{\mathcal{S}}(\mathbf{x}) - \hat{\tau}_{n,k}(\mathbf{x}) \right\|_{\mathcal{H}}^2 \xrightarrow[W]{D} \|\xi\|^2,$$
$$\frac{1}{\sigma_n^2} \left\| \hat{\tau}_{n,k}(\mathbf{x}) - \tau_k(\mathbf{x}) \right\|_{\mathcal{H}}^2 \xrightarrow{D} \|\xi\|^2, \tag{16}$$

with $\xi \sim N(0, \mathbf{\Sigma_x})$. Thus if we consider $H_0\colon \mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$, then $\tau_k(\mathbf{x}) = 0$ and $\frac{1}{\sigma_n^2} \left\| \hat{\tau}_{n,k}(\mathbf{x}) \right\|_{\mathcal{H}}^2$ has the same limiting distribution as its resampling bootstrap version

$$\frac{1}{\sigma_n^2} \left\| \hat{\tau}_{n,k}^{\mathcal{S}}(\mathbf{x}) - \hat{\tau}_{n,k}(\mathbf{x}) \right\|_{\mathcal{H}}^2 \tag{17}$$

given the data. Regardless of whether $\mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ holds, we can approximate the distribution under H_0 by sampling from $\mathcal{S}$. Consider choosing $c_{n,\alpha}$ to be the smallest value obtained from B draws, with B large enough to ensure that we have

$$\mathbb{P}\left(\frac{1}{\sigma_n^2} \left\| \hat{\tau}_{n,k}^{\mathcal{S}}(\mathbf{x}) - \hat{\tau}_{n,k}(\mathbf{x}) \right\|_{\mathcal{H}}^2 > c_{n,\alpha} \mid \mathcal{Z}_n \right) \leq \alpha. \tag{18}$$

The associated test $\phi(\mathcal{Z}_n)$ is given by

$$\phi(\mathcal{Z}_n) = \mathbf{1}\left\{ \frac{1}{\sigma_n^2} \|\hat{\tau}_{n,k}(\mathbf{x})\|_{\mathcal{H}}^2 > c_{n,\alpha} \right\}. \tag{19}$$

Note that (19) is a slight variation of (5), better suited for theoretical analysis. That is, (19) uses the theoretical version of $\hat{\tau}_{n,k}(\mathbf{x})$, with $N \to \infty$ and similarly the quantile $q_{n,\alpha}$ replaced by $\sigma_n^2 c_{n,\alpha}$, with $c_{n,\alpha}$ obtained as in (18). Though σ_n^2 could be estimated as well, we directly consider the resampled unscaled statistic (4) and identify its $1-\alpha$ quantile, which corresponds to $\sigma_n^2 c_{n,\alpha}$. The following theorem shows that ϕ is of level α with power converging to 1.

Theorem 4.4. *Assume conditions **(F1)**–**(F6)**, **(F1*)**, **(D1)**–**(D5)**, **(K1)**–**(K4)** and **(C1)**–**(C2)** hold. Then, as $n \to \infty$,*

(i) ϕ has a valid type-I error. That is, if $\mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$,

$$\limsup_{n \to \infty} \mathbb{P}\left(\frac{1}{\sigma_n^2} \|\hat{\tau}_{n,k}(\mathbf{x})\|_{\mathcal{H}}^2 > c_{n,\alpha} \right) \leq \alpha.$$

(ii) ϕ has power going to 1. That is, if $\mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} \neq \mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$,

$$\lim_{n \to \infty} \mathbb{P}\left(\frac{1}{\sigma_n^2} \|\hat{\tau}_{n,k}(\mathbf{x})\|_{\mathcal{H}}^2 > c_{n,\alpha} \right) = 1.$$

A confidence band for the conditional witness function $\mathbf{y} \mapsto \tau_k(\mathbf{x})(\mathbf{y})$ that is jointly valid for all $\mathbf{y}$-values can be obtained straightforwardly by inverting the testing procedure. Let $c_{n,\alpha}$ be as in (18). We propose constructing an interval given by

$$\mathcal{B}(\mathbf{y}) = [\hat{\tau}_{n,k}(\mathbf{x})(\mathbf{y}) - \sqrt{c_{n,\alpha} C}\sigma_n, \hat{\tau}_{n,k}(\mathbf{x})(\mathbf{y}) + \sqrt{c_{n,\alpha} C}\sigma_n], \tag{20}$$

where $C = \sup_{\mathbf{y}} k(\mathbf{y}, \mathbf{y})$. Note that the boundedness of the reproducing kernel ensures that C is finite; see Assumption (**K2**). In particular, $C = 1$ for the Gaussian kernel. In the following theorem we show that $\mathcal{B}(\mathbf{y})$ is a uniform $1-\alpha$ confidence band for the conditional witness function as a function of $\mathbf{y}$.

Theorem 4.5. *Assume conditions **(F1)**–**(F5)**, **(D1)**–**(D5)**, **(K1)**–**(K4)** and **(C1)**–**(C2)** hold. Then, as $n \to \infty$,*

$$\liminf_{n \to \infty} \mathbb{P}\left(\bigcap_{\mathbf{y}} \{\tau_k(\mathbf{x})(\mathbf{y}) \in \mathcal{B}(\mathbf{y})\} \right) \geq 1 - \alpha. \tag{21}$$

The confidence band in (20) is adapted analogously from the version in (6). We note that this band has uniform coverage in $\mathbf{y}$, thus the reason why any crossing with zero is a rejection, but pointwise coverage in $\mathbf{x}$. Thus, the test will not be valid over different test point $\mathbf{x}$ without a correction for multiple testing.

5 EMPIRICAL STUDY: IMPACT OF 401(k) ELIGIBILITY

We study the impact of 401(k) eligibility on wealth, based on a sample of households from Wave 4 of the 1990 Survey of Income and Program Participation (SIPP). This dataset was used in several studies of the relationship between 401(k) plans and wealth, e.g., (Benjamin, 2003; Chernozhukov and Hansen, 2004; Kallus and Oprescu, 2023), and includes 9,915 observations. The variables of interest $\mathbf{Y}$ are three free wealth measures: "Net Financial Assets", "Net Non-401(k) Financial Assets" and "Total Wealth". Following (Benjamin, 2003; Chernozhukov and Hansen, 2004; Kallus and Oprescu, 2023), we consider 9 covariates $\mathbf{X}$ that include age, income, education, family size, marital status, and IRA participation, as detailed in Table 1 in Appendix A. Finally, the treatment W is taken to be 401(k) eligibility. While 401(k) eligibility is not randomly assigned, prior studies (see, e.g., (Chernozhukov and Hansen, 2004; Kallus and Oprescu, 2023) and the references therein) have commonly assumed unconfoundedness conditional on these nine covariates, supported by subject-matter expertise. The three wealth measures we consider are the same as in (Chernozhukov and Hansen, 2004). While they use separate quantile regressions, we utilize the ability of Causal-DRF to estimate the CKTE of the three wealth measures jointly. For this reason, after taking the first 10 households as a test set, we train one Causal-DRF on the remaining 9,905 observations, taking the three variables of financial wealth as one response vector $\mathbf{Y}$ in $\mathbb{R}^3$. As one would expect, the three measures are highly skewed. We begin by scaling the three measures of wealth by subtracting the mean and dividing by the standard deviation. We then fit Causal-DRF with $N = 4,000$ trees with 100 trees per group ($B = 40$). We focus on two test points $\mathbf{x}_1$, $\mathbf{x}_2$ out of the test set of 10 households. The two households were chosen to be relatively dissimilar to each other; the reference persons in each household have an age gap of 19 years and the two households have a large difference in family size and income. A more detailed description of the characteristics of the two households is given in the captions of Figures 3 and 4. Using the test statistic in (5), a test of $H_0 : \mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}_j} = \mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}_j}$ is rejected for both points $j = 1, 2$. This indicates a significant impact of eligibility on wealth for these two points, though we did not account for multiple testing. To better understand the effect, Figures 3 and 4 show the (univariate) witness function for Net Financial Assets (Y_1), Net Non-401(k) Financial Assets (Y_2) and Total Wealth (Y_3). First, the significant result appears to stem mostly from the influence of eligibility on Net Financial Assets, while the other two wealth measures appear not significantly impacted by eligibility. This is in particular true for the older household, where there is barely an effect of eligibility on Net Non-401(k) Financial Assets. We thus focus on the effect on Net Financial Assets and Total Wealth. For Net Financial Assets the negative values for small amounts of assets indicate that we are more likely to obtain small values for people who are not eligible for 401(k) (the control group). On the other hand, high values of Net Financial Assets are more likely to come from people eligible for 401(k). The picture is not perfectly symmetric, which indicates a higher skewness of the distribution for $W = 1$. As skewness in wealth distributions is usually oriented towards higher values, this is not entirely surprising. Interestingly, though there is a slightly different magnitude, the picture for Net Financial Assets is roughly the same for both test points. Though these are only two test households, it might be that for this variable alone, there is less heterogeneity over different households. On the other side of the spectrum, the picture for Total Wealth of the two test points is quite different. While for the older household, the witness function looks similar to the one of Net Financial assets, albeit with a higher skewness in the eligible group, Total wealth for the younger household in Figure 3 has another dip in the witness function for higher values of income. This indicates there is a certain spectrum of higher incomes where it is more likely to observe people who are not eligible. These discussions should illustrate the more comprehensive analysis of the treatment effect that are possible with Causal-DRF.

6 CONCLUSION

In this paper, we introduced Causal-DRF, a forest-based method to obtain the conditional kernel treatment effect under the assumption of no confounding and overlap. We discussed how to estimate the CKTE using a single estimated forest fit. On the theoretical side, we derive the consistency and asymptotic normality of the Hilbert space-valued estimate for fixed $\mathbf{x}$. A computationally efficient method to approximate the asymptotic sampling distribution was also proposed. The latter was used to build tests of $H_0 \colon \mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ and confidence bands around $\tau(\mathbf{x})(\mathbf{y})$. We also discussed how the shared-partition, one-fit design differs from using two separate kernel mean embedding estimates, and why it can be beneficial when the two

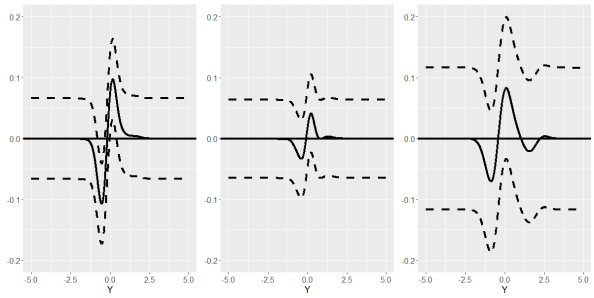

Figure 3: Witness functions with confidence bands for Net Financial Assets (Left), Net Non-401(k) Financial Assets (Middle), and Total Wealth (Right) for a person of age 31, 28'146 income, family size of 5, 12 years of education, married, single earner, homeowner, no defined benefit pension status, and no IRA participation.

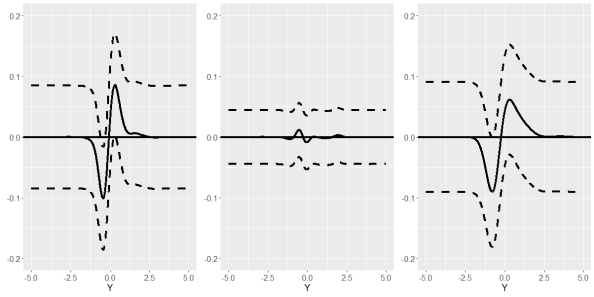

Figure 4: Witness functions with confidence bands for Net Financial Assets (Left), Net Non-401(k) Financial Assets (Middle), and Total Wealth (Right) for a person of age 50, 5'766 income, family size of 1, 14 years of education, unmarried, single earner, no defined benefit pension status, and IRA participation.

groups share learnable structure.

Our work also motivates several interesting research directions. First, while we obtain elegant asymptotically valid uncertainty quantification for fixed $\mathbf{x}$, uniform in $\mathbf{y}$, we lack uniform guarantees over $\mathbf{x}$. The same limitation holds for causal random forests. As such, when testing $H_0 \colon \mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ for several test points $\mathbf{x}$ one quickly encounters a multiple testing issue. A uniform bound in $\mathbf{x}$ would allow for a test of $H_0 \colon \mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}} = \mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}}$ with asymptotic guarantees. This is already possible, in the sense that rejecting $H_0 \colon \mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ for a test point $\mathbf{x}$ implies that $\mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}} = \mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}}$ does not hold. On the other hand, even if $\mathbb{P}^0_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ holds, it could be because they are the same or because we chose a test point $\mathbf{x}$ where they coincide. Second, since we were mainly interested in $\tau_k(\mathbf{x})$, we did not discuss potential estimators that could be derived from the estimate $\hat{\tau}_{n,k}(\mathbf{x})$, such as the CATE, and what guarantees could be derived for this estimator. Finally, we note that the methodology might also be used to study the difference in two conditional distributions more generally, not necessarily with a causal interpretation.

ACKNOWLEDGMENTS

This work is part of the DIGPHAT project which was supported by a grant from the French government, managed by the National Research Agency (ANR), under the France 2030 program, with reference ANR-22-PESN-0017. The computational requirements for this work were supported in part by the NYU Langone High Performance Computing (HPC) Core's resources and personnel. JP was supported by SNSF Grant 218343.

References

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

C. Bénard, J. Näf, and J. Josse. MMD-based variable importance for distributional random forest. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pages 1324–1332, 2024.

D. J. Benjamin. Does 401(k) eligibility increase saving?: Evidence from propensity score subclassification. *Journal of Public Economics*, 87(5):1259–1290, 2003.

L. Breiman. Random forests. *Machine learning*, 45(1): 5–32, 2001.

D. Ćevid, L. Michel, J. Näf, N. Meinshausen, and P. Bühlmann. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022.

X. Chen and H. White. Central limit and functional central limit theorems for hilbert-valued dependent heterogeneous arrays with applications. *Econometric Theory*, 14(2):260–284, 1998.

V. Chernozhukov and C. Hansen. The Effects of 401(K) Participation on the Wealth Distribution: An Instrumental Quantile Regression Analysis. *The Review of Economics and Statistics*, 86(3):735–751, 2004.

V. Chernozhukov, I. Fernández-Val, and B. Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

V. Chernozhukov, I. Fernandez-Val, and M. Weidner. Network and panel quantile effects via distribution regression. *Journal of Econometrics*, 2020.

A. Curth and M. van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From

theory to learning algorithms. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 1810–1818, 13–15 Apr 2021.

R. M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002.

J. Fawkes, R. Hu, R. J. Evans, and D. Sejdinovic. Doubly robust kernel statistics for testing distributional treatment effects even under one sided overlap. *Preprint arXiv:1510.04342*, 2022.

G. González-Rodríguez and A. Colubi. On the consistency of bootstrap methods in separable Hilbert spaces. *Econometrics and Statistics*, 1:118–127, 2017.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pages 513–520, 2007.

A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, pages 1205–1213, 2012.

M. Hohberg, P. Pütz, and T. Kneib. Treatment effects beyond the mean using distributional regression: Methods and guidance. *PLOS ONE*, 15(2): 1–29, 02 2020.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley, 2015.

N. Kallus and M. Oprescu. Robust and agnostic learning of conditional distributional treatment effects. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 6037–6060, 2023.

M. R. Kosorok. Bootstraps of sums of independent but not identically distributed stochastic processes. *Journal of Multivariate Analysis*, 84(2):299–318, 2003.

M. R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer New York, 2008.

D. Martinez Taboada, A. Ramdas, and E. Kennedy. An efficient doubly-robust test for the kernel treatment effect. In *Advances in Neural Information Processing Systems*, volume 36, pages 59924–59952, 2023.

K. Muandet, M. Kanagawa, S. Saengkyongam, and S. Marukatat. Counterfactual mean embeddings. *Journal of Machine Learning Research*, 22(162):1–71, 2021.

J. Näf, C. Emmenegger, P. Bühlmann, and N. Meinshausen. Confidence and uncertainty assessment for distributional random forests. *Journal of Machine Learning Research*, 24(366):1–77, 2023.

J. Park, U. Shalit, B. Schölkopf, and K. Muandet. Conditional distributional treatment effect with kernel conditional mean embeddings and U-statistic regression. In *Proceedings of 38th International Conference on Machine Learning (ICML)*, volume 139, pages 8401–8412, 2021.

G. Pisier. *Martingales in Banach Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2016.

J. Praestgaard and J. A. Wellner. Exchangeably Weighted Bootstraps of the General Empirical Process. *The Annals of Probability*, 21(4):2053–2086, 1993.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

D. B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100 (469):322–331, 2005.

C.-J. Simon-Gabriel, A. Barp, B. Schölkopf, and L. Mackey. Metrizing weak convergence with maximum mean discrepancies. *Journal of Machine Learning Research*, 24(184):1–20, 2023.

B. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.

H. Umegaki and A. Bharucha-Reid. Banach space-valued random variables and tensor products of banach spaces. *Journal of Mathematical Analysis and Applications*, 31(1):49–67, 1970.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Preprint arXiv:1510.04342*, 2017.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113 (523):1228–1242, 2018.

S. Wager and J. Taylor. *Causal Inference: A Statistical Learning Approach*. Draft, 2023. URL `https://web.stanford.edu/~swager/causal_inf_book.pdf`. Draft version available online.

H. Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004. doi: 10.1017/CBO9780511617539.

J. Zhao and D. Meng. FastMMD: Ensemble of circular discrepancy for efficient two-sample test. *Neural computation*, 27(6):1345–1372, 2015.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] See Section 3 and Appedix C.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] Appendix C.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Section 4.
   (b) Complete proofs of all theoretical results. [Yes] See Appendices D and E.
   (c) Clear explanations of any assumptions. [Yes] See Section 4 and Appendix D.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See Section 5 and Appendices A and B.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] See Section 5 and Appendices A and B.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A VARIABLES USED IN THE 401(k) ELIGIBILITY STUDY

In table 1, we summarize the covariates used in the 401(k) eligibility study, investigated in Section 5.

Name	**Description**	**Type**
inc	income	continuous
age	age	continuous
hown	home ownership	binary
db	defined benefit pension status	binary
pira	IRA participation	binary
educ	years of completed education	continuous
fsize	family size	continuous
two_earn	whether dual-earning household	binary
marr	marital status	binary

Table 1: Description of Variables in $\mathbf{X}$.

B SIMULATION STUDY

Inspired by (Wager and Athey, 2018; Ćevid et al., 2022; Näf et al., 2023) we define a general simulation framework which we use to define data generating processes that combine conditional treatment effects and confounding. Let $p = 5$ be the dimension of the baseline covariates. The generic variable $O = (\mathbf{X}, W, Y)$ is drawn according to the following laws:

$$\begin{aligned} \mathbf{X} &\sim \text{Unif}(0,1)^K, \\ W|\mathbf{X} &\sim \text{Bernoulli}(e(\mathbf{X})), \\ Y|W,\mathbf{X} &\sim \text{Normal}(2X_3 - 1 + (W - 0.5)t(\mathbf{X}), 1), \end{aligned}$$

where e is a function defining the probability of treatment and t defines the conditional treatment effect. Next we define four regimes, with the choices of e and t reflecting Näf et al. (2023):

1. No confounding and no conditional treatment effect: $e(\mathbf{X}) = 0.5$ and $t(\mathbf{X}) = 0$.
2. Confounding: $e(\mathbf{X}) = 0.25(1 + \beta_{2,4}(X_3))$, where $\beta_{a,b}$ is the density of the Beta distribution with parameters a and b, and $t(\mathbf{X}) = 0$.
3. Effect: $e(\mathbf{X}) = 0.5$ and $t(\mathbf{X}) = \eta(X_1) \cdot \eta(X_2)$, where $\eta(x) = 1 + (1 + \exp(-20(x - 1/3)))^{-1}$.
4. Confounding and effect: $e(\mathbf{X})$ is defined as in the confounding regime and $t(\mathbf{X})$ is defined as in the effect regime.

In addition, we investigated an alternative outcome model in which there was no shared structure between treatment arms, and the errors were heavy-tailed:

$$\begin{aligned} &Y \mid W, \mathbf{X} = 2X_3 - 1 - 0.5\eta(X_1) \cdot \eta(X_2) + \epsilon, \text{ when } W = 0, \\ &Y \mid W, \mathbf{X} = X_3^2 - 2X_1 \cdot X_2 + 2\epsilon, \text{ when } W = 1, \\ &\quad \text{and } \epsilon \sim \text{StudentT}(3), \end{aligned} \tag{22}$$

where StudentT(3) denotes Student's t distribution with 3 degrees of freedom. The alternative outcome model was used to define alternative versions of simulation regimes 3 and 4.

The simulation datasets were constructed by drawing n iid copies of O from one of the above regimes. We present empirical results based on 500 simulation datasets constructed for each regime and for $n \in \{250, 500, 1000, 5000\}$. Causal DRF was run with $N = 2500$ trees with 50 trees per group ($B = 50$). As a comparison, we apply the method described by (Näf et al., 2023) as a benchmark in which two DRFs are fit separately to treatment and control observations, again each with 2500 trees and 50 trees per group.

Both methods were used to estimate the conditional witness function evaluated at a test point $(0.7, 0.3, 0.5, 0.68, 0.43)^T$. In addition, we draw $8'000$ observations from the conditional distribution $\mathbf{Y} \mid \mathbf{X} = \mathbf{x}$ to obtain an approximation to the true underlying witness function $\tau_k(\mathbf{x})$. The mean absolute error

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{\tau}_{n,k}(\mathbf{x})(\mathbf{Y}_i) - \tau_k(\mathbf{x})(\mathbf{Y}_i)|$$

and the empirical coverage of the estimated 95% confidence bands were then calculated across all simulations for each sample size. The results are shown in Table 2. For the case of the shared-strcture, light tail outcome model, both Causal-DRF and the benchmark method achieve similar MAE across Regime 1, although under the confounding regime (Regime 2), fitting two DRFs appears to be more accurate in terms of MAE. Although this is surprising, the reverse is true as soon as there is an effect, such that in Regimes 3 and 4, the MAE is up to 30% lower for Causal-DRF. Crucially, while the empirical coverage of both methods tends to be conservative in scenarios without a treatment effect (Regimes 1 and 2), in regimes with a treatment effect (Regimes 3 and 4), Causal-DRF has near-optimal 95% empirical coverage at all sample sizes considered, while the benchmark method significantly undercovers for sample sizes less than 5000. Thus, while there is a small loss in accuracy for the two cases without effect, Causal-DRF provides accurate coverage for all cases and even for small sample sizes. Figure 5 provides intuition for these results in the case of $n = 1000$ and with confounding and treatment effect. It can be seen that the witness function estimated by fitting two DRFs, though also relatively accurate, is further away from the true line in gray overall. In addition, the confidence bands are narrower. This combination of less accurate estimation and narrower bands leads to a higher likelihood of the confidence bands not capturing the true witness function.

For the heavy-tailed case with no shared structure in (22), DRF somewhat outperforms Causal-DRF, as one might expect. However, the difference is relatively small in terms of MAE and mostly leads to an overcoverage of Causal-DRF, while the coverage for DRF is very close to the nominal level.

The performance of Causal-DRF with regard to the CATE depended on the outcome model. For the settings with a null outcome model, Causal-DRF exhibited performance comparable to Causal-RF. Interestingly, Causal-DRF struggled in the first (Gaussian) outcome model setting for CATE estimation, exhibiting poor coverage and higher MAE than Causal-RF; however, it performed better than Causal-RF for the second (heavy-tailed) outcome model.

	Witness Function				CATE					
	MAE		95% Coverage		MAE			95% Coverage		
n	DRF	Causal-DRF	DRF	Causal-DRF	DRF	Causal-DRF	Causal-RF	DRF	Causal-DRF	Causal-RF
no confounding, no effect (1)										
250	0.035	0.041	100.0%	100.0%	0.134	0.147	0.125	97.1%	97.0%	94.9%
500	0.030	0.033	99.4%	100.0%	0.120	0.125	0.103	97.5%	97.5%	94.8%
1000	0.027	0.029	100.0%	100.0%	0.103	0.108	0.089	97.8%	97.8%	94.8%
5000	0.019	0.022	100.0%	100.0%	0.072	0.083	0.065	98.9%	98.9%	97.1%
confounding (2)										
250	0.058	0.060	94.8%	99.6%	0.298	0.224	0.136	71.3%	87.5%	94.7%
500	0.035	0.048	99.6%	99.2%	0.194	0.172	0.109	85.3%	92.3%	94.6%
1000	0.027	0.039	99.8%	100.0%	0.144	0.143	0.091	91.0%	94.3%	95.3%
5000	0.020	0.027	100.0%	100.0%	0.082	0.101	0.067	97.5%	97.7%	97.0%
effect (3)										
250	0.066	0.065	78.2%	97.0%	0.858	0.705	0.483	14.8%	27.2%	55.3%
500	0.060	0.059	80.2%	96.4%	0.726	0.643	0.307	16.9%	26.4%	74.6%
1000	0.054	0.053	84.4%	97.4%	0.594	0.590	0.214	23.4%	27.6%	82.0%
5000	0.040	0.041	95.8%	96.0%	0.270	0.484	0.118	59.9%	33.1%	90.7%
confounding and effect (4)										
250	0.072	0.070	77.6%	97.4%	0.851	0.748	0.485	27.3%	31.9%	57.6%
500	0.062	0.061	86.0%	96.8%	0.709	0.688	0.315	28.7%	30.0%	74.7%
1000	0.055	0.053	91.8%	96.8%	0.583	0.641	0.220	32.6%	29.1%	82.7%
5000	0.039	0.037	97.6%	99.2%	0.279	0.529	0.124	56.8%	32.8%	90.4%
effect (3), heavy-tailed (22)										
250	0.043	0.048	97.6%	99.2%	0.410	0.437	0.435	86.9%	92.7%	86.9%
500	0.039	0.042	98.8%	99.6%	0.325	0.376	0.370	91.1%	93.9%	86.8%
1000	0.033	0.037	99.0%	100.0%	0.270	0.331	0.321	93.4%	94.7%	87.5%
5000	0.026	0.031	99.8%	99.6%	0.203	0.263	0.246	96.5%	96.8%	91.1%
confounding and effect (4), heavy-tailed (22)										
250	0.050	0.054	95.0%	99.2%	0.440	0.452	0.438	84.3%	93.8%	88.2%
500	0.041	0.044	97.2%	100.0%	0.346	0.396	0.378	88.6%	94.8%	87.6%
1000	0.034	0.038	99.0%	100.0%	0.296	0.355	0.331	91.2%	95.3%	87.6%
5000	0.027	0.032	99.2%	100.0%	0.217	0.290	0.251	95.8%	96.7%	90.7%

Table 2: Simulation study results showing the Mean Absolute Error (MAE) and 95% empirical coverage for the conditional witness function evaluated at the test point $(0.7, 0.3, 0.5, 0.68, 0.43)^\top$.

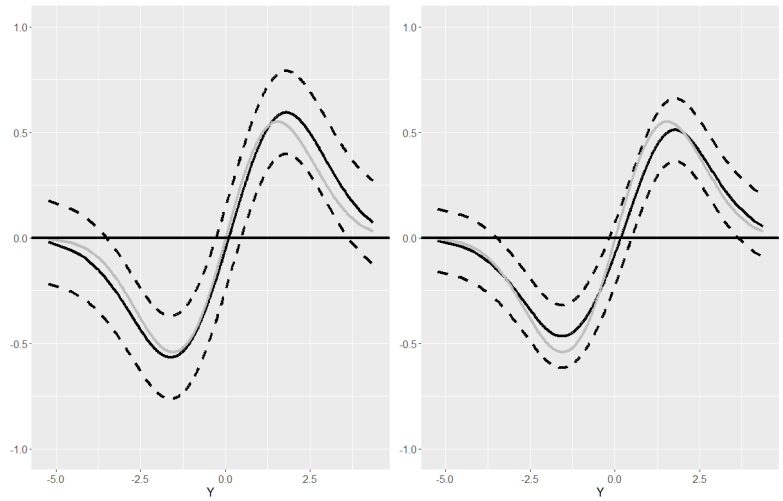

Figure 5: Simulation setting of both confounding and effect for $n = 1000$. The black solid lines are the estimated witness functions, while the dashed lines are the estimated 95% confidence bands. The true witness function is given in gray. Left: Causal-DRF, Right: DRF.

C ALGORITHM AND DETAILS ON THE SPLITTING CRITERION

Following (Ćevid et al., 2022), we adopt an approximate kernel $\tilde{k}$ that is a fast random approximation of the MMD statistic (Zhao and Meng, 2015). Such an approximation allows for the computational efficiency necessary to allow the number of kernel computations that are required when constructing a large forest.

By Bochner's theorem (see e.g. (Wendland, 2004, Theorem 6.6)), any bounded shift-invariant kernel can be written as

$$k(\mathbf{u}, \mathbf{v}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^T(\mathbf{u}-\mathbf{v})} d\nu(\boldsymbol{\omega}). \tag{23}$$

This can be interpreted as a Fourier transform of some measure ν. By randomly sampling the frequency vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_B$ from normalized ν, we can approximate our kernel k up to a scaling factor by the approximate kernel $\tilde{k}$ via:

$$k(\mathbf{u}, \mathbf{v}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^T(\mathbf{u}-\mathbf{v})} d\nu(\boldsymbol{\omega}) \approx \frac{1}{B} \sum_{b=1}^{B} e^{i\boldsymbol{\omega}_b^T(\mathbf{u}-\mathbf{v})} = \tilde{k}(\mathbf{u}, \mathbf{v}),$$

where $\tilde{k}(\mathbf{u}, \mathbf{v}) = \langle \widetilde{\boldsymbol{\varphi}}(\mathbf{u}), \widetilde{\boldsymbol{\varphi}}(\mathbf{v}) \rangle_{\mathbb{C}^B}$ is the kernel function with the feature map given by

$$\widetilde{\boldsymbol{\varphi}}(\mathbf{u}) = \frac{1}{\sqrt{B}} \left(\tilde{\varphi}_{\boldsymbol{\omega}_1}(\mathbf{u}), \ldots, \tilde{\varphi}_{\boldsymbol{\omega}_B}(\mathbf{u}) \right)^T = \frac{1}{\sqrt{B}} \left(e^{i\boldsymbol{\omega}_1^T \mathbf{u}}, \ldots, e^{i\boldsymbol{\omega}_B^T \mathbf{u}} \right)^T,$$

a random vector of the Fourier features $\widetilde{\varphi}_{\boldsymbol{\omega}}(\mathbf{u}) = e^{i\boldsymbol{\omega}^T \mathbf{u}} \in \mathbb{C}$ (Rahimi and Recht, 2008). Our default choice of kernel k is the Gaussian kernel with bandwidth σ, as there is a straightforward expression for ν. We can then sample $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_B \sim N_d(\mathbf{0}, \sigma^{-2} I_d)$. The bandwidth σ is chosen as the median pairwise distance between all training responses $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$(the 'median heuristic'; Gretton et al. 2012).

Our splitting criterion seeks to maximize

$$\begin{aligned}
&\frac{|\mathcal{I}_L||\mathcal{I}_R|}{(|\mathcal{I}_R| + |\mathcal{I}_L|)^2} \Big\| \sum_{i \in \mathcal{I}_L} \nu_{i,R} k(\mathbf{Y}_i, \cdot) - \sum_{i \in \mathcal{I}_R} \nu_{i,R} k(\mathbf{Y}_i, \cdot) \Big\|_{\mathcal{H}}^2 = \\
&\frac{|\mathcal{I}_L||\mathcal{I}_R|}{(|\mathcal{I}_R| + |\mathcal{I}_L|)^2} \Big(\sum_{i \in \mathcal{I}_L, j \in \mathcal{I}_L} v_{i,L} v_{j,L} k(\mathbf{Y}_i, \mathbf{Y}_j) + \sum_{i \in \mathcal{I}_R, j \in \mathcal{I}_R} v_{i,R} v_{j,R} k(\mathbf{Y}_i, \mathbf{Y}_j) \\
&- 2 \sum_{i \in \mathcal{I}_R, j \in \mathcal{I}_L} v_{i,R} v_{j,L} k(\mathbf{Y}_i, \mathbf{Y}_j) \Big),
\end{aligned} \tag{24}$$

Simulation Regime	n	Runtime (seconds) DRF		Causal-DRF		Causal-RF
no confounding, no effect (1)	250	1.8	(4.6)	1.1	(2.7)	0.4
	500	3.1	(4.5)	2.6	(3.8)	0.7
	1000	10.2	(7.0)	6.3	(4.4)	1.4
	5000	219.4	(21.7)	60.4	(6.0)	10.1
confounding (2)	250	1.9	(4.5)	1.1	(2.5)	0.4
	500	3.2	(4.7)	2.5	(3.7)	0.7
	1000	10.4	(7.4)	6.2	(4.4)	1.4
	5000	217.8	(23.6)	57.0	(6.2)	9.2
effect (3)	1.7	(4.1)	1.1	(2.5)	0.4	
	500	3.3	(4.9)	2.6	(3.9)	0.7
	1000	10.2	(7.5)	6.6	(4.8)	1.4
	5000	214.8	(23.2)	57.5	(6.2)	9.2
confounding and effect (4)	250	1.7	(4.0)	1.1	(2.5)	0.4
	500	3.4	(4.7)	2.5	(3.6)	0.7
	1000	10.3	(7.3)	6.5	(4.6)	1.4
	5000	268.9	(26.5)	66.7	(6.6)	10.1
effect (3), heavy tailed (22)	250	1.8	(4.0)	1.1	(2.5)	0.4
	500	3.2	(4.3)	2.6	(3.5)	0.7
	1000	10.5	(6.9)	6.6	(4.3)	1.5
	5000	274.4	(25.6)	65.9	(6.2)	10.7
confounding and effect (4), heavy tailed (22)	250	1.7	(3.9)	1.1	(2.6)	0.4
	500	3.4	(4.5)	2.6	(3.5)	0.7
	1000	10.8	(7.5)	6.5	(4.6)	1.4
	5000	316.1	(29.7)	71.1	(6.7)	10.7

Table 3: Simulation study results comparing the average runtime in seconds of each algorithm. The ratios of the runtimes of the DRF and Causal-DRF algorithms compared to that of the Causal-RF algorithm are shown in parentheses.

which is a weighted version of the MMD statistic used in (Ćevid et al., 2022), with weights

$$\nu_{i,L} = \frac{W_i}{|\mathcal{I}_L, W_i = 1|} - \frac{1 - W_i}{|\mathcal{I}_L, W_i = 0|}$$

and similarly $\nu_{i,R}$. Using the approximate $\tilde{k}$ in (24) results in:

$$\frac{1}{B}\sum_{b=1}^{B} \frac{|\mathcal{I}_L||\mathcal{I}_R|}{(|\mathcal{I}_R| + |\mathcal{I}_L|)^2} \left| \sum_{i \in \mathcal{I}_L} \nu_{i,L} \tilde{\varphi}_{\boldsymbol{\omega}_b}(\mathbf{Y}_i) - \sum_{i \in \mathcal{I}_R} \nu_{i,R} \tilde{\varphi}_{\boldsymbol{\omega}_b}(\mathbf{Y}_i) \right|^2, \tag{25}$$

which is the splitting criterion used in the forest.

The overall algorithm is summarized in Algorithm 1.

D PRELIMINARIES

The following assumptions are placed on the choice of kernel.

(**K1**) k is bounded, i.e., $\sup_{\mathbf{y}_1, \mathbf{y}_2} k(\mathbf{y}_1, \mathbf{y}_2) < \infty$.

(**K2**) $(\mathbf{y}_1, \mathbf{y}_2) \mapsto k(\mathbf{y}_1, \mathbf{y}_2)$ is (jointly) continuous.

(**K3**) The kernel is translation-invariant, i.e. $k(\mathbf{y}_1, \mathbf{y}_2) = k_0(\mathbf{y}_1 - \mathbf{y}_2)$ for some function k_0.

(**K4**) k is integrally strictly positive definite (denoted by $\int$spd), that is

$$\|\Phi(Q_1) - \Phi(Q_2)\|_{\mathcal{H}} = 0 \implies Q_1 = Q_2, \text{ for all } Q_1, Q_2 \in \mathcal{M}_b(\mathbb{R}^d);$$

Algorithm 1 Pseudocode for Causal-DRF. The functions BuildSubForest and GetWeights are defined in Algorithm 2.

procedure BuildForest2(set of samples $\mathcal{D} = \{(\mathbf{x}_i, w_i, \mathbf{y}_i)\}_{i=1}^n$, number of trees N, number of groups B)
  $L \leftarrow$ round(N/B)
  for $b = 1, \dots, B$ **do**
    $\mathcal{S} \leftarrow$ Random Subsample from $\mathcal{D}$
    $\mathcal{F}_b \leftarrow$ BuildSubForest($\mathcal{S}$, L) ▷ Build bth forest
  end for
  return $\mathcal{F} = \{\mathcal{F}_1, \dots, \mathcal{F}_B\}$
end procedure

procedure GetWeights2(forests $\mathcal{F}$, test point $\mathbf{x}$) ▷ Computes the weighting function with uncertainty
  for $b = 1, \dots, |\mathcal{F}|$ **do**
    $w_b \leftarrow$ GetWeights($\mathcal{F}_b$, $\mathbf{x}$)
  end for
  $w = \frac{1}{B}\sum_{b=1}^B w_b$
  return $w, w_1, \dots, w_B$
end procedure

see for instance (Simon-Gabriel et al., 2023; Sriperumbudur, 2016).

All of these assumptions are met by the Gaussian kernel

$$k(\mathbf{y}, \cdot) = \exp\left(-\frac{\|\mathbf{y} - \cdot\|}{2\sigma^2}\right). \tag{26}$$

The forest assumptions with a refinement are discussed below in Section D.1.

First, we reproduce notation, definitions, and elementary results from (Näf et al., 2023). Let $(\Omega, \mathcal{A}, \mathbb{P})$ denote the underlying probability space. Throughout, let $(\mathcal{H}, \langle, \cdot, \rangle)$ denote the RKHS associated with the kernel k. We assume that k is *bounded* and *continuous* in its two arguments. Boundedness of k ensures that μ is indeed defined on all of $\mathcal{M}_b(\mathbb{R}^d)$, and continuity of $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ ensures that $\mathcal{H}$ is *separable*. Thus, we assume throughout that **(K2)** holds, such that measurability issues can be avoided. Let us denote by $\xi\colon (\Omega, \mathcal{A}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ a map from Ω to $\mathcal{H}$. Separability implies that such a map ξ is measurable if and only if $\langle \xi, f\rangle$ is measurable for all $f \in \mathcal{H}$. Moreover, it can easily be checked that $\Phi(P)$ is linear on $\mathcal{M}_b(\mathbb{R}^d)$. Separability of $\mathcal{H}$ and $\mathbb{E}[\|\xi\|_{\mathcal{H}}] < \infty$ mean that the integral

$$\mathbb{E}[\xi] = \int_\Omega \xi d\mathbb{P},$$

is well defined and that

$$F(\mathbb{E}[\xi]) = \mathbb{E}[F(\xi)],$$

for any continuous linear function $F\colon \mathcal{H} \to \mathbb{R}$.[1] In particular, $\mathbb{E}[\langle \xi, f\rangle] = \langle \mathbb{E}[\xi], f\rangle$ for all $f \in \mathcal{H}$. Moreover, for $q \geq 1$, denote by

$$\begin{aligned}
\mathcal{L}^q(\Omega, \mathcal{A}, \mathcal{H}) &= \{\xi\colon (\Omega, \mathcal{F}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H})) \text{ measurable, with } \mathbb{E}[\|\xi\|^q] < \infty]\} \\
\mathbb{L}^q(\Omega, \mathcal{A}, \mathcal{H}) &= \text{Set of equivalence classes in } \mathcal{L}^q(\Omega, \mathcal{A}, \mathcal{H}) \\
\mathrm{Var}(\xi) &= \mathbb{E}[\|\xi - \mathbb{E}[\xi]\|_{\mathcal{H}}^2] = \mathbb{E}[\|\xi\|_{\mathcal{H}}^2] - \|\mathbb{E}[\xi]\|_{\mathcal{H}}^2, \quad \xi \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}) \\
\mathrm{Cov}(\xi_1, \xi_2) &= \mathbb{E}[\langle \xi_1 - \mathbb{E}[\xi_1], \xi_2 - \mathbb{E}[\xi_2]\rangle].
\end{aligned}$$

Furthermore, it is well-known that $(\mathbb{L}^q, \|\cdot\|_{\mathbb{L}^q(\mathcal{H})})$ is a Banach space with

$$\|\xi\|_{\mathbb{L}^q(\mathcal{H})} = \mathbb{E}[\|\xi\|_{\mathcal{H}}^q]^{1/q}.$$

This allows us to also define *conditional* expectations. For a sub σ-algebra $\mathcal{F} \subset \mathcal{A}$ and an element $\xi \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H})$, the conditional expectation $\mathbb{E}[\xi \mid \mathcal{F}]$ is the (a.s.) unique element such that

[1] Here and below, $F(\xi)$ is meant to denote $F(\xi(\omega))$ for all $\omega \in \Omega$.

Algorithm 2 Pseudocode for building Subforests

procedure BUILDSUBFOREST(set of samples $\mathcal{S} = \{(\mathbf{x}_i, w_i, \mathbf{y}_i)\}_{i=1}^{n}$, number of trees N)
  for $i = 1, \ldots, N$ **do**
    $\mathcal{S}_{\text{subsample}} \leftarrow$ Subsample $\mathcal{S}$ as in **(F5)**
    $\mathcal{S}_{\text{build}}, \mathcal{S}_{\text{populate}} \leftarrow$ Split $\mathcal{S}_{\text{subsample}}$ as in (**F1**) ▷ $\mathcal{S}_{\text{build}}$ to determine tree splits, $\mathcal{S}_{\text{populate}}$ to populate the leaves
    $\mathcal{N}_i \leftarrow$ initialize root node using $\mathcal{S}_{\text{build}}$
    $\mathcal{T}_i \leftarrow$ BUILDTREE($\mathcal{N}_i$) ▷ Start recursion from the root node
    Populate leaves with $\mathcal{S}_{\text{populate}}$
  end for
  return $\mathcal{F} = \{\mathcal{T}_1, \ldots, \mathcal{T}_N\}$
end procedure

procedure BUILDTREE(current node $\mathcal{N}$) ▷ Recursively constructs the trees
  $\mathcal{S} \leftarrow$ Extract the samples in the node
  $\mathcal{I} \leftarrow$ Random set of candidate variables to perform a split on
  $\mathcal{C} \leftarrow$ Initialize list ▷ Here we store info about candidate splits
  for idx $\in \mathcal{I}$, level l **do** ▷ l iterates over all values of variable X_{idx}
    $\mathcal{S}_L, \mathcal{S}_R \leftarrow$ Split samples based on $(\mathbf{x}_i)_{i \in \text{idx}} \leq l$
    $v \leftarrow$ Calculate splitting criterion in (2) using $(\mathcal{S}_L, \mathcal{S}_R)$
    Add $(v, \mathcal{S}_L, \mathcal{S}_R, \text{idx}, l)$ to $\mathcal{C}$
  end for
  $\mathcal{S}_L, \mathcal{S}_R, \text{idx}, l \leftarrow$ find the best split in $\mathcal{C}$
  $\mathcal{N}_L \leftarrow$ Create new node with set of samples $\mathcal{S}_L$
  $\mathcal{N}_R \leftarrow$ Create new node with set of samples $\mathcal{S}_R$
  BUILDTREE($\mathcal{N}_L$), BUILDTREE($\mathcal{N}_R$) ▷ Proceed building recursively
  CHILDREN($\mathcal{N}$) $\leftarrow \mathcal{N}_L, \mathcal{N}_R$
  SPLIT($\mathcal{N}$) $\leftarrow$ idx, l ▷ Store the split
  return CHILDREN($\mathcal{N}$), SPLIT($\mathcal{N}$)
end procedure

procedure GETWEIGHTS(forest $\mathcal{F}$, test point $\mathbf{x}$) ▷ Computes the weighting function
  vector of weights $w =$ ZEROS(n) ▷ n is the training set size
  for $i = 1, \ldots, |\mathcal{F}|$ **do**
    $\mathcal{L} \leftarrow$ Get indices of training samples in same leaf as $\mathbf{x}$
    for idx $\in \mathcal{L}$ **do**
      $w[\text{idx}] = w[\text{idx}] + 1/(|\mathcal{L}| \cdot |\mathcal{F}|)$
    end for
  end for
  return w
end procedure

(CE1) $\mathbb{E}[\xi \mid \mathcal{F}]\colon (\Omega, \mathcal{F}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ is measurable and $\mathbb{E}[\xi \mid \mathcal{F}] \in \mathbb{L}^1(\Omega, \mathcal{F}, \mathcal{H})$,

(CE2) $\mathbb{E}[\xi \mathbf{1}_F] = \mathbb{E}[\mathbb{E}[\xi \mid \mathcal{F}]\mathbf{1}_F]$ for all $F \in \mathcal{F}$;

see for instance (Umegaki and Bharucha-Reid, 1970) or (Pisier, 2016, Chapter 1). Particularly, condition (CE2) implies that $\mathbb{E}[\mathbb{E}[\xi \mid \mathcal{F}]] = \mathbb{E}[\mathbb{E}[\xi \mid \mathcal{F}]\mathbf{1}_\Omega] = \mathbb{E}[\xi]$ due to $\Omega \in \mathcal{F}$ for any σ-algebra. It can also be shown that $F(\mathbb{E}[\xi \mid \mathcal{F}]) = \mathbb{E}[F(\xi) \mid \mathcal{F}]$ for all linear and continuous $F\colon \mathcal{H} \to \mathbb{R}$ and that $\|\mathbb{E}[\xi \mid \mathcal{F}]\|_{\mathcal{H}} \leq \mathbb{E}[\|\xi\|_{\mathcal{H}} \mid \mathcal{F}]$ (Pisier, 2016, Chapter 1). Moreover, it can be shown that

(CE3) For $\xi \in \mathbb{L}^2(\Omega, \mathcal{A}, \mathcal{H})$, $\mathbb{E}[\xi \mid \mathcal{F}]$ is the orthogonal projection onto $\mathbb{L}^2(\Omega, \mathcal{F}, \mathcal{H})$;

see (Umegaki and Bharucha-Reid, 1970). Although the conditional expectation $\mathbb{E}[\xi \mid \mathcal{F}]$, similarly to real-valued conditional expectations, is only defined a.s., we do not explicitly state this in our developments below. Denote $\mathbb{E}[\xi \mid \mathbf{X}] = \mathbb{E}[\xi \mid \sigma(\mathbf{X})]$. Proposition 6 and 7 in (Ćevid et al., 2022) show that this notion is well-defined and establishes further properties of Hilbert space-valued conditional expectations.

For two functions f and g with $\liminf_{s\to\infty} g(s) > 0$, we denote $f(s) = \mathcal{O}(g(s))$ if

$$\limsup_{s\to\infty} \frac{|f(s)|}{g(s)} \leq C$$

for some $C > 0$. If $C = 1$, we write $f(s) \precsim g(s)$. For a sequence of random variables $X_n\colon \Omega \to \mathbb{R}$ and a sequence of real numbers $a_n \in (0, +\infty)$, $n \in \mathbb{N}$, we write $X_n = \mathcal{O}_p(a_n)$ if

$$\lim_{M\to\infty} \sup_n \mathbb{P}(a_n^{-1}|X_n| > M) = 0,$$

that is, X_n is bounded in probability. We write $X_n = o_p(a_n)$ if $a_n^{-1}X_n$ converges to zero in probability. Similarly, for (S, d) a separable metric space, $\mathbf{X}_n\colon (\Omega, \mathcal{A}) \to (S, \mathcal{B}(S))$, $n \in \mathbb{N}$ and $\mathbf{X}\colon (\Omega, \mathcal{A}) \to (S, \mathcal{B}(S))$ measurable, we write $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$, if $d(\mathbf{X}_n, \mathbf{X}) = o_p(1)$. By separability, every random element ξ with values in $\mathcal{H}$ is tight (Dudley, 2002, Chapter 7.1). That is, for all $\epsilon > 0$, there is a compact $K_\epsilon \subset \mathcal{H}$ such that $\mathbb{P}(\xi \in K_\epsilon) \geq 1 - \epsilon$. More generally, uniform tightness of a sequence ξ_n, $n \in \mathbb{N}$ means that for all $\epsilon > 0$, there is a compact $K_\epsilon \subset \mathcal{H}$ such that

$$\inf_n \mathbb{P}(\xi_n \in K_\epsilon) \geq 1 - \epsilon.$$

Finally, let $\mathbf{X} \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}_1)$ and $\xi \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}_2)$, and assume that $A \subset \Omega$ depends on $\mathbf{X}$, $A = A(\mathbf{X})$. Thus, for $\mathbf{X}$ fixed to a certain value, A is a fixed set. If $\mathbb{P}(A \mid \mathbf{X}) > 0$ almost everywhere, we define

$$\mathbb{E}[\xi \mid \mathbf{X}, A] = \frac{\mathbb{E}[\xi \mathbf{1}_A \mid \mathbf{X}]}{\mathbb{P}(A \mid \mathbf{X})} \in \mathcal{L}^2(\Omega, \sigma(\mathbf{X}), \mathcal{H}_2).$$

Then, we have by construction that

$$\mathbb{E}[\xi \mathbf{1}_A \mid \mathbf{X}] = \mathbb{E}[\xi \mid \mathbf{X}, A] \cdot \mathbb{P}(A \mid \mathbf{X}). \tag{27}$$

For a more compact notation in the following Lemma, let $N = \{1, \ldots, n\}$, and let for $A \subset N$ and $k \leq |A|$, let $C_k(A)$ be the set of all subsets of size k drawn from A without replacement, with $C_0 = \emptyset$. The following lemma presents a U-statistic expansion that we afterwards apply to an individual tree of our DRF forest.

Lemma D.1 (Lemma 9 in (Ćevid et al., 2022))**.** *Let $(\mathcal{H}_1, \langle\cdot,\cdot\rangle_1)$ and $(\mathcal{H}_2, \langle\cdot,\cdot\rangle_2)$ be two separable Hilbert spaces, and let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be i.i.d. copies of a random element $\mathbf{Z}\colon (\Omega, \mathcal{A}) \to (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1))$. Write $\mathcal{Z}_n = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$, and let $T\colon (\mathcal{H}_1^n, \mathcal{B}(\mathcal{H}_1^n)) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ measurable with $\mathbb{E}[\|T(\mathcal{Z}_n)\|_{\mathcal{H}_2}^2] < \infty$. If T is symmetric, there exist functions T_j, $j = 1, \ldots, n$, such that*

$$T(\mathcal{Z}_n) = \mathbb{E}[T(\mathcal{Z}_n)] + \sum_{i=1}^n T_1(\mathbf{Z}_i) + \sum_{i_1<i_2} T_2(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}) + \cdots + T_n(\mathcal{Z}_n), \tag{28}$$

and it holds that

$$\mathrm{Var}(T(\mathcal{Z}_n)) = \sum_{i=1}^n \binom{n}{i} \mathrm{Var}(T_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i)) \tag{29}$$

and

$$T_1(\mathbf{Z}_i) = \mathbb{E}[T(\mathcal{Z}_n) \mid \mathbf{Z}_i] - \mathbb{E}[T(\mathcal{Z}_n)].$$

Recall from the main text that we use $\mathbf{Z}_i = (Y_i, W_i, \mathbf{X}_i)$, for $i = 1, \ldots, n$ and $\mathcal{Z}_n = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ and similarly $\mathcal{Z}_{s_n} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_{s_n}\}$.

D.1 Forest Assumptions and a Refinment

We first list the forest assumptions taken in (Wager and Athey, 2018):

(**F1**) (*Honesty*) The data used for constructing each tree is split into two halves; the first is used for determining the splits and the second for populating the leaves and thus for estimating the response. In addition, $(W_i, \mathbf{X}_i)$ from the second sample may be used to define the splits.

(**F2**) (*Random-split*) At every split point and for all feature dimensions $j = 1, \ldots, p$, the probability that the split occurs along the feature X_j is bounded from below by π/p for some $\pi > 0$.

(**F3**) (*Symmetry*) The (randomized) output of a tree does not depend on the ordering of the training samples.

(**F4**) (α*-regularity*) For the second sample in (**F1**), after splitting a parent node, each child node contains at least a fraction $\alpha \leq 0.2$ of the parent's training samples. Moreover, the trees are grown until every leaf contains between κ and $2\kappa - 1$ many observations *from each treatment group*, for some fixed tuning parameter $\kappa \in \mathbb{N}$.

(**F5**) (*Data sampling*) To grow a tree, a subsample of size s_n out of the n training data points is sampled. We consider $s_n = n^\beta$ with

$$1 > \beta > \left(1 + \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})} \frac{\pi}{p}\right)^{-1},$$

where α is chosen in (**F4**).

These conditions can be ensured by forest construction, although, as mentioned in (Wager and Athey, 2018), care must be taken for different test points $\mathbf{x}$, as (**F4**) must be ensured for each test point.

In the following, denote the indices of the first sample used to build the trees in (**F1**) as $\mathcal{I}_1$, and the second sample used to populate the leaves as $\mathcal{I}_2$. (Näf et al., 2023) assumed an equivalent of (**F1**)–(**F5**) to derive analogous results for DRF. Unfortunately, a crucial proof in (Näf et al., 2023), corresponding to Theorem E.5 below, only goes through under the assumption that for each $i \in \mathcal{I}_2$, $\mathbf{X}_i$ is independent of $\mathcal{L}(\mathbf{x})$, the leaf $\mathbf{x}$ falls into. However, while this is approximately true, since an independent sample is used for each tree to build $\mathcal{L}(\mathbf{x})$, there can be a certain dependence, as $\mathcal{L}(\mathbf{x})$ might be adjusted to ensure **(F4)**. That is, while a first $\mathcal{L}^0(\mathbf{x})$ is built completely independently of $(\mathbf{X}_i, W_i)$ from the second sample, $\mathcal{L}(\mathbf{x})$ is then adjusted based on $(\mathbf{X}_i, W_i)$ from the second sample to meet **(F4)**. Although the nature of this dependence is hard to control, we now put some mild assumptions to limit the dependence between $(\mathbf{X}_i, W_i)$ and $\mathcal{L}(\mathbf{x})$ in any forest construction meeting (**F1**)–(**F5**). In particular, we assume:

(**F1***) (*Honesty**) The second sample is used such that: for all $i \in \mathcal{I}_2$, for all f bounded,

$$\mathrm{Var}(\mathbb{E}[f(\mathcal{L}(\mathbf{x})) \mid \mathbf{X}_i, W_i]) = o(s_n^{-(1+\epsilon)}), \tag{30}$$

and

$$\begin{aligned}\mathbb{E}[\|\mathbb{E}[f((\mathbf{X}_j, W_j, \mathbf{1}\{\mathbf{X}_j \in \mathcal{L}(\mathbf{x})\})_{j \in \mathcal{I}_2, j \neq i}) \mid \mathbf{1}\{\mathbf{X}_i \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x}), \mathbf{X}_i, W_i] - \\ \mathbb{E}[f((\mathbf{X}_j, W_j, \mathbf{1}\{\mathbf{X}_j \in \mathcal{L}(\mathbf{x})\})_{j \in \mathcal{I}_2, j \neq i}) \mid \mathbf{1}\{\mathbf{X}_i \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2] = o(s_n^{-(1+\epsilon)}),\end{aligned} \tag{31}$$

for some $\epsilon > 0$.

Finally, for Causal-DRF we in addition need an assumption of the behavior of the weight in each class:

(**F6**) (*Group Variance*) There exists some $c \in [0, \infty]$, such that for all $i \in \mathcal{I}_2$,

$$\frac{\mathrm{Var}(\mathbb{E}[S_i \mid \mathbf{X}_i, W_i = 1])}{\mathrm{Var}(\mathbb{E}[S_i \mid \mathbf{X}_i, W_i = 0])} \to c. \tag{32}$$

Remark D.2. Both Assumption (**F1***) and (**F6**) are needed to prove Theorem E.5, which considers the asymptotic ratio of variances of the projected trees:

$$\lim_{n\to\infty} \frac{\mathrm{Var}(\langle \Gamma_n(\mathbf{Z}_1), f\rangle)}{\mathrm{Var}(\Gamma_n(\mathbf{Z}_1))}$$

With (30), we assume that the adaptation of $\mathcal{L}(\mathbf{x})$ by $(\mathbf{X}_i, W_i)$, $i \in \mathcal{I}_2$ is done in such a way that the dependence is negligible for each individual $(\mathbf{X}_i, W_i)$. Intuitively, one might expect this variance to be $o(s_n^{-1})$, even if the second sample was freely used to build $\mathcal{L}(\mathbf{x})$, as $(\mathbf{X}_i, W_i)$ is one of s_n points used to obtain $\mathcal{L}(\mathbf{x})$. The additional ϵ encodes that $\mathcal{L}(\mathbf{x})$ is originally build from an independent set of observations and $(\mathbf{X}_i, W_i)$, $i \in \mathcal{I}_2$ is only used to adapt $\mathcal{L}(\mathbf{x})$ such that **(F4)** holds. Similarly, (31) demands that, given $\mathcal{L}(\mathbf{x})$ and $\mathbf{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x})\}$, the dependence of $(\mathbf{X}_i, W_i)$ on all other observations be negligible. Although $(\mathbf{X}_i, W_i)_{i\in\mathcal{I}_2}$ are independent, conditioning on $\mathcal{L}(\mathbf{x})$ and $\mathbf{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x})\}$ will induce (conditional) dependence in general. With (31) we assume that this dependence decreases sufficiently fast as $n \to \infty$.

As we are only interested in the ratio of variances with identical trees, the denominator and numerator deal with the same dependence of $(\mathbf{X}_i, W_i)$ on $\mathcal{L}(\mathbf{x})$. Thus, it might be possible to further weaken assumption (**F1***). However, (**F1***) is much more plausible for a forest construction satisfying (**F1**)–(**F5**) than the full independence of $\mathcal{L}(\mathbf{x})$ and the second sample implicitly assumed by (Näf et al., 2023). Both assumptions (without W_i) are also needed for the original DRF, as we discuss in a correction of a key proof argument of (Näf et al., 2023) in Section D.2.

Assumption (**F6**) is the same as Assumption (30) in (Näf et al., 2023), ensuring the convergence of the ratio by restricting how the variances can behave in each group. Without it can be shown that there exists constants $0 \le c_1 \le c_2 < \infty$, such that

$$c_1 \le \lim_{n\to\infty} \frac{\mathrm{Var}(\langle \Gamma_n(\mathbf{Z}_1), f\rangle)}{\mathrm{Var}(\Gamma_n(\mathbf{Z}_1))} \le c_2.$$

(**F6**) is a sufficient condition that ensures that the ratio does not indefinitely fluctuate between these values as $n \to \infty$.

Remark D.3. Owing to Assumption (**F1**), one should write:

$$\sum_{i\in\mathcal{I}_2} S_i k(\mathbf{Y}_i, \cdot)$$

instead of

$$\sum_{i=1}^{s_n} S_i k(\mathbf{Y}_i, \cdot).$$

However, following (Wager and Athey, 2018) and (Näf et al., 2023) we only acknowledge the sample splitting in the tree when it is necessary, such as in the proof of Theorem E.5.

Before we continue with the main proofs, it is instructive to present the corrected argument of a crucial part of the proof of Theorem 5 in (Näf et al., 2023). These arguments can then directly be reused to obtain important claims to ultimately proof Theorem E.5 below.

D.2 Improving upon the proof of Theorem 5 in [Näf et al., 2023]

We restate here the assumptions of (Näf et al., 2023): We refer here to the assumptions (**F1**)–(**F5**) from (Näf et al., 2023) as (**DRF-F1**)–(**DRF-F5**). These are essentially the assumptions (**F1**)–(**F5**) shown here when W_i is seen as part of $\mathbf{X}_i$. We repeat here the assumptions on the data taken in (Näf et al., 2023):

(**DRF-D1**) The covariates $\mathbf{X}_1, \dots, \mathbf{X}_n$ are independent and identically distributed on $[0,1]^p$ with a density bounded away from 0 and infinity.

(**DRF-D2**) The mapping $\mathbf{x} \mapsto \mu(\mathbf{x}) = \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}] \in \mathcal{H}$ is Lipschitz.

(**DRF-D3**) The mapping $\mathbf{x} \mapsto \mathbb{E}[\|k(\mathbf{Y}, \cdot)\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}]$ is Lipschitz.

(**DRF-D4**) $\mathrm{Var}(k(\mathbf{Y},\cdot) \mid \mathbf{X}=\mathbf{x}) = \mathbb{E}[\|k(\mathbf{Y},\cdot)\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}] - \|\mathbb{E}[k(\mathbf{Y},\cdot) \mid \mathbf{X}=\mathbf{x}]\|_{\mathcal{H}}^2 > 0$.

(**DRF-D5**) $\mathbb{E}[\|k(\mathbf{Y},\cdot) - \mu(\mathbf{x})\|_{\mathcal{H}}^{2+\delta} \mid \mathbf{X}=\mathbf{x}] \leq M$, for some constants δ, M uniformly over $\mathbf{x} \in [0,1]^d$.

(**DRF-D6**) For all $f \in \mathcal{H} \setminus \{0\}$, $\mathrm{Var}(\langle k(\mathbf{Y},\cdot), f\rangle \mid \mathbf{X}=\mathbf{x}) = \mathrm{Var}(f(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}) > 0$.

(**DRF-D7**) For all $f \in \mathcal{H} \setminus \{0\}$, $\mathbf{x} \mapsto \mathbb{E}[|f(\mathbf{Y})|^2 \mid \mathbf{X}=\mathbf{x}]$ is Lipschitz.

Instead of demanding (finite-sample) independence, we ask that the dependence between $\mathbf{X}_i$, $i \in \mathcal{I}_2$ and the leaf $\mathcal{L}(\mathbf{x})$ decays at a certain rate, namely:

(**DRF-F1***) (*Honesty**) The second sample is used such that: for all $i \in \mathcal{I}_2$, for all f bounded,

$$\mathrm{Var}(\mathbb{E}[f(\mathcal{L}(\mathbf{x})) \mid \mathbf{X}_i]) = o(s_n^{-(1+\epsilon)}). \tag{33}$$

and

$$\begin{aligned}\mathbb{E}[\|\mathbb{E}[f((\mathbf{X}_j, \mathbf{1}\{\mathbf{X}_j \in \mathcal{L}(\mathbf{x})\})_{j\in\mathcal{I}_2, j\neq i}) \mid \mathbf{1}\{\mathbf{X}_i \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x}), \mathbf{X}_i] - \\ \mathbb{E}[f((\mathbf{X}_j, \mathbf{1}\{\mathbf{X}_j \in \mathcal{L}(\mathbf{x})\})_{j\in\mathcal{I}_2, j\neq i}) \mid \mathbf{1}\{\mathbf{X}_i \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2] = o(s_n^{-(1+\epsilon)})\end{aligned} \tag{34}$$

With (33), we assume that the adaptation of $\mathcal{L}(\mathbf{x})$ by $\mathbf{X}_i$, $i \in \mathcal{I}_2$ is done in such a way that the dependence is negligible for each individual $\mathbf{X}_i$. Intuitively we might expect this variance to be $o(s_n^{-1})$, even if the second sample was freely used to build $\mathcal{L}(\mathbf{x})$, as $\mathbf{X}_i$ is one of s_n points used to obtain $\mathcal{L}(\mathbf{x})$. The additional ϵ encodes that $\mathcal{L}(\mathbf{x})$ is originally build from an independent set of observations and $\mathbf{X}_i$, $i \in \mathcal{I}_2$ is only used to adapt $\mathcal{L}(\mathbf{x})$ such that **(F4)** holds. This is a convenient and mild assumption that allows for an elegant correction of the proof in (Näf et al., 2023), although it might be possible to weaken it. This assumption holds in particular, if $\mathcal{L}(\mathbf{x})$ is independent of $\mathbf{X}_i$, $i \in \mathcal{I}_2$.

Similarly (34) is a mild assumption that holds in particular in the following: Assume the sample in $\mathcal{I}_2$ is used to adapt an original $\mathcal{L}^0(\mathbf{x})$ build from independent data in $\mathcal{I}_1$, based solely on whether $\mathbf{X}_i \in \mathcal{L}(\mathbf{x})$, as for instance shown in Figure 6. Then it holds that,

$$(\mathbf{X}_j, \mathbf{1}\{\mathbf{X}_j \in \mathcal{L}(\mathbf{x})\})_{j\in\mathcal{I}_2, j\neq i} \perp\!\!\!\perp \mathbf{X}_i \mid \mathcal{L}(\mathbf{x}), \mathbf{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x})\}. \tag{35}$$

Thus in this plausible scenario, a condition much stronger than (34) holds.

To establish the setting of (Näf et al., 2023), define in the following the number of data points belonging to the same leaf as $\mathbf{x}$ as $N_{\mathbf{x}} = |\{j \colon \mathbf{X}_j \in \mathcal{L}(\mathbf{x})\}|$ and let

$$S_i = \frac{\mathbf{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x})\}}{N_{\mathbf{x}}}, \tag{36}$$

be the weight associated with each observation i in a tree $T(\mathcal{Z}_{s_n})$ with $\mathcal{E}$ integrated out, i.e. $T(\mathcal{Z}_{s_n}) = \mathbb{E}[T(\mathbf{x}, \mathcal{E}; \mathcal{Z}_{s_n})$, such that

$$T(\mathcal{Z}_{s_n}) = \sum_{i=1}^{s_n} S_i k(\mathbf{Y}_i, \cdot).$$

We will make use the following property of the S_i:

$$1 = \mathbb{E}\left[\sum_{i=1}^{s_n} S_i\right] = \sum_{i=1}^{s_n} \mathbb{E}[S_i] = s_n \mathbb{E}[S_1]. \tag{37}$$

In particular,

$$\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1]) \leq \mathbb{E}[\mathbb{E}[S_1|\mathbf{X}_1]^2] \leq \mathbb{E}[\mathbb{E}[S_1|\mathbf{X}_1]] = \mathbb{E}[S_1] = \mathcal{O}(s_n^{-1}) \tag{38}$$

We define

$$T'(\mathcal{Z}_{s_n}) = T(\mathcal{Z}_{s_n}) \mathbf{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w}\}, \tag{39}$$

$$S_i' = S_i \mathbf{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w}\}, \text{ where } w = \frac{1}{2}\frac{\pi}{p}\frac{\log\left((1-\alpha)^{-1}\right)}{\log(\alpha^{-1})}, \tag{40}$$

so that $T'(\mathcal{Z}_{s_n}) = \sum_{i=1}^{s_n} S_i' k(\mathbf{Y}_i, \cdot)$. We now formulate the corrected version of Theorem 5 in (Näf et al., 2023) with (**DRF-F1***) added:

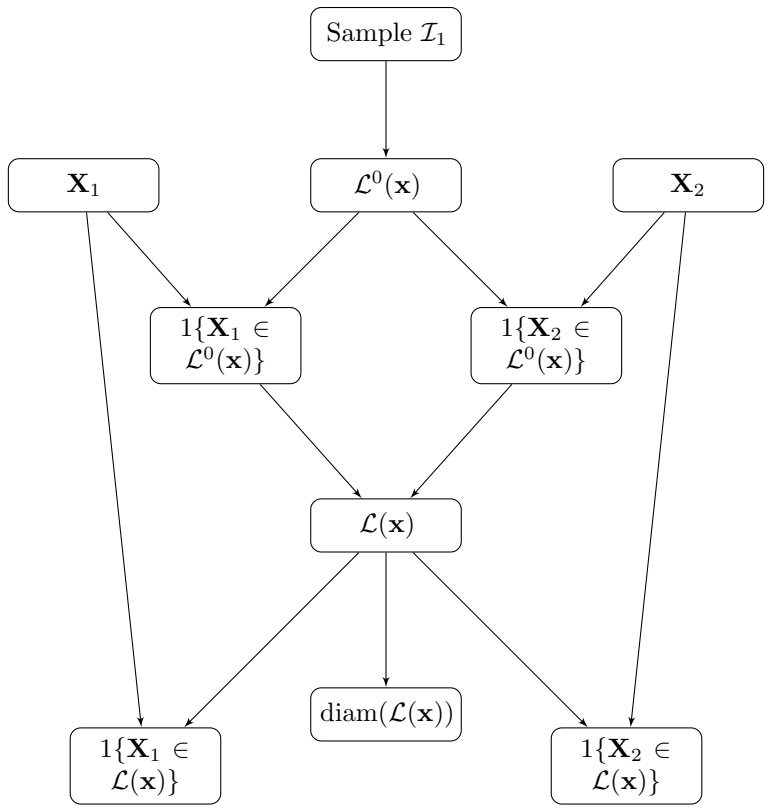

Figure 6: Potential Tree building approach for $n = 2$ that satisfies (34). Crucially, $\mathcal{L}(\mathbf{x})$ is an adapted version of $\mathcal{L}^0(\mathbf{x})$ based only on whether $\mathbf{X}_i$, $i \in \mathcal{I}_2$ is included in $\mathcal{L}^0(\mathbf{x})$.

Theorem D.4. *Assume conditions* ***(DRF-F1)–(DRF-F5)****,* ***(DRF-F1*)****,* ***(DRF-D1)–(DRF-D7)****,* ***(K1)****, and* ***(K2)*** *hold. Then, for all $f \in \mathcal{H} \setminus \{0\}$, we have*

$$\lim_{n\to\infty} \frac{\text{Var}(\langle T_n(\mathbf{Z}_1), f\rangle)}{\text{Var}(T_n(\mathbf{Z}_1))} = \frac{\text{Var}(\langle k(\mathbf{Y}, \cdot), f\rangle | \mathbf{X} = \mathbf{x})}{\text{Var}(k(\mathbf{Y}, \cdot)|\mathbf{X} = \mathbf{x})} = \sigma^2(f) > 0. \tag{41}$$

Proof. We only consider and correct the proof of the crucial Claim (90) in (Näf et al., 2023), that is we prove:

Claim:

$$\begin{aligned} \text{Var}(\mathbb{E}[\langle T'(\mathcal{Z}_{s_n}), f\rangle \mid \mathbf{X}_1]) &= o(s_n^{-(1+\epsilon)}) \\ \text{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n})|\mathbf{X}_1]) &= o(s_n^{-(1+\epsilon)}). \end{aligned} \tag{42}$$

Proof: First, due to honesty (**DRF-F1**), we have

$$\mathbb{E}[\langle T'(\mathcal{Z}_{s_n}), f\rangle \mid \mathbf{X}_1] = \mathbb{E}[S_1' \mid \mathbf{X}_1]\mathbb{E}[\langle k(\mathbf{Y}_1, \cdot), f\rangle \mid \mathbf{X}_1] + \sum_{i=2}^{s_n} \mathbb{E}[S_i' \langle k(\mathbf{Y}_i, \cdot), f\rangle \mid \mathbf{X}_1]$$

and

$$\mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1] = \mathbb{E}[S_1' \mid \mathbf{X}_1]\mathbb{E}[k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1] + \sum_{i=2}^{s_n} \mathbb{E}[S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1].$$

Subsequently, we consider the variance of the two terms and their covariance individually. First, we study the variance of the first terms.

Claim: For some $\epsilon > 0$,

$$\text{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X}_1]) = \text{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\|\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X} = \mathbf{x}]\|_{\mathcal{H}}^2 + o(s_n^{-(1+\epsilon)}) \tag{43}$$

and

$$\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[\langle k(\mathbf{Y}_1,\cdot),f\rangle|\mathbf{X}_1]) = \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\mathbb{E}[\langle k(\mathbf{Y}_1,\cdot),f\rangle|\mathbf{X}=\mathbf{x}]^2 + o(s_n^{-(1+\epsilon)}). \tag{44}$$

<u>Proof:</u>

We only show (43) because (44) follows analogously. We have

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}_1]) &= \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}_1]-\mu(\mathbf{x})+\mu(\mathbf{x})))\\ &= \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x})) + \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}_1]-\mu(\mathbf{x})))\\ &\quad + \mathrm{Cov}\left(\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x}), \mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}_1]-\mu(\mathbf{x}))\right).\end{aligned} \tag{45}$$

Because

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x})) =& \mathbb{E}[\|(\mathbb{E}[S_1'|\mathbf{X}_1]-\mathbb{E}[S_1'])\mu(\mathbf{x})\|_{\mathcal{H}}^2]\\ =& \mathbb{E}[(\mathbb{E}[S_1'|\mathbf{X}_1]-\mathbb{E}[S_1'])^2]\|\mu(\mathbf{x})\|_{\mathcal{H}}^2\\ =& \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\|\mu(\mathbf{x})\|_{\mathcal{H}}^2,\end{aligned}$$

it follows that

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}_1]) &= \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\|\mu(\mathbf{x})\|_{\mathcal{H}}^2\\ &\quad + \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}_1]-\mu(\mathbf{x})))\\ &\quad + \mathrm{Cov}\left(\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x}), \mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}_1]-\mu(\mathbf{x}))\right).\end{aligned}$$

Because $\mathbb{E}[S_1'|\mathbf{X}_1]$ maps into $\mathbb{R}_{\geq 0}$, we have

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}_1]-\mu(\mathbf{x}))) &\leq \mathbb{E}[\|\mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}_1])-\mu(\mathbf{x}))\|_{\mathcal{H}}^2]\\ &= \mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1]^2\|(\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}_1]-\mu(\mathbf{x}))\|_{\mathcal{H}}^2]\\ &\leq \mathbb{E}[\mathbb{E}[S_1'^2|\mathbf{X}_1]C^2\|\mathbf{X}_1-\mathbf{x}\|_{\mathbb{R}^p}^2]\\ &\leq \mathbb{E}[S_1'^2]C^2 s_n^{-2w},\end{aligned}$$

where the last step followed because $\mathbb{E}[S_1'^2|\mathbf{X}_1]=0$, for $\|\mathbf{X}_1-x\|_{\mathbb{R}^p} > s_n^{-w}$ by definition of $S_1' = S_1\mathbf{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w}\}$. Since $\mathbb{E}[S_1'^2] \leq \mathbb{E}[S_1'] \leq \mathbb{E}[S_1] = \mathcal{O}(s_n^{-1})$, we have

$$\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}_1]-\mu(\mathbf{x}))) = \mathcal{O}(s_n^{-(1+2w)}).$$

Finally, we infer from the Cauchy-Schwarz inequality,

$$\begin{aligned}&|\mathrm{Cov}\left(\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x}), \mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}_1]-\mu(\mathbf{x}))\right)|\\ \leq& \sqrt{\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x}))}\sqrt{\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}_1]-\mu(\mathbf{x})))}\\ =& \mathcal{O}(s_n^{-(1+w)}),\end{aligned}$$

due to

$$\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x})) = \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\cdot\mathcal{O}(1) = \mathcal{O}(s_n^{-1}). \tag{46}$$

Thus, choosing $0 < \epsilon < w$, Claim (43) holds. □

Before we continue proving the theorem, we note that, due to honesty (**DRF-F1**), we have

$$\begin{aligned}\sum_{i=2}^{s_n}\mathbb{E}[S_i'k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_1] &= \sum_{i=2}^{s_n}\mathbb{E}[\mathbb{E}[S_i'k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_i,\mathbf{X}_1]\mid\mathbf{X}_1]\\ &= \sum_{i=2}^{s_n}\mathbb{E}[\mathbb{E}[S_i'\mid\mathbf{X}_i,\mathbf{X}_1]\mathbb{E}[k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_i,\mathbf{X}_1]\mid\mathbf{X}_1]\\ &= \sum_{i=2}^{s_n}\mathbb{E}[\mathbb{E}[S_i'\mathbb{E}[k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_i]\mid\mathbf{X}_i,\mathbf{X}_1]\mid\mathbf{X}_1]\\ &= \sum_{i=2}^{s_n}\mathbb{E}[S_i'\mathbb{E}[k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_i]\mid\mathbf{X}_1].\end{aligned} \tag{47}$$

Now, we consider the variance of the sum in (47):

Claim: For some $\epsilon > 0$,

$$\mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1]\right) = \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\|\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X} = \mathbf{x}]\|_{\mathcal{H}}^2 + o(s_n^{-(1+\epsilon)}) \tag{48}$$

and

$$\mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i' \langle k(\mathbf{Y}_i, \cdot), f\rangle \mid \mathbf{X}_1]\right) = \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\mathbb{E}[\langle k(\mathbf{Y}_1, \cdot), f\rangle|\mathbf{X} = \mathbf{x}]^2 + o(s_n^{-(1+\epsilon)}). \tag{49}$$

Proof:

We only show (48), because (49) follows analogously. First we note that, using the definition of S_i', it holds that

$$\begin{aligned}\sum_{i=1}^{s_n} \mathbb{E}[S_i' \mid \mathbf{X}_1] &= \mathbb{E}[\sum_{i=1}^{s_n} S_i' \mid \mathbf{X}_1] \\ &= \mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1\right).\end{aligned}$$

It follows that

$$\sum_{i=2}^{s_n} \mathbb{E}[S_i' \mid \mathbf{X}_1] = \mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1\right) - \mathbb{E}[S_1' \mid \mathbf{X}_1] \tag{50}$$

By (50)

$$\begin{aligned}\mathrm{Var}\left(\mu(\mathbf{x})\sum_{i=2}^{s_n} \mathbb{E}[S_i' \mid \mathbf{X}_1]\right) &= \|\mu(\mathbf{x})\|_{\mathcal{H}}^2 \mathrm{Var}\left(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1\right) - \mathbb{E}[S_1' \mid \mathbf{X}_1]\right) \\ &= \|\mu(\mathbf{x})\|_{\mathcal{H}}^2 \mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1]) \\ &\quad + \|\mu(\mathbf{x})\|_{\mathcal{H}}^2 \mathrm{Var}(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1\right)) \\ &\quad - 2\|\mu(\mathbf{x})\|_{\mathcal{H}}^2 \mathrm{Cov}(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1\right), \mathbb{E}[S_1' \mid \mathbf{X}_1])\end{aligned}$$

Using Assumption (**DRF-F1***), it holds that

$$\mathrm{Var}(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) > s_n^{-w} \mid \mathbf{X}_1\right)) = o(s_n^{-(1+\epsilon)}),$$

where $\mathrm{diam}(\mathcal{L}(\mathbf{x}))$ is bounded because $[0,1]^p$ is a bounded set. Combining this with the fact that $\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1]) = \mathcal{O}(s_n^{-1})$, we get

$$\begin{aligned}&\mathrm{Cov}(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1\right), \mathbb{E}[S_1' \mid \mathbf{X}_1]) \\ &\leq (\mathrm{Var}(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1\right)))^{1/2}(\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1]))^{1/2} \\ &= o(s_n^{-(1+\epsilon/2)}).\end{aligned}$$

Thus,

$$\mathrm{Var}\left(\mu(\mathbf{x})\sum_{i=2}^{s_n} \mathbb{E}[S_i' \mid \mathbf{X}_1]\right) = \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\|\mu(\mathbf{x})\|_{\mathcal{H}}^2 + o(s_n^{-(1+\epsilon)}). \tag{51}$$

Thus, we need to show that

Claim: For some $\epsilon > 0$,

$$\mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1]\right) = \mathrm{Var}\left(\mu(\mathbf{x})\sum_{i=2}^{s_n} \mathbb{E}[S_i' \mid \mathbf{X}_1]\right) + o(s_n^{-(1+\epsilon)}), \tag{52}$$

Proof: Claim (52) is implied by

$$\mathrm{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_1]-\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1]\right)=\mathcal{O}(s_n^{-(1+2w)}), \tag{53}$$

$$\mathrm{Var}\left(\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1]\right)=\mathcal{O}(s_n^{-1}), \tag{54}$$

Indeed it holds that,

$$\begin{aligned}
&\mathrm{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_1]\right)\\
&=\mathrm{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_1]-\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1]+\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1]\right)\\
&=\mathrm{Var}(\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1])+\mathrm{Var}(\sum_{i=2}^{s_n}\mathbb{E}[S_i'k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_1]-\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1])\\
&+2\mathrm{Cov}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_1]-\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1],\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1]\right).
\end{aligned}$$

Thus (53) and (54), combined with the fact that

$$\begin{aligned}
&\mathrm{Cov}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_1]-\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1],\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1]\right)\leq\\
&\left(\mathrm{Var}(\sum_{i=2}^{s_n}\mathbb{E}[S_i'k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_1]-\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1])\right)^{1/2}\left(\mathrm{Var}\left(\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1]\right)\right)^{1/2},
\end{aligned}$$

gives (52). Condition (54) is an immediate consequence of (51) and the fact that $\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])=\mathcal{O}(s_n^{-1})$. Subsequently, we establish (53). Now, with (47), we have

$$\begin{aligned}
&\mathrm{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_1]-\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1]\right)\\
&=\mathrm{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mathbb{E}[k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_i]\mid\mathbf{X}_1]-\mathbb{E}[S_i'\mu(\mathbf{x})\mid\mathbf{X}_1]\right)\\
&=\mathrm{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'(\mathbb{E}[k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_i]-\mu(\mathbf{x}))\mid\mathbf{X}_1]\right)\\
&=\mathrm{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\Delta(\mathbf{X}_i)\mid\mathbf{X}_1]\right),
\end{aligned} \tag{55}$$

with $\Delta(\mathbf{X}_i)=\mathbb{E}[k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_i]-\mu(\mathbf{x})$.

Claim:

$$\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\mathcal{L}(\mathbf{x}),\mathbf{X}_1]-\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2]=o(s_n^{-(1+\epsilon)}). \tag{56}$$

Proof: Before we go on, note that for any conditioning set $\mathcal{F}$,

$$\begin{aligned}\left\|\sum_{i=2}^{s_n}\mathbb{E}[S_i'\Delta(\mathbf{X}_i)\mid\mathcal{F}]\right\|_{\mathcal{H}} &\leq \mathbb{E}\left[\sum_{i=2}^{s_n}S_i'\left\|\Delta(\mathbf{X}_i)\right\|_{\mathcal{H}}\mid\mathcal{F}\right]\\ &\leq \mathbb{E}\left[\sum_{i=2}^{s_n}S_i'C\left\|\mathbf{X}_i-\mathbf{x}\right\|_{\mathbb{R}^p}\mid\mathcal{F}\right]\\ &\leq Cs_n^{-w}\mathbb{E}\left[\sum_{i=2}^{s_n}S_i'\mid\mathcal{F}\right]\\ &\leq Cs_n^{-w},\end{aligned}\tag{57}$$

We now show that:

Claim: For some $\epsilon>0$,

$$\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\mathcal{L}(\mathbf{x}),\mathbf{X}_1]-\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\{\mathbf{X}_1\notin\mathcal{L}(\mathbf{x})\},\mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2]=o(s_n^{-(1+\epsilon)}).\tag{58}$$

Proof: We first note that

$$\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\mathcal{L}(\mathbf{x}),\mathbf{X}_1]-\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\mathbf{1}\{\mathbf{X}_1\notin\mathcal{L}(\mathbf{x})\},\mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2]=o(s_n^{-(1+\epsilon)}),\tag{59}$$

for some $\epsilon>0$. Indeed, since $\mathbf{1}\{\mathbf{X}_1\notin\mathcal{L}(\mathbf{x})\}$ is a deterministic function of $\mathbf{X}_1$, $\mathcal{L}(\mathbf{x})$, we have that

$$\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\mathcal{L}(\mathbf{x}),\mathbf{X}_1]=\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\mathbf{1}\{\mathbf{X}_1\notin\mathcal{L}(\mathbf{x})\},\mathcal{L}(\mathbf{x}),\mathbf{X}_1],\tag{60}$$

and since $f((\mathbf{X}_j,\mathbf{1}\{\mathbf{X}_j\in\mathcal{L}(\mathbf{x})\})_{j\in\mathcal{I}_2,j\neq i})=\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)$ is bounded, we can use (34) in (**DRF-F1***):

$$\begin{aligned}&\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\mathcal{L}(\mathbf{x}),\mathbf{X}_1]-\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\mathbf{1}\{\mathbf{X}_1\notin\mathcal{L}(\mathbf{x})\},\mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2]=\\ &\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\mathbf{1}\{\mathbf{X}_1\notin\mathcal{L}(\mathbf{x})\},\mathcal{L}(\mathbf{x}),\mathbf{X}_1]-\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\mathbf{1}\{\mathbf{X}_1\notin\mathcal{L}(\mathbf{x})\},\mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2]\\ &=o(s_n^{-(1+\epsilon)}),\end{aligned}$$

showing (59). Starting from (59), we note that

$$\begin{aligned}\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\mathbf{1}\{\mathbf{X}_1\notin\mathcal{L}(\mathbf{x})\},\mathcal{L}(\mathbf{x})]=\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\{\mathbf{X}_1\notin\mathcal{L}(\mathbf{x})\},\mathcal{L}(\mathbf{x})]\mathbf{1}\{\mathbf{X}_1\notin\mathcal{L}(\mathbf{x})\}+\\ \mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\{\mathbf{X}_1\in\mathcal{L}(\mathbf{x})\},\mathcal{L}(\mathbf{x})]\mathbf{1}\{\mathbf{X}_1\in\mathcal{L}(\mathbf{x})\}\end{aligned}$$

and thus,

$$\begin{aligned}&\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\mathbf{1}\{\mathbf{X}_1\notin\mathcal{L}(\mathbf{x})\},\mathcal{L}(\mathbf{x})]-\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\{\mathbf{X}_1\notin\mathcal{L}(\mathbf{x})\},\mathcal{L}(\mathbf{x})]\mathbf{1}\{\mathbf{X}_1\notin\mathcal{L}(\mathbf{x})\}\|_{\mathcal{H}}^2]\\ &=\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n}S_i'\Delta(\mathbf{X}_i)\mid\{\mathbf{X}_1\in\mathcal{L}(\mathbf{x})\},\mathcal{L}(\mathbf{x})]\mathbf{1}\{\mathbf{X}_1\in\mathcal{L}(\mathbf{x})\}\|_{\mathcal{H}}^2]\\ &\leq\mathbb{E}[\mathbb{E}[\sum_{i=2}^{s_n}S_i'\|\Delta(\mathbf{X}_i)\|_{\mathcal{H}}\mid\{\mathbf{X}_1\in\mathcal{L}(\mathbf{x})\},\mathcal{L}(\mathbf{x})]^2\mathbf{1}\{\mathbf{X}_1\in\mathcal{L}(\mathbf{x})\}]\\ &\leq Cs_n^{-2w}\mathbb{P}(\mathbf{X}_1\in\mathcal{L}(\mathbf{x}))\\ &=\mathcal{O}(s_n^{-(1+2w)}),\end{aligned}\tag{61}$$

utilizing (57). Similarly,

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\mathbf{1}\{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\} - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2] \\
&= \mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\mathbf{1}\{\mathbf{X}_1 \in \mathcal{L}(\mathbf{x})\}\|_{\mathcal{H}}^2] \\
&\leq C s_n^{-2w}\mathbb{P}(\mathbf{X}_1 \in \mathcal{L}(\mathbf{x})) \\
&= \mathcal{O}(s_n^{-(1+2w)}).
\end{aligned} \tag{62}$$

Repeatedly using $\|\xi_1 + \xi_2\|_{\mathcal{H}}^2 \leq 2\|\xi_1\|_{\mathcal{H}}^2 + 2\|\xi_2\|_{\mathcal{H}}^2$ and then utilizing (59), (61) and (62), we obtain

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x}), \mathbf{X}_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2] \leq \\
&2\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x}), \mathbf{X}_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathbf{1}\{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2] + \\
&4\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathbf{1}\{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\mathbf{1}\{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\}\|_{\mathcal{H}}^2] + \\
&4\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\mathbf{1}\{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\} - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2] \\
&= \mathcal{O}(s_n^{-(1+2w)}),
\end{aligned}$$

showing (58). □

Similarly, we have that

$$\begin{aligned}
\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] = \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\mathbb{P}(\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x}) \mid \mathcal{L}(\mathbf{x})) + \\
\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \{\mathbf{X}_1 \in \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\mathbb{P}(\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}) \mid \mathcal{L}(\mathbf{x}))
\end{aligned}$$

and, since $\mathbb{E}[\mathbb{P}(\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}) \mid \mathcal{L}(\mathbf{x}))^2] \leq \mathbb{E}[\mathbb{P}(\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}) \mid \mathcal{L}(\mathbf{x}))]$, we can use the same approach as in (61) and (62) to get:

$$\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2] = o(s_n^{-(1+\epsilon)}). \tag{63}$$

Using (58) and (63), we get:

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x}), \mathbf{X}_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2] \\
&\leq 2\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x}), \mathbf{X}_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2] \\
&+ 2\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2] \\
&= o(s_n^{-(1+\epsilon)}),
\end{aligned}$$

proving Claim (56). □

Thus it follows that

$$\begin{aligned}
&\mathrm{Var}(\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathbf{X}_1]) \\
&= \mathrm{Var}(\mathbb{E}[\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] \mid \mathbf{X}_1]) + \mathrm{Var}(\mathbb{E}[\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] \mid \mathbf{X}_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathbf{X}_1]) \\
&+ 2\mathrm{Cov}(\mathbb{E}[\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] \mid \mathbf{X}_1], \mathbb{E}[\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] \mid \mathbf{X}_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathbf{X}_1])
\end{aligned}$$

First, note that by assumption (**DRF-F1***), and the boundedness of $\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})]$,

$$\mathrm{Var}(\mathbb{E}[\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] \mid \mathbf{X}_1]) = o(s_n^{-(1+\epsilon)}).$$

Moreover, using Jensen's inequality and (56),

$$\begin{aligned}
&\mathrm{Var}(\mathbb{E}[\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] \mid \mathbf{X}_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathbf{X}_1]) \\
&\leq \mathbb{E}[\|\mathbb{E}[\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathbf{X}_1, \mathcal{L}(\mathbf{x})] \mid \mathbf{X}_1]\|_{\mathcal{H}}^2] \\
&\leq \mathbb{E}[\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathbf{X}_1, \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2 \mid \mathbf{X}_1]] \\
&= \mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathbf{X}_1, \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2] \\
&= o(s_n^{-(1+\epsilon)}).
\end{aligned}$$

Adding the Cauchy-Schwarz inequality to bound the covariances, we obtain

$$\mathrm{Var}(\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta(\mathbf{X}_i) \mid \mathbf{X}_1]) = o(s_n^{-(1+\epsilon)}),$$

showing (53). As outlined above, (53) and (54) imply (52). □

In turn, (52) combined with (51) proves Claim (48). □

Finally, we consider the covariance between $\mathbb{E}[S_1'k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1]$ and $\sum_{i=2}^{s_n} \mathbb{E}[S_i'k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1]$.

<u>Claim:</u> For some $\epsilon > 0$, we have

$$\begin{aligned}
&\mathrm{Cov}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i'k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1], \mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X}_1]\right) \\
&= -\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\|\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X} = \mathbf{x}]\|_{\mathcal{H}}^2 + o(s_n^{-(1+\epsilon)}) \qquad (64)
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{Cov}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i'\langle k(\mathbf{Y}_i, \cdot), f\rangle \mid \mathbf{X}_1], \mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[\langle k(\mathbf{Y}_1, \cdot), f\rangle|\mathbf{X}_1]\right) \\
&= -\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\mathbb{E}[\langle k(\mathbf{Y}_1, \cdot), f\rangle|\mathbf{X} = \mathbf{x}]^2 + o(s_n^{-(1+\epsilon)}). \qquad (65)
\end{aligned}$$

<u>Proof:</u> Again, we only show (64), because (65) follows analogously. Using (47), we can subtract and add

$\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'|\mathbf{X}_1]$ and $\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x})$ to obtain

$$\begin{aligned}
&\mathrm{Cov}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_1],\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}_1]\right)\\
=&\mathrm{Cov}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mathbb{E}[k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_i]\mid\mathbf{X}_1],\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}_1]\right)\\
=&\mathrm{Cov}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\Delta(\mathbf{X}_i)\mid\mathbf{X}_1]+\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1],\mathbb{E}[S_1'\Delta(\mathbf{X}_1)|\mathbf{X}_1]+\mu(\mathbf{x})\mathbb{E}[S_1'|\mathbf{X}_1]\right)\\
=&\mathrm{Cov}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\Delta(\mathbf{X}_i)\mid\mathbf{X}_1],\mathbb{E}[S_1'\Delta(\mathbf{X}_1)|\mathbf{X}_1]\right)+\mathrm{Cov}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\Delta(\mathbf{X}_i)\mid\mathbf{X}_1],\mu(\mathbf{x})\mathbb{E}[S_1'|\mathbf{X}_1]\right)\\
&+\mathrm{Cov}\left(\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1],\mathbb{E}[S_1'\Delta(\mathbf{X}_1)|\mathbf{X}_1]\right)+\mathrm{Cov}\left(\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1],\mu(\mathbf{x})\mathbb{E}[S_1'|\mathbf{X}_1]\right)\\
=&(I)+(II)+(III)+(IV),
\end{aligned}$$

where again $\Delta(\mathbf{X}_i)=\mathbb{E}[k(\mathbf{Y}_i,\cdot)\mid\mathbf{X}_i]-\mu(\mathbf{x})$. Since from (50),

$$\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'|\mathbf{X}_1]=\mu(\mathbf{x})(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x}))\le s_n^{-w}\mid\mathbf{X}_1\right)-\mathbb{E}[S_1'\mid\mathbf{X}_1]),\tag{66}$$

it holds that

$$\begin{aligned}
(IV)&=\mathrm{Cov}\left(\mu(\mathbf{x})(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x}))\le s_n^{-w}\mid\mathbf{X}_1\right)-\mathbb{E}[S_1'|\mathbf{X}_1]),\mu(\mathbf{x})\mathbb{E}[S_1'|\mathbf{X}_1]\right)\\
&=\|\mu(\mathbf{x})\|_{\mathcal{H}}^2\left(\mathrm{Cov}(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x}))\le s_n^{-w}\mid\mathbf{X}_1\right),\mathbb{E}[S_1'|\mathbf{X}_1])-\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\right)\\
&=-\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\|\mu(\mathbf{x})\|_{\mathcal{H}}^2+o(s_n^{-(1+\epsilon)}),
\end{aligned}$$

where the last step again followed by Cauchy–Schwarz and (**DRF-F1***):

$$\begin{aligned}
&\mathrm{Cov}(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x}))\le s_n^{-w}\mid\mathbf{X}_1\right),\mathbb{E}[S_1'|\mathbf{X}_1])\\
&\le\mathrm{Var}(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x}))\le s_n^{-w}\mid\mathbf{X}_1\right))^{1/2}\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])^{1/2}\\
&=o(s_n^{-(1+\epsilon)}),
\end{aligned}$$

as $\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])=\mathcal{O}(s_n^{-1})$.

Subsequently, we show that the remaining terms are negligible. Due to the Cauchy–Schwarz inequality, we have

$$|(I)|\le\mathrm{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\Delta(\mathbf{X}_i)\mid\mathbf{X}_1]\right)^{1/2}\mathrm{Var}(\mathbb{E}[S_1'\Delta(\mathbf{X}_1)|\mathbf{X}_1])^{1/2}.$$

As proven above (combining (53) and (55)), $\mathrm{Var}(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\Delta(\mathbf{X}_i)\mid\mathbf{X}_1])=o(s_n^{-(1+\epsilon)})$, and it holds that

$$\mathrm{Var}(\mathbb{E}[S_1'\Delta(\mathbf{X}_1)|\mathbf{X}_1])\le\mathbb{E}[\mathbb{E}[S_1'\|\Delta(\mathbf{X}_1)\|_{\mathcal{H}}|\mathbf{X}_1]^2]=\mathcal{O}(s_n^{-(1+2w)})$$

holds. Consequently, $(I)=o(s_n^{-(1+\epsilon)})$. Similarly,

$$|(II)|\le\mathrm{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\Delta(\mathbf{X}_i)\mid\mathbf{X}_1]\right)^{1/2}\mathrm{Var}(\mu(\mathbf{x})\mathbb{E}[S_1'|\mathbf{X}_1])^{1/2}=o(s_n^{-(1+\epsilon)}),$$

as $\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\le\mathbb{E}[(S_1')^2]=\mathcal{O}(s_n^{-1})$. Finally,

$$\begin{aligned}
|(III)|&=|\mathrm{Cov}\left(\mu(\mathbf{x})(1-\mathbb{E}[S_1'\Delta(\mathbf{X}_1)|\mathbf{X}_1]),\mathbb{E}[S_1'\Delta(\mathbf{X}_1)|\mathbf{X}_1]\right)|\\
&=|-\|\mu(\mathbf{x})\|_{\mathcal{H}}^2\mathrm{Var}(\mathbb{E}[S_1'\Delta(\mathbf{X}_1)|\mathbf{X}_1])|\\
&=\mathcal{O}(s_n^{-(1+2w)})
\end{aligned}$$

as above. Thus (I), (II) and (III) are all $o(s_n^{-(1+\epsilon)})$, for (potentially different) $\epsilon > 0$. Combining this with (IV), we obtain Claim (64).

□

Combining (43), (48), and (64), we obtain

$$\begin{aligned}
&\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n})|\mathbf{X}_1]) \\
&= 2\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\|\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}=\mathbf{x}]\|_{\mathcal{H}}^2 - 2\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\|\mathbb{E}[k(\mathbf{Y}_1,\cdot)|\mathbf{X}=\mathbf{x}]\|_{\mathcal{H}}^2 + o(s_n^{-(1+\epsilon)}) \\
&= o(s_n^{-(1+\epsilon)})
\end{aligned}$$

proving (42).

□

□

E MAIN PROOFS

We recall from the main text that

$$\hat{\tau}_{n,k}(\mathbf{x}) = \binom{n}{s_n}^{-1} \sum_{i_1<i_2<\cdots<i_{s_n}} \mathbb{E}_{\mathcal{E}}\left[\Gamma(\mathbf{x};\mathcal{E},\{\mathbf{Z}_{i_1},\ldots,\mathbf{Z}_{i_{s_n}}\})\right],$$

whereby

$$\begin{aligned}
\mathbb{E}_{\mathcal{E}}\left[\Gamma(\mathbf{x};\mathcal{E},\{\mathbf{Z}_{i_1},\ldots,\mathbf{Z}_{i_{s_n}}\})\right] =& \frac{1}{|\{j:\mathbf{X}_j\in\mathcal{L}(\mathbf{x}),W_j=1\}|}\sum_{\mathbf{X}_i\in\mathcal{L}(\mathbf{x}),W_i=1} k(\mathbf{Y}_i,\cdot) - \\
& \frac{1}{|\{j:\mathbf{X}_i\in\mathcal{L}(\mathbf{x}),W_j=0\}|}\sum_{\mathbf{X}_i\in\mathcal{L}(\mathbf{x}),W_i=0} k(\mathbf{Y}_i,\cdot).
\end{aligned}$$

To abbreviate this, we will just write

$$\Gamma(\mathcal{Z}_{s_n}) = \mathbb{E}_{\mathcal{E}}\left[\Gamma(\mathbf{x};\mathcal{E},\{\mathbf{Z}_{i_1},\ldots,\mathbf{Z}_{i_{s_n}}\})\right] = \sum_{i=1}^{s_n} S_i k(\mathbf{Y}_i,\cdot) \tag{67}$$

with weights

$$S_{i1} = \frac{\mathbf{1}\{\mathbf{X}_i\in\mathcal{L}(\mathbf{x}),W_i=1\}}{|\{j:\mathbf{X}_j\in\mathcal{L}(\mathbf{x}),W_j=1\}|} \tag{68}$$

$$S_{i0} = -\frac{\mathbf{1}\{\mathbf{X}_i\in\mathcal{L}(\mathbf{x}),W_i=0\}}{|\{j:\mathbf{X}_j\in\mathcal{L}(\mathbf{x}),W_j=0\}|} \tag{69}$$

$$S_i = S_{i1} + S_{i0}. \tag{70}$$

We note that $\hat{\tau}_{n,k}(\mathbf{x})$ takes the sum over all possible subsets of size s_n. As such, the chosen subsets of size s_n might not have the indices $\mathbf{Z}_1,\ldots,\mathbf{Z}_{s_n}$ in general. To simplify notation we usually only consider the indices $i=1,\ldots,s_n$, appealing to the assumption of i.i.d. data.

Lemma E.1. *Suppose **(D1)** holds for $\mathbf{X}_1,\mathbf{X}_2,\ldots$ and **(F3)**, **(F4)** for $\Gamma(\mathcal{Z}_{s_n})$. Moreover, assume **(C1)** holds. Then, there is a constant $C_{f,p}$ depending on f and p such that,*

$$s_n\mathrm{Var}(\mathbb{E}[S_1|\mathbf{Z}_1]) \gtrsim \frac{\varepsilon}{\kappa}\frac{C_{f,p}}{\log(s_n)^p}, \tag{71}$$

*where $\varepsilon > 0$ is from the overlap condition in **(C1)**. When f is uniform over $[0,1]^p$, the bound holds with $C_{f,p} = 2^{-(p+1)}(p-1)!$. Moreover, it holds that for $j\in\{0,1\}$*

$$\begin{aligned}
&\sum_{i=1}^{s_n}|S_i| = 2, \sum_{i:W_i=j}|S_{ij}| = 1 \\
&\mathbb{E}[|S_1|] = \mathcal{O}(s_n^{-1}), \mathbb{E}[|S_{1j}|] = \mathcal{O}(s_n^{-1}), \mathbb{E}[|S_{1j}| \mid W_1=j] = \mathcal{O}(s_n^{-1}), \\
&\mathrm{Var}(\mathbb{E}[S_1\mid\mathbf{Z}_1]) = \mathcal{O}(s_n^{-1}), \mathbb{E}[S_1^2] = \mathcal{O}(s_n^{-1}), \mathrm{Var}(\mathbb{E}[S_{1j}\mid\mathbf{Z}_1]) = \mathcal{O}(s_n^{-1}), \mathbb{E}[S_{1j}^2] = \mathcal{O}(s_n^{-1}).
\end{aligned} \tag{72}$$

Proof. The proof of (71) is the same as in (Wager and Athey, 2017, p. 47) when proving incrementally.

We next note that for $i = 1, \dots, s_n$

$$|S_i| = \begin{cases} \frac{\mathbf{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x})\}}{|\{j : \mathbf{X}_j \in \mathcal{L}(\mathbf{x}), W_j = 1\}|} & \text{if } W_i = 1 \\ \frac{\mathbf{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x})\}}{|\{j : \mathbf{X}_j \in \mathcal{L}(\mathbf{x}), W_j = 0\}|} & \text{if } W_i = 0, \end{cases}$$

and that by (**F4**), $|\{j : \mathbf{X}_j \in \mathcal{L}(\mathbf{x}), W_j = l\}|$ is not empty, for $l \in \{0, 1\}$. Thus indeed $\sum_{i:W_i=j} |S_{ij}| = 1$, so that $\sum_{i=1}^{s_n} |S_i| = 2$. Thus it follows by the i.i.d assumption that

$$2 = \mathbb{E}[\sum_{i=1}^{s_n} |S_i|] = s_n \mathbb{E}[|S_1|]$$

showing $\mathbb{E}[|S_1|] = \mathcal{O}(s_n^{-1})$. Since $\mathbb{E}[|S_1|] = \mathbb{E}[|S_{11}|] + \mathbb{E}[|S_{10}|]$, this also shows that $\mathbb{E}[|S_{1j}|] = \mathcal{O}(s_n^{-1})$. By the overlap conditon (**C1**), we moreover have:

$$\mathbb{E}[|S_{1j}| \mid W_1 = j] = \frac{\mathbb{E}[|S_{1j}|]}{\mathbb{P}(W_1 = j)} \leq \frac{\mathbb{E}[|S_{1j}|]}{\min(\varepsilon, 1 - \varepsilon)},$$

showing that $\mathbb{E}[|S_{1j}| \mid W_1 = j] = \mathcal{O}(s_n^{-1})$ as well. Finally, using the fact that $|S_{ij}| \in [0, 1]$,

$$0 \leq \mathrm{Var}(\mathbb{E}[S_{1j} \mid \mathbf{Z}_1]) \leq \mathbb{E}[\mathbb{E}[S_{1j} \mid \mathbf{Z}_1]^2] \leq \mathbb{E}[\mathbb{E}[|S_{1j}| \mid \mathbf{Z}_1]] = \mathbb{E}[|S_{1j}|] = \mathcal{O}(s_n^{-1}),$$

and similarly for $\mathbb{E}[S_1^2]$. □

We note that by honesty (**F1**), $\mathbb{E}[S_1|\mathbf{Z}_1] = \mathbb{E}[S_1|\mathbf{X}_1, W_1]$, and so the same properties also hold, for $\mathbb{E}[S_1|\mathbf{X}_1, W_1]$, in particular

$$s_n \mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1, W_1]) \gtrsim \frac{\varepsilon}{\kappa} \frac{C_{f,p}}{\log(s_n)^p}. \tag{73}$$

This also implies that

$$\begin{aligned} \mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1, W_1]) &= \mathbb{E}[\mathbb{E}[S_1|\mathbf{X}_1, W_1]^2] - \mathbb{E}[S_1]^2 \\ &= \mathbb{E}[\mathbb{E}[S_1|\mathbf{X}_1, W_1]^2] + \mathcal{O}(s_n^{-2}). \end{aligned} \tag{74}$$

With (73), this means that the difference between $\mathrm{Var}(\mathbb{E}[S_1|\mathbf{Z}_1])$ and $\mathbb{E}[\mathbb{E}[S_1|\mathbf{Z}_1]^2]$ is negligible. Let

$$\mathrm{Var}(\mathbb{E}[S_{1j}|\mathbf{X}_1, W_1 = j]) = \mathbb{E}[\mathbb{E}[S_{1j}|\mathbf{X}_1, W_1]^2 \mid W_1 = j] - \mathbb{E}[S_{1j}|W_1 = j]^2. \tag{75}$$

Then,

$$\begin{aligned} &\mathbb{E}[\mathbb{E}[S_1|\mathbf{X}_1, W_1]^2] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{E}[\mathbb{E}[S_1|\mathbf{X}_1, W_1]^2 \mid W_1]]] \\ &= \mathbb{E}[\mathbb{E}[S_1|\mathbf{X}_1, W_1]^2 \mid W_1 = 1]\mathbb{P}(W_1 = 1) + \mathbb{E}[\mathbb{E}[S_1|\mathbf{X}_1, W_1]^2 \mid W_1 = 0]\mathbb{P}(W_1 = 0) \\ &= \mathrm{Var}(\mathbb{E}[S_{11}|\mathbf{X}_1, W_1 = 1])\mathbb{P}(W_1 = 1) + \mathrm{Var}(\mathbb{E}[S_{10}|\mathbf{X}_1, W_1 = 0])\mathbb{P}(W_1 = 0) + \mathcal{O}(s_n^{-2}) \end{aligned} \tag{76}$$

as for $j \in \{0, 1\}$

$$\mathbb{E}[S_{1j} \mid W_1 = j]^2 \mathbb{P}(W_1 = j) = \mathcal{O}(s_n^{-2}).$$

Combining (74) and (76), we get

$$\begin{aligned} \mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1, W_1]) &= \mathrm{Var}(\mathbb{E}[S_{11}|\mathbf{X}_1, W_1 = 1])\mathbb{P}(W_1 = 1) \\ &\quad + \mathrm{Var}(\mathbb{E}[S_{10}|\mathbf{X}_1, W_1 = 0])\mathbb{P}(W_1 = 0) + \mathcal{O}(s_n^{-2}). \end{aligned} \tag{77}$$

This helps to relate the behavior of $\mathrm{Var}(\mathbb{E}[S_{1j}|\mathbf{X}_1, W_1 = j])$ to the overall variance $\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1, W_1])$. In particular, either $\mathrm{Var}(\mathbb{E}[S_{11}|\mathbf{X}_1, W_1 = 1]) = \Omega((\log(s_n)^p s_n)^{-1})$ or $\mathrm{Var}(\mathbb{E}[S_{10}|\mathbf{X}_1, W_1 = 0]) = \Omega((\log(s_n)^p s_n)^{-1})$ or both. These arguments will be crucial for the proof of Theorem E.5 below. We now turn to the behavior of the bias:

Lemma E.2. *Assume **(F2)–(F4)** that is trained on data $\mathcal{Z}_{s_n}$. Suppose that assumption **(D1)** holds for $\mathbf{X}_1, \dots, \mathbf{X}_{s_n}$ and that **(C1)** holds. Then,*

$$\mathbb{P}\left(\operatorname{diam}(\mathcal{L}(\mathbf{x})) \geq \sqrt{p}\left(\frac{\varepsilon s_n}{2\kappa - 1}\right)^{-0.51\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}\right) \leq p\left(\frac{\varepsilon s_n}{2\kappa - 1}\right)^{-1/2\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}. \tag{78}$$

Proof. This follows directly from the arguments in (Wager and Athey, 2017, pages 34 and 46), in addition to a union bound over all dimensions of $\mathbf{X}$. In particular, instead of s_n we need to consider the minimal number of observations of the smaller class Γ receives, say $s_{n,min}$. By (**C1**), it holds a.s. that $s_{n,min} \succsim \varepsilon s_n$, such that $s_{n,min}$ can be replaced by εs_n. □

Corollary E.3. *In addition to the conditions of Lemma E.2, assume **(D2)**, **(C2)** and that the trees $\Gamma(\mathcal{Z}_{s_n})$ in the forest satisfy **(F1)**. Then, we have*

$$\|\mathbb{E}[\hat{\tau}_{n,k}(\mathbf{x})] - \tau_k(\mathbf{x})\|_{\mathcal{H}} = \mathcal{O}\left((\varepsilon s_n)^{-1/2\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}\right). \tag{79}$$

Proof. We recall that

$$\begin{aligned} S_i &= \frac{\mathbf{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x}), W_i = 1\}}{|\{j : \mathbf{X}_j \in \mathcal{L}(\mathbf{x}), W_j = 1\}|} - \frac{\mathbf{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x}), W_i = 0\}}{|\{j : \mathbf{X}_j \in \mathcal{L}(\mathbf{x}), W_j = 0\}|} \\ &= S_{i1} + S_{i0} \end{aligned}$$

Analogous to (Wager and Athey, 2017), we may write

$$\begin{aligned} &\|\mathbb{E}[\Gamma(\mathcal{Z}_{s_n})] - \tau_k(\mathbf{x})\|_{\mathcal{H}} \\ &= \|\mathbb{E}\left[\sum_{W_i=1} S_{i1} k(\mathbf{Y}_i, \cdot)\right] - \mathbb{E}\left[\sum_{W_i=0} (-S_{i0}) k(\mathbf{Y}_i, \cdot)\right] - \tau_k(\mathbf{x})\|_{\mathcal{H}} \\ &\leq \|\mathbb{E}[\sum_{W_i=1} S_{i1} k(\mathbf{Y}_i, \cdot)] - \mathbb{E}[k(\mathbf{Y}^{(1)}, \cdot) \mid \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}} + \\ &\|\mathbb{E}[\sum_{W_i=0} (-S_{i0}) k(\mathbf{Y}_i, \cdot)] - \mathbb{E}[k(\mathbf{Y}^{(0)}, \cdot) \mid \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}}. \end{aligned}$$

Adapting the arguments of (Ćevid et al., 2022, Lemma 12), using that $\sum_{i:W_i=1} S_{i1} = \sum_{i:W_i=0}(-S_{i0}) = 1$,

$$\begin{aligned} \mathbb{E}[\sum_{W_j=1} S_{i1} k(\mathbf{Y}_i, \cdot)] &= \mathbb{E}[\mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X} \in \mathcal{L}(\mathbf{x}), W = 1, \mathcal{L}(\mathbf{x})]] \\ &= \mathbb{E}[\mathbb{E}[k(\mathbf{Y}^{(1)}, \cdot) \mid \mathbf{X} \in \mathcal{L}(\mathbf{x}), \mathcal{L}(\mathbf{x})]], \end{aligned}$$

where the last argument follows by (**C2**). The same applies to the control group, and with (**D2**) thus we can directly apply the arguments in (Ćevid et al., 2022, Corollary 13) for each term separately to obtain

$$\|\mathbb{E}[\Gamma(\mathcal{Z}_{s_n})] - \tau_k(\mathbf{x})\|_{\mathcal{H}} = \mathcal{O}\left((\varepsilon s_n)^{-1/2\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}\right).$$

Since the bias of a tree is the same as the bias of the overall forest, we obtain (79). □

Denote by

$$\Gamma_n(\mathbf{Z}_i) = \mathbb{E}[\Gamma(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_i] - \mathbb{E}[T(\mathcal{Z}_{s_n})], \tag{80}$$

and by

$$\tilde{\Gamma}(\mathcal{Z}_{s_n}) = \sum_{i=1}^{s_n} \Gamma_n(\mathbf{Z}_i), \tag{81}$$

the first order approximation of $\Gamma(\mathcal{Z}_{s_n})$ as in Lemma D.1. Recall from the main text that $\mu^{(0)}(\mathbf{x}) = \mathbb{E}[k(\mathbf{Y}^0, \cdot) \mid \mathbf{X} = \mathbf{x}]$ and $\mu^{(1)}(\mathbf{x}) = \mathbb{E}[k(\mathbf{Y}^1, \cdot) \mid \mathbf{X} = \mathbf{x}]$. Moreover, we define the random element

$$\mu^{(W)}(\mathbf{x}) = \mu^{(0)}(\mathbf{x})(1 - W) + \mu^{(1)}(\mathbf{x})W, \tag{82}$$

and

$$\sigma^2(\mathbf{x}, W) = \mathbb{E}[\|k(\mathbf{Y}, \cdot) - \mu^{(W)}(\mathbf{x})\|^2_{\mathcal{H}} \mid \mathbf{X} = \mathbf{x}, W]. \tag{83}$$

Finally, for any $f \in \mathcal{H} \setminus \{0\}$, we also define

$$\sigma^2_f(\mathbf{x}, W) = \mathbb{E}[|\langle f, k(\mathbf{Y}, \cdot)\rangle - \langle f, \mu^{(W)}(\mathbf{x})\rangle|^2 \mid \mathbf{X} = \mathbf{x}, W]. \tag{84}$$

Finally we define

$$\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1, W_1] \mid W_1) = \mathrm{Var}(\mathbb{E}[S_{11}|\mathbf{X}_1, W_1 = 1])W_1 + \mathrm{Var}(\mathbb{E}[S_{10}|\mathbf{X}_1, W_1 = 0])(1 - W_1), \tag{85}$$

such that,

$$\mathbb{E}\Big[\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1, W_1] \mid W_1)\sigma^2(\mathbf{x}, W_1)\Big] \tag{86}$$

$$= \mathrm{Var}(\mathbb{E}[S_{11}|\mathbf{X}_1, W_1 = 1])\sigma^2(\mathbf{x}, 1)\mathbb{P}(W_1 = 1) + \mathrm{Var}(\mathbb{E}[S_{10}|\mathbf{X}_1, W_1 = 0])\sigma^2(\mathbf{x}, 0)\mathbb{P}(W_1 = 0), \tag{87}$$

where $\mathrm{Var}(\mathbb{E}[S_{1j}|\mathbf{X}_1, W_1 = j])$ is defined in (75).

In the following, we give our crucial result, Theorem E.5, which studies the behavior of $\mathrm{Var}(\mathbb{E}[\langle \Gamma(\mathcal{Z}_{s_n}), f\rangle \mid \mathbf{Z}_1])$, for any $f \in \mathcal{H}$, and $\mathrm{Var}(\mathbb{E}[\Gamma(\mathcal{Z}_{s_n})|\mathbf{Z}_1])$. In particular Theorem E.5 implies that:

1. The ratio of $\mathrm{Var}(\mathbb{E}[\langle \Gamma(\mathcal{Z}_{s_n}), f\rangle \mid \mathbf{Z}_1])$ to $\mathrm{Var}(\mathbb{E}[\Gamma(\mathcal{Z}_{s_n})|\mathbf{Z}_1])$ converges to a real number under (**F6**). This will be crucial in showing univariate convergence, which in turn is a crucial ingredient in showing total convergence.
2. Together with Lemma E.1, (73) and (77), it shows that

$$\mathrm{Var}(\tilde{\Gamma}(\mathcal{Z}_{s_n})) = s_n \mathrm{Var}(\mathbb{E}[\Gamma(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) = \Omega\left(\frac{1}{\log(s_n)^p}\right). \tag{88}$$

   This will be essential, as it means that the variance of the first-order approximation decays slow enough relative to the variance of the actual tree $\Gamma(\mathcal{Z}_{s_n})$.

Remark E.4. Proving Theorem E.5 boils down to controlling the variance of $\mathbb{E}[\Gamma(\mathcal{Z}_{s_n})|\mathbf{X}_1, W_1]$. In Näf et al. (2023) a similar argument had to be made only conditioning on $\mathbf{X}_1$, which allowed to employ successive Lipschitz arguments. However, here the conditioning on W_1 adds variance that does not vanish as $n \to \infty$. Thus, compared to DRF for Causal-DRF we need to differentiate the case $W_1 = 1$ and $W_1 = 0$. Although this may seem benign, it requires a more careful treatment. This may best be seen in the somewhat awkward looking term $\mathbb{E}\Big[\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1, W_1] \mid W_1)\sigma^2(\mathbf{x}, W_1)\Big]$. Its analogue in the proof of (Näf et al., 2023, Theorem 5) was simply $\mathrm{Var}(\mathbb{E}[S_1 \mid \mathbf{X}_1])\mathrm{Var}(k(\mathbf{Y}_1, \cdot) \mid \mathbf{X} = \mathbf{x})$.

Theorem E.5. *Assume conditions (**F1**)–(**F6**), (**F1***), (**D1**)–(**D4**), (**K1**)–(**K2**) and (**C1**)–(**C2**) hold. Then, for all $f \in \mathcal{H} \setminus \{0\}$, we have*

$$\begin{aligned}\mathrm{Var}(\langle \Gamma_n(\mathbf{Z}_1), f\rangle) &= \mathbb{E}\Big[\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1, W_1] \mid W_1)\sigma^2_f(\mathbf{x}, W_1)\Big] + o(s_n^{-(1+\epsilon)}),\\ \mathrm{Var}(\Gamma_n(\mathbf{Z}_1)) &= \mathbb{E}\Big[\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1, W_1] \mid W_1)\sigma^2(\mathbf{x}, W_1)\Big] + o(s_n^{-(1+\epsilon)})\end{aligned} \tag{89}$$

*for some $\epsilon > 0$. In addition, if (**F6**) holds,*

$$\lim_{n\to\infty} \frac{\mathrm{Var}(\langle \Gamma_n(\mathbf{Z}_1), f\rangle)}{\mathrm{Var}(\Gamma_n(\mathbf{Z}_1))} = \frac{c\sigma^2_f(\mathbf{x}, 1)\mathbb{P}(W = 1) + \sigma^2_f(\mathbf{x}, 0)\mathbb{P}(W = 0)}{c\sigma^2(\mathbf{x}, 1)\mathbb{P}(W = 1) + \sigma^2(\mathbf{x}, 0)\mathbb{P}(W = 0)}, \tag{90}$$

with $\Gamma_n(\mathbf{Z}_1) = \mathbb{E}[\Gamma(\mathcal{Z}_{s_n})|\mathbf{Z}_1] - \mathbb{E}[\Gamma(\mathcal{Z}_{s_n})]$.

Proof. Throughout we focus on proving (89) for $\mathrm{Var}(\mathbb{E}[\Gamma(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1])$, as the proof for $\mathrm{Var}(\mathbb{E}[\langle \Gamma(\mathcal{Z}_{s_n}), f\rangle \mid \mathbf{Z}_1])$ works analogously.

By the same arguments as in (Näf et al., 2023, Theorem 5) and (Wager and Athey, 2017), we can ignore the double-sampling aspect in (**F1**) and condition on a point $\mathbf{Z}_1$ used in populating the trees (i.e. $1 \in \mathcal{I}_2$) and use s_n rather than $|\mathcal{I}_2|$ elements in the tree predictions to simplify notation.

Recall that $\Gamma_n(\mathbf{Z}_1) = \mathbb{E}[\Gamma(\mathcal{Z}_{s_n})|\mathbf{Z}_1] - \mathbb{E}[\Gamma(\mathcal{Z}_{s_n})]$, such that,

$$\frac{\mathrm{Var}(\langle \Gamma_n(\mathbf{Z}_1), f\rangle)}{\mathrm{Var}(\Gamma_n(\mathbf{Z}_1))} = \frac{\mathrm{Var}(\mathbb{E}[\langle \Gamma(\mathcal{Z}_{s_n}), f\rangle \mid \mathbf{Z}_1])}{\mathrm{Var}(\mathbb{E}[\Gamma(\mathcal{Z}_{s_n})|\mathbf{Z}_1])}.$$

Let $\mathbf{1}_{w,s_n} = \mathbf{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w}\}$, $S_i' = S_i \mathbf{1}_{w,s_n}$ and

$$\Gamma'(\mathcal{Z}_{s_n}) = \Gamma(\mathcal{Z}_{s_n})\mathbf{1}_{w,s_n} = \sum_{i=1}^{s_n} S_i' k(\mathbf{Y}_i, \cdot)$$

Crucially, w is chosen such that

$$\mathbb{P}(\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}) = \mathcal{O}(s_n^{-w}), \tag{91}$$

following from Lemma E.2, similar to (Wager and Athey, 2017).

First, using the fact that conditional expectations are projections in $\mathbb{L}^2(\Omega, \mathcal{A}, \mathcal{H})$, we can write

$$\mathrm{Var}(\mathbb{E}[\Gamma'(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) = \mathrm{Var}(\mathbb{E}[\Gamma'(\mathcal{Z}_{s_n})|\mathbf{X}_1, W_1]) + \mathrm{Var}(\mathbb{E}[\Gamma'(\mathcal{Z}_{s_n})|\mathbf{X}_1, W_1, \mathbf{Y}_1] - \mathbb{E}[\Gamma'(\mathcal{Z}_{s_n})|\mathbf{X}_1, W_1]), \tag{92}$$

as in (Näf et al., 2023, Claim (69)). We start by focusing on the second variance in this sum:

<u>Claim:</u>

$$\mathrm{Var}(\mathbb{E}[\Gamma'(\mathcal{Z}_{s_n})|\mathbf{X}_1, W_1, \mathbf{Y}_1] - \mathbb{E}[\Gamma'(\mathcal{Z}_{s_n})|\mathbf{X}_1, W_1]) = \mathbb{E}\Big[\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1] \mid W_1)\sigma^2(\mathbf{x}, W_1)\Big] + o(s_n^{-(1+\epsilon)}), \tag{93}$$

for some $\epsilon > 0$.

<u>Proof:</u> We adapt the expansion of (Näf et al., 2023): By honesty (i) $\mathbf{Y}_i$ is independent of S_i' conditional on $\mathbf{X}_i, W_i$, and more generally, (ii) $\mathbf{Y}_i$ is independent of S_j', $j = 1, \ldots, n$, conditional on $\mathbf{X}_i, W_i$. Thus, using (i), (ii), and the independence of $\mathbf{Y}_1$ from $\mathbf{Y}_j$, $j > 1$, we have

$$\begin{aligned}
\mathbb{E}[\Gamma'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1, W_1, \mathbf{Y}_1] &= \mathbb{E}[S_1' k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1, \mathbf{Y}_1] + \sum_{i=2}^{s_n} \mathbb{E}[S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1, \mathbf{Y}_1] \\
&= \mathbb{E}[S_1' \mid \mathbf{X}_1, W_1] k(\mathbf{Y}_1, \cdot) + \sum_{i=2}^{s_n} \mathbb{E}[S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1]
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}[\Gamma'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1, W_1] &= \mathbb{E}[S_1' k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1] + \sum_{i=2}^{s_n} \mathbb{E}[S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1] \\
&= \mathbb{E}[S_1' \mid \mathbf{X}_1, W_1]\mathbb{E}[k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1] + \sum_{i=2}^{s_n} \mathbb{E}[S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1]
\end{aligned}$$

and consequently

$$\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1, W_1, \mathbf{Y}_1] - \mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1, W_1]) = \mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](k(\mathbf{Y}_1, \cdot) - \mathbb{E}[k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1])). \tag{94}$$

We note that by (**C2**),

$$\begin{aligned}
\mathbb{E}[k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1] &= \mathbb{E}[k(\mathbf{Y}_1^0, \cdot) \mid \mathbf{X}_1]\mathbf{1}\{W_1 = 0\} + \mathbb{E}[k(\mathbf{Y}_1^1, \cdot) \mid \mathbf{X}_1]\mathbf{1}\{W_1 = 1\} \\
&= \mu^{(0)}(\mathbf{X}_1)\mathbf{1}\{W_1 = 0\} + \mu^{(1)}(\mathbf{X}_1)\mathbf{1}\{W_1 = 1\} \\
&= \mu^{(W_1)}(\mathbf{X}_1).
\end{aligned}$$

Note that by the Lipschitz assumption on $\mu^{(0)}(\mathbf{X}_1)$, $\mu^{(1)}(\mathbf{X}_1)$, i.e., **(D2)**, it holds that

$$\begin{aligned}\|\mu^{(W_1)}(\mathbf{X}_1) - \mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}} &= \|(\mu^{(0)}(\mathbf{X}_1) - \mu^{(0)}(\mathbf{x}))\mathbf{1}\{W_1 = 0\} + (\mu^{(1)}(\mathbf{X}_1) - \mu^{(1)}(\mathbf{x}))\mathbf{1}\{W_1 = 1\}\|_{\mathcal{H}} \\ &\leq \|(\mu^{(0)}(\mathbf{X}_1) - \mu^{(0)}(\mathbf{x}))\|_{\mathcal{H}}\mathbf{1}\{W_1 = 0\} + \|(\mu^{(1)}(\mathbf{X}_1) - \mu^{(1)}(\mathbf{x}))\|_{\mathcal{H}}\mathbf{1}\{W_1 = 1\} \\ &\leq C\|\mathbf{X}_1 - \mathbf{x}\|_{\mathcal{H}}. \end{aligned} \tag{95}$$

Then

$$\begin{aligned}&\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](k(\mathbf{Y}_1, \cdot) - \mathbb{E}[k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1])) \\ =&\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{X}_1) + \mu^{(W_1)}(\mathbf{x}) - \mu^{(W_1)}(\mathbf{x}))) \\ =&\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x}))) + \mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](\mu^{(W_1)}(\mathbf{X}_1) - \mu^{(W_1)}(\mathbf{x}))) \\ &- \mathrm{Cov}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})), \mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](\mu^{(W_1)}(\mathbf{X}_1) - \mu^{(W_1)}(\mathbf{x}))). \end{aligned} \tag{96}$$

For the second variance we have:

$$\begin{aligned}&\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](\mu^{(W_1)}(\mathbf{X}_1) - \mu^{(W_1)}(\mathbf{x}))) \\ &\leq \mathbb{E}[\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1]^2\|\mu^{(W_1)}(\mathbf{X}_1) - \mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2] \\ &\leq \mathbb{E}[\mathbb{E}[S_1'^2 \mid \mathbf{X}_1, W_1]C^2\|\mathbf{X}_1 - \mathbf{x}\|_{\mathbb{R}^p}^2] \\ &\leq \mathbb{E}[S_1'^2]C^2 s_n^{-2w}, \end{aligned}$$

where we used assumption **(D2)** for the second inequality and where the last step followed because $\mathbb{E}[S_1'^2 \mid \mathbf{X}_1, W_1] = 0$, for $\|\mathbf{X}_1 - \mathbf{x}\|_{\mathbb{R}^p} > s_n^{-w}$ by definition of $S_1' = S_1\mathbf{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w}\}$. Finally, because $\mathbb{E}[S_1'^2] \leq \mathbb{E}[|S_1'|] = \mathcal{O}(s_n^{-1})$ by Lemma E.1, we have that

$$\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](\mu^{(W_1)}(\mathbf{X}_1) - \mu^{(W_1)}(\mathbf{x}))) = o(s_n^{-(1+2w)}). \tag{97}$$

Similarly, for the covariance, it holds that

$$\begin{aligned}&\mathrm{Cov}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})), \mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](\mu^{(W_1)}(\mathbf{X}_1) - \mu^{(W_1)}(\mathbf{x}))) \\ &\leq \mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})))^{1/2}\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](\mu^{(W_1)}(\mathbf{X}_1) - \mu^{(W_1)}(\mathbf{x})))^{1/2} \\ &\leq \mathbb{E}\left[\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1]^2\|k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}\right]^{1/2} o(s_n^{-(1/2+w)}) \\ &\leq C\mathbb{E}\left[S_1'\right]^{1/2} o(s_n^{-(1/2+w)}) \\ &= o(s_n^{-(1+w)}), \end{aligned} \tag{98}$$

where the second last step followed because

$$\|k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}} \leq \|k(\mathbf{Y}_1, \cdot)\|_{\mathcal{H}} + \|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}} \leq C < \infty,$$

by the boundedness of k (**K1**). Combining (96), (97) and (98), we obtain

$$\begin{aligned}&\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](k(\mathbf{Y}_1, \cdot) - \mathbb{E}[k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1])) \\ &= \mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x}))) + o(s_n^{-(1+\epsilon)}). \end{aligned} \tag{99}$$

Focussing on the first term, we have:

$$\begin{aligned}&\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x}))) \\ =&\mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1, W_1]^2\|k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2] - \mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1, W_1]\|k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}]^2 \\ =&\mathbb{E}\Big[\mathbb{E}[S_1'|\mathbf{X}_1, W_1]^2\mathbb{E}[\|k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})\|^2|\mathbf{X}_1, W_1]\Big] - \mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1, W_1]\|k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}]^2. \end{aligned} \tag{100}$$

For the second part, note that because $\|k(\mathbf{Y}, \cdot)\|_{\mathcal{H}}$ and $\|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}$ are bounded by **(K2)**, it holds that

$$\mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1, W_1]\|k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}]^2 \leq \mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1, W_1]\|k(\mathbf{Y}_1, \cdot)\|_{\mathcal{H}} + \|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}]^2 \tag{101}$$

$$\leq C\mathbb{E}[S_1']^2 \tag{102}$$

$$= \mathcal{O}(s_n^2). \tag{103}$$

Define $\sigma^2(\mathbf{x}, W_1)$ as in (83) and note that by (**C2**),

$$\begin{aligned}\sigma^2(\mathbf{x}, W_1) &= \mathbb{E}[\|k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})\|^2 | W_1, \mathbf{X} = \mathbf{x}] \\ &= \mathrm{Var}(k(\mathbf{Y}_1^0, \cdot) \mid \mathbf{X} = \mathbf{x})(1 - W_1) + \mathrm{Var}(k(\mathbf{Y}_1^1, \cdot) \mid \mathbf{X} = \mathbf{x}) W_1.\end{aligned}$$

The first term in (100) can be bounded by

$$\begin{aligned}&\mathbb{E}\Big[\mathbb{E}[S_1' | \mathbf{X}_1, W_1]^2 \mathbb{E}[\|k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})\|^2 | \mathbf{X}_1, W_1]\Big] \\ =&\mathbb{E}\Big[\mathbb{E}[S_1' | \mathbf{X}_1, W_1]^2 \Big(\mathbb{E}[\|k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})\|^2 | \mathbf{X}_1, W_1] - \sigma^2(\mathbf{x}, W_1) + \sigma^2(\mathbf{x}, W_1)\Big)\Big] \\ =&\mathbb{E}\Big[\mathbb{E}[S_1' | \mathbf{X}_1, W_1]^2 \Big(\mathbb{E}[\|k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})\|^2 | \mathbf{X}_1, W_1] - \sigma^2(\mathbf{x}, W_1) + \sigma^2(\mathbf{x}, W_1)\Big)\Big] \\ =&\mathbb{E}\Big[\mathbb{E}[S_1' | \mathbf{X}_1, W_1]^2 \Big(\mathbb{E}[\|k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})\|^2 | \mathbf{X}_1, W_1] - \sigma^2(\mathbf{x}, W_1)\Big)\Big] + \mathbb{E}[\mathbb{E}[S_1' | \mathbf{X}_1, W_1]^2 \sigma^2(\mathbf{x}, W_1)] \qquad (104)\end{aligned}$$

By the same argument, as before, it can be shown that

$$\begin{aligned}&|\mathbb{E}[\|k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})\|^2 | \mathbf{X}_1, W_1] - \sigma^2(\mathbf{x}, W_1)| \\ &\leq |\mathbb{E}[\|k(\mathbf{Y}_1^1, \cdot) - \mu^{(1)}(\mathbf{x})\|^2 | \mathbf{X}_1] - \mathbb{E}[\|k(\mathbf{Y}_1^1, \cdot) - \mu^{(1)}(\mathbf{x})\|^2 | \mathbf{X} = \mathbf{x}]| \\ &+ |\mathbb{E}[\|k(\mathbf{Y}_1^0, \cdot) - \mu^{(0)}(\mathbf{x})\|^2 | \mathbf{X}_1] - \mathbb{E}[\|k(\mathbf{Y}_1^0, \cdot) - \mu^{(0)}(\mathbf{x})\|^2 | \mathbf{X} = \mathbf{x}]| \\ &\leq C\|\mathbf{X}_1 - \mathbf{x}\|,\end{aligned}$$

using the fact that $\mathbf{x} \mapsto \mu^{(j)}(\mathbf{x})$ and $\mathbf{x} \mapsto \mathbb{E}[\|k(\mathbf{Y}_1^j, \cdot) - \mu^{(j)}(\mathbf{x})\|^2 | \mathbf{X} = \mathbf{x}]$, for $j \in \{0, 1\}$, are all Lipschitz (see e.g., the argument in (Näf et al., 2023, Equation (78))). Thus,

$$\begin{aligned}&\mathbb{E}\Big[\mathbb{E}[S_1' | \mathbf{X}_1, W_1]^2 \Big(\mathbb{E}[\|k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})\|^2 | \mathbf{X}_1, W_1] - \sigma^2(\mathbf{x}, W_1)\Big)\Big] \\ &\leq \mathbb{E}\Big[\mathbb{E}[S_1' | \mathbf{X}_1, W_1]^2 | \mathbb{E}[\|k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x})\|^2 | \mathbf{X}_1, W_1] - \sigma^2(\mathbf{x}, W_1)|\Big] \\ &= \mathbb{E}[\mathbb{E}[S_1' | \mathbf{X}_1, W_1]^2] C s_n^w \\ &= o(s_n^{-(1+w)}), \qquad (105)\end{aligned}$$

using again that $\mathbb{E}[S_1'^2 \mid \mathbf{X}_1, W_1] = 0$, for $\|\mathbf{X}_1 - \mathbf{x}\|_{\mathbb{R}^p} > s_n^{-w}$ by definition of S_1'. Combining (100), (101), (104) and (105) it holds that

$$\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1](k(\mathbf{Y}_1, \cdot) - \mu^{(W_1)}(\mathbf{x}))) = \mathbb{E}\Big[\mathbb{E}[S_1' | \mathbf{X}_1, W_1]^2 \sigma^2(\mathbf{x}, W_1)\Big] + o(s_n^{-(1+w)}). \qquad (106)$$

We can further rewrite this:

$$\begin{aligned}&\mathbb{E}\Big[\mathbb{E}[S_1' | \mathbf{X}_1, W_1]^2 \sigma^2(\mathbf{x}, W_1)\Big] \\ &= \mathbb{E}\Big[\mathbb{E}[\mathbb{E}[S_1' | \mathbf{X}_1, W_1]^2 \mid W_1] \sigma^2(\mathbf{x}, W_1)\Big]\end{aligned}$$

Moreover, since $\sigma^2(\mathbf{x}, W_1)$ is again bounded and $\mathbb{E}[S_{1j}' \mid W_1 = j] = \mathcal{O}(s_n^{-1})$ by Lemma E.1,

$$\begin{aligned}\mathbb{E}\Big[\mathbb{E}[\mathbb{E}[S_1' | \mathbf{X}_1, W_1] \mid W_1]^2 \sigma^2(\mathbf{x}, W_1)\Big] &= \mathbb{E}[\mathbb{E}[\mathbb{E}[S_1' \mid W_1]^2 \sigma^2(\mathbf{x}, W_1)] \\ &= \mathbb{E}[S_1' \mid W_1 = 1]^2 * P(W_1 = 1) + \mathbb{E}[S_1' \mid W_1 = 0]^2 * P(W_1 = 0) \\ &= \mathcal{O}(s_n^{-2}). \qquad (107)\end{aligned}$$

Consequently,

$$\mathbb{E}\Big[\mathbb{E}[S_1' | \mathbf{X}_1, W_1]^2 \sigma^2(\mathbf{x}, W_1)\Big] = \mathbb{E}\Big[\mathrm{Var}(\mathbb{E}[S_1' | \mathbf{X}_1, W_1] \mid W_1) \sigma^2(\mathbf{x}, W_1)\Big] + \mathcal{O}(s_n^{-2}). \qquad (108)$$

Combining (94), (99), (106) and (108), we obtain Claim (93).

□

Combining (92) with (93), we have

$$\mathrm{Var}(\mathbb{E}[\langle \Gamma'(\mathcal{Z}_{s_n}), f\rangle \mid \mathbf{Z}_1]) = \mathrm{Var}(\mathbb{E}[\langle \Gamma'(\mathcal{Z}_{s_n}), f\rangle \mid \mathbf{X}_1, W_1]) + \mathbb{E}\Big[\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1] \mid W_1)\sigma^2(\mathbf{x}, W_1)\Big] + o(s_n^{-(1+\epsilon)}),$$

$$\mathrm{Var}(\mathbb{E}[\Gamma'(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) = \mathrm{Var}(\mathbb{E}[\Gamma'(\mathcal{Z}_{s_n})|\mathbf{X}_1, W_1]) + \mathbb{E}\Big[\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1] \mid W_1)\sigma_f^2(W_1, \mathbf{x})\Big] + o(s_n^{-(1+\epsilon)}) \quad (109)$$

for some $\epsilon > 0$. We now show that

<u>Claim:</u> For some $\epsilon > 0$,

$$\mathrm{Var}(\mathbb{E}[\Gamma'(\mathcal{Z}_{s_n})|\mathbf{X}_1, W_1]) = o(s_n^{-(1+\epsilon)}). \quad (110)$$

<u>Proof:</u> First, due to honesty (**F1**), we have

$$\mathbb{E}[\langle \Gamma'(\mathcal{Z}_{s_n}), f\rangle \mid \mathbf{X}_1, W_1] = \mathbb{E}[S_1' \mid \mathbf{X}_1, W_1]\mathbb{E}[\langle k(\mathbf{Y}_1, \cdot), f\rangle \mid \mathbf{X}_1, W_1] + \sum_{i=2}^{s_n} \mathbb{E}[S_i'\langle k(\mathbf{Y}_i, \cdot), f\rangle \mid \mathbf{X}_1, W_1]$$

and

$$\mathbb{E}[\Gamma'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1, W_1] = \mathbb{E}[S_1' \mid \mathbf{X}_1, W_1]\mathbb{E}[k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1] + \sum_{i=2}^{s_n} \mathbb{E}[S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1].$$

Subsequently, we consider the variance of the two terms and their covariance individually. That is, we study:

$$\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1]\mathbb{E}[k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1]) + \mathrm{Var}(\sum_{i=2}^{s_n} \mathbb{E}[S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1])$$
$$+2\mathrm{Cov}(\mathbb{E}[S_1' \mid \mathbf{X}_1, W_1]\mathbb{E}[k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1], \sum_{i=2}^{s_n} \mathbb{E}[S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1]). \quad (111)$$

First, we study the variance of the first terms. The variances satisfy

<u>Claim:</u>

$$\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1]\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X}_1, W_1]) = \mathbb{E}\left[\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1] \mid W_1)\|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2\right] + \mathcal{O}(s_n^{-(1+\epsilon)}) \quad (112)$$

<u>Proof:</u>

We have

$$\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1]\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X}_1, W_1]) \quad (113)$$
$$= \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1](\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X}_1, W_1] - \mu^{(W_1)}(\mathbf{x}) + \mu^{(W_1)}(\mathbf{x})))$$
$$= \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1]\mu^{(W_1)}(\mathbf{x})) + \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1](\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X}_1, W_1] - \mu^{(W_1)}(\mathbf{x})))$$
$$+ \mathrm{Cov}\left(\mathbb{E}[S_1'|\mathbf{X}_1, W_1]\mu^{(W_1)}(\mathbf{x}), \mathbb{E}[S_1'|\mathbf{X}_1, W_1](\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X}_1, W_1] - \mu^{(W_1)}(\mathbf{x}))\right). \quad (114)$$

For the first term it holds that

$$\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1]\mu^{(W_1)}(\mathbf{x})) = \mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1, W_1]^2\|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2] + O(s_n^{-2}), \quad (115)$$

with

$$\mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1, W_1]^2\|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2] = \mathbb{E}\Big[\mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1, W_1]^2\|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2 \mid W_1]\Big]$$
$$= \mathbb{E}\Big[\mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1, W_1]^2 \mid W_1]\|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2\Big]$$
$$= \mathbb{E}\Big[\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1] \mid W_1)\|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2\Big] + O(s_n^{-2}),$$

since,$\mathbb{E}\Big[\mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1, W_1] \mid W_1]^2 \|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2\Big] = O(s_n^{-2})$, using the same arguments as in (107). As was shown in (95),

$$\|\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X}_1, W_1] - \mu^{(W_1)}(\mathbf{x}))\|_{\mathcal{H}} \leq C\|\mathbf{X} - \mathbf{x}\|,$$

so that,

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1](\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X}_1, W_1] - \mu^{(W_1)}(\mathbf{x}))) &\leq \mathbb{E}[\|\mathbb{E}[S_1'|\mathbf{X}_1, W_1](\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X}_1, W_1]) - \mu^{(W_1)}(\mathbf{x}))\|_{\mathcal{H}}^2] \\
&= \mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1, W_1]^2 \|(\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X}_1, W_1] - \mu^{(W_1)}(\mathbf{x}))\|_{\mathcal{H}}^2] \\
&\leq \mathbb{E}[\mathbb{E}[S_1'^2|\mathbf{X}_1, W_1] C^2 \|\mathbf{X}_1 - \mathbf{x}\|_{\mathbb{R}^p}^2] \\
&= \mathcal{O}(s_n^{-(1+2w)}).
\end{aligned} \tag{116}$$

where the last step followed because $\mathbb{E}[S_1'^2|\mathbf{X}_1, W_1] = 0$, for $\|\mathbf{X}_1 - \mathbf{x}\|_{\mathbb{R}^p} > s_n^{-w}$ by definition of S_1' and the fact that $\mathbb{E}[S_1'^2] \leq \mathbb{E}[|S_1'|] \leq \mathbb{E}[|S_1|] = \mathcal{O}(s_n^{-1})$. Finally, we infer

$$\begin{aligned}
&\Big|\mathrm{Cov}\left(\mathbb{E}[S_1'|\mathbf{X}_1, W_1]\mu^{(W_1)}(\mathbf{x}), \mathbb{E}[S_1'|\mathbf{X}_1, W_1](\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X}_1, W_1] - \mu^{(W_1)}(\mathbf{x}))\right)\Big| \\
\leq &\sqrt{\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1]\mu^{(W_1)}(\mathbf{x}))}\sqrt{\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1](\mathbb{E}[k(\mathbf{Y}_1, \cdot)|\mathbf{X}_1, W_1] - \mu^{(W_1)}(\mathbf{x})))} \\
= &\mathcal{O}(s_n^{-(1+w)}),
\end{aligned}$$

due to (116), and,

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1]\mu^{(W_1)}(\mathbf{x})) &= \mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1, W_1]^2 \|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2] + \mathcal{O}(s_n^{-2}) \\
&= \mathbb{E}\Big[\mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1, W_1]^2 \mid W_1]\|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2\Big] + \mathcal{O}(s_n^{-2}) \\
&= \mathbb{E}[S_1'|\mathbf{X}_1, W_1 = 1]^2 \|\mu^{(1)}(\mathbf{x})\|_{\mathcal{H}}^2 \mathbb{P}(W_1 = 1) \\
&+ \mathbb{E}[S_1'|\mathbf{X}_1, W_1 = 0]^2 \|\mu^{(0)}(\mathbf{x})\|_{\mathcal{H}}^2 \mathbb{P}(W_1 = 0) + \mathcal{O}(s_n^{-2}) \\
&= \mathcal{O}(s_n^{-1}).
\end{aligned}$$

Thus, our Claim (112) holds. □

Before we continue proving the theorem, we note that, due to honesty, we have

$$\begin{aligned}
\sum_{i=2}^{s_n} \mathbb{E}[S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1] &= \sum_{i=2}^{s_n} \mathbb{E}[\mathbb{E}[S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_i, W_i, \mathbf{X}_1, W_1] \mid \mathbf{X}_1, W_1] \\
&= \sum_{i=2}^{s_n} \mathbb{E}[\mathbb{E}[S_i' \mid \mathbf{X}_i, W_i, \mathbf{X}_1, W_1]\mathbb{E}[k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_i, W_i, \mathbf{X}_1, W_1] \mid \mathbf{X}_1, W_1] \\
&= \sum_{i=2}^{s_n} \mathbb{E}[\mathbb{E}[S_i' \mathbb{E}[k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_i, W_i] \mid \mathbf{X}_i, W_i, \mathbf{X}_1, W_1] \mid \mathbf{X}_1, W_1] \\
&= \sum_{i=2}^{s_n} \mathbb{E}[S_i' \mathbb{E}[k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_i, W_i] \mid \mathbf{X}_1, W_1].
\end{aligned} \tag{117}$$

Now, we consider the variance of the sum in (117):

<u>Claim:</u>

$$\mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1]\right) = \mathbb{E}\left[\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1] \mid W_1)\|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2\right] + o(s_n^{-(1+\epsilon)}) \tag{118}$$

<u>Proof:</u> We start by showing the following crucial claim:

Claim: For some $\epsilon > 0$,

$$\mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1] - \sum_{i=2}^{s_n} \mathbb{E}[S_i' \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) = \mathcal{O}(s_n^{-(1+\epsilon)}), \tag{119}$$

Proof: We first note that, by (**C2**),

$$\begin{aligned}\mathbb{E}[\sum_{i=2}^{s_n} S_i' \mathbb{E}[k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_i, W_i] \mid \mathbf{X}_1, W_1] &= \mathbb{E}[\sum_{i=2}^{s_n} S_i' \mathbb{E}[k(\mathbf{Y}_i^{(W_i)}, \cdot) \mid \mathbf{X}_i] \mid \mathbf{X}_1, W_1] \\ &= \mathbb{E}[\sum_{i=2}^{s_n} S_i' \mu^{(W_i)}(\mathbf{X}_i) \mid \mathbf{X}_1, W_1].\end{aligned}$$

Combining this with (117), we have,

$$\begin{aligned}&\mathrm{Var}\left(\mathbb{E}[\sum_{i=2}^{s_n} S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i' \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) \\ &= \mathrm{Var}\left(\mathbb{E}[\sum_{i=2}^{s_n} S_i' \mu^{(W_i)}(\mathbf{X}_i) \mid \mathbf{X}_1, W_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i' \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) \\ &= \mathrm{Var}\left(\mathbb{E}[\sum_{i=2}^{s_n} S_i' (\mu^{(W_i)}(\mathbf{X}_i) - \mu^{(W_i)}(\mathbf{x})) \mid \mathbf{X}_1, W_1]\right) \\ &= \mathrm{Var}\left(\mathbb{E}[\sum_{i=2}^{s_n} S_i' \Delta^{(W_i)}(\mathbf{X}_i) \mid \mathbf{X}_1, W_1]\right),\end{aligned} \tag{120}$$

with $\Delta^{(W_i)}(\mathbf{X}_i) = \mu^{(W_i)}(\mathbf{X}_i) - \mu^{(W_i)}(\mathbf{x})$. We can now proceed to show (119) in the same way we showed (53) in Section D.2:

Claim:

$$\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i' \Delta^{(W_i)}(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x}), \mathbf{X}_1, W_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i' \Delta^{(W_i)}(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2] = o(s_n^{-(1+\epsilon)}). \tag{121}$$

Proof: First, using (31) in place of (34), the same proof as in Claim (59) shows that,

$$\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i' \Delta^{(W_i)}(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x}), \mathbf{X}_1, W_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i' \Delta^{(W_i)}(\mathbf{X}_i) \mid \mathbf{1}\{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x})\}, \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2] = o(s_n^{-(1+\epsilon)}), \tag{122}$$

for some $\epsilon > 0$. The remainder of the proof then works analogously as for the claim (56) in case of a regression tree. □

We can then split as follows:

$$\begin{aligned}&\mathrm{Var}(\mathbb{E}[\sum_{i=2}^{s_n} S_i' \Delta^{(W_i)}(\mathbf{X}_i) \mid \mathbf{X}_1, W_1]) \\ &= \mathrm{Var}\Big(\mathbb{E}[\mathbb{E}[\sum_{i=2}^{s_n} S_i' \Delta^{(W_i)}(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] \mid \mathbf{X}_1, W_1]\Big) \\ &+ \mathrm{Var}\Big(\mathbb{E}[\mathbb{E}[\sum_{i=2}^{s_n} S_i' \Delta^{(W_i)}(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] \mid \mathbf{X}_1, W_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i' \Delta^{(W_i)}(\mathbf{X}_i) \mid \mathbf{X}_1, W_1]\Big) \\ &- 2\mathrm{Cov}\Big(\mathbb{E}[\mathbb{E}[\sum_{i=2}^{s_n} S_i' \Delta^{(W_i)}(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] \mid \mathbf{X}_1, W_1], \\ &\mathbb{E}[\mathbb{E}[\sum_{i=2}^{s_n} S_i' \Delta^{(W_i)}(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] \mid \mathbf{X}_1, W_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i' \Delta^{(W_i)}(\mathbf{X}_i) \mid \mathbf{X}_1, W_1]\Big).\end{aligned}$$

For the first term, by assumption (30), and the boundedness of $\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta^{(W_i)}(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})]$,

$$\mathrm{Var}(\mathbb{E}[\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta^{(W_i)}(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] \mid \mathbf{X}_1, W_1]) = o(s_n^{-(1+\epsilon)}).$$

Moreover, using Jensen's inequality and (121),

$$\begin{aligned}
&\mathrm{Var}(\mathbb{E}[\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta^{(W_i)}(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] \mid \mathbf{X}_1, W_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta^{(W_i)}(\mathbf{X}_i) \mid \mathbf{X}_1, W_1]) \\
&\leq \mathbb{E}[\|\mathbb{E}[\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta^{(W_i)}(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta^{(W_i)}(\mathbf{X}_i) \mid \mathbf{X}_1, W_1, \mathcal{L}(\mathbf{x})] \mid \mathbf{X}_1, W_1]\|_{\mathcal{H}}^2] \\
&\leq \mathbb{E}[\mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta^{(W_i)}(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta^{(W_i)}(\mathbf{X}_i) \mid \mathbf{X}_1, W_1, \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2 \mid \mathbf{X}_1, W_1]] \\
&= \mathbb{E}[\|\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta^{(W_i)}(\mathbf{X}_i) \mid \mathcal{L}(\mathbf{x})] - \mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta^{(W_i)}(\mathbf{X}_i) \mid \mathbf{X}_1, W_1, \mathcal{L}(\mathbf{x})]\|_{\mathcal{H}}^2] \\
&= o(s_n^{-(1+\epsilon)}).
\end{aligned}$$

Adding the Cauchy-Schwarz inequality to bound the covariances, we obtain

$$\mathrm{Var}(\mathbb{E}[\sum_{i=2}^{s_n} S_i'\Delta^{(W_i)}(\mathbf{X}_i) \mid \mathbf{X}_1, W_1]) = o(s_n^{-(1+\epsilon)}),$$

showing (119).

□

Now, we further split

$$\mathbb{E}[\sum_{i=1}^{s_n} S_i' \mid \mathbf{X}_1, W_1] = \mathbb{E}[\sum_{W_i=1} S_{i1}' \mid \mathbf{X}_1, W_1] - \mathbb{E}[\sum_{W_i=0} (-S_{i0}') \mid \mathbf{X}_1, W_1].$$

and note that

$$\begin{aligned}
\mathbb{E}[\sum_{W_i=1} S_{i1}' \mid \mathbf{X}_1, W_1] &= \mathbb{E}[\sum_{W_i=0} (-S_{i0}') \mid \mathbf{X}_1, W_1] \\
&= \mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1, W_1\right),
\end{aligned}$$

so that in particular,

$$\begin{aligned}
\mathbb{E}[\sum_{W_i=1, i\geq 2} S_{i1}' \mid \mathbf{X}_1, W_1] &= \mathbb{E}[(\mathbf{1}_{w,s_n} - S_{11}')W_1 + \mathbf{1}_{w,s_n}(1-W_1) \mid \mathbf{X}_1, W_1] \\
&= \mathbb{E}[\mathbf{1}_{w,s_n} - S_{11}'W_1 \mid \mathbf{X}_1, W_1],
\end{aligned}$$

and

$$\mathbb{E}[\sum_{W_i=0, i\geq 2} (-S_{i0}') \mid \mathbf{X}_1, W_1] = \mathbb{E}[\mathbf{1}_{w,s_n} - S_{10}'(1-W_1) \mid \mathbf{X}_1, W_1].$$

Thus, if follows that:

$$\begin{aligned}
\mathbb{E}[\sum_{W_i=0, i\geq 2} (-S_{i0}') \mid \mathbf{X}_1, W_1] &= \mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1, W_1\right) - \mathbb{E}[(-S_{10}') \mid \mathbf{X}_1, W_1](1-W_1) \\
\mathbb{E}[\sum_{W_i=1, i\geq 2} S_{i1}' \mid \mathbf{X}_1, W_1] &= \mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1, W_1\right) - \mathbb{E}[S_{11}' \mid \mathbf{X}_1, W_1]W_1
\end{aligned} \tag{123}$$

Following this logic and splitting (117) according to W_i, we obtain

$$\begin{aligned}
&\mathrm{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mu^{(W_i)}(\mathbf{x})\mid \mathbf{X}_1,W_1]\right)\\
&=\mathrm{Var}\left(\mathbb{E}[\sum_{i\geq 2,W_i=1}S_{i1}'\mu^{(1)}(\mathbf{x})\mid \mathbf{X}_1,W_1]\right)+\mathrm{Var}\left(\mathbb{E}[\sum_{i\geq 2,W_i=0}(-S_{i0}')\mu^{(0)}(\mathbf{x})\mid \mathbf{X}_1,W_1]\right) \qquad (124)\\
&-2\mathrm{Cov}\left(\mathbb{E}[\sum_{i\geq 2,W_i=1}S_{i1}'\mu^{(1)}(\mathbf{x})\mid \mathbf{X}_1,W_1],\mathbb{E}[\sum_{i\geq 2,W_i=0}(-S_{i0}')\mu^{(0)}(\mathbf{x})\mid \mathbf{X}_1,W_1]\right)\\
&=(I)+(II)-(III) \qquad (125)
\end{aligned}$$

and

$$\mathrm{Var}\left(\mathbb{E}[\sum_{i\geq 2,W_i=1}S_{i1}'\mu^{(1)}(\mathbf{x})\mid \mathbf{X}_1,W_1]\right)=\|\mu^{(1)}(\mathbf{x})\|_{\mathcal{H}}^2\mathrm{Var}\left(\mathbb{E}[S_{11}'\mid \mathbf{X}_1,W_1=1]\right)\mathbb{P}(W_1=1)+\mathcal{O}(s_n^{-(1+2w)}) \quad (126)$$

$$\mathrm{Var}\left(\mathbb{E}[\sum_{i\geq 2,W_i=0}(-S_{i0}')\mu^{(0)}(\mathbf{x})\mid \mathbf{X}_1,W_1]\right)=\|\mu^{(0)}(\mathbf{x})\|_{\mathcal{H}}^2\mathrm{Var}\left(\mathbb{E}[S_{10}'\mid \mathbf{X}_1,W_1=0]\right)\mathbb{P}(W_1=0)+\mathcal{O}(s_n^{-(1+2w)}) \quad (127)$$

$$\mathrm{Cov}\Big(\mu^{(1)}(\mathbf{x})\mathbb{E}[\sum_{i\geq 2,W_i=1}S_{i1}'\mid \mathbf{X}_1,W_1],\mu^{(0)}(\mathbf{x})\mathbb{E}[\sum_{i\geq 2,W_i=0}(-S_{i0}')\mid \mathbf{X}_1,W_1]\Big)=\mathcal{O}(s_n^{-(1+2w)}). \quad (128)$$

<u>Claim:</u> For some $\epsilon>0$, (126), (127), and (128) hold.

<u>Proof:</u>

We first show (126). By (123),

$$\begin{aligned}
&\mathrm{Var}\left(\mathbb{E}[\sum_{i\geq 2,W_i=1}S_{i1}'\mu^{(1)}(\mathbf{x})\mid \mathbf{X}_1,W_1]\right)\\
&=\|\mu^{(1)}(\mathbf{x})\|_{\mathcal{H}}^2\mathrm{Var}\left(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x}))\leq s_n^{-w}\mid \mathbf{X}_1,W_1\right)-\mathbb{E}[S_{11}'\mid \mathbf{X}_1,W_1]W_1\right)\\
&=\|\mu^{(1)}(\mathbf{x})\|_{\mathcal{H}}^2(\mathrm{Var}\left(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x}))\leq s_n^{-w}\mid \mathbf{X}_1,W_1\right)\right)+\mathrm{Var}(\mathbb{E}[S_{11}'\mid \mathbf{X}_1,W_1]W_1)\\
&-2\mathrm{Cov}(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x}))\leq s_n^{-w}\mid \mathbf{X}_1,W_1\right),\mathbb{E}[S_{11}'\mid \mathbf{X}_1,W_1]W_1))
\end{aligned}$$

By assumption (**F1***), $\mathrm{Var}\left(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x}))\leq s_n^{-w}\mid \mathbf{X}_1,W_1\right)\right)=o(s_n^{-(1+\epsilon)})$. Moreover,

$$\mathrm{Var}(\mathbb{E}[S_{11}'\mid \mathbf{X}_1,W_1]W_1)\leq \mathbb{E}[(\mathbb{E}[S_{11}'\mid \mathbf{X}_1,W_1])^2W_1]\leq \mathbb{E}[S_{11}']=\mathcal{O}(s_n^{-1}),$$

so that by Cauchy-Schwarz also $\mathrm{Cov}(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x}))\leq s_n^{-w}\mid \mathbf{X}_1,W_1\right),\mathbb{E}[S_{11}'\mid \mathbf{X}_1,W_1]W_1)=o(s_n^{-(1+\epsilon)})$. Thus,

$$\mathrm{Var}\left(\mathbb{E}[\sum_{i\geq 2,W_i=1}S_{i1}'\mu^{(1)}(\mathbf{x})\mid \mathbf{X}_1,W_1]\right)=\|\mu^{(1)}(\mathbf{x})\|_{\mathcal{H}}^2\mathrm{Var}(\mathbb{E}[S_{11}'\mid \mathbf{X}_1,W_1]W_1)+o(s_n^{-(1+\epsilon)}).$$

Since in addition, $\mathbb{E}[\mathbb{E}[S_{11}'\mid \mathbf{X}_1,W_1]W_1]^2=\mathbb{E}[S_{11}'W_1]^2=\mathbb{E}[S_{11}']^2=\mathcal{O}(s_n^{-2})$, it further holds that

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}[S_{11}'\mid \mathbf{X}_1,W_1]W_1)&=\mathbb{E}[(\mathbb{E}[S_{11}'\mid \mathbf{X}_1,W_1]W_1)^2]+\mathcal{O}(s_n^{-2})\\
&=\mathbb{E}[\mathbb{E}[S_{11}'\mid \mathbf{X}_1,W_1]^2W_1]+\mathcal{O}(s_n^{-2})\\
&=\mathbb{E}[\mathbb{E}[\mathbb{E}[S_{11}'\mid \mathbf{X}_1,W_1]^2\mid W_1]W_1]+\mathcal{O}(s_n^{-2})\\
&=\mathbb{E}[\mathrm{Var}(\mathbb{E}[S_{11}'\mid \mathbf{X}_1,W_1]\mid W_1)W_1]+\mathcal{O}(s_n^{-2}), \qquad (129)
\end{aligned}$$

where the last step again followed because $\mathbb{E}[\mathbb{E}[\mathbb{E}[S'_{11} \mid \mathbf{X}_1, W_1] \mid W_1]^2 W_1] = \mathcal{O}(s_n^{-2})$. Finally,

$$\mathbb{E}[\mathrm{Var}(\mathbb{E}[S'_{11} \mid \mathbf{X}_1, W_1] \mid W_1) W_1] = \mathrm{Var}(\mathbb{E}[S'_{11} \mid \mathbf{X}_1, W_1] \mid W_1 = 1)\mathbb{P}(W = 1), \tag{130}$$

showing (126). Condition (127) can be shown analogously.

Finally, we need to show (128). By (123),

$$\begin{aligned}
&\mathrm{Cov}\Big(\mu^{(1)}(\mathbf{x})\mathbb{E}[\sum_{i\geq 2, W_i=1} S'_{i1} \mid \mathbf{X}_1, W_1], \mu^{(0)}(\mathbf{x})\mathbb{E}[\sum_{i\geq 2, W_i=0} (-S'_{i0}) \mid \mathbf{X}_1, W_1]\Big) \\
&= \langle \mu^{(1)}(\mathbf{x}), \mu^{(0)}(\mathbf{x})\rangle \mathrm{Cov}\Big(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1, W_1\right) - \mathbb{E}[S'_{11} \mid \mathbf{X}_1, W_1] W_1, \\
&\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1, W_1\right) + \mathbb{E}[S'_{10} \mid \mathbf{X}_1, W_1](1 - W_1)\Big) \\
&= -\langle \mu^{(1)}(\mathbf{x}), \mu^{(0)}(\mathbf{x})\rangle \mathrm{Cov}(\mathbb{E}[S'_{11} \mid \mathbf{X}_1, W_1] W_1, \mathbb{E}[S'_{10} \mid \mathbf{X}_1, W_1](1 - W_1)) + o(s_n^{-(1+\epsilon)},
\end{aligned} \tag{131}$$

where the last step again followed because $\mathrm{Var}(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1, W_1\right)) = o(s_n^{-(1+\epsilon)})$ by Assumption (**F1***), $\mathrm{Var}(\mathbb{E}[S'_{11} \mid \mathbf{X}_1, W_1] W_1) = \mathcal{O}(s_n^{-1})$, $\mathrm{Var}(\mathbb{E}[S'_{10} \mid \mathbf{X}_1, W_1](1 - W_1)) = \mathcal{O}(s_n^{-1})$, combined with Cauchy-Schwarz. Now since,

$$\mathbb{E}[\mathbb{E}[S'_{11} \mid \mathbf{X}_1, W_1] W_1]\mathbb{E}[\mathbb{E}[S'_{10} \mid \mathbf{X}_1, W_1](1 - W_1)] = \mathcal{O}(s_n^{-2}),$$

as before, we arrive at

$$\begin{aligned}
&\mathrm{Cov}(\mathbb{E}[S'_{11} \mid \mathbf{X}_1, W_1] W_1, \mathbb{E}[S'_{10} \mid \mathbf{X}_1, W_1](1 - W_1)) \\
&= \mathbb{E}[\mathbb{E}[S'_{11} \mid \mathbf{X}_1, W_1]\mathbb{E}[S'_{10} \mid \mathbf{X}_1, W_1] W_1(1 - W_1)] + \mathcal{O}(s_n^{-2}) \\
&= \mathcal{O}(s_n^{-2}).
\end{aligned} \tag{132}$$

Combining, (131) with (132) gives (128).

□

Combining (126)-(128) with (124), we get:

$$\begin{aligned}
&\mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S'_i \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) \\
&= \|\mu^{(1)}(\mathbf{x})\|_{\mathcal{H}}^2 \mathrm{Var}\left(\mathbb{E}[S'_{11} \mid \mathbf{X}_1, W_1 = 1]\right)\mathbb{P}(W_1 = 1) \\
&+ \|\mu^{(0)}(\mathbf{x})\|_{\mathcal{H}}^2 \mathrm{Var}\left(\mathbb{E}[S'_{10} \mid \mathbf{X}_1, W_1 = 0]\right)\mathbb{P}(W_1 = 0) + o(s_n^{-(1+\epsilon)}) \\
&= \mathbb{E}\left[\mathrm{Var}(\mathbb{E}[S'_1 | \mathbf{X}_1, W_1] \mid W_1)\|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2\right] + o(s_n^{-(1+\epsilon)}).
\end{aligned} \tag{133}$$

In addition, as $\mathrm{Var}\left(\mathbb{E}[S'_{11} \mid \mathbf{X}_1, W_1 = 1]\right) = \mathcal{O}(s_n^{-1})$, $\mathrm{Var}\left(\mathbb{E}[S'_{10} \mid \mathbf{X}_1, W_1 = 0]\right) = \mathcal{O}(s_n^{-1})$ and since $\|\mu^{(j)}(\mathbf{x})\|_{\mathcal{H}}^2$ is bounded for $j \in \{0, 1\}$ by (**K1**), it follows that

$$\mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S'_i \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) = \mathcal{O}(s_n^{-1}). \tag{134}$$

Combining (119) with (134) and Cauchy Schwarz, we get

$$\begin{aligned}
&\mathrm{Var}(\sum_{i=2}^{s_n} \mathbb{E}[S'_i k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1]) \\
&= \mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S'_i \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) + \mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S'_i k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1] - \sum_{i=2}^{s_n} \mathbb{E}[S'_i \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) \\
&+ \mathrm{Cov}(\sum_{i=2}^{s_n} \mathbb{E}[S'_i \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \sum_{i=2}^{s_n} \mathbb{E}[S'_i k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1] - \sum_{i=2}^{s_n} \mathbb{E}[S'_i \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]) \\
&= \mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S'_i \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) + o(s_n^{-(1+\epsilon)}).
\end{aligned} \tag{135}$$

Combining (133) and (135) gives Claim (118).

□

Finally, we consider the covariances between $\mathbb{E}[S_1' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1]$ and $\mathbb{E}[\sum_{i=2}^{s_n} S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1]$:

Claim: For some $\epsilon > 0$,

$$\begin{aligned}&\mathrm{Cov}\left(\mathbb{E}[S_1' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1], \mathbb{E}[\sum_{i=2}^{s_n} S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1]\right)\\ &= -\mathbb{E}\left[\mathrm{Var}(\mathbb{E}[S_1' | \mathbf{X}_1, W_1] \mid W_1) \|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2\right] + o(s_n^{-(1+\epsilon)}).\end{aligned} \tag{136}$$

Proof:

Similar to above, we first simplify the expression:

Claim: For some $\epsilon > 0$, we have

$$\begin{aligned}&\mathrm{Cov}\left(\mathbb{E}[S_1' k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1], \mathbb{E}[\sum_{i=2}^{s_n} S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1]\right)\\ &= \mathrm{Cov}\left(\mathbb{E}[S_1' \mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \mathbb{E}[\sum_{i=2}^{s_n} S_i' \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) + o(s_n^{-(1+\epsilon)}).\end{aligned} \tag{137}$$

Proof: It holds that:

$$\begin{aligned}&\mathrm{Cov}\left(\mathbb{E}[S_1' k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1], \mathbb{E}[\sum_{i=2}^{s_n} S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1]\right)\\ &= \mathrm{Cov}\left(\mathbb{E}[S_1' k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1], \mathbb{E}[\sum_{i=2}^{s_n} S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i' \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right)\\ &+ \mathrm{Cov}\left(\mathbb{E}[S_1' k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1] - \mathbb{E}[S_1' \mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \mathbb{E}[\sum_{i=2}^{s_n} S_i' \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right)\\ &+ \mathrm{Cov}\left(\mathbb{E}[S_1' \mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \mathbb{E}[\sum_{i=2}^{s_n} S_i' \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right)\\ &= (C1) + (C2) + \mathrm{Cov}\left(\mathbb{E}[S_1' \mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \mathbb{E}[\sum_{i=2}^{s_n} S_i' \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right).\end{aligned}$$

We now use the Cauchy-Schwarz inequality to bound the first two covariances by the product of the square root of the variances of the respective terms. For (C1), as k is bounded (**K1**),

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[S_1' k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1]) &\leq \mathbb{E}[\|\mathbb{E}[S_1' k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1]\|_{\mathcal{H}}^2]\\ &\leq \mathbb{E}[\mathbb{E}[S_1' \|k(\mathbf{Y}_1, \cdot)\|_{\mathcal{H}} \mid \mathbf{X}_1, W_1]^2]\\ &\leq C\mathbb{E}[S_1']\\ &= \mathcal{O}(s_n^{-1}),\end{aligned}$$

and, as shown in (119),

$$\mathrm{Var}\left(\mathbb{E}[\sum_{i=2}^{s_n} S_i' k(\mathbf{Y}_i, \cdot) \mid \mathbf{X}_1, W_1] - \mathbb{E}[\sum_{i=2}^{s_n} S_i' \mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) = o(s_n^{-(1+\epsilon)}), \tag{138}$$

so that $|(C1)| \leq \mathcal{O}(s_n^{-1/2}) o(s_n^{-(1+\epsilon)/2}) = o(s_n^{-(1+\epsilon/2)})$. Similarly, by (116), it holds that

$$\begin{aligned}&\mathrm{Var}(\mathbb{E}[S_1' k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1] - \mathbb{E}[S_1' \mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1])\\ &= \mathrm{Var}(\mathbb{E}[S_1' | \mathbf{X}_1, W_1](\mathbb{E}[k(\mathbf{Y}_1, \cdot) | \mathbf{X}_1, W_1] - \mu^{(W_1)}(\mathbf{x})))\\ &= o(s_n^{-(1+\epsilon)}),\end{aligned}$$

whereby we recall that $\mathbb{E}[S_1' k(\mathbf{Y}_1, \cdot) \mid \mathbf{X}_1, W_1] = \mathbb{E}[S_1' | \mathbf{X}_1, W_1]\mathbb{E}[k(\mathbf{Y}_1, \cdot) | \mathbf{X}_1, W_1]$ by honesty (**F1**). Combining this with (134), we obtain $|(C2)| \leq \mathcal{O}(s_n^{-1/2}) o(s_n^{-(1+\epsilon)/2}) = o(s_n^{-(1+\epsilon/2)})$, thus showing (137).

□

We now again split up the term into the part for control and treatment groups:

$$\begin{aligned}
&\mathrm{Cov}\left(\mathbb{E}[S_1'\mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \mathbb{E}[\sum_{i=2}^{s_n} S_i'\mu^{(W_i)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) \\
&= \mathrm{Cov}\left(\mathbb{E}[S_1'\mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \mathbb{E}[\sum_{i\geq 2, W_i=1} S_{i1}'\mu^{(1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) \\
&+ \mathrm{Cov}\left(\mathbb{E}[S_1'\mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \mathbb{E}[\sum_{i\geq 2, W_i=0} S_{i0}'\mu^{(0)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) \qquad (139)
\end{aligned}$$

We then study each term separately:

<u>Claim:</u> For some $\epsilon > 0$,

$$\mathrm{Cov}\left(\mathbb{E}[S_1'\mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \mathbb{E}[\sum_{i\geq 2, W_i=1} S_{i1}'\mu^{(1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) \qquad (140)$$

$$= -\|\mu^{(1)}(\mathbf{x})\|_{\mathcal{H}}^2 \mathrm{Var}\left(\mathbb{E}[S_{11}' \mid \mathbf{X}_1, W_1 = 1]\right) \mathbb{P}(W_1 = 1) + o(s_n^{-(1+\epsilon)}) \qquad (141)$$

$$\mathrm{Cov}\left(\mathbb{E}[S_1'\mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \mathbb{E}[\sum_{i\geq 2, W_i=0} S_{i0}'\mu^{(0)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) \qquad (142)$$

$$= -\|\mu^{(0)}(\mathbf{x})\|_{\mathcal{H}}^2 \mathrm{Var}\left(\mathbb{E}[S_{10}' \mid \mathbf{X}_1, W_1 = 0]\right) \mathbb{P}(W_1 = 0) + o(s_n^{-(1+\epsilon)}) \qquad (143)$$

<u>Proof:</u> Following (123), it holds that

$$\begin{aligned}
&\mathrm{Cov}\left(\mathbb{E}[S_1'\mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \mathbb{E}[\sum_{i\geq 2, W_i=1} S_{i1}'\mu^{(1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]\right) \\
&= \mathrm{Cov}\left(\mathbb{E}[S_1'\mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \mu^{(1)}(\mathbf{x})(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1, W_1\right) - \mathbb{E}[S_{11}' \mid \mathbf{X}_1, W_1]W_1)\right) \qquad (144) \\
&= \mathrm{Cov}\Big(\mathbb{E}[S_1'\mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \mu^{(1)}(\mathbf{x})\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1, W_1\right)\Big) \\
&- \mathrm{Cov}\Big(\mathbb{E}[S_1'\mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \mathbb{E}[S_{11}'\mu^{(1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]W_1\Big). \qquad (145)
\end{aligned}$$

The first element is $o(s_n^{-(1+\epsilon)})$, as

$$\begin{aligned}
&|\mathrm{Cov}\Big(\mathbb{E}[S_1'\mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \mu^{(1)}(\mathbf{x})\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1, W_1\right)\Big)| \\
&\leq \mathrm{Var}(\mathbb{E}[S_1'\mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1])^{1/2} \|\mu^{(1)}(\mathbf{x})\|_{\mathcal{H}} \mathrm{Var}(\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x})) \leq s_n^{-w} \mid \mathbf{X}_1, W_1\right))^{1/2} \\
&= \mathcal{O}(s_n^{-1/2}) o(s_n^{-(1+\epsilon)/2}), \qquad (146)
\end{aligned}$$

where the last step follows because of (**F1***), the fact that $\mathrm{Var}(\mathbb{E}[S_1'\mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1]) = \mathcal{O}(s_n^{-1})$ and $\|\mu^{(1)}(\mathbf{x})\|_{\mathcal{H}} \leq$

C by boundedness of k (**K1**). In addition, we write

$$\begin{aligned}
&\mathrm{Cov}\Big(\mathbb{E}[S_1'\mu^{(W_1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1], \mathbb{E}[S_{11}' \mid \mathbf{X}_1, W_1]W_1\Big) \\
&= \mathrm{Cov}\Big(\mathbb{E}[S_{11}'\mu^{(1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1 = 1]W_1, \mathbb{E}[S_{11}'\mu^{(1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1 = 1]W_1\Big) \\
&+ \mathrm{Cov}\Big(\mathbb{E}[S_{10}'\mu^{(0)}(\mathbf{x}) \mid \mathbf{X}_1, W_1 = 0](1 - W_1), \mathbb{E}[S_{11}'\mu^{(1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1 = 1]W_1\Big). \qquad (147)
\end{aligned}$$

As

$$\mathbb{E}[\mathbb{E}[S_{10}' \mid \mathbf{X}_1, W_1 = 0](1 - W_1)] \cdot \mathbb{E}[\mathbb{E}[S_{11}' \mid \mathbf{X}_1, W_1 = 1]W_1] = \mathcal{O}(s_n^{-2}),$$

the second covariance is $\mathcal{O}(s_n^{-2})$:

$$\begin{aligned}
&\mathrm{Cov}\Big(\mathbb{E}[S_{10}'\mu^{(0)}(\mathbf{x}) \mid \mathbf{X}_1, W_1 = 0](1 - W_1), \mathbb{E}[S_{11}'\mu^{(1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1 = 1]W_1\Big) \\
&= \mathbb{E}[\mathbb{E}[S_{10}'\mu^{(0)}(\mathbf{x}) \mid \mathbf{X}_1, W_1 = 0](1 - W_1)\mathbb{E}[S_{11}'\mu^{(1)}(\mathbf{x}) \mid \mathbf{X}_1, W_1 = 1]W_1] + \mathcal{O}(s_n^{-2}) \\
&= \mathcal{O}(s_n^{-2}). \qquad (148)
\end{aligned}$$

Similarly, the first covariance can be rewritten as

$$\begin{aligned}
&\|\mu^{(1)}(\mathbf{x})\|_{\mathcal{H}}^2 \mathrm{Cov}\Big(\mathbb{E}[S_{11}' \mid \mathbf{X}_1, W_1 = 1]W_1, \mathbb{E}[S_{11}' \mid \mathbf{X}_1, W_1 = 1]W_1\Big) \\
&= \|\mu^{(1)}(\mathbf{x})\|_{\mathcal{H}}^2 \mathrm{Var}(\mathbb{E}[S_{11}' \mid \mathbf{X}_1, W_1] \mid W_1 = 1)\mathbb{P}(W = 1), \qquad (149)
\end{aligned}$$

using the same steps as in (129) and (130). Plugging (146), (147), (148) and (149) into (144) gives (140). Using analogous arguments, we also obtain (142).

□

Plugging (137), (140), (142) into (139), we obtain (136). □

Plugging (112), (118), (136) into (111), shows

$$\begin{aligned}
&\mathbb{E}[\Gamma'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1, W_1] \\
&= 2\mathbb{E}\left[\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1] \mid W_1)\|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2\right] - 2\mathbb{E}\left[\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1] \mid W_1)\|\mu^{(W_1)}(\mathbf{x})\|_{\mathcal{H}}^2\right] + o(s_n^{-(1+\epsilon)}) \\
&= o(s_n^{-(1+\epsilon)}),
\end{aligned}$$

showing (110). □

Combining (109) with (110), we obtain

$$\mathrm{Var}(\mathbb{E}[\Gamma'(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) = \mathbb{E}\Big[\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1, W_1] \mid W_1)\sigma_f^2(W_1, \mathbf{x})\Big] + o(s_n^{-(1+\epsilon)}). \qquad (150)$$

We now replace S_1' with S_1:

<u>Claim:</u> For $\epsilon > 0$ and $j \in \{0, 1\}$,

$$\big|\mathrm{Var}(\mathbb{E}[S_{1j}'|\mathbf{X}_1, W_1 = j]) - \mathrm{Var}(\mathbb{E}[S_{1j}|\mathbf{X}_1, W_1 = j])\big| = o(s_n^{-(1+\epsilon)}). \qquad (151)$$

<u>Proof:</u> We have

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}[S_{1j}|\mathbf{X}_1, W_1 = j] - \mathbb{E}[S_{1j}'|\mathbf{X}_1, W_1 = j]) &= \mathrm{Var}(\mathbb{E}[S_{1j} - S_{1j}'|\mathbf{X}_1, W_1 = j]) \\
&= \mathrm{Var}(\mathbb{E}[S_{1j}\mathbf{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}|\mathbf{X}_1, W_1 = j]) \\
&\leq \mathbb{E}[\mathbb{E}[S_{1j}\mathbf{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}|\mathbf{X}_1, W_1 = j]^2] \\
&\leq \mathbb{E}[|S_{1j}|\mathbf{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\} \mid W_1 = j] \\
&= \frac{1}{\mathbb{P}(W_1 = j)}\mathbb{E}[|S_{1j}|\mathbf{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}] \\
&\leq \frac{1}{\min(\varepsilon, 1 - \varepsilon)}\mathbb{E}[|S_{1j}|\mathbf{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}],
\end{aligned}$$

where the last step followed due to (**C1**). Since moreover, $|S_{1j}| \leq |S_1| = |S_{11}| + |S_{10}|$, we further have

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[S_{1j}|\mathbf{X}_1, W_1 = j] - \mathbb{E}[S'_{1j}|\mathbf{X}_1, W_1 = j]) &\leq \frac{1}{\min(\varepsilon, 1-\varepsilon)}\mathbb{E}[|S_1|\mathbf{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}] \\ &\leq \frac{1}{\min(\varepsilon, 1-\varepsilon)s_n}\sum_{i=1}^{s_n}\mathbb{E}[|S_i|\mathbf{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}] \\ &= \frac{2}{\min(\varepsilon, 1-\varepsilon)s_n}\mathbb{P}(\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}) \\ &= \mathcal{O}(s_n^{-(1+w)}), \end{aligned} \tag{152}$$

due to $\sum_{i=1}^{s_n} S_i = 2$ and where the last step followed due to (91). As $\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1, W_1 = j]) = \mathcal{O}(s_n^{-1})$ from Lemma E.1 and analogously $\mathrm{Var}(\mathbb{E}[S'_1|\mathbf{X}_1, W_1 = j]) = \mathcal{O}(s_n^{-1})$, it holds that

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[S'_{1j}|\mathbf{X}_1, W_1 = j]) &= \mathrm{Var}(\mathbb{E}[S_{1j}|\mathbf{X}_1, W_1 = j]) + \mathrm{Var}(\mathbb{E}[S_{1j}|\mathbf{X}_1, W_1 = j] - \mathbb{E}[S'_{1j}|\mathbf{X}_1, W_1 = j]) \\ &\quad + 2\mathrm{Cov}(\mathbb{E}[S_{1j}|\mathbf{X}_1, W_1 = j], \mathbb{E}[S_{1j}|\mathbf{X}_1, W_1 = j] - \mathbb{E}[S'_{1j}|\mathbf{X}_1, W_1 = j]) \\ &= \mathrm{Var}(\mathbb{E}[S_{1j}|\mathbf{X}_1, W_1 = j]) + \mathcal{O}(s_n^{-(1+w/2)}),\end{aligned}$$

again making use of Cauchy-Schwarz to bound the covariances. □

Combining Claim (151) with (150), we have:

$$\mathrm{Var}(\mathbb{E}[\Gamma'(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) = \mathbb{E}\Big[\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1, W_1] \mid W_1)\sigma_f^2(W_1, \mathbf{x})\Big] + o(s_n^{-(1+\epsilon)}). \tag{153}$$

It remains to show that:

<u>Claim:</u> For some $\epsilon > 0$,

$$|\mathrm{Var}(\mathbb{E}[\Gamma(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) - \mathrm{Var}(\mathbb{E}[\Gamma'(\mathcal{Z}_{s_n})|\mathbf{Z}_1])| = o(s_n^{-(1+\epsilon)}). \tag{154}$$

<u>Proof:</u> This can be proven in exactly the same way as Claim (85) in Theorem 24 of (Näf et al., 2023) with the regression trees T, T' exchanged by the causal trees Γ, Γ'. □

Thus, as $\Gamma_n(\mathbf{Z}_1) = \mathbb{E}[\Gamma(\mathcal{Z}_{s_n})|\mathbf{Z}_1]$, combining (153) and (154), we arrive at (89).

We recall that $\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1, W_1 = j]) = \Omega((s_n \log(s_n))^{-1})$, by Lemma E.1 and thus the $o(s_n^{-(1+\epsilon)})$ terms can be ignored. This together with assumption (**F6**) establishes (90). □

Having established these basics, we can show asymptotic normality of the estimator in $\mathcal{H}$.

Theorem 4.1. *Assume conditions (**F1**)–(**F6**), (**F1***), (**D1**)–(**D5**), (**K1**)–(**K4**) and (**C1**)–(**C2**) hold. Then, there exists $\sigma_n \to 0$, such that, for some self-adjoint HS operator $\mathbf{\Sigma_x}$, we have*

$$\xi_n(\mathbf{x}) = \frac{1}{\sigma_n}\left(\hat{\tau}_{n,k}(\mathbf{x}) - \tau_k(\mathbf{x})\right) \xrightarrow{D} N(0, \mathbf{\Sigma_x}). \tag{10}$$

Proof. Recall that

$$\tilde{\Gamma}(\mathcal{Z}_{s_n}) = \sum_{i=1}^{s_n}\mathbb{E}[\Gamma(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_i] - \mathbb{E}[T(\mathcal{Z}_{s_n})] = \sum_{i=1}^{s_n}\Gamma_n(\mathbf{Z}_i),$$

the first order approximation of $\Gamma(\mathcal{Z}_{s_n})$ as in Lemma D.1.

<u>Claim:</u>

$$\hat{\tau}_{n,k}(\mathbf{x}) - \tau_k(\mathbf{x}) = \frac{s_n}{n}\sum_{i=1}^{n}\Gamma_n(\mathbf{Z}_i) + o_p(\sigma_n), \tag{155}$$

with

$$\sigma_n^2 = \frac{s_n^2}{n}\mathrm{Var}(\Gamma_n(\mathbf{Z}_1)) = \frac{s_n}{n}\mathrm{Var}(\tilde{\Gamma}(\mathcal{Z}_{s_n})), \tag{156}$$

Proof: We first note that by boundedness of the kernel, $\mathrm{Var}(\Gamma(\mathcal{Z}_{s_n})) \leq C$ for all n, for some $C < \infty$. Second, as a consequence of Theorem E.5, Lemma E.1, (73) and (77), (88) holds, that is:

$$\mathrm{Var}(\tilde{\Gamma}(\mathcal{Z}_{s_n})) = s_n\mathrm{Var}(\Gamma_n(\mathbf{Z}_1)) = \Omega\left(\frac{1}{\log(s_n)^p}\right).$$

Thus it follows that:

$$\frac{\mathrm{Var}(\tilde{\Gamma}(\mathcal{Z}_{s_n}))}{\mathrm{Var}(\Gamma(\mathcal{Z}_{s_n}))} = \Omega\left(\frac{1}{\log(s_n)^p}\right). \tag{157}$$

The same steps as in (Näf et al., 2023, Theorem 3) in turn implies that

$$\begin{aligned}\frac{1}{\sigma_n^2}\mathbb{E}[\|\hat{\tau}_{n,k}(\mathbf{x}) - \mathbb{E}[\hat{\tau}_{n,k}(\mathbf{x})] - \frac{s_n}{n}\sum_{i=1}^n \Gamma_n(\mathbf{Z}_i)\|_{\mathcal{H}}^2] \leq& \frac{s_n^2}{n^2}\frac{\mathrm{Var}(\Gamma(\mathcal{Z}_{s_n}))}{\sigma_n^2}\\ =&\frac{s_n}{n}\frac{\mathrm{Var}(\Gamma(\mathcal{Z}_{s_n}))}{\mathrm{Var}(\tilde{\Gamma}(\mathcal{Z}_{s_n}))}\\ =&\mathcal{O}\left(\frac{s_n}{n}\frac{\log(s_n)^p}{\varepsilon C_{f,p}}\right),\end{aligned}$$

the latter going to zero as $s_n = n^\beta$, with $\beta < 1$ by **(F5)**.[2] By Theorem E.3, we also have that

$$\frac{1}{\sigma_n^2}\|\mathbb{E}[\hat{\tau}_{n,k}(\mathbf{x})] - \tau_k(\mathbf{x})\|_{\mathcal{H}}^2 = \frac{1}{\sigma_n^2}\mathcal{O}\left((\varepsilon s_n)^{-\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}\right).$$

As, with (88),

$$\sigma_n^2 = \frac{s_n}{n}\mathrm{Var}(\tilde{\Gamma}(\mathcal{Z}_{s_n})) = \Omega\left(\frac{s_n}{n}\frac{1}{\log(s_n)^p}\right)$$

it holds with $C_\alpha = \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}$, that,

$$\frac{1}{\sigma_n^2}\|\mathbb{E}[\hat{\tau}_{n,k}(\mathbf{x})] - \tau_k(\mathbf{x})\|_{\mathcal{H}}^2 = \frac{1}{\sigma_n^2}\mathcal{O}\left(s_n^{-C_\alpha\frac{\pi}{p}}\right) = \mathcal{O}\left(\frac{n\log(s_n)^p}{s_n^{1+C_\alpha\pi/p}}\right) = \mathcal{O}\left(\frac{n\log(s_n)^p}{n^{\beta(1+C_\alpha\pi/p)}}\right).$$

This goes to zero using the scaling of β in **(F5)**. □

Claim (155) is the first key to show asymptotic normality, as it shows that $1/\sigma_n(\hat{\tau}_{n,k}(\mathbf{x}) - \tau_k(\mathbf{x}))$ is asymptotically equivalent to a sum of independent elements in the Hilbert space $\mathcal{H}$. Thus we can use CLT results for $\mathcal{H}$, and in particular, we will show (I) marginal convergence and (II) uniform tightness:

Claim: (I) For all $f \in \mathcal{H}$,

$$\frac{s_n}{n\sigma_n}\sum_{i=1}^n \langle f, \Gamma_n(\mathbf{Z}_i)\rangle \xrightarrow{D} N\left(0, \frac{c\sigma_f^2(\mathbf{x},1)\mathbb{P}(W=1) + \sigma_f^2(\mathbf{x},0)\mathbb{P}(W=0)}{c\sigma^2(\mathbf{x},1)\mathbb{P}(W=1) + \sigma^2(\mathbf{x},0)\mathbb{P}(W=0)}\right) \tag{158}$$

Proof:

For $f \in \mathcal{H}$, consider the univariate marginal $\frac{s_n}{n}\sum_{i=1}^n\langle\Gamma_n(\mathbf{Z}_i), f\rangle$. Assumptions **(F1)**–**(F5)**, **(C1)**–**(C2)** and Assumptions **(D1)**–**(D5)** verify all assumptions of Theorem 11 of (Wager and Athey, 2017). Consequently, there exists a $\sigma_n(f) > 0$ converging to zero with n such that

$$\left\langle \frac{1}{\sigma_n(f)}\left(\frac{s_n}{n}\sum_{i=1}^n \Gamma_n(\mathbf{Z}_i)\right), f\right\rangle \xrightarrow{D} N(0,1). \tag{159}$$

[2]We note that (Näf et al., 2023) uses the abbreviation $k(\mathbf{Y}_i, \cdot) = \xi_i$.

Unfortunately, the scaling factor

$$\sigma_n^2(f) = \frac{s_n^2}{n}\mathrm{Var}(\langle \Gamma_n(\mathbf{Z}_1), f\rangle)$$

obtained from (Wager and Athey, 2018) depends on f. However, from Theorem E.5, it follows that

$$\frac{\sigma_n(f)}{\sigma_n} = \frac{\sqrt{\mathrm{Var}(\langle \Gamma_n(\mathbf{Z}_1), f\rangle)}}{\sqrt{\mathrm{Var}(\Gamma_n(\mathbf{Z}_1))}} \to \frac{c\sigma_f^2(\mathbf{x},1)\mathbb{P}(W=1) + \sigma_f^2(\mathbf{x},0)\mathbb{P}(W=0)}{c\sigma^2(\mathbf{x},1)\mathbb{P}(W=1) + \sigma^2(\mathbf{x},0)\mathbb{P}(W=0)},$$

which is finite and strictly larger zero by (**D3**), (**D4**). Thus

$$\begin{aligned}\left\langle \frac{1}{\sigma_n}\left(\frac{s_n}{n}\sum_{i=1}^n \Gamma_n(\mathbf{Z}_i)\right), f\right\rangle &= \frac{\sigma_n(f)}{\sigma_n}\frac{1}{\sigma_n(f)}\frac{s_n}{n}\sum_{i=1}^n \langle \Gamma_n(\mathbf{Z}_i), f\rangle \\ &\xrightarrow{D} N\left(0, \frac{c\sigma_f^2(\mathbf{x},1)\mathbb{P}(W=1) + \sigma_f^2(\mathbf{x},0)\mathbb{P}(W=0)}{c\sigma^2(\mathbf{x},1)\mathbb{P}(W=1) + \sigma^2(\mathbf{x},0)\mathbb{P}(W=0)}\right),\end{aligned}$$

proving the claim. □

Before continuing, by the definition of σ_n, we define

$$\xi_n^0(\mathbf{x}) := \sum_{i=1}^n \frac{s_n}{n\sigma_n}\Gamma_n(\mathbf{Z}_i) = \sum_{i=1}^n \frac{\Gamma_n(\mathbf{Z}_i)}{\sqrt{n\mathrm{Var}(\Gamma_n(\mathbf{Z}_1))}}.$$

<u>Claim:</u> (II) $\xi_n^0(\mathbf{x})$ is uniformly tight.

<u>Proof:</u> Because $\mathcal{H}$ is separable due to our assumptions on the kernel, there exists a complete orthogonal basis $(e_j)_{j\in\mathbb{N}}$ of $\mathcal{H}$; see for instance (Hsing and Eubank, 2015). Let P_k be the projection operator onto the linear span of the first k elements of $(e_j)_{j\in\mathbb{N}}$, $S_k = \mathrm{span}(e_1, \ldots, e_k)$. Given the tools from above, we can use the exact same arguments as in (Näf et al., 2023, Theorem 6) to verify the condition $\limsup_n \mathbb{E}[\|\xi_n^0(\mathbf{x}) - P_k(\xi_n^0(\mathbf{x}))\|_{\mathcal{H}}^2] \to 0$, as $k \to \infty$, which is sufficient for tightness (Chen and White, 1998, Lemma 3.2). □

Univariate convergence together with tightness imply $\xi_n^0(\mathbf{x}) \xrightarrow{D} N(0, \mathbf{\Sigma_x})$; see for example (Chen and White, 1998, Lemma 3.1/3.2) or (Hsing and Eubank, 2015, Chapter 7). Since by Equation (155) we have

$$\frac{1}{\sigma_n}(\hat{\tau}_{n,k}(\mathbf{x}) - \tau_k(\mathbf{x})) = \xi_n^0(\mathbf{x}) + o_p(1),$$

the result follows.

□

Finally, having obtained these prior results Theorems 4.3, 4.4 and 4.5 can be proven using the same approach as in (Näf et al., 2023), making use of the general results given in Praestgaard and Wellner (1993); Kosorok (2003); González-Rodríguez and Colubi (2017). The proofs are thus omitted.