Variable importance for causal forests: breaking down the heterogeneity of treatment effects

Clément Bénard[1] and Julie Josse[2]

[1]Safran Tech, Digital Sciences & Technologies, 78114 Magny-Les-Hameaux, France
[2]PreMeDICaL, Inria-Inserm, Montpellier, France

February 27, 2023

Abstract

Causal learning provides powerful algorithms to estimate heterogeneous treatment effects, such as causal random forests or double robust methods. However, these procedures are black boxes, and do not identify the factors involved in the heterogeneity of treatment effects. This limitation has strong practical consequences. In healthcare for example, a treatment may not provide any benefit to a large portion of patients presenting a given disease. Therefore, the identification of factors responsible for treatment heterogeneity is a promising route for better treatment targeting, with a potential high impact on patient care. In this article, we develop a new importance variable algorithm based on causal forests, and inspired from the drop and relearn principle, widely used for regression tasks. In particular, we show that causal forests trained without a given confounding variable can still lead to consistent estimates, thus providing theoretical guaranties of the proposed algorithm. Additionally, experiments show the good performance of our importance measure, which outperforms competitors on several test cases.

1 Introduction

1.1 Context and Objectives

Treatment effects have recently raised a strong interest in the machine learning community, especially for medical applications (Obermeyer and Emanuel, 2016). This field provides a clear example of the high potential impact of variable importance measures for heterogeneous treatment effects. In healthcare, it is critical to assess that a given treatment provides significant benefits before its wide distribution. In particular, Randomized Controlled Trials (RCTs) are extensively used for this purpose (Imbens and Rubin, 2015; Deaton and Cartwright, 2018). The main principle of RCT is to randomize the treatment assignment in a given group of individuals to annihilate confounding effects, and thus estimate average treatment effects (ATEs) with quite straightforward statistical procedures. However, RCTs suffer from several limitations. For example, the population involved in the trial may be different from the ultimate target

population (Rothwell, 2005), and the sample size is often quite small. Therefore, there is also a strong interest to use observational data to overcome these limitations to quantify treatment effects. Nevertheless, the mathematical analysis becomes more complicated in this case, because of confounding factors. Indeed, without the proper randomization of RCTs, both the measured response and the treatment assignment may depend on the same variables. Consequently, the treated and non-treated populations are intrinsically different, and their different behaviors are entangled with the treatment effect. The core of the problem is thus to estimate the causal treatment effect, which is really the quantity of interest to assess the treatment benefits. The causal inference theory ensures that such quantity is identifiable, provided that all confounding factors are observed. Using this causal setting, several procedures have been proposed to estimate ATEs, such as matching, inverse propensity weighting (Hirano and Imbens, 2001; Austin and Stuart, 2015, IPW), or augmented IPW (Robins et al., 1994; Imbens and Rubin, 2015, AIPW). This article focuses on the critical property of heterogeneity of treatment effects. To continue with the healthcare example, it is well known that a large portion of patients with a specific disease may not benefit from a given treatment, which has shown a significant impact in average. Therefore, it is of high interest to estimate the heterogeneity of treatment effects with respect to the patient characteristics, in order to better target the relevant patients. Although efficient algorithms have been recently developed to estimate heterogeneous treatment effects, such as double robust estimates (Kennedy, 2020; Nie and Wager, 2021), causal forests (Wager and Athey, 2018; Athey et al., 2019), the lasso (Kosuke and Marc, 2013), BART (Künzel et al., 2019), or neural networks (Shalit et al., 2017), little effort has been made to quantify the variables involved in the heterogeneity. We can mention the importance measure of the causal forest package `grf`, the double robust approach of Hines et al. (2022), and the algorithm from Boileau et al. (2022) for high dimensional cases. Besides, let us also mention policy learning, which aims at selecting relevant patients to treat (Athey and Wager, 2021). However, these procedures are often black boxes. This property strongly limits their practical use by medical doctors, since they require interpretability of the policy for both efficiency and ethical reasons.

The main purpose of this article is to introduce a variable importance measure, improving over the existing algorithms, to better identify the sources of heterogeneity in treatment effects. We take advantage of causal forests (Athey et al., 2019) to build such a variable importance algorithm, as explained in Section 2. Then, theoretical properties are proved in Section 3, and finally, the experiments of Section 4 show the good practical performances of the proposed method. We first formalize the problem in the following subsection.

1.2 Definitions

To deepen the discussion, we need to formalize heterogeneous treatment effects. We first introduce a standard causal setting with an input vector $\mathbf{X} = (X^{(1)}, \dots, X^{(p)}) \in \mathbb{R}^p$ with $p \in \mathbb{N}^\star$, the binary treatment assignment $W \in \{0, 1\}$, the potential outcome $Y(1) \in \mathbb{R}$ for the subject receiving the treatment, and the potential outcome without treatment $Y(0) \in \mathbb{R}$. We denote by $\mathbf{X}^{(\mathcal{C})}$ the subvector with only the components in $\mathcal{C} \subset \{1, \dots, p\}$, and $\mathbf{X}^{(-j)}$ the vector $\mathbf{X}$ with the j-th component removed. The observed outcome is given by $Y = WY(1) + (1 - W)Y(0)$, which is known as the SUTVA assumption in the literature. More precisely, the potential

outcomes are defined by

$$Y(0) = \mu(\mathbf{X}) + \varepsilon(0),$$
$$Y(1) = \mu(\mathbf{X}) + \tau(\mathbf{X}^{(\mathcal{C})}) + \varepsilon(1),$$

where $\mu(\mathbf{X})$ is a noise parameter, $\tau(\mathbf{X}^{(\mathcal{C})})$ is the conditional average treatment effect (CATE) only depending on variables in $\mathcal{C} \subset \{1, \dots, p\}$, and $\varepsilon(0), \varepsilon(1)$ are some noise variables satisfying $\mathbb{E}[\varepsilon(0) \mid \mathbf{X}] = \mathbb{E}[\varepsilon(1) \mid \mathbf{X}] = 0$. Notice that the CATE is also defined as the mean difference between potential outcomes, conditional on $\mathbf{X}$, i.e., $\mathbb{E}[Y(1) - Y(0) \mid \mathbf{X}] = \tau(\mathbf{X}^{(\mathcal{C})})$, by construction. Overall, the observed outcome Y also writes

$$Y = \mu(\mathbf{X}) + \tau(\mathbf{X}^{(\mathcal{C})}) \times W + \varepsilon(W).$$

The cornerstone of causal treatment effect estimates is the assumption of unconfoundedness given below, which states that all confounding variables are observed in the data. By definition, the responses $Y(0)$, $Y(1)$, and the treatment assignment W simultaneously depend on the confounding variables. If all confounding variables are observed, then the responses and the treatment assignment are independent conditional on the inputs. Consequently, the treatment effect is identifiable, as stated in the following proposition. Notice that all proofs of propositions and theorems stated throughout the article are gathered in Appendix A.

Assumption 1. *Potential outcomes are independent of the treatment assignment conditional on the observed input variables, i.e.,* $Y(0), Y(1) \perp\!\!\!\perp W \mid \boldsymbol{X}$.

Proposition 1. *If the unconfoundedness Assumption 1 is satisfied, then we have*

$$\tau(\boldsymbol{X}^{(\mathcal{C})}) = \mathbb{E}[Y \mid \boldsymbol{X}, W = 1] - \mathbb{E}[Y \mid \boldsymbol{X}, W = 0].$$

Importantly, notice that we define above the treatment effect as the expected difference between between potential outcomes, conditional on the input variables. However, the heterogeneity properties strongly depends on the treatment effect definition (Rothman, 2012; VanderWeele and Robins, 2007). Indeed, the ratio between potential outcome means may also define a treatment effect, which can be heterogeneous in this case, but constant with our original definition of the outcome difference. A thorough discussion of this topic is out of scope of this article, and we take the difference of potential outcomes as treatment effect, the widely used metric for healthcare applications (VanderWeele and Robins, 2007). Next, we mathematically specify that the treatment effect τ is heterogeneous with respect to all variables in $\mathcal{C}$. Intuitively, it means that τ varies with respect to each of these variables, which is formalized in the following assumption.

Assumption 2. *The treatment effect* $\tau(\boldsymbol{X}^{(\mathcal{C})})$ *is heterogeneous with respect to all variables in* $\mathcal{C}$*, i.e., for any variable* $j \in \mathcal{C}$, $\mathbb{V}[\tau(\boldsymbol{X}^{(\mathcal{C})}) | \boldsymbol{X}^{(-j)}] > 0$ *with a strictly positive probability.*

Finally, as already mentioned, our objective is to quantify the influence of the input variables $\mathbf{X}$ on the treatment heterogeneity τ, using an available sample $\mathscr{D}_n = \{(\mathbf{X}_i, Y_i, W_i)\}_{i=1}^n$, made of $n \in \mathbb{N}^\star$ independent and identically distributed (iid) observations.

2 Variable Importance for Heterogeneous Treatment Effects

2.1 Variable Importance Definition

To propose a variable importance measure, we build on Sobol (1993) and Williamson et al. (2021), which define variable importance in the case of regression as the proportion of output explained variance lost when a given input variable is removed. Hines et al. (2022) extend this idea to treatment effects, and introduce the theoretical importance measure $\mathrm{I}^{(j)}$ of $X^{(j)}$, defined by

$$\mathrm{I}^{(j)} = \frac{\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{C})})] - \mathbb{V}[\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{C})})|\mathbf{X}^{(-j)}]]}{\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{C})})]},$$

which is well-defined under Assumption 2, since $\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{C})})] > 0$. In fact, this importance measure gives the proportion of treatment effect variance lost when a given input variable is removed. Additionally, the following proposition shows that $\mathrm{I}^{(j)}$ properly identifies variables in $\mathcal{C}$, which have an impact on treatment heterogeneity, where the proof in Appendix A is a direct consequence of Assumption 2.

Proposition 2. *Let Assumption 2 be satisfied. If $j \notin \mathcal{C}$, then we have $\mathrm{I}^{(j)} = 0$. Otherwise, if $j \in \mathcal{C}$, we have $0 < \mathrm{I}^{(j)} \leq 1$.*

By definition of $\mathrm{I}^{(j)}$, a variable strongly correlated to the other inputs, has a low importance value. Indeed, because of this strong dependence, little information is lost about the treatment effect heterogeneity by removing such a variable. As suggested by Williamson et al. (2021) and Hines et al. (2022), a possible approach is to extend the importance measure to a group of variables, where strongly dependent variables are grouped together, or using expert knowledge. In the sequel, we focus on the case of a single variable for the sake of clarity, but the extension to groups of variables is straightforward. More importantly, Hines et al. (2022) highlight that a key problem to estimate the above quantity $\mathrm{I}^{(j)}$, is that the unconfoundedness Assumption 1 does not imply unconfoundedness for the reduce set of input variables $\mathbf{X}^{(-j)}$, i.e., we may have $Y(0), Y(1) \not\perp\!\!\!\perp W \mid \mathbf{X}^{(-j)}$. Hines et al. (2022) overcome this issue using double robust approaches (Kennedy, 2020; Nie and Wager, 2021) to estimate τ with all input variables in a first step, and then regress the obtained treatment effect on $\mathbf{X}^{(-j)}$ to estimate $\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{C})})|\mathbf{X}^{(-j)}]$. Actually, the generalized random forest framework from Athey et al. (2019) enables to get closer to the original proposal of Williamson et al. (2021) by retraining the causal forest without variable $X^{(j)}$ and still get consistent estimates of $\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{C})})|\mathbf{X}^{(-j)}]$. Therefore, we focus on causal forests (Wager and Athey, 2018; Athey et al., 2019), one of the state-of-the-art algorithm to estimate heterogeneous treatment effects, to propose efficient estimates of $\mathrm{I}^{(j)}$.

2.2 Algorithm

Causal forests. Generalized random forests (Athey et al., 2019) are a generic framework to build efficient estimates of quantities defined as solutions of local moment equations. As opposed to original Breiman's forests, generalized forests are not the average of tree outputs.

Instead, trees are aggregated to generate weights for each observation of the training data, used in a second step to build a weighted estimate of the target quantity. Causal forests are a specific case of generalized forest, where the following local moment equation identifies the treatment effect under the unconfoundedness Assumption 1,

$$\tau(\mathbf{X}^{(\mathcal{C})}) \times \mathbb{V}[W \mid \mathbf{X}] - \mathrm{Cov}[W, Y \mid \mathbf{X}] = 0. \tag{1}$$

The local moment equation (1) is thus used to define the causal forest estimate $\tau_{M,n}(\mathbf{x})$ at a new query point $\mathbf{x}$, built from the data $\mathscr{D}_n$, with $M \in \mathbb{N}^\star$ trees randomized by $\mathbf{\Theta}_M$, and formally defined in Athey et al. (2019, Section 6.1) by

$$\tau_{M,n}(\mathbf{x}) = \frac{\sum_{i=1}^n \alpha_i(\mathbf{x}) W_i Y_i - \overline{W}_\alpha \overline{Y}_\alpha}{\sum_{i=1}^n \alpha_i(\mathbf{x})(W_i - \overline{W}_\alpha)^2}, \tag{2}$$

where $\overline{Y}_\alpha = \sum_{i=1}^n \alpha_i(\mathbf{x}) Y_i$, $\overline{W}_\alpha = \sum_{i=1}^n \alpha_i(\mathbf{x}) W_i$, and the weights $\alpha_i(\mathbf{x})$ are generated by the forest to quantify the frequency of $\mathbf{x}$ and the training observation $\mathbf{X}_i$ both falling in the same terminal leaves of trees. Besides, notice that the local moment equation (1) is also used to define an efficient splitting criterion of the tree nodes.

Local centering. The causal forest algorithm first performs a local centering step, by regressing Y and W on $\mathbf{X}$ using regression forests, fit with $\mathscr{D}_n$. The obtained out-of-bag forest estimates of $m(\mathbf{X}_i) = \mathbb{E}[Y_i \mid \mathbf{X}_i]$ and $\pi(\mathbf{X}_i) = \mathbb{E}[W_i \mid \mathbf{X}_i]$ are denoted by $\hat{m}_n(\mathbf{X}_i)$ and $\hat{\pi}_n(\mathbf{X}_i)$. Then, these quantities are subtracted to get the centered outcome $\tilde{Y}_i = Y_i - \hat{m}_n(\mathbf{X}_i)$, and centered treatment $\tilde{W}_i = W_i - \hat{\pi}_n(\mathbf{X}_i)$, used to fit the causal forest $\tau_{M,n}(\mathbf{x})$. Next, to estimate $\mathrm{I}^{(j)}$, the causal forest is retrained dropping the j-th variable, and thus using the observations $\{(\mathbf{X}_i^{(-j)}, \tilde{Y}_i, \tilde{W}_i)\}_{i=1}^n$ to generate new weights $\alpha'(\mathbf{x}^{(-j)})$ and build $\tau_{M,n}^{(-j)}(\mathbf{x})$ through equation (2). As we deepen below, a critical feature of this procedure is that all input variables are used in the local centering of Y_i and W_i before the j-th variable is dropped to build $\tau_{M,n}^{(-j)}(\mathbf{x})$.

The above local moment equation (1) identifies the treatment effect, provided that the unconfoundedness Assumption 1 is satisfied. Unfortunately, when $X^{(j)}$ is removed from the input variables, this moment equation does not hold anymore, since unconfoundedness may be violated with a reduced set of inputs. However, an important feature of causal forests is the preliminary step of local centering of the observed outcome and treatment assignment, mentioned above. The following proposition shows that the treatment effect is well identified by the local moment equation of causal forests including only variables in $\mathcal{C}$, provided that the data is centered with all inputs. We recall that $m(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$ and $\pi(\mathbf{X}) = \mathbb{E}[W \mid \mathbf{X}]$.

Proposition 3. *If Assumption 1 is satisfied, we have*

$$\tau(\boldsymbol{X}^{(\mathcal{C})}) \times \mathbb{V}[W - \pi(\boldsymbol{X}) \mid \boldsymbol{X}^{(\mathcal{C})}] - \mathrm{Cov}[W - \pi(\boldsymbol{X}), Y - m(\boldsymbol{X}) \mid \boldsymbol{X}^{(\mathcal{C})}] = 0,$$

which is the local moment equation defining causal forests, with input variables $\boldsymbol{X}^{(\mathcal{C})}$, centered outcome $Y - m(\boldsymbol{X})$, and centered treatment assignment $W - \pi(\boldsymbol{X})$.

On the other hand, removing an influential variable $j \in \mathcal{C}$ to learn a causal forest is more delicate. Indeed, a local moment equation to identify the mean CATE over $X^{(j)}$ exists if the treatment effect is uncorrelated to the squared centered treatment assignment.

Proposition 4. *If Assumption 1 is satisfied, then we have for $j \in \mathcal{C}$*

$$\mathbb{E}[\tau(\boldsymbol{X}^{(\mathcal{C})}) \mid \boldsymbol{X}^{(-j)}] \times \mathbb{V}[W - \pi(\boldsymbol{X}) \mid \boldsymbol{X}^{(-j)}] - \mathrm{Cov}[W - \pi(\boldsymbol{X}), Y - m(\boldsymbol{X}) \mid \boldsymbol{X}^{(-j)}] + \mathrm{Cov}[\tau(\boldsymbol{X}^{(\mathcal{C})}), (W - \pi(\boldsymbol{X}))^2 \mid \boldsymbol{X}^{(-j)}] = 0.$$

Then, for a query point $\boldsymbol{x}^{(-j)} \in [0,1]^{p-1}$, if $\mathrm{Cov}[\tau(\boldsymbol{X}^{(\mathcal{C})}), (W - \pi(\boldsymbol{X}))^2 \mid \boldsymbol{X}^{(-j)} = \boldsymbol{x}^{(-j)}] = 0$, $\mathbb{E}[\tau(\boldsymbol{X}^{(\mathcal{C})}) \mid \boldsymbol{X}^{(-j)} = \boldsymbol{x}^{(-j)}]$ is identified by the original local moment equation of causal forests, with $\boldsymbol{X}^{(-j)}$ as input variables, centered outcome $Y - m(\boldsymbol{X})$, and centered treatment assignment $W - \pi(\boldsymbol{X})$.

Corrected causal forests. In the general case, the local moment equation of causal forests does not identify the treatment effect with a confounding variable removed. However, the additional covariance term involved in the moment equation of Proposition 4 takes small values in practice, as we will see in the experimental Section 4. Therefore, we introduce the corrected causal forest estimate when a confounding variable $X^{(j)}$ with $j \in \mathcal{C}$ is removed. Recall that the weights $\alpha'(\mathbf{x}^{(-j)})$ are generated by the causal forest using centered data and dropping variable $X^{(j)}$, to define $\tau^{(-j)}_{M,n}(\mathbf{x})$. We define the corrected causal forest estimate $\theta^{(-j)}_{M,n}(\mathbf{x})$ as

$$\theta^{(-j)}_{M,n}(\mathbf{x}) = \tau^{(-j)}_{M,n}(\mathbf{x}) - \frac{\sum_{i=1}^n \alpha'_i(\mathbf{x}^{(-j)}) \tilde{W}_i^2 \tau_{M,n}(\mathbf{X}_i) - \overline{W^2_{\alpha'}}\overline{\tau}_{\alpha'}}{\overline{W^2_{\alpha'}} - (\overline{W}_{\alpha'})^2}, \tag{3}$$

where $\overline{W^2_{\alpha'}} = \sum_{i=1}^n \alpha'_i(\mathbf{x}^{(-j)}) \tilde{W}_i^2$, $\overline{W}_{\alpha'} = \sum_{i=1}^n \alpha'_i(\mathbf{x}^{(-j)}) \tilde{W}_i$, and the mean treatment effect is $\overline{\tau}_{\alpha'} = \sum_{i=1}^n \alpha'_i(\mathbf{x}^{(-j)}) \tau_{M,n}(\mathbf{X}_i)$. With such correction, the causal forest retrained without a confounding variable is consistent, as we will show in the next section.

Variable importance estimate. Now, we can state our variable importance estimate, based on causal forests. Using $\mathscr{D}'_n = \{(\mathbf{X}'_i, Y'_i, W'_i)\}_{i=1}^n$ an independent copy of $\mathscr{D}_n$, we define

$$\mathrm{I}^{(j)}_n = \frac{\sum_{i=1}^n \left[\tau_{M,n}(\mathbf{X}'_i) - \theta^{(-j)}_{M,n}(\mathbf{X}'_i)\right]^2}{\sum_{i=1}^n \left[\tau_{M,n}(\mathbf{X}'_i) - \overline{\tau_{M,n}}\right]^2},$$

where $\overline{\tau_{M,n}} = \sum_{i=1}^n \tau_{M,n}(\mathbf{X}'_i)/n$. Notice that the above definition is formalized with $\mathscr{D}'_n$ for the sake of clarity, but that such additional data is usually not available in practice. Instead, out-of-bag causal forest estimates are rather used to define $\mathrm{I}^{(j)}_n$.

3 Theoretical Properties

Propositions 3 and 4 are the cornerstones of the consistency of our variable importance algorithm. This result relies on the asymptotic analysis of Athey et al. (2019), which states the consistency of causal forests in Theorem 1. Several mild assumptions are required, mainly about the input distribution, the regularity of the involved functions, and the forest growing. Then, the core of our mathematical analysis is the extension to the case of a causal forest fit without a given input variable. When the removed input is a confounding variable, consistency

is obtained thanks to the corrective term introduced in equation (3) of the previous section. Then, the convergence of our variable importance algorithm follows using a standard asymptotic analysis. We first formalize the required assumptions and specifications on the tree growing from Athey et al. (2019), that are frequently used in the theoretical analysis of random forests (Meinshausen, 2006; Scornet et al., 2015; Wager and Athey, 2018).

Assumption 3. *The input $\boldsymbol{X}$ takes value in $[0,1]^p$, and admits a density bounded from above and below by strictly positive constants.*

Assumption 4. *The functions π, m, and τ are Lipschitz, $0 < \pi(\boldsymbol{x}) < 1$ for $\boldsymbol{x} \in [0,1]^p$, and μ and τ are bounded.*

Specification 1. *Tree splits are constrained to put at least a fraction $\gamma > 0$ of the parent node observations in each child node. The probability to split on each input variable at every tree node is greater than $\delta > 0$. The forest is honest, and built via subsampling with subsample size a_n, satisfying $a_n/n \to 0$ and $a_n \to \infty$.*

The first part of Specification 1 is originally introduced by Meinshausen (2006). The idea is to enforce the diameter of each cell of the trees to vanish as the sample size increases, by adding a constraint on the minimum size of children nodes, and slightly increasing the randomization of the variable selection for the split at each node. Then, vanishing cell diameters combined to Lipschitz functions lead to the forest convergence. Additionally, honesty is a key property of the tree growing, extensively discussed in Wager and Athey (2018), where half of the data is used to optimize the splits, and the other half to estimate the cell outputs. With these assumptions satisfied, we state below the causal forest consistency proved in Athey et al. (2019). Notice that the original proof is conducted for generalized forests, for any local moment equations satisfying regularity assumptions, automatically fulfilled for the moment equation (1) involved in our analysis. In Appendix A, we give a specific proof of Theorem 1 in the case of causal forests. We will built on this proof to further extend the consistency result when a confounding variable is removed.

Theorem 1 (Theorem 3 from Athey et al. (2019))**.** *If Assumptions 1-4 and Specification 1 are satisfied, and the causal forest $\tau_{M,n}(\boldsymbol{x})$ is built with $\mathscr{D}_n$ without local centering, then we have for $\boldsymbol{x} \in [0,1]^p$,*

$$\tau_{M,n}(\boldsymbol{x}) \xrightarrow{p} \tau(\boldsymbol{x}^{(\mathcal{C})}).$$

Next, we need a slight simplification of our variable importance algorithm to alleviate the mathematical analysis. We assume that a centered dataset $\mathscr{D}_n^\star = \{(\mathbf{X}_i, W_i^\star, Y_i^\star)\}$ is directly available, where $W_i^\star = W_i - \pi(\mathbf{X}_i)$ and $Y_i^\star = Y_i - m(\mathbf{X}_i)$. A causal forest grown with this dataset where a given input variable $j \in \{1, \dots, p\} \setminus \mathcal{C}$ is dropped, consistently estimates the treatment effect as stated below. Consistency also holds for variables $j \in \mathcal{C}$ in specific cases, whereas in the general case, the corrected term introduced in equation (3) is required. Notice that we may extend the following theorems without such centered dataset (not available in practice), but when m and π are estimated by $\mathbb{L}^2$-consistent learning algorithms, and a local centering step is performed prior to the causal forest fit. However, it is out of scope of this article. Theorem 2 states the consistency of causal forests when an input variable is removed.

Theorem 2. *If Assumptions 1-4 and Specification 1 are satisfied, and the causal forest* $\tau_{M,n}^{(-j)}(\boldsymbol{x})$ *is fit with the centered data* $\mathscr{D}_n^{\star(-j)}$ *without the* j*-th variable,*

(i) for $j \in \{1, \dots, p\} \setminus \mathcal{C}$ *and* $\boldsymbol{x} \in [0,1]^p$*, we have*

$$\tau_{M,n}^{(-j)}(\boldsymbol{x}) \xrightarrow{p} \tau(\boldsymbol{x}^{(\mathcal{C})}),$$

(ii) for $j \in \mathcal{C}$ *and* $\boldsymbol{x} \in [0,1]^p$*, if* $\mathrm{Cov}[\tau(\boldsymbol{X}^{(\mathcal{C})}), (W - \pi(\boldsymbol{X}))^2 \mid \boldsymbol{X}^{(-j)} = \boldsymbol{x}^{(-j)}] = 0$*, we have*

$$\tau_{M,n}^{(-j)}(\boldsymbol{x}) \xrightarrow{p} \mathbb{E}[\tau(\boldsymbol{X}^{(\mathcal{C})}) \mid \boldsymbol{X}^{(-j)} = \boldsymbol{x}^{(-j)}].$$

Theorem 2 is a direct consequence of Propositions 3 and 4 combined with Theorem 1. Indeed, provided that the outcome and treatment assignment are centered, if the removed variable j is not involved in the treatment heterogeneity, i.e. $j \notin \mathcal{C}$, consistency holds. On the other hand, if $j \in \mathcal{C}$, we need an additional assumption that $\tau(\mathbf{X}^{(\mathcal{C})})$ and $(W - \pi(\mathbf{X}))^2$ are not correlated conditional on $\mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}$, where $\mathbf{x}^{(-j)}$ is the new query point. Otherwise, consistency is obtained with a corrective term defined in equation (3), as we will see. However, we need an additional small modification of causal forests to enforce the generated estimates to be bounded, as stated in the specification below, and then avoid cumbersome technical issues in the mathematical analysis. We also need an additional assumption to strengthen the convergence of causal forests, to get the consistency of the corrected causal forest.

Specification 2. *The causal forest estimates are truncated from below and above by* $-K$ *and* K*, where* $K \in \mathbb{R}$ *is an arbitrarily large constant.*

Assumption 5. *The causal forest estimate converges uniformly, i.e.,*

$$\sup_{\boldsymbol{x} \in [0,1]} |\tau_{M,n}(\boldsymbol{x}) - \tau(\boldsymbol{x}^{(\mathcal{C})})| \xrightarrow{p} 0.$$

Theorem 3. *Let the initial causal forest* $\tau_{M,n}(\boldsymbol{x})$ *fit with the centered data* $\mathscr{D}_n^{\star}$*, and the corrected causal forest* $\theta_{M,n}^{(-j)}(\boldsymbol{x})$ *fit using* $\tau_{M,n}(\boldsymbol{x})$ *and* $\mathscr{D}_n^{\star(-j)}$*, an independent copy of the centered data with the* j*-th variable dropped. If Assumptions 1-5, and Specifications 1 and 2 are satisfied, then for* $j \in \{1, \dots, p\}$ *and* $\boldsymbol{x} \in [0,1]^p$*, we have*

$$\theta_{M,n}^{(-j)}(\boldsymbol{x}) \xrightarrow{p} \mathbb{E}[\tau(\boldsymbol{X}^{(\mathcal{C})}) \mid \boldsymbol{X}^{(-j)} = \boldsymbol{x}^{(-j)}].$$

Since Theorems 1 and 3 give the consistency of causal forests respectively fit with all input variables, and when a given variable is removed, we can deduce the consistency of our variable importance algorithm from standard asymptotic arguments.

Theorem 4. *Under the same assumptions than Theorem 3, we have for all* $j \in \{1, \dots, p\}$

$$\mathrm{I}_n^{(j)} \xrightarrow{p} \mathrm{I}^{(j)}.$$

Theorem 4 states that the introduced variable importance algorithm gets arbitrarily close to the true theoretical value, provided that the sample size is large enough. In particular, for $j \notin \mathcal{C}$, we have $\mathrm{I}_n^{(j)} \xrightarrow{p} 0$, which means that the variables not involved in the treatment heterogeneity by construction, get a null importance. Besides, notice that for $j \notin \mathcal{C}$, Theorem 4 holds without Assumption 5, only involved for the consistency of the corrected forest, required for the case where $j \in \mathcal{C}$.

4 Experiments

We assess the performance of the introduced algorithm through two batches of experiments. First, we use simulated data, where the theoretical importance values are known by construction, to compare our algorithm to the existing competitors. Secondly, we test our procedure with the semi-synthetic cases of the ACIC data challenge 2019, where the variables involved in the heterogeneity are known, but not the importance value. Our approach is compared to the importance of the `grf` package and TE-VIM, the double robust approach of Hines et al. (2022). For TE-VIM, any learning method can be used, and we report the performance of GAM models, which outperform regression forests in the presented experiments. When reading the results, recall that TE-VIM targets the same theoretical quantities $\mathrm{I}^{(j)}$ as our algorithm, whereas the `grf` importance is the frequency of variable occurrence in tree splits. Besides, the algorithm of Boileau et al. (2022) is designed for high dimensional cases and linear treatment effects, and is thus not appropriate to our goal of precisely quantifying variable importance in non-linear settings.

4.1 Simulated Data

Experiment 1. We consider a first example of simulated data to highlight the good performance of the proposed importance measure. The input is of dimension $p = 8$, and is defined by $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, with Σ the identity matrix except that $\mathrm{Cov}(X^{(1)}, X^{(5)}) = 0.9$. The treatment assignment is given by $W \sim \text{Bernouilli}(0.4 + 0.2\mathbb{1}_{X^{(1)}>0})$, and the response Y follows

$$Y = \left(X^{(1)}\mathbb{1}_{X^{(1)}>0} + 0.6X^{(2)}\mathbb{1}_{X^{(2)}>0}\right) \times W + (X^{(3)} \times X^{(4)})^2 + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, 0.1)$. In practice, we take a sample size $n = 3000$, and the causal forest is fit with $M = 2000$ trees. Notice that the ratio $\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{C})})]/\mathbb{V}[Y]$ is about 5% in this setting. Results are averaged over 10 repetitions, and are reported in Table 1 (30 repetition for grf-vimp to stabilize the ranking). Additionally, the standard deviation of the mean importance for each variable is displayed in brackets, except for negligible values (< 0.005).

I		I_n		TE-VIM		grf-vimp	
$X^{(2)}$	0.26	$X^{(2)}$	0.24 (0.02)	$X^{(1)}$	0.42 (0.07)	$X^{(1)}$	0.49 (0.02)
$X^{(1)}$	0.18	$X^{(1)}$	0.19 (0.01)	$X^{(2)}$	0.40 (0.08)	$X^{(3)}$	0.13 (0.01)
$X^{(3)}$	0	$X^{(3)}$	0.05 (0.01)	$X^{(4)}$	0.19 (0.32)	$X^{(4)}$	0.12 (0.01)
$X^{(4)}$	0	$X^{(4)}$	0.04 (0.01)	$X^{(8)}$	0.14 (0.16)	$X^{(5)}$	0.11 (0.01)
$X^{(5)}$	0	$X^{(5)}$	0.01	$X^{(5)}$	0.14 (0.15)	$X^{(2)}$	0.10 (0.01)
$X^{(6)}$	0	$X^{(6)}$	0.01	$X^{(3)}$	0.12 (0.19)	$X^{(6)}$	0.02
$X^{(7)}$	0	$X^{(7)}$	0.01	$X^{(6)}$	0.05 (0.15)	$X^{(7)}$	0.02
$X^{(8)}$	0	$X^{(8)}$	0.01	$X^{(7)}$	-0.01 (0.17)	$X^{(8)}$	0.02

Table 1: Variable importance ranking of Experiment 1 for $\mathrm{I}_n^{(j)}$, the importance measure of `grf` package, and TE-VIM. Standard deviations are displayed in brackets when greater than 0.005.

The results displayed in Table 1 show that our algorithm is the only one to provide the accurate variable ranking, where $X^{(2)}$ is the most important variable, and $X^{(1)}$ the second

most important one. TE-VIM accurately identifies these two variables as the most influential, with a similar importance. On the other hand, the importance measure from the `grf` package underestimates the importance of variable $X^{(2)}$, and identifies $X^{(3)}$, $X^{(4)}$, and $X^{(5)}$ as slightly more important than $X^{(2)}$, although these three variables are not involved in the treatment heterogeneity by construction. In particular, $X^{(5)}$ is not involved at all in the response Y, but is strongly correlated to the influential input $X^{(1)}$. Because of this dependence, $X^{(5)}$ is frequently used in the causal forests splits, leading to this quite high importance given by the `grf` package. On the other hand, $\mathrm{I}_n^{(j)}$ gives an importance close to 0 for $X^{(5)}$. This result is expected, since the removal of $X^{(5)}$ does not lead to any loss of information regarding the treatment heterogeneity, by definition. An additional interesting phenomenon is the significant non-null importance for variables $X^{(3)}$ and $X^{(4)}$ given by all procedures. In fact, the interaction term in μ, which takes the form of a squared product, is rather difficult to estimate by regression forests. Then, the local centering of Y is only partial, and $X^{(3)}$ and $X^{(4)}$ still have impact on the variance of treatment estimates. Besides, notice that the corrective term of equation (3) is negligible in this experiment, and that using the original causal forest retrained with one variable removed, gives the same result as in Table 1 for $\mathrm{I}_n^{(j)}$.

Experiment 2. This second experiment has the same setting than Experiment 1, except that variable $X^{(1)}$ is not involved in the treatment effect anymore, but only in μ. Now, the response writes

$$Y = \left(0.6X^{(2)}\mathbb{1}_{X^{(2)}>0}\right) \times W + \mathbf{X^{(1)}}\mathbb{1}_{\mathbf{X^{(1)}>0}} + (X^{(3)} \times X^{(4)})^2 + \varepsilon.$$

The results are provided in Table 2. Clearly, $\mathrm{I}_n^{(j)}$ outperforms the competitors. Indeed, $X^{(2)}$ is well-identified by $\mathrm{I}_n^{(j)}$ as responsible for most of the heterogeneity of the treatment effect, whereas TE-VIM is strongly biased, and the importance procedure of the `grf` package outputs quite close values for $X^{(2)}$, $X^{(4)}$, and $X^{(3)}$. As expected, the importance of these last two variables is relatively larger than in Experiment 1, since the ratio $\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{C})})]/\mathbb{V}[Y]$ drops to 1% in this case, explaining that the importance of $I_n^{(3)}$ and $I_n^{(4)}$ is multiplied by about 5 with respect to Experiment 1.

I	
$X^{(2)}$	1
$X^{(1)}$	0
$X^{(3)}$	0
$X^{(4)}$	0
$X^{(5)}$	0
$X^{(6)}$	0
$X^{(7)}$	0
$X^{(8)}$	0

I_n	
$X^{(2)}$	0.82 (0.03)
$X^{(3)}$	0.20 (0.03)
$X^{(4)}$	0.19 (0.03)
$X^{(1)}$	0.02
$X^{(5)}$	0.02
$X^{(6)}$	0.02
$X^{(7)}$	0.02
$X^{(8)}$	0.02

TE-VIM	
$X^{(2)}$	1.76 (0.11)
$X^{(4)}$	1.65 (0.04)
$X^{(3)}$	1.03 (0.02)
$X^{(8)}$	0.99
$X^{(1)}$	0.96 (0.02)
$X^{(5)}$	0.88 (0.02)
$X^{(6)}$	0.71 (0.03)
$X^{(7)}$	0.57 (0.04)

grf-vimp	
$X^{(2)}$	0.36 (0.01)
$X^{(4)}$	0.24 (0.01)
$X^{(3)}$	0.23 (0.01)
$X^{(1)}$	0.03
$X^{(5)}$	0.03
$X^{(6)}$	0.03
$X^{(7)}$	0.03
$X^{(8)}$	0.03

Table 2: Variable importance ranking of Experiment 2 for $\mathrm{I}_n^{(j)}$, the importance measure of `grf` package, and TE-VIM. Standard deviations are displayed in brackets when greater than 0.005.

4.2 ACIC Data Challenge 2019

We run a second batch of experiments using the data from the ACIC data challenge 2019 (https://sites.google.com/view/acic2019datachallenge/data-challenge), where the goal was to estimate ATEs in various settings. The input data is taken from real datasets available online on the UCI repository. Next, outcomes are simulated with different scenarios, and the associated code scripts were released after the challenge. Since the data generating mechanism is available, we have access to the variables involved in the heterogeneous treatment effect. In each scenario, a hundred datasets were randomly sampled.

I_n		grf-vimp	
$X^{(3)}$	0.81 (0.03)	$X^{(3)}$	0.60 (0.02)
$X^{(29)}$	0.01 (0.002)	$X^{(29)}$	0.05 (0.009)
$X^{(28)}$	0.01 (0.003)	$X^{(11)}$	0.03 (0.01)
$X^{(27)}$	0.01 (0.002)	$X^{(14)}$	0.02 (0.01)
$X^{(4)}$	0.01 (0.002)	$X^{(25)}$	0.02 (0.002)

Table 3: Top 5 variables for "Student performance 2 (Scenario 4)" dataset using $\mathrm{I}_n^{(j)}$ and the importance measure of `grf` package. Standard deviations are displayed in brackets.

We first use the "student performance 2" data with 31 input variables, considering Scenario 4, involving heterogeneity of the treatment effect. We merge datasets two by two to get a sample size of $n = 1298$, and run 10 repetitions for uncertainties. Table 3 gives the top 5 variables ranked by $\mathrm{I}_n^{(j)}$, which accurately identifies $X^{(3)}$ as the only variable involved in the treatment heterogeneity. The `grf` importance gives similar results, except that the importance of the irrelevant variables $X^{(29)}$, $X^{(11)}$ is relatively higher than for $\mathrm{I}_n^{(j)}$.

Secondly, we use the "spam email" data, made of 22 input variables. We also consider Scenario 4, where variables $X^{(8)}$ and $X^{(19)}$ are involved in the heterogeneous treatment effect. In this case, we merge 20 datasets to get a quite large sample of size $n = 10000$, and run 5 repetitions to compute standard deviations. The two relevant variables are properly identified by the two tested algorithms. Notice that the impact of $X^{(19)}$ is small, and a large sample size is required to detect its influence. Again, the `grf` importance gives higher values to irrelevant variables than $\mathrm{I}_n^{(j)}$.

I_n		grf-vimp	
$X^{(8)}$	0.82 (9.10^{-4})	$X^{(8)}$	0.85 (4.10^{-3})
$X^{(19)}$	0.012 (2.10^{-3})	$X^{(19)}$	0.064 (6.10^{-3})
$X^{(12)}$	0.004 (6.10^{-4})	$X^{(1)}$	0.013 (3.10^{-3})
$X^{(22)}$	0.004 (3.10^{-4})	$X^{(22)}$	0.013 (1.10^{-3})
$X^{(15)}$	0.002 (4.10^{-4})	$X^{(15)}$	0.010 (8.10^{-4})
$X^{(17)}$	0.002 (3.10^{-4})	$X^{(17)}$	0.009 (2.10^{-3})

Table 4: Top 6 variables for "Spam email (Scenario 4)" dataset using $\mathrm{I}_n^{(j)}$ and the importance measure of `grf` package. Standard deviations are displayed in brackets.

5 Conclusion

We introduced a variable importance algorithm based on causal forests, following the drop and retrain principle, which is well-established for regression problems. In the context of causal inference, the main obstacle of this approach is to retrain the forest without a confounding variable. However, we have shown that the local centering of the outcome and treatment assignment leads to consistent estimates, provided that the removed variable is not involved in the treatment heterogeneity. To handle the remaining cases, we introduced a corrective term in the retrained forest. Overall, our proposed variable importance algorithm is shown to be consistent, under standard assumptions in the literature about the theoretical analysis of random forests. Next, we have run several batches of experiments on simulated and semi-synthetic data to show the good performance of the introduced method compared to the existing competitors. Finally, this variable importance measure for heterogeneous treatment effects provides a break down of the heterogeneity with respect to all inputs variables, which opens the route to promising applications, especially in healthcare as discussed in the introduction.

References

S. Athey and S. Wager. Policy learning with observational data. *Econometrica*, 89:133–161, 2021.

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

P.C. Austin and E.A. Stuart. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34:3661–3679, 2015.

P. Boileau, N.T. Qi, M.J. van der Laan, S. Dudoit, and N. Leng. A flexible approach for predictive biomarker discovery. *arXiv preprint arXiv:2205.01285*, 2022.

A. Deaton and N. Cartwright. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210:2–21, 2018.

O. Hines, K. Diaz-Ordaz, and S. Vansteelandt. Variable importance measures for heterogeneous causal effects. *arXiv preprint arXiv:2204.06030*, 2022.

K. Hirano and G.W. Imbens. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2:259–278, 2001.

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

E.H. Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

I. Kosuke and R. Marc. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7:443 – 470, 2013.

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

N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108:299–319, 2021.

Z. Obermeyer and E.J. Emanuel. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375:1216, 2016.

J.M. Robins, A. Rotnitzky, and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89: 846–866, 1994.

K.J. Rothman. *Epidemiology: an introduction*. Oxford university press, 2012.

P.M. Rothwell. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet*, 365:82–93, 2005.

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

U. Shalit, F.D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

I.M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1:407–414, 1993.

T.J. VanderWeele and J.M. Robins. Four types of effect modification: a classification based on directed acyclic graphs. *Epidemiology*, 18:561–568, 2007.

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

B.D. Williamson, P.B. Gilbert, N.R. Simon, and M. Carone. A unified approach for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, pages 1–14, 2021.

A Proofs of Propositions 1-4 and Theorems 1-4

Proof of Proposition 1. Using the observed outcome definition with SUTVA (line 1), and the unconfoundedness Assumption 1 (line 2 to 3), we have

$$\begin{aligned}
\mathbb{E}[Y \mid \mathbf{X}, W] &= \mathbb{E}[WY(1) + (1-W)Y(0) \mid \mathbf{X}, W] \\
&= W\mathbb{E}[Y(1) \mid \mathbf{X}, W] + (1-W)\mathbb{E}[Y(0) \mid \mathbf{X}, W] \\
&= W\mathbb{E}[Y(1) \mid \mathbf{X}] + (1-W)\mathbb{E}[Y(0) \mid \mathbf{X}] \\
&= \mathbb{E}[Y(0) \mid \mathbf{X}] + W(\mathbb{E}[Y(1) \mid \mathbf{X}] - \mathbb{E}[Y(0) \mid \mathbf{X}]) \\
&= \mathbb{E}[Y(0) \mid \mathbf{X}] + W\mathbb{E}[Y(1) - Y(0) \mid \mathbf{X}]) \\
&= \mathbb{E}[\mu(\mathbf{X}) + \varepsilon(0) \mid \mathbf{X}] + W\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{C})}) + \varepsilon(1) - \varepsilon(0) \mid \mathbf{X}]) \\
&= \mu(\mathbf{X}) + W\tau(\mathbf{X}^{(\mathcal{C})}),
\end{aligned}$$

and the final result follows. □

Proof of Proposition 2. Assumption 2 implies that $\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{C})})] > 0$. By definition,

$$\mathrm{I}^{(j)} = \frac{\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{C})})] - \mathbb{V}[\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{C})})|\mathbf{X}^{(-j)}]]}{\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{C})})]},$$

which also writes using the law of total variance

$$\mathrm{I}^{(j)} = \frac{\mathbb{E}[\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{C})})|\mathbf{X}^{(-j)}]]}{\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{C})})]} = \frac{\mathbb{E}[(\tau(\mathbf{X}^{(\mathcal{C})}) - E[\tau(\mathbf{X}^{(\mathcal{C})})|\mathbf{X}^{(-j)}])^2]}{\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{C})})]}. \tag{4}$$

If $j \notin \mathcal{C}$, we clearly have $E[\tau(\mathbf{X}^{(\mathcal{C})})|\mathbf{X}^{(-j)}] = \tau(\mathbf{X}^{(\mathcal{C})})$, and then equation (4) gives that $\mathrm{I}^{(j)} = 0$.

We now consider the case where $j \in \mathcal{C}$. According to Assumption 2, $\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{C})})|\mathbf{X}^{(-j)}] > 0$ with a strictly positive probability. It directly implies that

$$\mathbb{E}[\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{C})})|\mathbf{X}^{(-j)}]] > 0,$$

and then, combined with equation (4), we obtain $\mathrm{I}^{(j)} > 0$. Since $\mathbb{V}[\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{C})})|\mathbf{X}^{(-j)}]] \geq 0$, we also have $\mathrm{I}^{(j)} \leq 1$. □

Proof of Proposition 3. We first expand the covariance term

$$\begin{aligned}
&\mathrm{Cov}[W - \pi(\mathbf{X}), Y - m(\mathbf{X}) \mid \mathbf{X}^{(\mathcal{C})}] \\
&\quad = \mathbb{E}[(W - \pi(\mathbf{X}))(Y - m(\mathbf{X})) \mid \mathbf{X}^{(\mathcal{C})}] - \mathbb{E}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(\mathcal{C})}]\mathbb{E}[Y - m(\mathbf{X}) \mid \mathbf{X}^{(\mathcal{C})}].
\end{aligned}$$

Notice that the second term is null since $\mathbb{E}[Y - m(\mathbf{X}) \mid \mathbf{X}^{(\mathcal{C})}] = \mathbb{E}[\mathbb{E}[Y - m(\mathbf{X}) \mid \mathbf{X}] \mid \mathbf{X}^{(\mathcal{C})}] = 0$. Additionally, by definition,

$$m(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}] = \mathbb{E}[\mu(\mathbf{X}) + \tau(\mathbf{X}^{(\mathcal{C})}) \times W + \varepsilon(W) \mid \mathbf{X}] = \mu(\mathbf{X}) + \tau(\mathbf{X}^{(\mathcal{C})})\pi(\mathbf{X}),$$

then $Y - m(\mathbf{X}) = (W - \pi(\mathbf{X}))\tau(\mathbf{X}^{(\mathcal{C})}) + \varepsilon(W)$, and we get

$$\begin{aligned}
&\mathrm{Cov}[W - \pi(\mathbf{X}), Y - m(\mathbf{X}) \mid \mathbf{X}^{(\mathcal{C})}] \\
&\qquad = \mathbb{E}[(W - \pi(\mathbf{X}))((W - \pi(\mathbf{X}))\tau(\mathbf{X}^{(\mathcal{C})}) + \varepsilon(W)) \mid \mathbf{X}^{(\mathcal{C})}] \\
&\qquad = \tau(\mathbf{X}^{(\mathcal{C})}) \times \mathbb{E}[(W - \pi(\mathbf{X}))^2 \mid \mathbf{X}^{(\mathcal{C})}] + \mathbb{E}[\varepsilon(W)(W - \pi(\mathbf{X})) \mid \mathbf{X}^{(\mathcal{C})}] \\
&\qquad = \tau(\mathbf{X}^{(\mathcal{C})}) \times \mathbb{E}[(W - \pi(\mathbf{X}))^2 \mid \mathbf{X}^{(\mathcal{C})}] + \mathbb{E}[(W - \pi(\mathbf{X}))\mathbb{E}[\varepsilon(W) \mid \mathbf{X}, W] \mid \mathbf{X}^{(\mathcal{C})}]] \\
&\qquad = \tau(\mathbf{X}^{(\mathcal{C})}) \times \mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(\mathcal{C})}],
\end{aligned}$$

which gives the final local moment equation in $\mathbf{X}^{(\mathcal{C})}$. □

Proof of Proposition 4. As in the proof of Proposition 3, we obtain

$$\mathrm{Cov}[W - \pi(\mathbf{X}), Y - m(\mathbf{X}) \mid \mathbf{X}^{(-j)}] = \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{C})})(W - \pi(\mathbf{X}))^2 \mid \mathbf{X}^{(-j)}].$$

Notice that

$$\begin{aligned}
\mathrm{Cov}[\tau(\mathbf{X}^{(\mathcal{C})}), (W - \pi(\mathbf{X}))^2 \mid \mathbf{X}^{(-j)}] = \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{C})})(W - \pi(\mathbf{X}))^2 \mid \mathbf{X}^{(-j)}] \\
- \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{C})}) \mid \mathbf{X}^{(-j)}]\mathbb{E}[(W - \pi(\mathbf{X}))^2 \mid \mathbf{X}^{(-j)}].
\end{aligned}$$

Combining the above two equations, we have

$$\begin{aligned}
\mathrm{Cov}[W - \pi(\mathbf{X}), Y - m(\mathbf{X}) \mid \mathbf{X}^{(-j)}] =& \mathrm{Cov}[\tau(\mathbf{X}^{(\mathcal{C})}), (W - \pi(\mathbf{X}))^2 \mid \mathbf{X}^{(-j)}] \\
&+ \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{C})}) \mid \mathbf{X}^{(-j)}] \times \mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(-j)}],
\end{aligned}$$

which gives the final result. □

Proof of Theorem 1. The result is obtained by applying Theorem 3 from Athey et al. (2019). The first paragraph of section 3 of Athey et al. (2019) provides conditions to apply Theorem 3, that are satisfied by our Assumptions 3 and 4: $\mathbf{X} \in [0,1]^p$, $\mathbf{X}$ admits a density bounded from below and above by strictly positive constants, and μ and τ are bounded.

Next, Assumptions 1-6 from Athey et al. (2019) must be verified. As stated at the end of Section 6.1, Assumptions 3-6 always hold for causal forests, the first assumption holds because the functions m, μ, and τ are Lispschitz from our Assumption 4 (the product of Lipschitz functions is Lipschitz), and Assumption 2 is satisfied because $0 < \mathbb{V}[W \mid \mathbf{X}] = \pi(\mathbf{X})(1 - \pi(\mathbf{X})) < 1$ from our Assumption 4.

Finally, the forest is grown from Specification 1, and the treatment effect is identified by equation (1) since Assumption 1 enforces unconfoundedness. Overall, we apply Theorem 3 from Athey et al. (2019) to get the consistency of the causal forest estimate, i.e., for $\mathbf{x} \in [0,1]^p$

$$\tau_{M,n}(\mathbf{x}) \xrightarrow{p} \tau(\mathbf{x}^{(\mathcal{C})}).$$

Notice that Theorem 3 from Athey et al. (2019) states the consistency of generalized forests. As it will be useful for further results, we give below a proof of the weak consistency in the specific case of causal forests, using arguments of Athey et al. (2019). In particular, we take

advantage of Specification 1, which enforces the honesty property, and that the diameters of tree cells vanish as the sample size n increases. First, in our case of binary treatment W, the causal forest estimate writes

$$\tau_{M,n}(\mathbf{x}) = \frac{\sum_{i=1}^n \alpha_i(\mathbf{x}) W_i Y_i - (\sum_{i=1}^n \alpha_i(\mathbf{x}) W_i)(\sum_{i=1}^n \alpha_i(\mathbf{x}) Y_i)}{\sum_{i=1}^n \alpha_i(\mathbf{x}) W_i^2 - (\sum_{i=1}^n \alpha_i(\mathbf{x}) W_i)^2},$$

where the weight $\alpha_i(\mathbf{x})$ is defined by equation (3) of Athey et al. (2019), as the weight associated to training observation $\mathbf{X}_i$ to form an estimate at the new query point $\mathbf{x}$. The weights $\alpha_i(\mathbf{x})$ sum to 1 over all observations, i.e., $\sum_{i=1}^n \alpha_i(\mathbf{x}) = 1$. Also notice that we alleviate notations of $\alpha_i(\mathbf{x})$ throughout the article, but the full expression with all dependencies is $\alpha_i(\mathbf{x}, \mathbf{X}_i, \mathbf{\Theta}_M, \mathscr{D}_n)$, where the causal forest is built with data $\mathscr{D}_n$, and trees are randomized with $\mathbf{\Theta}_M$. Now, we denote by $\Delta_{1,n}(\mathbf{x}) = \sum_{i=1}^n \alpha_i(\mathbf{x}) W_i Y_i$ the first term of the numerator of $\tau_{M,n}(\mathbf{x})$, and derive its convergence. Since the weights sum to 1,

$$\Delta_{1,n}(\mathbf{x}) - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}] = \sum_{i=1}^n \alpha_i(\mathbf{x})(W_i Y_i - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]),$$

and then,

$$\mathbb{E}[\Delta_{1,n}(\mathbf{x}) - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]] = \sum_{i=1}^n \mathbb{E}[\mathbb{E}[\alpha_i(\mathbf{x})(W_i Y_i - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]) \mid \mathbf{X}_i]].$$

Here, we use a key property of the forest growing given by Specification 1 : honesty. Indeed, it enforces that $\mathscr{D}_n$ is randomly split in two halves for each tree, where one part is used to build the splits, and the other half to compute the weights. Therefore, $\alpha_i(\mathbf{x}, \mathbf{X}_i, \mathbf{\Theta}_M, \mathscr{D}_n)$ and $W_i Y_i$ are independent conditional on $\mathbf{X}_i$, for all $\{i, \dots, n\}$. Then, we have

$$\begin{aligned}
\mathbb{E}[\Delta_{1,n}(\mathbf{x}) - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]] &= \sum_{i=1}^n \mathbb{E}[\mathbb{E}[\alpha_i(\mathbf{x}) \mid \mathbf{X}_i]\mathbb{E}[W_i Y_i - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}] \mid \mathbf{X}_i]] \\
&= \sum_{i=1}^n \mathbb{E}[\mathbb{E}[\alpha_i(\mathbf{x}) \mid \mathbf{X}_i](\mathbb{E}[W_i Y_i \mid \mathbf{X}_i] - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}])].
\end{aligned}$$

Since W and Y are independent conditional on $\mathbf{X}$ from the unconfoundedness Assumption 1, $\mathbb{E}[W_i Y_i \mid \mathbf{X}_i] = \mathbb{E}[W_i \mid \mathbf{X}_i]\mathbb{E}[Y_i \mid \mathbf{X}_i]$. Additionally, Assumption 4 states that the functions π and m are Lipschitz, and since the product of two Lipschitz functions is Lipschitz, $\mathbb{E}[W_i Y_i \mid \mathbf{X}_i]$ is Lipschitz, with a constant $C > 0$. Therefore, we obtain

$$\begin{aligned}
\mathbb{E}[\Delta_{1,n}(\mathbf{x}) - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]] &\leq \sum_{i=1}^n \mathbb{E}[\mathbb{E}[\alpha_i(\mathbf{x}) \mid \mathbf{X}_i] C \|\mathbf{X}_i - \mathbf{x}\|_2] \\
&\leq C\mathbb{E}\Big[\sum_{i=1}^n \alpha_i(\mathbf{x}) \|\mathbf{X}_i - \mathbf{x}\|_2\Big] \\
&\leq C\mathbb{E}\Big[\sup_i \|\mathbf{X}_i - \mathbf{x}\|_2 \mathbb{1}_{\alpha_i(\mathbf{x})>0} \sum_{i=1}^n \alpha_i(\mathbf{x})\Big] \\
&\leq C\mathbb{E}\Big[\sup_i \|\mathbf{X}_i - \mathbf{x}\|_2 \mathbb{1}_{\alpha_i(\mathbf{x})>0}\Big].
\end{aligned}$$

Since Assumptions 3 and 4 and Specification 1 are satisfied, equation (26) in the Supplementary Material of Athey et al. (2019) states that

$$\mathbb{E}\big[\sup_i \|\mathbf{X}_i - \mathbf{x}\|_2 \mathbb{1}_{\alpha_i(\mathbf{x})>0}\big] \longrightarrow 0,$$

which gives that

$$\mathbb{E}[\Delta_{1,n}(\mathbf{x})] \longrightarrow \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]. \tag{5}$$

Next, we use equation (24) in Lemma 7 of the Supplementary Material of Athey et al. (2019), to get that $\mathbb{V}[\Delta_{1,n}(\mathbf{x})] = O(a_n/n)$. Since $a_n/n \longrightarrow 0$ by Specification 1, we finally have $\mathbb{V}[\Delta_{1,n}(\mathbf{x})] \longrightarrow 0$. Finally, this last limit combined with equation (5), states that $\Delta_{1,n}(\mathbf{x}) - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]$ is asymptotically unbiased and of null variance. Using the bias-variance decomposition, we obtain the $\mathbb{L}^2$-consistency of $\Delta_{1,n}(\mathbf{x})$ towards $\mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]$, which implies the weak consistency

$$\sum_{i=1}^{n} \alpha_i(\mathbf{x}) W_i Y_i \xrightarrow{p} \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}].$$

Identically, we obtain the weak consistency of the other terms involved in $\tau_{M,n}(\mathbf{x})$, i.e., $\sum_{i=1}^n \alpha_i(\mathbf{x}) W_i \xrightarrow{p} \pi(\mathbf{x})$, $\sum_{i=1}^n \alpha_i(\mathbf{x}) Y_i \xrightarrow{p} m(\mathbf{x})$, and $\sum_{i=1}^n \alpha_i(\mathbf{x}) W_i^2 \xrightarrow{p} \mathbb{E}[W^2 \mid \mathbf{X} = \mathbf{x}]$. The continuous mapping theorem gives for the last term that $\big(\sum_{i=1}^n \alpha_i(\mathbf{x}) W_i\big)^2 \xrightarrow{p} \mathbb{E}[W \mid \mathbf{X} = \mathbf{x}]^2$. Finally, using Slutsky's Lemma, we obtain

$$\begin{aligned}
\tau_{M,n}(\mathbf{x}) &\xrightarrow{p} \frac{\mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}] - \mathbb{E}[W \mid \mathbf{X} = \mathbf{x}]\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]}{\mathbb{E}[W^2 \mid \mathbf{X} = \mathbf{x}] - \mathbb{E}[W \mid \mathbf{X} = \mathbf{x}]^2} \\
&= \frac{\mathrm{Cov}[W, Y \mid \mathbf{X} = \mathbf{x}]}{\mathbb{V}[W \mid \mathbf{X} = \mathbf{x}]} \\
&= \tau(\mathbf{x}^{(\mathcal{C})}),
\end{aligned}$$

where the last line is given by the local moment equation (1), which identifies the treatment effect. Finally, notice that this proof applies to any linear local moment equation defining a generalized random forest. □

Proof of Theorem 2. We consider $j \notin \mathcal{C}$, and follow the same proof as Theorem 1, to show that the causal forest $\tau_{M,n}^{(-j)}(\mathbf{x})$ fit with $\mathscr{D}_n^{\star(-j)}$ converges as

$$\tau_{M,n}^{(-j)}(\mathbf{x}) \xrightarrow{p} \theta(\mathbf{x}^{(-j)}),$$

where $\theta(\mathbf{x}^{(-j)})$ satisfies the following equation by definition of causal forests,

$$\theta(\mathbf{x}^{(-j)}) \times \mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}] - \mathrm{Cov}[W - \pi(\mathbf{X}), Y - m(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}] = 0.$$

Then, according to Proposition 3, the above moment equation identifies the treatment effect under Assumptions 1 and 2, and we obtain

$$\theta(\mathbf{x}^{(-j)}) = \tau(\mathbf{x}^{(\mathcal{C})}),$$

which gives (i). For (ii), we apply the same proof, except that the obtained local moment equation identifies $\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{C})}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]$ according to Proposition 4. □

Proof of Theorem 3. With $j \in \{1, \dots, p\}$, recall that the causal forest $\tau_{M,n}(\mathbf{x})$ is fit with a centered dataset $\mathscr{D}_n^\star$, and the corrected causal forest estimate $\theta_{M,n}^{(-j)}(\mathbf{x})$ is fit with $\mathscr{D}_n^{\star(-j)}$, an independent copy of the centered dataset with the j-th variable dropped, and is formally defined as

$$\theta_{M,n}^{(-j)}(\mathbf{x}) = \tau_{M,n}^{(-j)}(\mathbf{x}) - \frac{\sum_{i=1}^n \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))^2 \tau_{M,n}(\mathbf{X}_i) - \overline{W^2_{\alpha'}}\overline{\tau}_{\alpha'}}{\sum_{i=1}^n \alpha_i'(\mathbf{x}^{(-j)})(W_i - \overline{W}_{\alpha'})^2},$$

where $\overline{W^2_{\alpha'}} = \sum_{i=1}^n \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))^2$, $\overline{\tau}_{\alpha'} = \sum_{i=1}^n \alpha_i'(\mathbf{x}^{(-j)})\tau_{M,n}(\mathbf{X}_i)$, and $\overline{W}_{\alpha'} = \sum_{i=1}^n \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))$. We first prove the convergence of the first term of the numerator,

$$\begin{aligned}
\Delta_n &= \sum_{i=1}^n \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))^2 \tau_{M,n}(\mathbf{X}_i) \\
&= \sum_{i=1}^n \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))^2 \tau(\mathbf{X}_i) + \sum_{i=1}^n \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))^2 (\tau_{M,n}(\mathbf{X}_i) - \tau(\mathbf{X}_i)).
\end{aligned}$$

Using the same proof as for Theorem 1, we get that

$$\sum_{i=1}^n \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))^2 \tau(\mathbf{X}_i) \xrightarrow{p} \mathbb{E}[(W - \pi(\mathbf{X}))^2 \tau(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}^{(-j)}].$$

For the second term involved in Δ_n,

$$\begin{aligned}
\Big| \sum_{i=1}^n \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))^2 (\tau_{M,n}(\mathbf{X}_i) - \tau(\mathbf{X}_i)) \Big| &\leq \sup_i |(\tau_{M,n}(\mathbf{X}_i) - \tau(\mathbf{X}_i))| \sum_{i=1}^n \alpha_i'(\mathbf{x}^{(-j)}) \\
&\leq \sup_i |(\tau_{M,n}(\mathbf{X}_i) - \tau(\mathbf{X}_i))| \longrightarrow 0,
\end{aligned}$$

where the limit is given by the uniform convergence of Assumption 5. Overall, we have

$$\Delta_n \xrightarrow{p} \mathbb{E}[(W - \pi(\mathbf{X}))^2 \tau(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}^{(-j)}].$$

Next, $\overline{\tau}_{\alpha'}$ is handled similarly, and we follow the same proof as for Theorem 1 to get the weak consistency of the remaining terms involved in $\theta_{M,n}^{(-j)}(\mathbf{x})$, and using Slutsky's lemma, we obtain

$$\frac{\sum_{i=1}^n \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))^2 \tau_{M,n}(\mathbf{X}_i) - \overline{W^2_{\alpha'}}\overline{\tau}_{\alpha'}}{\sum_{i=1}^n \alpha_i'(\mathbf{x}^{(-j)})(W_i - \overline{W}_{\alpha'})^2} \xrightarrow{p} \frac{\mathrm{Cov}[\tau(\mathbf{X}^{(\mathcal{C})}), (W - \pi(\mathbf{X}))^2 \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]}{\mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]}.$$

Then, following the case (ii) of Theorem 2, we get

$$\tau_{M,n}^{(-j)}(\mathbf{x}) \xrightarrow{p} \frac{\mathrm{Cov}[W - \pi(\mathbf{X}), Y - m(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]}{\mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]},$$

which gives the final result

$$\begin{aligned}
\theta_{M,n}^{(-j)}(\mathbf{x}) \xrightarrow{p} &\frac{\mathrm{Cov}[W - \pi(\mathbf{X}), Y - m(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]}{\mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]} \\
&- \frac{\mathrm{Cov}[\tau(\mathbf{X}^{(\mathcal{C})}), (W - \pi(\mathbf{X}))^2 \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]}{\mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]} \\
&= \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{C})}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}],
\end{aligned}$$

where the last equality is given by Proposition 4. □

Proof of Theorem 4. We first consider the case $j \in \{1, \dots, p\} \setminus \mathcal{C}$ for the sake of clarity. We assume that Assumptions 1-5, and Specifications 1 and 2 are satisfied, and causal forests are trained as specified in Theorem 3. Then, we can apply Theorems 1 and 3 to get that

$$\tau_{M,n}(\mathbf{X}) - \theta_{M,n}^{(-j)}(\mathbf{X}) \xrightarrow{p} 0.$$

According to Specification 2, $\tau_{M,n}(\mathbf{X}) - \theta_{M,n}^{(-j)}(\mathbf{X})$ is bounded, and therefore convergence in probability implies $\mathbb{L}^2$-convergence, i.e.,

$$\mathbb{E}[(\tau_{M,n}(\mathbf{X}) - \theta_{M,n}^{(-j)}(\mathbf{X}))^2] \longrightarrow 0. \tag{6}$$

Next, recall that

$$\mathrm{I}_n^{(j)} = \frac{\sum_{i=1}^n \left[\tau_{M,n}(\mathbf{X}_i') - \theta_{M,n}^{(-j)}(\mathbf{X}_i')\right]^2}{\sum_{i=1}^n \left[\tau_{M,n}(\mathbf{X}_i') - \overline{\tau_{M,n}}\right]^2}.$$

We first consider

$$\Delta_{n,1} = \frac{1}{n}\sum_{i=1}^n \left[\tau_{M,n}(\mathbf{X}_i') - \theta_{M,n}^{(-j)}(\mathbf{X}_i')\right]^2,$$

and then

$$\mathbb{E}[\Delta_{n,1}] = \mathbb{E}\left[\left(\tau_{M,n}(\mathbf{X}_1') - \theta_{M,n}^{(-j)}(\mathbf{X}_1')\right)^2\right].$$

Since $|\Delta_{n,1}| = \Delta_{n,1}$, according to equation (6), we have

$$\mathbb{E}[|\Delta_{n,1}|] \longrightarrow 0,$$

which also implies the convergence in probability of $\Delta_{n,1}$.

Similarly for the denominator, we write

$$\Delta_{n,2} = \frac{1}{n}\sum_{i=1}^n \tau_{M,n}(\mathbf{X}_i')^2 - \overline{\tau_{M,n}}^2$$

We first show the convergence of $\overline{\tau_{M,n}}$. Hence,

$$\mathbb{E}[\overline{\tau_{M,n}}] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n \tau_{M,n}(\mathbf{X}_i')] = \mathbb{E}[\tau_{M,n}(\mathbf{X})] \longrightarrow \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{C})})],$$

where the limit is obtained because Theorem 1 gives the weak consistency of $\tau_{M,n}(\mathbf{X})$, which implies the convergence of the first moment since $\tau_{M,n}(\mathbf{X})$ is bounded from Specification 2. Next, we show that the variance of $\overline{\tau_{M,n}}$ vanishes. We use the law of total variance to get

$$\mathbb{V}[\overline{\tau_{M,n}}] = \mathbb{V}[\mathbb{E}[\overline{\tau_{M,n}} \mid \mathbf{\Theta}_M, \mathscr{D}_n]] + \mathbb{E}[\mathbb{V}[\overline{\tau_{M,n}} \mid \mathbf{\Theta}_M, \mathscr{D}_n]].$$

For $\mathbb{E}[\mathbb{V}[\overline{\tau_{M,n}} \mid \mathbf{\Theta}_M, \mathscr{D}_n]]$, notice that $\tau_{M,n}(\mathbf{X}'_i)$ are iid conditional on $\mathbf{\Theta}_M$ and $\mathscr{D}_n$. Therefore,

$$\mathbb{V}[\overline{\tau_{M,n}} \mid \mathbf{\Theta}_M, \mathscr{D}_n] = \frac{\mathbb{V}[\tau_{M,n}(\mathbf{X}) \mid \mathbf{\Theta}_M, \mathscr{D}_n]}{n} < \frac{K^2}{n},$$

since $\tau_{M,n}(\mathbf{X})$ is bounded by K from Specification 2. We thus obtain $\mathbb{E}[\mathbb{V}[\overline{\tau_{M,n}} \mid \mathbf{\Theta}_M, \mathscr{D}_n]] \longrightarrow 0$. For the first term, notice that

$$\mathbb{V}[\mathbb{E}[\overline{\tau_{M,n}} \mid \mathbf{\Theta}_M, \mathscr{D}_n]] = \mathbb{V}[\mathbb{E}[\tau_{M,n}(\mathbf{X}) \mid \mathbf{\Theta}_M, \mathscr{D}_n]] < \mathbb{V}[\tau_{M,n}(\mathbf{X})],$$

where this upper bound converges to 0, since $\tau_{M,n}(\mathbf{X})$ converges towards $\tau(\mathbf{X}^{(\mathcal{C})})$ in $\mathbb{L}^2$. Overall, $\overline{\tau_{M,n}}$ is asymptotically unbiased and its variance vanishes, and therefore converges towards 0 in $\mathbb{L}^2$, and the weak consistency follows, i.e.,

$$\overline{\tau_{M,n}} \xrightarrow{p} \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{C})})].$$

Using the continuous mapping theorem, we conduct the same analysis to get that $\frac{1}{n}\sum_{i=1}^{n} \tau_{M,n}(\mathbf{X}'_i)^2 \xrightarrow{p} \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{C})})^2]$, and then

$$\Delta_{n,2} \xrightarrow{p} \mathbb{V}[\tau(\mathbf{X}^{(\mathcal{C})})],$$

with $\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{C})})] > 0$ from Assumption 2. Finally, both the numerator $\Delta_{n,1}$ and denominator $\Delta_{n,2}$ of $\mathrm{I}_n^{(j)}$ converge in probability, and we can apply Slutsky's Lemma to obtain

$$\mathrm{I}_n^{(j)} \xrightarrow{p} 0.$$

The proof is similar for the case where $j \notin \mathcal{C}$. □