Heterogeneous Treatment Effect based Random Forest: HTERF

Bérénice-Alexia Jocteur, Véronique Maume-Deschamps, Pierre Ribereau

Heterogeneous Treatment Effect based Random Forest: HTERF

Bérénice-Alexia Jocteur[a,b], Véronique Maume-Deschamps[a], Pierre Ribereau[a]

[a] *Institut Camille Jordan, Université de Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, Villeurbanne, F-69622, France*
[b] *Enterprise Risk Management, Natixis, Paris, 75013, France*

Abstract

Estimates of causal impacts can be needed to answer what-if questions about shifts in policy, such as new treatments in pharmacology or new pricing strategies for a business owner. In this paper we propose a non-parametric approach to estimate heterogeneous treatment effect based on random forests: HTERF. In the potential outcome framework with unconfoundedness we show that HTERF is pointwise a.s.-consistent to the true treatment effect. An interpretability result is also presented. A software implementation, `CausalForest` for `Julia` is available on the general repository of `Julia`.

Keywords: causal forest, causal inference, heterogeneous treatment effect, potential outcomes

1. Introduction

The automation of decision-making across a wide range of application domains is one of machine learning's goals. The estimation of heterogeneous treatment effect or, more specifically, how to determine how an intervention will affect a particular outcome in relation to a variety of observable characteristics of the treated sample presents a fundamental challenge in the majority of data-driven personalized decision scenarios. It occurs in clinical studies when the aim is to evaluate how a pharmacological treatment affects a patient's clinical response in relation to patient variables. It also occurs in empirical research in economics and related fields when the goal is to determine the impact of realized or hypothetical interventions in order to assess theories and improve policies.

Two classes of statistical methods can be identified for causal inference: metalearners and tree based methods. The main contribution of this paper is the introduction of a new algorithm for CATE (Conditional Average Treatment Effect) estimation: HTERF (Heterogeneous Treatment Effect based Random Forest). This new algorithm uses a random forest with a new splitting criterion specifically designed for binary treatments and is improved by a preliminary step in the metalearners spirit (see Section 5). Our aim is to put the emphasis on interpretability of the algorithm. Indeed for regression random forests, under certain assumptions it is proven that the most informative variables appear more often in probability in the tree construction, see Scornet et al. (2015). We obtain an almost sure result for HTERF.

Finally we compare our approach with previously developed ones on simulated data inspired from ones presented in the causal treatment effect litterature. We compare the performance of HTERF algorithm against GRF (Generalized Random Forest) causal forest (Athey et al. (2019)) using simulations, finding that HTERF dominates both in term of CATE RMSE and interpretability in different settings.

The paper is organised as follows. Section 2 introduces the potential outcomes framework and CATE estimations methods. Section 3 describes our method HTERF and Section 4 presents consistency results. Section 5 evaluates performances of HTERF. Finally Section 6 present our conclusions.

2. Inference for treatment effect

In this section the potential outcome framework is presented as well as state of the art methods for CATE estimation.

2.1. The causal framework

Following the potential outcomes framework as presented in Imbens and Rubin (2015), we posit the potential outcomes respectively $Y(1)$ and $Y(0)$ corresponding to the outcome we would have observed, had we assigned respectively control or treatment to the quantity of interest Y. Assume that we observe $Y = Y(W)$, where W is a binary treatment. We also consider a set of covariates $\mathbf{X} \in \mathbb{R}^d$. The conditional average treatment effect (CATE) at $\mathbf{x}$ is defined as:

$$\tau(\mathbf{x}) = \mathbb{E}\left[Y(1) - Y(0) | \mathbf{X} = \mathbf{x}\right]. \tag{2.1}$$

A standard assumption for identifiability of CATE is unconfoundedness (Rosenbaum and Rubin (1983)), meaning that conditionally on $\mathbf{X}$ the treatment assignment W is independent of the potential outcomes for Y :

$$\{Y(1), Y(0)\} \perp\!\!\!\perp W|\mathbf{X}. \tag{2.2}$$

We consider n independent and identically distributed training individuals labeled $i = 1, \ldots, n$. Each of them is constituted of a feature vector $\mathbf{X}_i \in \mathbb{R}^d$, an outcome $Y_i \in \mathbb{R}$ and a treatment indicator $W_i \in \{0, 1\}$. We denote the observed data as

$$\mathscr{D}_n = (Y_i, \mathbf{X}_i, W_i)_{1 \leq i \leq n}.$$

The distribution of $\mathscr{D}_n$ is specified by distribution $\mathscr{P}$.

In this work we are interested in consistent estimators $\hat{\tau}(.)$ of τ. The difficulty to evaluate the function $\tau(.)$ is that we only observe one of the two potential outcomes for a given training example, so we cannot directly train a classical machine learning method on the difference $Y_i(1) - Y_i(0)$.

2.2. Methods for causal effect estimation

We can categorize the methods for evaluating CATE in two groups. On one hand methods using classical machine learning methods (random forest, boosting...), these estimators cannot evaluate CATE directly and are usually called metalearners. On the other hand there exist machine learning methods designed to estimate CATE directly, examples are causal forests or Bayesian regression tree models for causal inference.

A review on metalearners can be found in Künzel et al. (2019). Metalearner combine base learners in a specific fashion to estimate CATE. Base learners are supervised learning or regression estimators, but they are not specified in the metalearner.

Two basic examples of metalearners are T- and S-learners. The T-learner estimates $Y(1)$ and $Y(0)$ separately, the estimated CATE is given by:

$$\hat{\tau}_T(\mathbf{x}) = \hat{\mu}_1(\mathbf{x}) - \hat{\mu}_0(\mathbf{x}), \tag{2.3}$$

where $\hat{\mu}_1(\mathbf{x})$ (respectively $\hat{\mu}_0(\mathbf{x})$) is an estimator of $\mu_1(\mathbf{x}) = \mathbb{E}[Y(1)|\mathbf{X} = \mathbf{x}]$ (resp. $\mu_0(\mathbf{x}) = \mathbb{E}[Y(0)|\mathbf{X} = \mathbf{x}]$) using the observations in $\{(\mathbf{X}_i, Y_i)\}_{W_i=1}$ (resp. $\{(\mathbf{X}_i, Y_i)\}_{W_i=0}$).

The S-learner uses a single base learner $\hat{\mu}$. It estimates the quantity $\mu(\mathbf{x}, w) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, W = w]$ with any base learner on the whole dataset.

The CATE estimator is then given by:

$$\hat{\tau}_S(\mathbf{x}) = \hat{\mu}(\mathbf{x}, 1) - \hat{\mu}(\mathbf{x}, 0). \tag{2.4}$$

These methods allow a full control on which estimation method to use at each stage. Moreover, they allow to perform cross-validation for more data-adaptive estimation at each stage. Hence, they allow the user to do model selection both for base learners and for the final CATE model. However in order to get good CATE estimations the base learners need to reach a matching of the features of the data and the underlying model.

In the literature the use of forest based algorithms to estimate heterogeneous treatment effect has been proposed. Some of these papers use the Bayesian Additive Regression Tree (BART) method from Chipman et al. (2010), such approaches can be seen in Hill (2011); Green and Kern (2012); Hill and Su (2013). Other approaches relying on tree-based methods have been developed, they modify the standard random forest algorithm to focus on estimating CATE directly. These methods are often called causal trees and causal forests. A first approach using random forest with custom splitting criterion is given by Su et al. (2009). An alternative criterion for causal trees is then proposed by Athey and Imbens (2016), this approach also allows the construction of confidence intervals for causal effect. It inspired the causal forest developed in Wager and Athey (2018) which introduced the idea of double sampling: using one sample to build trees and another one for the CATE estimation. Finally GRF causal forests introduced in Athey et al. (2019) are a special case of the previous causal forest.

Generalized random forests, is a method for non parametric estimation that applies to a wide variety of quantities of interest: quantile regression, CATE estimation, instrumental variable regression. We will focus on CATE estimation. We will compare our method with GRF since it improves previous ones. We can decompose the GRF algorithm in two parts: the growing of the trees and the quantity of interest estimation.

A random forest as presented in Breiman (2001), consists in trees $T_1, \ldots, T_B$. To obtain a prediction for a test point $\mathbf{x}$, this point is pushed down in each of the trees until it reaches a leaf, a prediction is associated to each leaf. Let $\hat{\mu}_b$ be the prediction from tree b, then the random forest prediction for $\mathbf{x}$ is: $\frac{1}{B}\sum_{b=1}^{B} \hat{\mu}_b(\mathbf{x})$.

In GRF the strategy is slightly different, the test $\mathbf{x}$ is still pushed down in the trees, but instead of looking for a prediction at each tree, we consider

$L_b(\mathbf{x})$ the set of elements in the training sample that fall into the same leaf as $\mathbf{x}$. For each $i = 1, \ldots, n$ define:

$$\alpha_{b,i}(\mathbf{x}) = \frac{\mathbb{1}\{\mathbf{X}_i \in L_b(\mathbf{x})\}}{|L_b(\mathbf{x})|}, \alpha_i(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \alpha_{b,i}(\mathbf{x}). \tag{2.5}$$

The $\alpha_i(\mathbf{x})$ can be seen as a weighting function that indicates how important each training sample is when trying to predict at $\mathbf{x}$. Indeed we also notice that $\sum_{i=1}^{n} \alpha_i(\mathbf{x}) = 1$.

Once all the weights have been calculated, CATE is estimated as follows:

$$\hat{\tau}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \alpha_i(\mathbf{x})(W_i - \overline{W_\alpha})(Y_i - \overline{Y_\alpha})}{\sum_{i=1}^{n} \alpha_i(\mathbf{x})(W_i - \overline{W_\alpha})^2}, \tag{2.6}$$

where $\overline{W_\alpha} = \sum_{i=1}^{n} \alpha_i(\mathbf{x}) W_i$ and $\overline{Y_\alpha} = \sum_{i=1}^{n} \alpha_i(\mathbf{x}) Y_i$. This estimation step is described in Algorithm 2.

This expression is an empirical version of $\frac{Cov(W,Y|\mathbf{X}=\mathbf{x})}{Var(Y|\mathbf{X}=\mathbf{x})}$. A simple computation shows that if W has a linear impact : $Y = \tau(\mathbf{X})W + \gamma(\mathbf{X})$ then $\tau(\mathbf{X}) = \frac{Cov(W,Y|\mathbf{X})}{Var(Y|\mathbf{X})}$.

Before studying the splitting criterion used in GRF, two differences with Breiman random forests can be mentioned, the trees are trained on subsamples of the training data and a subsampling technique named honesty is also used. These strategies are used to obtain a good theoretical statistical behavior. The idea of honesty is to split the training subsample in two subsets before building each tree, the first one is used to build the nodes of the tree and the second one is used to fill the tree and will be used to estimate the quantity of interest.

Let P be some parent node, $\mathcal{J}$ the elements of the sample belonging to P and let C_1 and C_2 be the two child nodes for a given split. A criterion similar to CART regression would be to minimize:

$$err(C_1, C_2) = \sum_{j=1}^{2} \mathbb{P}(\mathbf{x} \in C_j | \mathbf{x} \in P) \mathbb{E}\left[(\hat{\tau}_{C_j}(\mathcal{J}) - \tau(\mathbf{x}))^2 | \mathbf{x} \in C_j\right], \tag{2.7}$$

where $\hat{\tau}_{C_j}(\mathcal{J})$ is the estimation of τ over child nodes C_j.

Unfortunately the true CATE is unknown, a calculable criterion would be to maximize the following quantity. It favores splits that increase the heterogeneity of the CATE estimates betwen children. This idea has been already proposed by Athey and Imbens (2016):

$$\Delta(C_1, C_2) = \frac{n_{C_1} n_{C_2}}{n_P^2} \left[\hat{\tau}_{C_1}(\mathcal{J}) - \hat{\tau}_{C_2}(\mathcal{J})\right]^2, \tag{2.8}$$

where n_{C_1}, n_{C_1} and n_P are the number of points that fall into node C_1, C_2 and P respectively.

We present GRF for the estimation of CATE but recall that GRF is applicable to a wide range of quantities of interest. Optimizing Equation (2.8) over all possible splits, would mean to estimate the quantity of interest for both children for each candidate split, which is too expensive in terms of complexity for most cases. Instead gradient-based approximations named pseudo-outcomes are used. For CATE estimation, the following pseudo-outcomes are computed:

$$\rho_i = A_P^{-1}(W_i - \overline{W}_P)(Y_i - \overline{Y}_P - (W_i - \overline{W}_P)\hat{\beta}_P), \tag{2.9}$$

where β_P is the least-squares regression solution of Y_i on W_i and:

$$A_P = \frac{1}{|\{i : X_i \in P\}|} \sum_{i:\mathbf{X}_i \in P} (W_i - \overline{W}_P). \tag{2.10}$$

Finally the chosen split is the one which maximizes, it is a classical CART regression split over pseudo-outcomes:

$$\tilde{\Delta}(C_1, C_2) = \frac{1}{n_{C_1}} \left(\sum_{i:\mathbf{X}_i \in C_1} \rho_i \right)^2 + \frac{1}{n_{C_2}} \left(\sum_{i:\mathbf{X}_i \in C_2} \rho_i \right)^2. \tag{2.11}$$

This tree building step is applied in Algorithm 1.

In practice a prior centering is applied before running the algorithm: it involves regressing out the effect of the features $\mathbf{X}_i$ on W_i and Y_i separately. It improves the performances on finite datasets.

Remark 2.1. *The pseudo-outcomes have been introduced for easier calculations, however for CATE estimation we obtain the same algorithmic complexity when computing* $\Delta(C_1, C_2)$ *or* $\tilde{\Delta}(C_1, C_2)$.

2.3. Limitations of the GRF approach

Random forests are effective regression algorithms and are quite interpretable, under the assumption that Y follows an additive regression model, Scornet et al. (2015) proved that the algorithm selects splits mostly along

Algorithm 1 Building random forest algorithm

Input: $B > 0$ number of trees, s subsampling rate, $\mathcal{S}$ set of examples
  for $b = 1$ to B **do**
    set of examples $\mathcal{I}_b \leftarrow \textsc{Subsample}(S, s)$
    ▷ Draw a subsample from $\mathcal{S}$ without replacement of size $s|\mathcal{S}|$
    sets of examples $\mathcal{I}_{b,1}, \mathcal{I}_{b,2} \leftarrow \textsc{SplitSample}(\mathcal{I}_b)$
    ▷ Randomly divides a set into two evenly-sized, non-overlapping halves
    node $P_0 \leftarrow \textsc{CreateNode}(\mathcal{I}_{b,1})$
    queue $\mathcal{Q} \leftarrow \textsc{InitializeQueue}(P_0)$
    while $\textsc{NotNull}(\text{node } P \leftarrow \textsc{Pop}(\mathcal{Q}))$ **do**
      vector $R_P \leftarrow \textsc{GetPseudoOutcomes}(P)$ ▷ Computes (2.9)
      split $\Sigma \leftarrow \textsc{MakeCartSplit}(P, R_P)$ ▷ Optimizes (2.11)
      if SplitSucceeded **then** ▷ If there is a legal split
        $\textsc{SetChildren}(P, \textsc{GetLeftChild}(\Sigma), \textsc{GetRightChild}(\Sigma))$
        $\textsc{AddToQueue}(\mathcal{Q}, \textsc{GetLeftChild}(\Sigma))$
        $\textsc{AddToQueue}(\mathcal{Q}, \textsc{GetRightChild}(\Sigma))$
      end if
    end while
    ▷ Tree $\mathcal{T}_b$ has been built
  end for
Output: A causal forest with trees $\mathcal{T}_1, \ldots, \mathcal{T}_B$

Algorithm 2 Estimation algorithm

Input: A causal forest with trees $\mathcal{T}_1, \ldots, \mathcal{T}_B$, a test point $\mathbf{x}$, the size of training set n.
  weight vector $\alpha \leftarrow \textsc{Zeros}(n)$ ▷ Create a vector of zeros of length n
  for $b = 1$ to B **do**
    $\mathcal{N} \leftarrow \textsc{Neighbors}(\mathbf{x}, \mathcal{T}_b, \mathcal{I}_{b,2})$
    ▷ Elements of $\mathcal{I}_{b,2}$ that fall into the same leaf as $\mathbf{x}$ in the tree $\mathcal{T}_b$
    for all example $e \in \mathcal{N}$ **do**
      $\alpha[e] \mathrel{+}= \frac{1}{|\mathcal{N}|}$
    end for
  end for
  $\alpha = \alpha / B$
Output: $\hat{\tau}(\mathbf{x})$ ▷ Uses (2.6)

informative variables. On simple linear regression examples, in the first stages of the regression forest only the significant variables are present. On the contrary when we consider a simple linear causal example, the overrepresentation of informative variables is not so striking (see Table 5.1), which limits interpretability of GRF causal forests. This is one of our motivations to propose a new splitting criterion for causal forests.

The GRF approach is a general framework not tailored for causal inference and the GRF causal forest is adapted to linear treatment effects. Our HTERF splitting criterion is specifically designed to assess CATE when the treatment is binary and could be adapted for multiple discrete treatments.

3. Estimation of causal effect with HTERF

The splitting criterion used in HTERF is based on the idea to maximise the difference on treatment effect between child nodes. In particular Equation (2.1) is used to define this splitting criterion.

3.1. Algorithm

We assume that we are given a training sample $\mathcal{D}_n = (Y_j, \mathbf{X}_j, W_j)_{j=1,\dots,n}$ of independent random variables distributed as the prototype triple $(Y, \mathbf{X}, W)$ which is a $(d+2)$-dimensional random vector. The purpose is to use the dataset $\mathcal{D}_n$ to construct an estimator $\hat{\tau}_{B,n} : \mathcal{X} \to \mathbb{R}$ of τ.

The tree building process of HTERF is the following. First of all prior to the construction of each tree, a subsampling and an honest splitting is done as in GRF. The optimisation of the splitting citerion is done over a subset of features $\mathcal{M}_{try}$. The features are selected randomly, with positive probability for each covariate to be selected, which includes the uniform selection. Then the best split is the one maximizing the splitting criterion $\Delta(A, j, z)$, where $A = \prod_{i=1}^{d} [a_i, b_i]$ is the current node, j is chosen in $\mathcal{M}_{try}$ and $z \in A^j = [a_j, b_j]$.

$$\Delta(A, j, z) = \frac{|A_L||A_R|}{|A|^2} \left(\left(\overline{Y}_{A_{L1}} - \overline{Y}_{A_{L0}} \right) - \left(\overline{Y}_{A_{R1}} - \overline{Y}_{A_{R0}} \right) \right)^2, \qquad (3.1)$$

where $A_{L1} = \{\mathbf{X}_i \in A | \mathbf{X}_i^{(j)} < z, W_i = 1\}$, $A_{L0} = \{\mathbf{X}_i \in A | \mathbf{X}_i^{(j)} < z, W_i = 0\}$, $A_{R1} = \{\mathbf{X}_i \in A | \mathbf{X}_i^{(j)} \geq z, W_i = 1\}$, $A_{R0} = \{\mathbf{X}_i \in A | \mathbf{X}_i^{(j)} \geq z, W_i = 0\}$, $A_L = A_{L1} \cup A_{L0}$ and $A_R = A_{R1} \cup A_{R0}$. For all sets B, we denote $\overline{Y}_B = \frac{1}{|B|} \sum_{i \in B} Y_i$. This splitting criterion is partially inspired by the ones used in Athey and Imbens (2016) and Athey et al. (2019).

For the estimation of τ we reuse the procedure of GRF with Algorithm 2, but the estimation is different, namely:

$$\hat{\tau}_{B,n}(\mathbf{x}) = \sum_{i:W_i=1} \alpha_i(\mathbf{x}) Y_i - \sum_{i:W_i=0} \alpha'_i(\mathbf{x}) Y_i, \tag{3.2}$$

where α (resp. α') is the weight vector defined as in Algorithm 2 associated to observations such as $W_i = 1$ (resp. $W_i = 0$), see below.

We have the following notations:

- $\Theta_\ell, \ell = 1, \ldots, B$ are independent random vectors, distributed as a generic random vector $\Theta = (\Theta^1, \Theta^2, \Theta^3)$ and independent of $\mathcal{D}_n$ and (Θ^1, Θ^2) is independent of Θ^3. Θ^1 contains indices of observations that are used to build each tree, i.e. the subsample $\mathcal{I}_1$, Θ^2 contains indices of observations that are used for estimations in each tree, i.e. the subsample $\mathcal{I}_2$ and Θ^3 contains indices of splitting candidate variables in each node, we assume that Θ^3 gives a positive probability to each co-variate, we need to consider both Θ^1 and Θ^2 because $\mathcal{I}_2$ is the complementary of $\mathcal{I}_1$ in $\mathcal{I}$ which is random itself,

- $\mathcal{D}^\star_{n,1}(\Theta_\ell)$ and $\mathcal{D}^\star_{n,2}(\Theta_\ell)$ are the disjoint subsamples selected prior to the tree construction, the first one is used to build the tree and the second one allow to build weights used during estimation step,

- $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ is the tree cell (subspace of $\mathcal{X}$) containing $\mathbf{x}$,

- $N_{n,1}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ (resp. $N_{n,0}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$) is the number of elements of $\mathcal{D}^\star_{n,2}(\Theta_\ell)$ that fall into $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ and such as $W_i = 1$ (resp. $W_i = 0$).

We define the weights:

$$\alpha_i(\mathbf{x}) = \frac{1}{B} \sum_{l=1}^{B} \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}; \Theta_l, \mathcal{D}_n) \cap W_i = 1 \cap i \in \mathcal{D}^\star_{n,2}(\Theta_\ell)}}{N_{n,1}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)}, \tag{3.3}$$

$$\alpha'_i(\mathbf{x}) = \frac{1}{B} \sum_{l=1}^{B} \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}; \Theta_l, \mathcal{D}_n) \cap W_i = 0 \cap i \in \mathcal{D}^\star_{n,2}(\Theta_\ell)}}{N_{n,0}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)}. \tag{3.4}$$

Remark 3.1. *The output $\hat{\tau}(x)$ can also be seen as an average of estimations obtained by several causal trees (as in Breiman random forest). Indeed:*

$$\begin{aligned}\hat{\tau}_{B,n}(\mathbf{x}) = &\sum_{i:W_i=1} \frac{1}{B}\sum_{l=1}^{B} \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x};\Theta_l,\mathcal{D}_n)\cap i\in\mathcal{D}^{\star}_{n,2}(\Theta_\ell)}}{N_{n,1}\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)} Y_i \\ &- \sum_{i:W_i=0} \frac{1}{B}\sum_{l=1}^{B} \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x};\Theta_l,\mathcal{D}_n)\cap i\in\mathcal{D}^{\star}_{n,2}(\Theta_\ell)}}{N_{n,0}\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)} Y_i \\ =&\frac{1}{B}\sum_{l=1}^{B} \sum_{\substack{i\in\mathcal{D}^{\star}_{n,2}(\Theta_\ell)\\ W_i=1\\ \mathbf{X}_i\in A_n(\mathbf{x};\Theta_l,\mathcal{D}_n)}} \frac{Y_i}{N_{n,1}\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)} \\ &- \frac{1}{B}\sum_{l=1}^{B} \sum_{\substack{i\in\mathcal{D}^{\star}_{n,2}(\Theta_\ell)\\ W_i=0\\ \mathbf{X}_i\in A_n(\mathbf{x};\Theta_l,\mathcal{D}_n)}} \frac{Y_i}{N_{n,0}\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)}.\end{aligned}$$

Following the `R` package `grf` which provides an implementation of Athey et al. (2019) algorithm, we define a notion of importance of variables. Let $freq$ be the matrix of split depth by feature index: $freq_{i,j}$ is the number of time (over the forest) the split has been done along X_i at depth j, divided by the total number of splits at depth j. It is the frequency of splits on each feature for a given depth. The importance of a feature can be defined as a weighted sum of how many times the feature was split on at each depth in the forest. It depends on two parameters: max_depth, the maximum depth considered to get the $freq$ matrix and $decay$, the decay exponent that controls the importance of split depth.

We now define the importance of the ith feature as:

$$Imp_i(max\_depth, decay) = \frac{\sum_{k=1}^{max\_depth} freq_{k,i} k^{-decay}}{\sum_{k=1}^{max\_depth} k^{-decay}}. \tag{3.5}$$

In the numerical applications that follow the values used for max_depth and $decay$ are respectively 4 and 2.

3.2. Theoretical tree

Similarly to what is done in Scornet et al. (2015), a random theoretical tree can be defined for HTERF. The theoretical equivalent of the empirical HTERF splitting criterion on a node A is:

$$\begin{aligned}\Delta^\star(A,j,z) =& \mathbb{P}[\mathbf{X}^{(j)} < z | \mathbf{X} \in A]\mathbb{P}[\mathbf{X}^{(j)} \geq z | \mathbf{X} \in A] \\ & (\mathbb{E}[Y(1) - Y(0) | \mathbf{X}^{(j)} < z, \mathbf{X} \in A] \\ & - \mathbb{E}[Y(1) - Y(0) | \mathbf{X}^{(j)} \geq z, \mathbf{X} \in A])^2.\end{aligned} \tag{3.6}$$

A theoretical tree is obtained by optimizing the best consecutive cuts $(j^\star, z^\star)$ optimizing the previous criterion $\Delta^\star(A, ., .)$.

4. Consistency of HTERF

4.1. Existing results

By construction of the HTERF algorithm, the following assumptions are made:

- Unconfoundedness as in Equation (2.2).
- Honesty: two different samples are used for constructing the splits and predicting the labels.
- The resampling is done by subsampling and not by boostrap as in Breiman forests.

With Remark 3.1 we can see that $\hat{\tau}_{B,n}$ is a U-statistic and under additional assumptions below a normality result for $\hat{\tau}_{B,n}$ follows from Wager and Athey (2018).

Assumption 4.1.

1. $\mathbf{X}$ *is a uniform random vector with independent coordinates:* $\mathbf{X} \sim U([0,1]^d)$.
2. $(\mathbf{X}, Y(u))$ *with* $u \in \{0,1\}$ *verifies* $\mathbf{x} \mapsto \mathbb{E}[Y(u)|\mathbf{X} = \mathbf{x}]$ *and* $\mathbf{x} \mapsto \mathbb{E}[Y(u)^2|\mathbf{X} = \mathbf{x}]$ *are Lipschitz-continuous,* $Var[Y(u)|\mathbf{X} = \mathbf{x}] > 0$ *and* $\mathbb{E}[|Y(u) - \mathbb{E}[Y(u)|\mathbf{X} = \mathbf{x}]|^{2+\delta}|\mathbf{X} = \mathbf{x}] \leq M$ *for some constants* $\delta, M > 0$ *uniformly over all* $\mathbf{x} \in [0,1]^d$.
3. *At every step of the tree building procedure, the probability that the next split is done along the* $j-th$ *feature is bounded below by* π/d *for some* $0 < \pi \leq 1$ *for all* $j = 1, \ldots, d$.

4. *A causal tree is $\alpha-$ regular at $\mathbf{x}$ for some $\alpha > 0$ if the sample $\mathcal{I}_2$ used for estimation verifies: (1) each split leaves at least a fraction α of the available training sample on each side of the split, (2) the leaf containing $\mathbf{x}$ has at least k observations for each treatment group for some $k \in \mathbb{N}$ and (3) the leaf containing $\mathbf{x}$ has either less than $2k-1$ observations with $W_i = 0$ or $2k-1$ observations with $W_i = 1$.*
5. *The subsample size s_n scales as: $s_n \asymp n^{\beta}$ for some $\beta_{min} := 1 - \left(1 + \frac{d}{\pi}\frac{\log(\alpha^{-1})}{\log((1-\alpha)^{-1})}\right) < \beta < 1$.*

Theorem (Athey et al. (2019))**.** *Under Assumptions 4.1 and that $Y = \tau(\mathbf{X})W + \gamma(\mathbf{X})$, we have:*

$$\frac{\hat{\tau}_{B,n}(\mathbf{x}) - \tau(\mathbf{x})}{\sqrt{Var(\hat{\tau}(\mathbf{x}))}} \longrightarrow N(0,1)$$

,

where the variance of the causal forest can be consistently estimated using the infinitesimal jackknife for random forests, see Wager and Athey (2018) for more details.

4.2. New consistency results

We propose a consistency result under assumptions weaker than Assumptions 4.1, based on Elie-Dit-Cosaque and Maume-Deschamps (2022). In what follows, $\mathcal{X}$ is a compact hyper-rectangle of $\mathbb{R}^d$:
$\mathcal{X} = \prod_{i=1}^{d}[u_i, v_i]$, $-\infty < u_i \leq v_i < \infty$ and we denote by $\mathcal{A}$ the set of hyper-rectangles in $\mathcal{X}$: $A \in \mathcal{A}$ writes $A = \prod_{i=1}^{d}[a_i, b_i]$ with $u_i \leq a_i \leq b_i \leq v_i$. Also, we denote by $A^{-j} = \prod_{k\neq j}[a_k, b_k]$ and $A^J = \prod_{k\in J}[a_k, b_k]$ for any $J \subset \{1, \ldots, d\}$. Given $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x}^{-j}$ is the vector of $\mathbb{R}^{d-1}$ where the j-th coordinate has been removed and $\mathbf{x}^J$ is the vector of $\mathbb{R}^J$ whose coordinates are $x^{(j)}$, $j \in J$.

Definition 4.1. *Let $f : \mathcal{X} \to \mathbb{R}$, it does NOT belong to the ♠-class if there exists a rectangle $A = \prod_{j=1}^{d}[a_j, b_j] \subset \mathcal{X}$, with $a_j \leq b_j$ such that for all $j = 1, \ldots, d$, $z \mapsto \mathbb{E}\left[f\left(z, \mathbf{X}^{-j}\right)\mathbb{1}_{\{\mathbf{X}^{-j}\in A^{-j}\}}\right]$ is constant on $[a_j, b_j]$ and f is not constant on A.*

Remark 4.1. *The ♠-class contains many functions such as additive functions, multiplicative functions. A more elaborated list can be found in Elie-Dit-Cosaque and Maume-Deschamps (2022). A noteworthy example is the set of linear combination of Gaussian radial basis functions on* $[0,1]^d$*, with positive weights:*

$$\mathcal{G} = \left\{ \sum_{i=1}^{p} a_i \exp[\sum_{j=1}^{d} (x_j - \mu_j)^2 \sigma_j^2],\ a_i \geq 0,\ \sigma_j \geq 0,\ \mu_j \in \mathbb{R} \right\}.$$

It is known that the class $\mathcal{G}$ *is dense in the set of non-negative continuous functions on* $[0,1]^d$ *(see Park and Sandberg (1991) and also Klusowski (2019) where the class* $\mathcal{G}$ *is also considered to study CART).*

Let $\tau_1(\mathbf{x}) = \mathbb{E}[Y(1)|\mathbf{X} = \mathbf{x}]$ and $\tau_0(\mathbf{x}) = \mathbb{E}[Y(0)|\mathbf{X} = \mathbf{x}]$ and similarly $\hat{\tau}_1(\mathbf{x}) = \sum_{i:W_i=1} \alpha_i(\mathbf{x}) Y_i$ and $\hat{\tau}_0(\mathbf{x}) = \sum_{i:W_i=0} \alpha_i'(\mathbf{x}) Y_i$. We shall make the following assumptions.

Assumption 4.2.

- $Y = \tau(\mathbf{X}) g(\mathbf{W}) + \gamma(\mathbf{X}) + \varepsilon$.
- $\mathbf{X} = (X_1, \ldots, X_d)$ *is a continuous random vector with independent coordinates.*
- ε *and* $\mathbf{X}$ *are independent,* ε *is a continuous, centered random variable with increasing distribution function and light tails i.e. there exists* $0 < \theta < 1$ *such that for any* $D > 0$, $\mathbb{P}(|\varepsilon| > D) \leq C\theta^D$.
- $\mathbf{X}$ *takes its values in* $\mathcal{X}$ *which is assumed to be a compact hyper-rectangle of* $\mathbb{R}^d$*:* $\mathcal{X} = \prod_{i=1}^{d} [u_i, v_i]$, $-\infty < u_i \leq v_i < \infty$.
- $\mathbf{x} \mapsto \gamma(\mathbf{x})$, $\mathbf{x} \mapsto \tau_1(\mathbf{x})$ *and* $\mathbf{x} \mapsto \tau_0(\mathbf{x})$ *are continuous. So in particular* $\mathbf{x} \mapsto \tau(\mathbf{x})$ *is continuous.*

Remark 4.2. $g(W) = W$ *is the linear treatment effect considered in Athey et al. (2019). As noticed in Hill (2011) non linear functions have practical interest.*

We denote:

- $\mathrm{CV}(X) = \sigma_X / \mathbb{E}[X]$
- $f(n) = \Omega(g(n)) \iff \exists k > 0, \exists n_0 > 0 \mid \forall n \geqslant n_0 \quad |f(n)| \geqslant k \cdot |g(n)|$

Assumption 4.3. *The following assumptions are made on B (number of trees), $N_{n,1}(\mathbf{x}; \Theta, \mathcal{D}_n)$ resp. $N_{n,0}(\mathbf{x}; \Theta, \mathcal{D}_n)$ (number of observations in a leaf node such as $W = 1$, resp. $W = 0$):*

1. $B = \mathcal{O}(n^{\alpha})$, with $\alpha > 0$.
2. $\forall \mathbf{x} \in \mathcal{X}, \quad \mathbb{E}[N_{n,1}(\mathbf{x}; \Theta, \mathcal{D}_n)] = \Omega\left(\sqrt{n}(\ln(n))^{\beta}\right)$, with $\beta > 1$, and
   $\forall \mathbf{x} \in \mathcal{X}, \quad \mathrm{CV}(N_{n,1}(\mathbf{x}; \Theta, \mathcal{D}_n)) = \mathcal{O}\left(\dfrac{1}{n^{(\alpha+1)/2}(\ln(n))^{\gamma/2}}\right)$, with $\gamma > 1$.
3. $\forall \mathbf{x} \in \mathcal{X}, \quad \mathbb{E}[N_{n,0}(\mathbf{x}; \Theta, \mathcal{D}_n)] = \Omega\left(\sqrt{n}(\ln(n))^{\beta}\right)$, and
   $\forall \mathbf{x} \in \mathcal{X}, \quad \mathrm{CV}(N_{n,0}(\mathbf{x}; \Theta, \mathcal{D}_n)) = \mathcal{O}\left(\dfrac{1}{n^{(\alpha+1)/2}(\ln(n))^{\gamma/2}}\right)$.

Remark 4.3. *Items 2 and 3 in Assumption 4.3 are easier to verify than the Assumption 4.1, because the number of observations in leaves can be controlled as a standard construction parameters of trees of the forest.*

Theorem 4.1. *Let Assumptions 4.2 and 4.3 be verified, with τ_1 and τ_0 belonging to the ♠-class, assume that for fixed $\beta > \frac{5}{2}$, $C > 0$, each constructed tree is the highest such that $C\sqrt{n}(\ln n)^{\beta} \leq N_{n,0}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n), N_{n,1}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$. Also assume that $\mathbb{E}[\max \varepsilon_i^2] \leq K(\ln n)^u$ with $\beta - u > \frac{1}{2}$ and K is a positive constant. Then*

$$\forall \mathbf{x} \in \mathcal{X}, |\hat{\tau}_{B,n}(\mathbf{x}) - \tau(\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0.$$

Remark 4.4. *The property on $\mathbb{E}[\max \varepsilon_i^2]$ is verified for subgaussian distributions (Boucheron et al. (2013)).*

The proof follows the lines of Elie-Dit-Cosaque and Maume-Deschamps (2022) and the main steps are described in Appendix A.

4.3. Interpretability

Using Proposition A.4, we can state an almost surely version of Proposition 1 in Scornet et al. (2015) (which gives an interpretability result in probability). Indeed Proposition A.4 gives that for any $h \in \mathbb{N}$ and

any empirical tree $\mathcal{T}_e$ satisfying Assumption 4.2 and the upper bounds on $N_{n,0}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n), N_{n,1}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ in Theorem 4.1, there exists a theoretical tree $\mathcal{T}_h$ as close as wanted to $\mathcal{T}_e$ until height h.

We can define the informative variables as the set of variables upon which τ depends. Given that τ belongs to the ♠-class, we have that in the theoretical tree the splits are only made along informative variables, indeed the theoretical splitting criterion equals zero along non informative variables. Noting $inf \subseteq \{1, \ldots, d\}$ the set of indices of informative variables, we have $\tau(\mathbf{X}) \perp\!\!\!\perp \mathbf{X}^{-inf}$. Thus up to height h, the empirical cuts are performed along the same coordinates as the theoretical tree $\mathcal{T}_h$. The following result is then straightforward.

Theorem 4.2. *Assume that Assumption 4.2 is verified and set $\mathcal{M}_{try} = d$, let $h \in \mathbb{N}$. Assume that τ belongs to the ♠-class and that τ is non constant in every node up to height h. Then for n large enough, all the cuts in an empirical tree up to height h are made along informative variables almost surely.*

5. Simulations results

Firstly we consider a simulation where $\gamma(W) = W$, then we consider a non linear case. We also study the same kind of examples studied in Athey et al. (2019). In a fourth example a modified version of HTERF is considered when the term γ is linear, finally to assess interpretabilty we study an example close to the Ishigami function adapted to a causal perspective.

In practice, a preliminary step called centering is applied, we estimate the quantity $\mu_0(\mathbf{x}) = \mathbb{E}[Y(0)|\mathbf{X} = \mathbf{x}]$ with $\hat{\mu}_0$ on observations such as $W_i = 0$. Then we consider the quantity $Y_i^e = Y_i - \hat{\mu}_0(\mathbf{X}_i)$. In the `Julia` implementation of HTERF, we use a cross-validated regression random forest to get $\hat{\mu}_0$. Cross-validation is used to optimise hyperparameters such as minimum sample size in nodes and leaves and the value of $\mathcal{M}_{try}$. In the same fashion as any algorithm can be used in metalearners, we could use any algorithm for the centering. The building and the estimating steps of HTERF are done with the centered data Y_i^e. Below is an example to motivate the prior centering. Let $Y = \tau(\mathbf{X})W + \gamma(\mathbf{X})$ where $\tau(\mathbf{X}) = X^{(1)}$ and $\gamma(\mathbf{X}) = X^{(2)}$. Assume that γ is perfectly known when calculating Y^e. Consider the following data set in Table 5.

$X^{(1)}$	$X^{(2)}$	W	Y	Y^e
5	5	1	10	5
5	5	0	5	0
10	5	1	15	10
10	10	1	20	10
10	10	0	10	0

Table 1: Motivational example for centering

With no centering, the criterion along $X^{(1)}$ is:

$$\frac{2 \times 3}{5^2}\left((10-5)-\left(\frac{15+20}{2}-10\right)\right)^2 = 1.5. \tag{5.1}$$

The criterion along $X^{(2)}$ is:

$$\frac{2 \times 3}{5^2}\left(\left(\frac{10+15}{2}-5\right)-(20-10)\right)^2 = 1.5. \tag{5.2}$$

The criterion is equal for both covariates so none of them seems more informative, which is unfortunate since only $X^{(1)}$ is informative.
But if we consider the criterion with the centered outcome Y^e, we get the following criterion along $X^{(1)}$:

$$\frac{2 \times 3}{5^2}\left((5-0)-\left(\frac{10+10}{2}-0\right)\right)^2 = 6. \tag{5.3}$$

The criterion along $X^{(2)}$ is:

$$\frac{2 \times 3}{5^2}\left(\left(\frac{5+10}{2}-0\right)-(10-0)\right)^2 = 1.5. \tag{5.4}$$

With the centered outcome, the criterion is larger when splitting along $X^{(1)}$ as intended.

Remark 5.1. *If the observations are unbalanced regarding the treatment distribution with a lot less untreated cases, then the estimator could not be as good as expected. In practice when more than 55% of the observations are treated, we estimate the quantity* $\mu_1(\mathbf{x}) = \mathbb{E}[Y(1)|\mathbf{X} = \mathbf{x}]$ *with and estimator* $\hat{\mu}_1$ *trained on observations such as* $W_i = 1$*. Then we define the quantity* $Y_i^e = Y_i - \hat{\mu}_1(\mathbf{X}_i)$ *and proceed as previously. In this case the quantity* $-\tau$ *that is estimated instead of* τ*.*

5.1. A first example

We consider simulated data close to causal frameworks previously studied (Athey et al. (2019)). Let $\mathbf{X}_i \sim U([0,1]^p)$, $W_i \sim Bern(0.5)$ and $Y_i = \tau(\mathbf{X}_i)W_i + \beta\gamma(\mathbf{X}_i)$. Where $p = 10, \tau(\mathbf{x}) = \sin(x^{(1)})$ and $\gamma(\mathbf{x}) = \cos(2x^{(2)} + 3x^{(3)})$. The underlying model for Y follows the causal framework presented in Athey et al. (2019), and the unconfoundedness hypothesis is respected. The scalar β allows to consider the impact of the magnitude of τ relative to γ.

β	GRF	HTERF
5	0.276	0.117
1	0.122	0.012
0.2	0.079	0.004

Table 2: Mean squared errors of GRF and HTERF methods that estimate heterogeneous treatment effect. All causal forests have 500 trees, both the centering forests for GRF and the forest of the first step in HTERF have 500 trees, the same subsampling rate (0.7) is used in both methods. The results are aggregated over 60 simulation replications with 1000 test points each. The mean square errors are multiplied by 1000.

We notice the influence of the order magnitude of $\beta\gamma$ term relative to τ. The bigger β is, the larger the RMSE is. It is true for both GRF and HTERF. We also notice that for a small β, the gain of HTERF relatively to GRF is more significant in term of RMSE.

	GRF				HTERF			
β	dep.3	dep.5	dep.10	imp.	dep.3	dep.5	dep.10	imp.
5	0.870	0.378	0.150	0.852	1	0.498	0.175	0.985
1	0.874	0.526	0.174	0.866	1	0.995	0.282	1
0.2	0.875	0.627	0.2	0.866	1	1	0.603	1

Table 3: Frequencies of splitting on $X^{(1)}$ at depths 3, 5 and 10 and importance of $X^{(1)}$.

τ only depends on the variable $X^{(1)}$, so it is expected that for small depths, the splits should be done only on this variable. Also the importance of $X^{(1)}$ should be high. These expected results are clearer for HTERF than GRF.

These observations regarding the relative magnitude of τ highlights the importance of the quality of fit of the model in the first step of HTERF. To illustrate this we considered a similar simulation, where the γ term is much simpler to estimate in Section 5.4.

5.2. Non linear framework

The BART algorithm (Hill (2011)) allows to estimate CATE in a more general context where $Y = f(W, X) + \varepsilon$ with ε being normal iid. In Athey et al. (2019) the relationship between Y and W needs to be affine, which is not the case in HTERF. Indeed in GRF the algorithm hugely relies on the relation $\tau(\mathbf{X}) = \frac{Cov(Y,W|\mathbf{X})}{Var(W|\mathbf{X})}$ which is only true in the affine case.

Let $\mathbf{X} \sim U([0,1]^p)$, $W \sim Bern(0.5)$ and $Y = \sin(X^{(1)})(W+2)^3 + \cos(X^{(2)})$, where $p = 3$. Hence we have CATE that satisfies: $\tau(\mathbf{x}) = 19 \sin(x^{(1)})$.

Method	RMSE	importance
GRF	0.321	0.777
HTERF	0.209	1

Table 4: Root mean squared errors of GRF and HTERF methods that estimate heterogeneous treatment effect. All causal forests have 500 trees, both the centering forests for GRF and the forest of the first step in HTERF have 500 trees, the same subsampling rate (0.7) is used in both methods. The results are aggregated over 60 simulation replications with 1000 test points each. We also consider the importance of $X^{(1)}$.

Better performances are observed for HTERF in term of estimation (lower RMSE for HTERF) and interpretability (see Table 4). Indeed $X^{(1)}$ is the only informative variable, so its importance is expected to be 1.

5.3. GRF example

To illustrate the performance of HTERF, we reuse a simulation from Athey et al. (2019). Let $\mathbf{X}_i \sim U([0,1]^p)$, $W_i|\mathbf{X}_i \sim Bern(e(\mathbf{X}_i))$ and $Y_i|\mathbf{X}_i, W_i \sim N(m(\mathbf{X}_i) + (W_i - 0.5)\tau(\mathbf{X}_i), 1)$, where $p = 10$ or $p = 20$ depending on the simulation considered. The authors considered three settings:

- No confounding, $m(\mathbf{x}) = 0$ and $e(\mathbf{x}) = 0.5$ but treatment heterogeneity $\tau(\mathbf{x}) = \varsigma(x^{(1)})\varsigma(x^{(2)}), \varsigma(u) = 1 + 1/(1 + e^{-20(u-1/3)})$.
- Confounding, $e(\mathbf{x}) = \frac{1}{4}(1 + \beta_{2,4}(x^{(3)}))$ and $m(\mathbf{x}) = 2x^{(3)} - 1$, where $\beta_{a,b}$ is beta distribution with shape parameters a and b, but no treatment heterogeneity, $\tau(\mathbf{x}) = 0$.
- Both confounding $e(\mathbf{x}) = \frac{1}{4}(1 + \beta_{2,4}(x^{(3)}))$ and $m(\mathbf{x}) = 2x^{(3)} - 1$, and treatment heterogeneity, $\tau(\mathbf{x}) = \varsigma(x^{(1)})\varsigma(x^{(2)})$.

conf.	heterog.	p	n	GRF	HTERF
no	yes	10	800	1.01	0.84
no	yes	10	1600	0.58	0.50
no	yes	20	800	1.07	0.92
no	yes	20	1600	0.65	0.55
yes	no	10	800	0.14	0.15
yes	no	10	1600	0.09	0.09
yes	no	20	800	0.10	0.11
yes	no	20	1600	0.08	0.08
yes	yes	10	800	1.16	1.12
yes	yes	10	1600	0.69	0.63
yes	yes	20	800	1.29	1.23
yes	yes	20	1600	0.74	0.63

Table 5: Mean squared errors of GRF and HTERF methods that estimate heterogeneous treatment effect. All causal forests have 500 trees, both the centering forests for GRF and the forest of the first step in HTERF have 500 trees, the same subsampling rate (0.7) is used in both methods. The results are aggregated over 60 simulation replications with 1000 test points each. The mean square errors are multiplied by 10.

		GRF			HTERF		
p	n	$X^{(1)}$	$X^{(2)}$	$\sum$	$X^{(1)}$	$X^{(2)}$	$\sum$
10	800	0.416	0.410	0.826	0.446	0.416	0.862
10	1600	0.413	0.431	0.844	0.440	0.447	0.887
20	800	0.398	0.413	0.811	0.410	0.427	0.837
20	1600	0.416	0.420	0.836	0.434	0.433	0.867

Table 6: Importances of $X^{(1)}, X^{(2)}$ and their sum ($\sum$) for the previous setting with treatment heterogeneity and unconfoundedness.

Results in Table 5 show that under the three configurations HTERF has similar or better performances than GRF. When there is no confounding HTERF is more performant.

In term of interpretability in the first setting with treatment heterogeneity and no confounding, we expect high and similar importances for $X^{(1)}$ and $X^{(2)}$, indeed these are the only informative covariates and their contributions are quite symetrical. In Table 5.3 we can see good results for both methods with an improvement for HTERF method. In the third setting with confounding we expect the same kind of results and no significant importance for the

		GRF				HTERF			
p	n	$X^{(1)}$	$X^{(2)}$	$\sum$	$X^{(3)}$	$X^{(1)}$	$X^{(2)}$	$\sum$	$X^{(3)}$
10	800	0.401	0.411	0.812	0.023	0.375	0.486	0.861	0.018
10	1600	0.403	0.435	0.838	0.021	0.381	0.510	0.891	0.014
20	800	0.404	0.395	0.799	0.012	0.359	0.478	0.837	0.010
20	1600	0.385	0.443	0.828	0.010	0.334	0.537	0.871	0.007

Table 7: Importances of $X^{(1)}, X^{(2)}$, their sum ($\sum$) and $X^{(3)}$ for the previous setting with treatment heterogeneity and confounding.

confounding variable $X^{(3)}$. In Table 5.3, for both methods we have no significant importance for $X^{(3)}$ (for example for the first line of GRF, the remaining importances for non informative variables is 0.188 so we could expect an importance of 0.024 for each non informative variable which is very close to the observed importance of 0.023). We have an improvement in term of sum of importances on informative variables with HTERF.

5.4. Linear γ function

We perform a simulation study in order to show the importance of an accurate estimation of μ_0 in the pre-processing step.

Let $\mathbf{X}_i \sim U([0,1]^p)$, $W_i \sim Bern(0.5)$ and $Y_i = \tau(\mathbf{X}_i)W_i + \gamma(\mathbf{X}_i)$. Where $p = 10, \tau(\mathbf{x}) = \sin(x^{(1)})$ and $\gamma(\mathbf{x}) = 2x^{(2)} + 3x^{(3)}$. We consider a new estimator HTERF-OLS where μ_0 is a linear regression instead of a random forest. Since γ is a simple linear function, μ_0 will fit γ better and we can expect better results on CATE estimation.

Method	RMSE	depth 3	depth 5	depth 10	importance
GRF	11.56	0.875	0.514	0.171	0.867
HTERF	4,35	1	0.954	0.239	1
HTERF-OLS	1,50	1	1	0.944	1

Table 8: Root mean squared errors of GRF, HTERF and HTERF-OLS methods that estimate heterogeneous treatment effect. All causal forests have 500 trees, both the centering forests for GRF and the forest of the first step in HTERF have 500 trees, the same subsampling rate (0.7) is used in both methods. The results are aggregated over 60 simulation replications with 1000 test points each. We also consider the frequency of split on $X^{(1)}$ at depth 3, 5 and 10 and the importance of this variable.

HTERF-OLS has the best results in term of quality of fit and in term of interpretability especially at deeper splits such as depth 10. A low quality of

the μ_0 estimator is a flaw for the overall HTERF algorithm. We propose the use of cross validated random forests in general but with external knowledge about the nature of γ better choices can be made.

5.5. Ishigami-like model

A last example is proposed, based on Ishigami functions, often used in sensitivity analysis (Ishigami and Homma (1990)). Let $\mathbf{X}_i \sim U([-\pi, \pi]^3)$, $W_i \sim Bern(0.5)$ and $Y_i = \tau(\mathbf{X}_i)W_i + \gamma(\mathbf{X}_i)$. Where $\tau(\mathbf{x}) = 0.3(x^{(3)})^4 \sin(x^{(1)})$ and $\gamma(\mathbf{x}) = \sin(x^{(1)}) + 7\sin(x^{(2)})^2$.

Method	RMSE	importance $X^{(1)}$	importance $X^{(2)}$	importance $X^{(3)}$
GRF	0.982	0.655	0.075	0.269
HTERF	0.766	0.763	0	0.237

Table 9: Root mean squared errors of GRF and HTERF methods that estimate heterogeneous treatment effect. All causal forests have 500 trees, both the centering forests for GRF and the forest of the first step in HTERF have 500 trees, the same subsampling rate (0.7) is used in both methods. The results are aggregated over 60 simulation replications with 1000 test points each. We also consider the importances of the three variables $X^{(1)}, X^{(2)}$ and $X^{(3)}$.

Once again HTERF is a better estimator for CATE in term of RMSE. Also since $\mathbf{X}^{(2)}$ does not appear in the expression of τ we expect a null importance for this variable. In regard of this remark, HTERF has more consistent results in term of interpretability.

6. Discussion

In this paper we proposed a novel causal forest based algorithm, namely HTERF to estimate CATE with a binary treatment. We have shown empirically that HTERF is more efficient in term of quality of estimation of CATE and in term of interpretability compared to GRF. We also proved an almost surely consistency result on HTERF under realistic assumptions.
Additional work could be done on the choice of the μ_0 estimator used in the centering process. Indeed when there are clues on the nature of γ, a well chosen estimator can improve drastically the performances of HTERF.

A. Proof of consistency

We follow a similar approach than in Elie-Dit-Cosaque and Maume-Deschamps (2022).
In what follows, C denotes any positive constant, allowing to write: $C+C=C$ or $uC=C$ where $u>0$.

We consider an intermediate result before proving Theorem 4.1.

Assumption A.1. *For all $\ell \in [1,B]$, we assume that the variation of CATE function within any cell goes to 0:*

$$\forall \mathbf{x} \in \mathcal{X}, \quad \sup_{\mathbf{z} \in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)} |\tau(\mathbf{z}) - \tau(\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0 \ .$$

Assumption A.1 is verified if there is a bounded probability to split on each variable even non informative ones as in Athey. In what follows, we prove that Assumption A.1 is satisfied under hypothesis closer to the random forest practice.

Theorem A.1. *Let Y satisfy Assumption 4.2, with τ belonging to the ♠-class, let $\beta > \frac{5}{2}$, $C>0$, let the constructed trees be the highest such that $C\sqrt{n}(\ln n)^\beta \leq N_{n,0}(\Theta_\ell,\mathcal{D}_n), N_{n,1}(\Theta_\ell,\mathcal{D}_n)$, then Assumption A.1 is verified.*

Lemma A.2. *Assume that Assumption 4.2 is satisfied with the function τ in the ♠-class, let $S^\infty = (s_j, j=1,\ldots)$ with $s_j \in \{L,R\}$, it describes an infinite path in a binary tree, let $S^h = (s_j, j=1,\ldots,h)$, it describes a path in a binary tree of height h. Let $A_h(S^h,\Theta)$ be the corresponding leaf in a theoretical tree. Then the variation of $\tau(\cdot)$ on $A_h(S^h,\Theta)$ goes to 0 a.s. as h goes to infinity.*

Proof.

The proof is quite similar to the proof of Lemma 5.3 in Elie-Dit-Cosaque and Maume-Deschamps (2022). One have to notice that:

$$\begin{aligned}
&\Delta^\star(A,j,z) = 0 \Leftrightarrow \\
&\mathbb{E}\left[Y(1)-Y(0)|\, X_{j^\infty} < z, \mathbf{X}^J \in A^J\right] - \mathbb{E}\left[Y(1)-Y(0)|\, X_{j^\infty} \geqslant z, \mathbf{X}^J \in A^J\right] = 0 \\
&\Leftrightarrow \mathbb{P}(\mathbf{X} \in A_\infty)\mathbb{E}\left[(Y(1)-Y(0))\mathbb{1}_{\{X_i \leq z,\ \mathbf{X}\in A_\infty\}}\right] \\
&= \mathbb{P}(X_i \leq z, \mathbf{X} \in A_\infty)\mathbb{E}\left[(Y(1)-Y(0))\mathbb{1}_{\{\mathbf{X}\in A_\infty\}}\right].
\end{aligned}$$

By derivating with respect to z, we may see that it is equivalent to $z \mapsto \mathbb{E}\left[\tau\left(z, \mathbf{X}^{-i}\right) \mathbb{1}_{\left\{\mathbf{X}^{-i} \in A_{\infty}^{-i}\right\}}\right]$ is constant for all $i = 1, \ldots, d$. Since we assumed that τ belongs to the ♠-class, either τ is constant on $A_{\infty}(S^{\infty})$ or the diameter of $A_{\infty}(S^{\infty})$ is zero. In both cases, we conclude that the variation of $\tau(\cdot)$ on $A_h(S^h, \Theta)$ goes to 0 as h goes to infinity, as in Elie-Dit-Cosaque and Maume-Deschamps (2022). □

Proposition A.3. *Let Assumption 4.2 be satisfied. Let $\beta > \frac{5}{2}$, let A be a rectangle in $\mathcal{X}$, we shall say that $(A, j, z) \in \mathcal{A}^n$ if $|A_{L0}|$,$|A_{L1}|$,$|A_{R0}|$ and $|A_{R1}|$ are greater than $C\sqrt{n}(\ln n)^{\beta}$ (so this bound is also true for $|A_L|$ and $|A_R|$). We have*

$$\sup_{(A,j,z)\in\mathcal{A}^n} |\Delta^{\star}(A, j, z) - \Delta(A, j, z)| \xrightarrow{a.s.} 0 \text{ as } n \to \infty.$$

Proof.

The proof follows closely Proposition 5.3. in Elie-Dit-Cosaque and Maume-Deschamps (2022). It makes use of the decomposition:

$$\begin{aligned}
&|\Delta^{\star}(A, j, z) - \Delta(A, j, z)| = |T_1 + T_2| \\
&=: \frac{|A_L||A_R|}{|A|^2}\big[\big(\overline{Y}_{A_{L1}} - \overline{Y}_{A_{L0}} - \overline{Y}_{A_{R1}} + \overline{Y}_{A_{R0}}\big)^2 \\
&- \big(\mathbb{E}[Y(1) - Y(0)|\mathbf{X}^{(j)} < z, \mathbf{X} \in A] - \mathbb{E}[Y(1) - Y(0)|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A]\big)^2\big] \\
&+ \left(\mathbb{E}[Y(1) - Y(0)|\mathbf{X}^{(j)} < z, \mathbf{X} \in A] - \mathbb{E}[Y(1) - Y(0)|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A]\right)^2 \\
&\Big(\frac{|A_L||A_R|}{|A|^2} - \mathbb{P}(\mathbf{X}^{(j)} < z|\mathbf{X} \in A)\mathbb{P}(\mathbf{X}^{(j)} \geq z|\mathbf{X} \in A)\Big).
\end{aligned}$$

In order to prove the proposition, we shall prove that $\sup_{A,j,z} T_1$ and $\sup_{A,j,z} T_2$ go to 0 a.s. The two main ingredient of the proof are Vapnik-Chervonenkis theory on rectangles in $\mathcal{A}$ which gives:

$$\mathbb{P}\left(\sup_{B\in\mathcal{A}}\left|\frac{|B|}{n} - \mathbb{P}(\mathbf{X} \in B)\right| > \kappa\right) \leq 8(n+1)^{2d}e^{-n\kappa^2/32}, \tag{A.1}$$

and Theorem 9.6 in Györfi et al. (2002) and Lemma A.2 in Elie-Dit-Cosaque

and Maume-Deschamps (2022), that lead for any $L > 0$ and $\frac{1}{p} + \frac{1}{q} = 1$:

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}_{\mathbf{X}_i \in A_L, W=1} - \mathbb{E}[Y \mathbb{1}_{\mathbf{X} \in A_{L1}}]\right| > \kappa\right) \\
\leq & 24\left(\frac{32eL}{\kappa}\log\left(\frac{48eL}{\kappa}\right)\right)^{2d} \exp\left(\frac{-n\kappa^2}{512L^2}\right) + C\frac{\mathbb{E}[Y^p]^{\frac{1}{p}}\mathbb{P}(Y > L)^{\frac{1}{q}}}{\kappa}.
\end{aligned}$$

□

Proposition A.4. *Let Assumption 4.2 be satisfied. Assume that for $\beta > \frac{5}{2}$, $N_{n,0}(\Theta_\ell, \mathcal{D}_n), N_{n,1}(\Theta_\ell, \mathcal{D}_n) \geq C\sqrt{n}(\ln n)^\beta$. For $h \in \mathbb{N}$, let $S \in \{L, R\}^h$ describe a path of length h in a binary tree, let $A^n(S)$ and $A(S)$ be corresponding nodes in empirical and theoretical trees. Denote*

$$A(S) = \prod_{j=1}^{d}[a_j, b_j] \text{ and } A^n(S) = \prod_{j=1}^{d}[a_j^n, b_j^n].$$

Denote $\mathcal{T}_h$ the set of theoretical trees of height h, then

$$\inf_{\mathcal{T}_h} \max_{j=1,\dots,d} \max\left(|a_j - a_j^n|, |b_j - b_j^n|\right) \longrightarrow 0 \text{ a.s. as } n \to \infty. \tag{A.2}$$

Proof.

The proof is the same that in Proposition 5.4 from Elie-Dit-Cosaque and Maume-Deschamps (2022). □

Proof of Theorem A.1.

It can be proved as Theorem 5.1 in Elie-Dit-Cosaque and Maume-Deschamps (2022). □

Following the lines of the proof of Theorem A.1, the same result for τ_1 and τ_0 can be obtained.

Assumption A.2. *For all $\ell \in [1, B]$, we assume that the variation of τ_1 and τ_0 within any cell goes to 0:*

$$\forall \mathbf{x} \in \mathcal{X}, \quad \sup_{\mathbf{z} \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)} |\tau_1(\mathbf{z}) - \tau_1(\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0 \ .$$

$$\forall \mathbf{x} \in \mathcal{X}, \quad \sup_{\mathbf{z} \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)} |\tau_0(\mathbf{z}) - \tau_0(\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0 \ .$$

Theorem A.5. *Let Y satisfy Assumption 4.2, with τ_0 and τ_1 belonging to the ♠-class, let $\beta > \frac{5}{2}$, $C > 0$, let the constructed trees be the highest such that $C\sqrt{n}(\ln n)^{\beta} \leq N_{n,0}\left(\Theta_\ell, \mathcal{D}_n\right), N_{n,1}\left(\Theta_\ell, \mathcal{D}_n\right)$, then A.2 is verified.*

Theorem 4.1 is a direct consequence of Theorem A.6 that we now state.

Theorem A.6. *Consider a random forest which satisfies Assumptions A.2, 4.3 and hypotheses of Theorem 4.1. Then,*

$$\forall \mathbf{x} \in \mathcal{X}, |\hat{\tau}_{B,n}(\mathbf{x}) - \tau(\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0 \ .$$

Proof of Theorem A.6.

The proof follows ideas in Elie-Dit-Cosaque and Maume-Deschamps (2022) but some differences arise due to the honest subsampling rather than bootstrap. We give it in detail for completeness.

The main ingredient of the proof is to use a second sample $\mathcal{D}_n^\diamond$ in order to deal with the data-dependent aspect. Thus, we first define a dummy estimator based on two samples $\mathcal{D}_n$ and $\mathcal{D}_n^\diamond$ which will be used below. The trees are grown using $\mathcal{D}_n$, but we consider another sample $\mathcal{D}_n^\diamond$ (independent of $\mathcal{D}_n$ and Θ) which is used to define a dummy estimator

$$\begin{aligned}
&\tau_{B,n}^\diamond\left(y|\, \mathbf{X}=\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathcal{D}_n^\diamond, \mathcal{D}_n\right) \\
=&\sum_{j=1}^{n} \alpha_{n,j}^\diamond\left(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}, W^{\diamond 1}, \ldots, W^{\diamond n}, \mathcal{D}_n\right) Y^{e\diamond j} \\
&-\sum_{j=1}^{n} \alpha_{n,j}^{'\diamond}\left(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}, W^{\diamond 1}, \ldots, W^{\diamond n}, \mathcal{D}_n\right) Y^{e\diamond j},
\end{aligned}$$

where the weights are

$$\begin{aligned}
&\alpha_{n,j}^\diamond\left(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}, W^{\diamond 1}, \ldots, W^{\diamond n}, \mathcal{D}_n\right) \\
=&\frac{1}{B}\sum_{\ell=1}^{B} \frac{\mathbb{1}_{\{\mathbf{X}^{\diamond j} \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)\} \cap W^{\diamond j}=1}}{N_{n,1}^\diamond\left(\mathbf{x}; \Theta_\ell, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}, W^{\diamond 1}, \ldots, W^{\diamond n}, \mathcal{D}_n\right)}, \ j = 1, \ldots, n.
\end{aligned}$$

with $N_{n,1}^\diamond\left(\mathbf{x}; \Theta_\ell, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}, W^{\diamond 1}, \ldots, W^{\diamond n}, \mathcal{D}_n\right)$, the number of elements of $\mathcal{D}_n^\diamond$ that fall into $A_n\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right)$ such as $W^\diamond = 1$. Throughout this section, we shall use the convention $\frac{0}{0} = 0$ in case $N_{n,1}^\diamond\left(\mathbf{x}; \Theta_\ell, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}, W^{\diamond 1}, \ldots, W^{\diamond n}, \mathcal{D}_n\right) =$

0 and thus $\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}\cap W^{\diamond j}=1}=0$ for $j=1,\dots n$.
Similarly we have:

$$\begin{aligned}&\alpha_{n,j}^{'\diamond}\left(\mathbf{x};\Theta_1,\dots,\Theta_k,\mathbf{X}^{\diamond 1},\dots,\mathbf{X}^{\diamond n},W^{\diamond 1},\dots,W^{\diamond n},\mathcal{D}_n\right)\\ =&\frac{1}{B}\sum_{\ell=1}^{B}\frac{\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}\cap W^{\diamond j}=0}}{N_{n,0}^{\diamond}\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\dots,\mathbf{X}^{\diamond n},W^{\diamond 1},\dots,W^{\diamond n},\mathcal{D}_n\right)},\ j=1,\dots,n.\end{aligned}$$

To lighten the notation in the sequel, we will simply write $\tau_{B,n}^{\diamond}\left(y|\,\mathbf{X}=\mathbf{x}\right)=\sum_{j=1}^{n}\alpha_j^{\diamond}\left(\mathbf{x}\right)Y^{\diamond j}-\sum_{j=1}^{n}\alpha_j^{'\diamond}\left(\mathbf{x}\right)Y^{\diamond j}$.

Let $\mathbf{x}\in\mathcal{X}$ and $y\in\mathbb{R}$, we have:

$$\begin{aligned}\left|\hat{\tau}\left(\mathbf{x}\right)-\tau\left(\mathbf{x}\right)\right|\leq&\left|\hat{\tau}\left(\mathbf{x}\right)-\tau^{\diamond}\left(\mathbf{x}\right)\right|\\ &+\left|\tau^{\diamond}\left(\mathbf{x}\right)-\tau\left(\mathbf{x}\right)\right|.\end{aligned}$$

Let $\mathbf{x}$ in $\mathcal{X}$: $|\tau^{\diamond}(\mathbf{x})-\tau(\mathbf{x})|\leq|\tau_1^{\diamond}(\mathbf{x})-\tau_1(\mathbf{x})|+|\tau_0^{\diamond}(\mathbf{x})-\tau_0(\mathbf{x})|$ Each of the two terms will be treated the same way.

$$\begin{aligned}\left|\tau_1^{\diamond}(\mathbf{x})-\tau_1(\mathbf{x})\right|\leq&\left|\sum_{\substack{i=1\\W_i^{\diamond}=1}}^{n}\alpha_i^{\diamond}(\mathbf{x})\left[(Y_i^{\diamond})-\mathbb{E}\left[Y(1)|\mathbf{X}_i^{\diamond}\right]\right]\right|\\ &+\left|\sum_{\substack{i=1\\W_i^{\diamond}=1}}^{n}\alpha_i^{\diamond}(\mathbf{x})\left[\mathbb{E}\left[Y(1)|\mathbf{X}_i^{\diamond}\right]-\mathbb{E}\left[Y(1)\mathbf{X}=\mathbf{x}\right]\right]\right|\\ \leq&U_n+V_n.\end{aligned}$$

The last term tends to 0 with Theorem A.5:

$$
\begin{aligned}
V_n &= \left| \sum_{\substack{i=1 \\ W_i^\diamond = 1}}^{n} \alpha_i(\mathbf{x}) \left[\mathbb{E}\left[Y(1)|\mathbf{X}_i^\diamond\right] - \mathbb{E}\left[Y(1)|\mathbf{X}=\mathbf{x}\right] \right] \right| \\
&= \left| \sum_{\substack{i=1 \\ W_i^\diamond = 1 \\ \exists l | \mathbf{X}_i^\diamond \in A_n(\mathbf{x}, \Theta_l)}}^{n} \alpha_i^\diamond(\mathbf{x}) \left[\mathbb{E}\left[Y(1)|\mathbf{X}_i^\diamond\right] - \mathbb{E}\left[Y(1)|\mathbf{X}=\mathbf{x}\right] \right] \right| \\
&\leq \sum_{\substack{i=1 \\ W_i^\diamond = 1 \\ \exists l | \mathbf{X}_i^\diamond \in A_n(\mathbf{x}, \Theta_l)}}^{n} \left| \alpha_i(\mathbf{x}) \left[\mathbb{E}\left[Y(1)|\mathbf{X}_i^\diamond\right] - \mathbb{E}\left[Y(1)|\mathbf{X}=\mathbf{x}\right] \right] \right| \\
&\leq \sup_{\mathbf{z} \in A_n(\mathbf{x})} \left| \tau_1(\mathbf{z}) - \tau_1(\mathbf{x}) \right| \xrightarrow[n \to +\infty]{} 0.
\end{aligned}
$$

For the first term we have:

$$
U_n = \left| \sum_{\substack{i=1 \\ W_i^\diamond = 1}}^{n} \alpha_i^\diamond \varepsilon_i^\diamond \right|.
$$

The following Lemma is useful:

Lemma A.7. *Let $u \in \{0,1\}$, as before, $N_{n,u}\left(A_n\left(\Theta\right)\right) = N_{n,u}\left(\mathbf{x}; \Theta, \mathcal{D}_n\right)$ is the number of observations of $\mathcal{D}_n$ such as $W = u$ that fall into in $A_n\left(\Theta\right) = A_n\left(\mathbf{x}; \Theta, \mathcal{D}_n\right)$ and $N_{n,u}^\diamond\left(A_n\left(\Theta\right)\right) = N_{n,u}^\diamond\left(\mathbf{x}; \Theta, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}, \mathcal{D}_n\right)$, the number of observations of $\mathcal{D}_n^\diamond$ such as $W = u$ that fall into in $A_n\left(\Theta\right)$. Then,*

$$
\forall \varepsilon > 0, \quad \mathbb{P}\left(\left| N_{n,u}^b\left(A_n\left(\Theta\right)\right) - N_{n,u}^\diamond\left(A_n\left(\Theta\right)\right) \right| > \varepsilon \right) \leqslant 16(n+1)^{2d} e^{-\varepsilon^2/128n} \ .
$$

Proof.

The proof is similar to Lemma 6.3 in Elie-Dit-Cosaque and Maume-Deschamps (2022) without the bootstrap considerations. □

So:

$$\begin{aligned}
\mathbb{E}\left[(U_n)^2\right] &= \mathbb{E}\left[\left(\sum_{j=1}^{n} \alpha_j^{\diamond}\varepsilon_j^{\diamond}\right)^2\right] \\
&= \sum_{j=1}^{n}\sum_{m=1}^{n} \mathbb{E}\left[\alpha_j^{\diamond}\alpha_m^{\diamond}\varepsilon_j^{\diamond}\varepsilon_m^{\diamond}\right] \\
&= \sum_{j=1}^{n} \mathbb{E}\left[\alpha_j^{\diamond 2}\varepsilon_j^{\diamond 2}\right] + \sum_{\substack{1\leqslant j,m\leqslant n \\ j\neq m}} \mathbb{E}\left[\alpha_j^{\diamond}\alpha_m^{\diamond}\varepsilon_j^{\diamond}\varepsilon_m^{\diamond}\right] \\
&\stackrel{\text{def}}{=} I_n + J_n.
\end{aligned}$$

For I_n:

$$\begin{aligned}
I_n =& \mathbb{E}\left[\sum_{j=1}^{n} \alpha_j^{\diamond 2}\varepsilon_j^{\diamond 2}\right] \\
\leq& \mathbb{E}\left[\max \alpha_j^{\diamond} \sum_{j=1}^{n} \alpha_j^{\diamond}\varepsilon_j^{\diamond 2}\right] \\
\leq& \mathbb{E}\left[\max \alpha_j^{\diamond} \max \varepsilon_j^{\diamond 2}\right] \\
\leq& \mathbb{E}\left[\max \alpha_j^{\diamond}\right] \mathbb{E}\left[\max \varepsilon_j^{\diamond 2}\right].
\end{aligned}$$

Let $\lambda = \frac{\mathbb{E}[N_{n,1}(A_n(\theta))]}{4}$:

$$\begin{aligned}
\mathbb{E}\left[\max \alpha_j^{\diamond}\right] &= \mathbb{E}\left[\max \frac{1}{B}\sum_{l=1}^{B} \frac{\mathbb{1}_{\mathbf{X}_i^{\diamond}\in A_n(\mathbf{x};\Theta_l,\mathcal{D}_n)\cap W_i^{\diamond}=1}}{N_{n,1}^{\diamond}(\mathbf{x};\Theta_\ell,\mathcal{D}_n)}\right] \\
&\leq \mathbb{E}\left[\frac{1}{\min N_{n,1}^{\diamond}(\mathbf{x};\Theta_\ell,\mathcal{D}_n)}\right] \\
&\leq \mathbb{E}\left[\frac{\mathbb{1}_{\{\forall \ell, N_{n,1}^{\diamond}(\mathbf{x};\Theta_\ell,\mathcal{D}_n)>\lambda\}}}{\min N_{n,1}^{\diamond}(\mathbf{x};\Theta_\ell,\mathcal{D}_n)}\right] + \mathbb{E}\left[\frac{\mathbb{1}_{\{\exists \ell, N_{n,1}^{\diamond}(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\leq\lambda\}}}{\min N_{n,1}^{\diamond}(\mathbf{x};\Theta_\ell,\mathcal{D}_n)}\right] \\
&\leq \frac{1}{\lambda} + \mathbb{P}(\exists \ell | N_{n,1}^{\diamond}(\Theta_\ell) \leq \lambda) \\
&\leq \frac{1}{\lambda} + B\mathbb{P}(N_{n,1}^{\diamond}(\Theta) \leq \lambda) \\
&\leq \frac{1}{\lambda} + B\mathbb{P}(N_{n,1}^{\diamond}(\Theta) \leq \lambda, \left|N_{n,1}(A_n(\theta)) - N_{n,1}^{\diamond}(A_n(\theta))\right| \leq \lambda) \\
&+ B\mathbb{P}(N_{n,1}^{\diamond}(\Theta) \leq \lambda, \left|N_{n,1}(A_n(\theta)) - N_{n,1}^{\diamond}(A_n(\theta))\right| > \lambda) \\
&\leq \frac{1}{\lambda} + B\mathbb{P}(N_{n,1}(\Theta) \leq 2\lambda) + B\mathbb{P}(\left|N_{n,1}(A_n(\theta)) - N_{n,1}^{\diamond}(A_n(\theta))\right| > \lambda).
\end{aligned}$$

Using Bienaimé-Tchebychev's inequality, Assumption 4.3 and Lemma A.7, there exists C, K and M positive constants such that:

$$\begin{aligned}
\mathbb{E}\left[\max \alpha_j^{\diamond}\right] \leq& \frac{4}{\mathbb{E}\left[N_{n,1}(A_n(\theta))\right]} + 4B\left(\mathrm{CV}\left(N_{n,1}(A_n(\theta))\right)\right)^2 + B\mathbb{P}(\left|N_{n,1}(A_n(\theta)) - N_{n,1}^{\diamond}(A_n(\theta))\right| > \lambda) \\
\leq& \frac{4}{K\sqrt{n}(\ln n)^{\beta}} + \frac{4CM^2}{n(\ln n)^{\gamma}} + 16Cn^{\alpha}(n+1)^{2d}e^{-K^2(\ln n)^{2\beta}/2048}.
\end{aligned}$$

Finally:

$$I_n \leq \frac{4}{K\sqrt{n}(\ln n)^{\beta-u}} + \frac{4CM^2}{n(\ln n)^{\gamma-u}} + 16Cn^{\alpha}(n+1)^{2d}(\ln n)^{u}e^{-K^2(\ln n)^{2\beta}/2048}. \tag{A.3}$$

For J_n:

$$\begin{aligned}
J_n &= \sum_{\substack{1\leqslant j,m\leqslant n\\ j\neq m}} \mathbb{E}\left[\alpha_j^\diamond \alpha_m^\diamond \varepsilon_j^\diamond \varepsilon_m^\diamond\right] \\
&= \sum_{\substack{1\leqslant j,m\leqslant n\\ j\neq m}} \mathbb{E}\left[\mathbb{E}\left[\alpha_j^\diamond \alpha_m^\diamond \varepsilon_j^\diamond \varepsilon_m^\diamond | \mathcal{D}_n, \mathbf{X}_1^\diamond, \ldots, \mathbf{X}_n^\diamond, \varepsilon_j^\diamond\right]\right] \\
&= \sum_{\substack{1\leqslant j,m\leqslant n\\ j\neq m}} \mathbb{E}\left[\alpha_j^\diamond \alpha_m^\diamond \mathbb{E}\left[\varepsilon_j^\diamond \varepsilon_m^\diamond | \mathcal{D}_n, \mathbf{X}_1^\diamond, \ldots, \mathbf{X}_n^\diamond, \varepsilon_j^\diamond\right]\right] \\
&\quad \text{because } \alpha_j^\diamond, \alpha_m^\diamond \text{ are } \mathcal{D}_n, \mathbf{X}_1^\diamond, \ldots, \mathbf{X}_n^\diamond \text{ measurable.} \\
&= \sum_{\substack{1\leqslant j,m\leqslant n\\ j\neq m}} \mathbb{E}\left[\alpha_j^\diamond \alpha_m^\diamond \varepsilon_j^\diamond \mathbb{E}\left[\varepsilon_m^\diamond | \mathcal{D}_n, \mathbf{X}_1^\diamond, \ldots, \mathbf{X}_n^\diamond, \varepsilon_j^\diamond\right]\right] \\
&= \sum_{\substack{1\leqslant j,m\leqslant n\\ j\neq m}} \mathbb{E}\left[\alpha_j^\diamond \alpha_m^\diamond \varepsilon_j^\diamond \mathbb{E}\left[\varepsilon_m^\diamond\right]\right] \\
&= 0.
\end{aligned}$$

Using Bienaimé-Tchebychev's inequality, we have:

$$\forall \varepsilon > 0, \mathbb{P}\left(|U_n| \geq \varepsilon\right) \leq \frac{I_n}{\varepsilon^2}. \tag{A.4}$$

Since $\sum_{n\geq 1} I_{n^2} < \infty$, with Borel-Cantelli lemma:

$$\forall \varepsilon > 0, \mathbb{P}\left(\overline{\lim_{n\to+\infty}} \{|U_{n^2}| \geq \varepsilon\}\right) = 0. \tag{A.5}$$

So $U_{n^2} \to 0$.

We now show that $(U_n)_{n\geq 1}$ converges almost surely to 0.

$$\begin{aligned}
\mathbb{P}\left(\overline{\lim_{n\to+\infty}}\left\{\left|U_n - U_{\lfloor\sqrt{n}\rfloor^2}\right| \geq \varepsilon\right\}\right) &= \mathbb{P}\left(\overline{\lim_{n\to+\infty}}\left\{\sum_{i=\lfloor\sqrt{n}\rfloor^2+1}^{n} |\alpha_i^\diamond \varepsilon_i^\diamond| \geq \varepsilon\right\}\right) \\
&= \mathbb{P}\left(\forall n, \exists N_0 > n, \left\{\sum_{i=\lfloor\sqrt{N_0}\rfloor^2+1}^{N_0} |\alpha_i^\diamond \varepsilon_i^\diamond| \geq \varepsilon\right\}\right) \\
&= \lim_{n\to+\infty} \mathbb{P}\left(\exists N_0 > n, \left\{\sum_{i=\lfloor\sqrt{N_0}\rfloor^2+1}^{N_0} |\alpha_i^\diamond \varepsilon_i^\diamond| \geq \varepsilon\right\}\right) \\
&= \lim_{n\to+\infty} \sum_{N_0>n} \mathbb{P}\left(\sum_{i=\lfloor\sqrt{N_0}\rfloor^2+1}^{N_0} |\alpha_i^\diamond \varepsilon_i^\diamond| \geq \varepsilon\right).
\end{aligned}$$

For a given N_0, let $D(N_0) = (\ln N_0)^\gamma$:

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=\lfloor\sqrt{N_0}\rfloor^2+1}^{N_0} |\alpha_i^\diamond \varepsilon_i^\diamond| \geq \varepsilon\right) &\leq \mathbb{P}\left(\exists i \in [\![\lfloor\sqrt{N_0}\rfloor^2 + 1, N_0]\!], |\varepsilon_i^\diamond| > D(N_0)\right) \\
&\quad + \mathbb{P}\left(D(N_0) \sum_{i=\lfloor\sqrt{N_0}\rfloor^2+1}^{N_0} |\alpha_i^\diamond| \geq \varepsilon\right) \\
&\leq \mathbb{P}\left(\exists i \in [\![\lfloor\sqrt{N_0}\rfloor^2 + 1, N_0]\!], |\varepsilon_i^\diamond| > D(N_0)\right) \\
&\quad + \mathbb{P}\left(D(N_0)\frac{2\sqrt{N_0}}{\min N_{n,1}^\diamond(\mathbf{x}; \Theta_l, \mathcal{D}_n)} \geq \varepsilon\right).
\end{aligned}$$

Let's treat the second term:

$$\begin{aligned}
\mathbb{P}\left(D(N_0)\frac{2\sqrt{N_0}}{\min N_{n,1}^\diamond(\mathbf{x}; \Theta_l, \mathcal{D}_n)} \geq \varepsilon\right) &\leq \mathbb{P}\left(D(N_0)\frac{2\sqrt{N_0}}{\min N_{n,1}^\diamond(\mathbf{x}; \Theta_l, \mathcal{D}_n)} \geq \varepsilon, \forall \ell | N^\diamond(A_n(\Theta_\ell)) > \lambda\right) \\
&\quad + \mathbb{P}\left(D(N_0)\frac{2\sqrt{N_0}}{\min N_{n,1}^\diamond(\mathbf{x}; \Theta_l, \mathcal{D}_n)} \geq \varepsilon, \exists \ell | N^\diamond(A_n(\Theta_\ell)) \leq \lambda\right).
\end{aligned}$$

The first term is zero since $\mathbb{E}\left[N_{n,1}\left(A_n(\theta)\right)\right] \geq \frac{8\sqrt{N_0}D(N_0)}{\epsilon}$ for N_0 large enough according to Assumption 4.3.

$$\begin{aligned}
\mathbb{P}\left(D(N_0)\frac{2\sqrt{N_0}}{\min N^{\diamond}_{n,1}(\mathbf{x};\Theta_l,\mathcal{D}_n)} \geq \varepsilon\right) &\leq \mathbb{P}\left(\exists \ell | N^{\diamond}(A_n(\Theta_\ell)) \leq \lambda\right) \\
&\leq B\mathbb{P}(N_{n,1}\left(\Theta\right) \leq 2\lambda) \\
&+ B\mathbb{P}(\left|N_{n,1}\left(A_n(\theta)\right) - N^{\diamond}_{n,1}\left(A_n(\theta)\right)\right| > \lambda) \\
&\leq 4B\left(\mathrm{CV}\left(N_{n,1}\left(A_n(\theta)\right)\right)\right)^2 \\
&+ B\mathbb{P}(\left|N_{n,1}\left(A_n(\theta)\right) - N^{\diamond}_{n,1}\left(A_n(\theta)\right)\right| > \lambda) \\
&\leq \frac{4CM^2}{n(\ln n)^{\gamma}} + 16Cn^{\alpha}(n+1)^{2d}e^{-K^2(\ln n)^{2\beta}/2048}.
\end{aligned}$$

Finally we have:

$$\mathbb{P}\left(\sum_{i=\lfloor\sqrt{N_0}\rfloor^2+1}^{N_0} |\alpha_i^{\diamond}\varepsilon_i^{\diamond}| \geq \varepsilon\right) \leq C(N_0 - \lfloor\sqrt{N_0}\rfloor^2)\theta^{D(N_0)} + \frac{4CM^2}{n(\ln n)^{\gamma}} + 16Cn^{\alpha}(n+1)^{2d}e^{-K^2(\ln n)^{2\beta}/2048}.$$

Using Borel-Cantelli lemma:

$$\forall \varepsilon > 0, \mathbb{P}\left(\overline{\lim_{n\to+\infty}}\left\{\left|U_n - U_{\lfloor\sqrt{n}\rfloor^2}\right| \geq \varepsilon\right\}\right) = 0. \tag{A.6}$$

Finally we have that $(U_n)_{n\geq 1}$ goes to 0 almost surely. Finally it gives that $|\tau^{\diamond}(\mathbf{x}) - \tau(\mathbf{x})|$ goes to 0.

The quantity $|\hat{\tau}(\mathbf{x}) - \tau^{\diamond}(\mathbf{x})|$ is now treated. We use the same decomposition and consider separately but in similar fashion $|\hat{\tau}_1(\mathbf{x}) - \tau_1^{\diamond}(\mathbf{x})|$ and $|\hat{\tau}_0(\mathbf{x}) - \tau_0^{\diamond}(\mathbf{x})|$:

$$|\hat{\tau}_1(\mathbf{x}) - \tau_1^{\diamond}(\mathbf{x})| = \left|\frac{1}{B}\sum_{l=1}^{B}\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j\in A_n(l)}\mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j - \frac{\mathbb{1}_{\mathbf{X}_j^{\diamond}\in A_n(l)}\mathbb{1}_{W_j^{\diamond}=1}}{N^{\diamond}_{n,1}(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j^{\diamond}\right|.$$

We decompose the following way:

$$\begin{aligned}
&\left|\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j\in A_n(l)}\mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j-\frac{\mathbb{1}_{\mathbf{X}_j^\diamond\in A_n(l)}\mathbb{1}_{W_j^\diamond=1}}{N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j^\diamond\right|\\
\leq&\left|\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j\in A_n(l)}\mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j-\mathbb{E}\left[Y|\mathbf{X}\in A_n(\Theta),W=1\right]\right|\\
&+\left|\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j^\diamond\in A_n(l)}\mathbb{1}_{W_j^\diamond=1}}{N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j^\diamond-\mathbb{E}\left[Y|\mathbf{X}\in A_n(\Theta),W=1\right]\right|.
\end{aligned}$$

Let Λ be the statement "$\left|\frac{|B|}{n}-\mathbb{P}(\mathbf{X}\in B)\right|\leq\frac{C}{2}\frac{(\ln n)^\beta}{\sqrt{n}}$ verified for $B=\{\mathbf{X}_i\in A_n(\Theta)|W_i=1\}$", for any $\varepsilon>0$ we have:

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j\in A_n(l)}\mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j-\frac{\mathbb{1}_{\mathbf{X}_j^\diamond\in A_n(l)}\mathbb{1}_{W_j^\diamond=1}}{N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j^\diamond\right|>\varepsilon\right)\\
\leq&\mathbb{P}\left(\left|\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j\in A_n(l)}\mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j-\mathbb{E}\left[Y|\mathbf{X}\in A_n(\Theta),W=1\right]\right|>\frac{\varepsilon}{2},\Lambda\right)\\
&+\mathbb{P}\left(\left|\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j^\diamond\in A_n(l)}\mathbb{1}_{W_j^\diamond=1}}{N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j^\diamond-\mathbb{E}\left[Y|\mathbf{X}\in A_n(\Theta),W=1\right]\right|>\frac{\varepsilon}{2},\Lambda\right)\\
&+8(n+1)^{2d}e^{\frac{-C(\ln n)^{2\beta}}{128}}\left(\text{ id est }\mathbb{P}\left(\Lambda^C\right)\right).
\end{aligned}$$

Noticing that:

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j\in A_n(l)}\mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j-\mathbb{E}\left[Y|\mathbf{X}\in A_n(\Theta),W=1\right]\right|>\frac{\varepsilon}{2},\Lambda\right)\\
\leq&\mathbb{P}\left(\frac{n}{N_{n,1}(\mathbf{x};\Theta_l,\mathcal{D}_n)}\left|\frac{1}{n}\sum_{j=1}^{n}Y_j\mathbb{1}_{\mathbf{X}_j\in A_n(l)}\mathbb{1}_{W_j=1}-\mathbb{E}[Y\mathbb{1}_{\mathbf{X}\in A_n(\Theta)}\mathbb{1}_{W=1}]\right|>\frac{\varepsilon}{4},\Lambda\right)\\
&+\mathbb{P}\left(\left|\mathbb{E}[Y\mathbb{1}_{\mathbf{X}\in A_n(\Theta)}\mathbb{1}_{W=1}]\right|\left|\frac{n}{N_{n,1}(\mathbf{x};\Theta_l,\mathcal{D}_n)}-\frac{1}{\mathbb{P}(\mathbf{X}\in A')}\right|>\frac{\varepsilon}{4},\Lambda\right).
\end{aligned}$$

We can treat this term the same way as $T_{1,1,1}$.

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j\in A_n(l)}\mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j-\mathbb{E}\left[Y|\mathbf{X}\in A_n(\Theta),W=1\right]\right|>\frac{\varepsilon}{2},\Lambda\right)\\
\leq&24\left(\frac{Ce(\ln n)^{\delta}\sqrt{n}}{\varepsilon(\ln n)^{\beta}}\log\left(\frac{Ce(\ln n)^{\delta}\sqrt{n}}{\varepsilon(\ln n)^{\beta}}\right)\right)^{2d}\exp\left(\frac{-C\varepsilon^2(\ln n)^{2\beta}}{(\ln n)^{2\delta}}\right)+C\frac{\sqrt{n}e^{(\ln n)^{\delta}\frac{\ln\theta}{q}}}{\varepsilon(\ln n)^{\beta}}\\
&+8(n+1)^{2d}e^{-\frac{\varepsilon^2C(\ln n)^{2\beta-2\delta}}{2048}}.
\end{aligned}$$

Second term is treated with the same idea but needs a bit more work:

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j^{\diamond}\in A_n(l)}\mathbb{1}_{W_j^{\diamond}=1}}{N_{n,1}^{\diamond}(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j^{\diamond}-\mathbb{E}\left[Y|\mathbf{X}\in A_n(\Theta),W=1\right]\right|>\frac{\varepsilon}{2},\Lambda\right)\\
\leq&\mathbb{P}\left(\left|\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j^{\diamond}\in A_n(l)}\mathbb{1}_{W_j^{\diamond}=1}}{N_{n,1}^{\diamond}(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j^{\diamond}-\mathbb{E}\left[Y|\mathbf{X}\in A_n(\Theta),W=1\right]\right|>\frac{\varepsilon}{2},N_{n,1}^{\diamond}(\mathbf{x};\Theta_l,\mathcal{D}_n)>\lambda,\Lambda\right)\\
&+\mathbb{P}\left(N_{n,1}^{\diamond}(\mathbf{x};\Theta_l,\mathcal{D}_n)\leq\lambda\right).
\end{aligned}$$

The term:

$$\mathbb{P}\left(\left|\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j^{\diamond}\in A_n(l)}\mathbb{1}_{W_j^{\diamond}=1}}{N_{n,1}^{\diamond}(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j-\mathbb{E}\left[Y|\mathbf{X}\in A_n(\Theta),W=1\right]\right|>\frac{\varepsilon}{2},N_{n,1}^{\diamond}(\mathbf{x};\Theta_l,\mathcal{D}_n)>\lambda,\Lambda\right) \tag{A.7}$$

is treated as previously. The last term is close to an expression already bounded:

$$\begin{aligned}
&\mathbb{P}\left(N_{n,1}^{\diamond}(\mathbf{x};\Theta_l,\mathcal{D}_n)\leq\lambda\right)\\
\leq&\mathbb{P}(N_{n,1}\left(\Theta\right)\leq 2\lambda)\\
&+\mathbb{P}(\left|N_{n,1}\left(A_n(\theta)\right)-N_{n,1}^{\diamond}\left(A_n(\theta)\right)\right|>\lambda)\\
\leq&4\left(\mathrm{CV}\left(N_{n,1}\left(A_n(\theta)\right)\right)\right)^2\\
&+\mathbb{P}(\left|N_{n,1}\left(A_n(\theta)\right)-N_{n,1}^{\diamond}\left(A_n(\theta)\right)\right|>\lambda)\\
\leq&\frac{4CM^2}{n^{1+\alpha}(\ln n)^{\gamma}}+16C(n+1)^{2d}e^{-K^2(\ln n)^{2\beta}/2048}.
\end{aligned}$$

Thanks to Borel-Cantelli, provided that $2\beta-2\delta>1$, we conclude that $|\hat{\tau}_1(\mathbf{x})-\tau_1^{\diamond}(\mathbf{x})|$ goes to 0.

Finally we have $|\hat{\tau}(\mathbf{x})-\tau(\mathbf{x})|$ goes to 0.

□

References

Athey, S., Imbens, G., 2016. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences 113, 7353–7360.

Athey, S., Tibshirani, J., Wager, S., 2019. Generalized random forests. The Annals of Statistics 47, 1148–1178.

Boucheron, S., Lugosi, G., Massart, P., 2013. Concentration inequalities: A nonasymptotic theory of independence. Oxford university press.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Chipman, H.A., George, E.I., McCulloch, R.E., 2010. Bart: Bayesian additive regression trees. The Annals of Applied Statistics 4, 266–298.

Elie-Dit-Cosaque, K., Maume-Deschamps, V., 2022. Random forest estimation of conditional distribution functions and conditional quantiles. Electronic Journal of Statistics 16, 6553–6583.

Green, D.P., Kern, H.L., 2012. Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. Public opinion quarterly 76, 491–511.

Györfi, L., Kohler, M., Krzyzak, A., Walk, H., et al., 2002. A distribution-free theory of nonparametric regression. volume 1. Springer.

Hill, J., Su, Y.S., 2013. Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. The Annals of Applied Statistics , 1386–1420.

Hill, J.L., 2011. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics 20, 217–240.

Imbens, G.W., Rubin, D.B., 2015. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

Ishigami, T., Homma, T., 1990. An importance quantification technique in uncertainty analysis for computer models, in: [1990] Proceedings. First International Symposium on Uncertainty Modeling and Analysis, IEEE. pp. 398–403.

Klusowski, J.M., 2019. Analyzing cart. arXiv preprint arXiv:1906.10086 .

Künzel, S.R., Sekhon, J.S., Bickel, P.J., Yu, B., 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the national academy of sciences 116, 4156–4165.

Park, J., Sandberg, I.W., 1991. Universal approximation using radial-basis-function networks. Neural computation 3, 246–257.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Scornet, E., Biau, G., Vert, J.P., 2015. Consistency of random forests. The Annals of Statistics 43, 1716–1741.

Su, X., Tsai, C.L., Wang, H., Nickerson, D.M., Li, B., 2009. Subgroup analysis via recursive partitioning. Journal of Machine Learning Research 10.

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