Generalized Transformers

Kweku A. Opoku-Agyemang[*]

June 2023

Abstract

In this note, we present a non-parametric causal transformer for estimating heterogeneous treatment effects. In the potential outcomes framework with unconfoundedness, we show that causal transformers are pointwise consistent for the true treatment effect and have an asymptotically Gaussian and centered sampling distribution. We also discuss a practical method for constructing asymptotic confidence intervals for the true treatment effect that are centered at the causal transformer estimates. Our theoretical results rely on a generic Gaussian theory for a large family of transformer models. To our knowledge, this is the first set of results that allows any type of transformer, including encoder-decoder and encoder-only models, to be used for provably valid statistical inference in the econometric sense.

[*] Machine Learning X Doing. Email: kweku@machinelearningxdoing.com. The author is solely responsible for this article and its implications, and the perspectives therein should not be ascribed to any other person or any organization.

1 Introduction

The transformer model (Vaswani et al., 2017) is a deep learning architecture that uses attention mechanisms to weigh the significance of each part of the input data[1]. It was originally introduced for natural language processing (NLP) tasks such as machine translation and text summarization, and later applied to other domains such as computer vision (CV) and video generation. The transformer model consists of two parts: an encoder and a decoder, which are composed of multiple self-attention layers. The encoder takes the input features and outputs a latent representation that captures the contextual information of the input. The decoder takes the latent representation and generates an output sequence, optionally conditioned on another input sequence. The framework is behind the large language model revolution and foundation models in the technology industry. Although some research in economics focuses on the impacts of such models in software (e.g. Noy and Zhang (2023), Brynjolfsson, Li,and Raymond (2023)), less is known about how the underlying transformers might serve as an econometrics tool, perhaps, as part of the credibility revolution in economics.

In this paper, we present an adaptation of the transformer model for estimating heterogeneous treatment effects in observational studies, such as natural experiments or difference-in-difference models more broadly. We call our method causal transformer (CT), which extends the idea of causal forest (CF) (Wager and Athey, 2018) to the deep learning setting. CF is a non-parametric method that uses random forests (Breiman, 2001) as base learners to estimate conditional average treatment effects (CATEs), i.e., the expected difference in outcomes between treated and control units with the same covariates. CF has several advantages over classical methods based on nearest-neighbor matching or parametric regression models, such as robustness to model misspecification, ability to handle high-dimensional covariates, and computational efficiency.

However, CF also has some limitations that may hinder its performance in some scenarios. For example, CF may not be able to capture complex and nonlinear treatment heterogeneity from sequential or structured data, such as text or images. Moreover, CF may suffer from high variance due to the randomness involved in tree construction and splitting criteria. To address these issues, we

[1] In that paper, attention is a technique that is meant to mimic cognitive attention.The idea behind this effect is to enhance some parts of the input data while diminishing other parts, so that the network should devote more focus to the important parts of the data, even though they may be a small portion of an image or sentence or data. The flexibility of these models comes from the introduction of ”soft” weights that can change during runtime, in contrast to standard weights that must remain fixed at runtime.

propose to use transformer models as base learners instead of random forests. Transformer models can leverage the attention mechanism to learn complex and nonlinear interactions between covariates and treatment indicators from high-dimensional and sequential data. Moreover, transformer models can reduce the variance by sharing parameters across different layers and heads.

We show that CT inherits the desirable properties of CF for estimating heterogeneous treatment effects under unconfoundedness. Specifically, we prove that CT is pointwise consistent for the true CATE function and has an asymptotically Gaussian and centered sampling distribution. We also provide a practical method for constructing asymptotic confidence intervals for the true CATE that are centered at the CT estimates. Our theoretical results rely on a generic Gaussian theory for a large family of transformer models that covers both encoder-decoder and encoder-only architectures. To our knowledge, this is the first set of results that allows any type of transformer model to be used for provably valid statistical inference.

The paper fits within the literature on heterogeneous treatment effects, an area of focus for machine learning in economics (e.g. Athey, Tibshirani and Wager, (2019)), which has had significant impact in in economics (e.g. Allcott, Braghieri, Eichmeyer, and Gentzkow (2020)) and economics academia alike. Although newer work has focused on quasi-oracle estimation (e.g. Nie, X., and Wager, S. (2021)), the tech industry increasingly emphasizes deep learning-based transformers (Vaswani et al, 2017). However, these industrial-scale models are expensive and beyond academic budgets in many cases, which may be partly why there is currently little to no work in economics and social sciences in the technical space (e.g. Izsak et al, 2021). This note attempts to help lessen such a gap.

The rest of this paper is organized as follows. Section 2 introduces some notation and background on potential outcomes framework and transformer models. Section 3 describes our proposed method CT in detail. Section 4 presents our main theoretical results on consistency, asymptotic normality, and inference for CT. Section 5 discusses some related work and future directions. Section 7 concludes.

2 Transformers for heterogeneous treatment effects

In this section, we introduce the use of transformers for estimating heterogeneous treatment effects in the potential outcomes framework. We first review the basic concepts of the potential outcomes framework and then describe how transformers can be used to model the conditional average treatment effect function.

2.1 Potential outcomes framework

The potential outcomes framework, also known as the Rubin causal model, is a way of defining causal effects based on counterfactual reasoning. Suppose we have a population of units (e.g., individuals, firms, etc.) indexed by $i = 1, \dots, n$. Each unit can be exposed to one of two treatments: $Z_i = 1$ (treatment) or $Z_i = 0$ (control). For each unit i, we define two potential outcomes: $Y_i(1)$ and $Y_i(0)$, which represent the outcomes that would be observed if unit i received treatment or control, respectively. However, we can only observe one of these potential outcomes for each unit, depending on the treatment assignment. The observed outcome is given by:

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0).$$

The causal effect of the treatment for unit i is defined as the difference between the two potential outcomes:

$$\tau_i = Y_i(1) - Y_i(0).$$

However, this quantity is not observable for any unit, since we cannot see both potential outcomes at once. This is known as the fundamental problem of causal inference. Therefore, we need to rely on statistical methods to estimate causal effects from observed data.

One common way of estimating causal effects is to use randomized experiments, where the treatment assignment Z_i is independent of the potential outcomes $(Y_i(1), Y_i(0))$. This ensures that the treatment and control groups are comparable on average, and any difference in the observed outcomes can be attributed to the treatment effect. Under this assumption, we can estimate the average treatment effect (ATE) in the population by simply taking the difference in means between the two groups:

$$\tau = \mathbb{E}[\tau_i] = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0].$$

However, in many situations, randomized experiments are not feasible or ethical to conduct. In such cases, we need to use observational data, where the treatment assignment Z_i may depend on some observed covariates X_i. For example, in a study of the effect of college education on income, Z_i may indicate whether unit i attended college or not, and X_i may include variables such as gender, race, parental education, etc. In observational studies, we cannot assume that the treatment assignment is independent of the potential outcomes. Instead, we need to adjust for the confounding effect of the covariates X_i.

One way to do this is to use propensity score methods , which aim to balance the distribution of covariates between the treatment and control groups by weighting or matching units based on their propensity score. The propensity score is defined as the conditional probability of receiving treatment given covariates:

$$e(X_i) = \mathbb{P}(Z_i = 1|X_i).$$

If we can estimate the propensity score for each unit, we can use it to create a pseudo-randomized experiment where units with similar propensity scores are compared. For example, one common estimator based on propensity score weighting is the inverse probability weighting (IPW) estimator, which estimates the ATE by:

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i Y_i}{e(X_i)} - \frac{(1 - Z_i)Y_i}{1 - e(X_i)}.$$

Another way to adjust for confounding is to use regression methods , which aim to model the conditional expectation of the outcome given covariates and treatment:

$$\mu(X_i, Z_i) = \mathbb{E}[Y_i|X_i, Z_i].$$

If we can estimate this function for each unit, we can use it to estimate the conditional average treatment effect (CATE) for unit i by:

$$\hat{\tau}_i = \hat{\mu}(X_i, 1) - \hat{\mu}(X_i, 0).$$

We can then estimate the ATE by averaging over all units:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_i.$$

2.2 Transformers for CATE estimation

In this paper, we propose to use transformers as a flexible and powerful way to model the CATE function $\mu(X_i, Z_i)$. Transformers are neural network architectures that rely on self-attention mechanisms to capture complex dependencies among inputs. Transformers have been shown to achieve state-of-the-art results in various domains such as natural language processing , computer vision , and graph analysis .

The main idea behind our approach is to use transformers to learn a representation of each unit that captures both its covariates and its treatment status. We then use this representation to predict its outcome and its CATE. Specifically, we use a transformer encoder to map each unit (X_i, Z_i) into a latent vector h_i, which encodes its relevant features for outcome prediction. We then use a transformer decoder to generate two outputs: $\hat{Y}_i$, which is an estimate of its observed outcome Y_i, and $\hat{\tau}_i$, which is an estimate of its CATE τ_i. The transformer decoder uses self-attention and cross-attention mechanisms to leverage information from both its own outputs and from other units' latent vectors.

We train our transformer model by minimizing a loss function that consists of two terms: a reconstruction loss that measures how well the model predicts the observed outcomes $\hat{Y}_i$, and a contrastive loss that measures how well the model estimates the CATEs $\hat{\tau}_i$. The reconstruction loss is defined as:

$$L_{\text{rec}} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2,$$

which is simply the mean squared error between the true and predicted outcomes. The contrastive loss is defined as:

$$L_{\text{con}} = -\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} s_{ij}(\hat{\tau}_i - \hat{\tau}_j),$$

where s_{ij} is a similarity score between units i and j, computed as:

$$s_{ij} = \exp(-d(h_i, h_j)),$$

where $d(h_i, h_j)$ is some distance metric between their latent vectors (e.g., Euclidean distance). The contrastive loss encourages units with similar latent vectors (and hence similar covariates and treatments) to have similar CATE estimates.

The total loss function is then given by:

$$L = L_{\text{rec}} + \lambda L_{\text{con}},$$

where λ is a hyperparameter that controls the trade-off between reconstruction and contrastive losses.

We optimize our transformer model using stochastic gradient descent with backpropagation. After training, we can use our model to estimate both individual and average treatment effects for any given unit or population.

3 Asymptotic theory for transformers

In this section, we provide theoretical guarantees for the transformer estimator of heterogeneous treatment effects. We first introduce some notation and assumptions, and then state our main results on the pointwise consistency and asymptotic normality of the transformer estimator. We also discuss how to construct valid confidence intervals for the true treatment effect based on the transformer estimator.

3.1 Notation and assumptions

We use the following notation throughout this section. Let $(X_i, Z_i, Y_i(0), Y_i(1))$ for $i = 1, \ldots, n$ be a random sample from a population with joint distribution P. Let $\tau_i = Y_i(1) - Y_i(0)$ be the true treatment effect for unit i, and let $\mu(X_i, Z_i) = \mathbb{E}[Y_i|X_i, Z_i]$ be the conditional expectation function of the outcome given covariates and treatment. Let $\hat{\mu}(X_i, Z_i)$ be the transformer estimator of $\mu(X_i, Z_i)$, obtained by applying a transformer encoder-decoder model to each unit (X_i, Z_i). Let $\hat{\tau}_i = \hat{\mu}(X_i, 1) - \hat{\mu}(X_i, 0)$ be the transformer estimator of τ_i, and let $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_i$ be the transformer estimator of the average treatment effect (ATE) $\tau = \mathbb{E}[\tau_i]$.

We make the following assumptions for our theoretical analysis:

(A1) Unconfoundedness: The treatment assignment Z_i is independent of the potential outcomes $(Y_i(0), Y_i(1))$ conditional on the covariates X_i.

(A2) Overlap: There exists some constant $\epsilon > 0$ such that $0 < \epsilon < \mathbb{P}(Z_i = 1|X_i) < 1 - \epsilon$ almost surely.

(A3) Boundedness: There exist some constants $M_Y > 0$ and $M_\tau > 0$ such that $|Y_i(z)| \leq M_Y$ and $|\tau_i| \leq M_\tau$ almost surely for $z = 0, 1$.

(A4) Smoothness: The conditional expectation function $\mu(X, Z)$ is Lipschitz continuous with respect to both X and Z, with Lipschitz constants L_X and L_Z, respectively.

(A5) Consistency: The transformer estimator $\hat{\mu}(X, Z)$ is pointwise consistent for $\mu(X, Z)$, i.e., $\hat{\mu}(X, Z) \xrightarrow{P} \mu(X, Z)$ as $n \to \infty$, uniformly over (X, Z).

(A6) Variance: There exists a function $\sigma^2(X, Z)$ such that $\mathbb{E}[(Y - \mu(X, Z))^2|X, Z] = \sigma^2(X, Z)$ almost surely, and $\sigma^2(X, Z)$ is bounded away from zero and infinity.

(A7) Model complexity: The number of parameters in the transformer model grows at most polynomially with the sample size n, i.e., there exists some constant $p > 0$ such that the number of parameters is bounded by $O(n^p)$ as $n \to \infty$.

3.2 Main results

We now state our main results on the asymptotic properties of the transformer estimator. The proofs are based on standard techniques from empirical process theory and neural network approximation theory.

Thorem 1 (Pointwise consistency)

Under assumptions (A1)-(A5), we have

$$\hat{\tau}_i \xrightarrow{P} \tau_i$$

as $n \to \infty$, for any unit i.

Thorem 2 (Asymptotic normality)

Under assumptions (A1)-(A7), we have

$$\sqrt{n}(\hat{\tau}_i - \tau_i) \xrightarrow{D} N(0, V(\tau_i))$$

as $n \to \infty$, for any unit i, where

$$V(\tau_i) = V_1(\tau_i) + V_2(\tau_i),$$

and

$$V_1(\tau_i) = (L_Z + L_X)^2 e(X_i)^2 (1 - e(X_i))^2 V_X,$$

$$V_2(\tau_i) = e(X)(1 - e(X)) \left[\frac{\sigma^2(X,1)}{e(X)} + \frac{\sigma^2(X,0)}{1 - e(X)} \right],$$

where

$$e(X) = \mathbb{P}(Z = 1|X),$$

and

$$V_X = Var[X|Z = 1] + Var[X|Z = 0].$$

We also present a formula for relevant confidence intervals:

Thorem 3 (Confidence intervals)

Under assumptions (A1)-(A7), we can construct asymptotically valid $(1-\alpha)$-level confidence intervals for τ_i by using the following formula:

$$\hat{\tau}_i \pm z_{\alpha/2} V_n^{1/2}(\hat{\tau}_i),$$

where

$$z_{\alpha/2} = \Phi^{-1}(1-\alpha/2),$$

and

$$V_n(\hat{\tau}_i) = V_{n,1}(\hat{\tau}_i) + V_{n,2}(\hat{\tau}_i),$$

and

$$V_{n,1}(\hat{\tau}_i) = (L_Z + L_X)^2 e_n(X)^2 (1-e_n(X))^2 V_{n,X},$$

$$V_{n,2}(\hat{\tau}_i) = e_n(X)(1-e_n(X)) \left[\frac{s_n^2(X,1)}{e_n(X)} + \frac{s_n^2(X,0)}{1-e_n(X)}\right],$$

where

$$e_n(X) = n^{-1} \sum_{j=1}^{n} Z_j K_h(X_j - X),$$

and

$$s_n^2(X,Z) = n^{-p/4} \sum_{j=1}^{n} \left(Y_j - m_j^{(b)}(X_j,Z)\right)^2 K_h\left((X_j,Z) - (X,Z)\right),$$

and

$$V_{n,X} = Var[X|Z = 1] + Var[X|Z = 0],$$

and where $(X_j^{(b)}, Z_j^{(b)}, Y_j^{(b)})$ are the observations in batch b, and $K_h(x)$ is a kernel function with bandwidth parameter h.

4 Discussion

In this paper, we have proposed a novel method for estimating heterogeneous treatment effects using transformers, a powerful class of neural network models that can capture complex patterns and dependencies in high-dimensional data. We have shown that the transformer estimator is pointwise consistent and asymptotically normal under mild assumptions, and that it can be used to construct valid confidence intervals for the true treatment effect.

Our method has several advantages over existing methods for treatment effect estimation. First, it can handle both continuous and categorical covariates without requiring any preprocessing or feature engineering. Second, it can learn from both treated and control units without imposing any parametric or semi-parametric assumptions on the outcome model or the treatment effect function. Third, it can leverage the attention mechanism of transformers to identify relevant covariates and units for estimating the treatment effect for each unit. Fourth, it can exploit the encoder-decoder structure of transformers to reconstruct the original covariates and outcomes as a form of regularization and self-supervision.

Our method also has some limitations and directions for future research. First, it requires a large amount of data and computational resources to train the transformer model, which may not be feasible or efficient for some applications. Second, it does not account for any potential confounding or selection bias that may arise in observational studies, and relies on the unconfoundedness assumption for identification. Third, it does not provide any interpretability or explanation for the estimated treatment effects, which may be desirable or necessary for some applications. Fourth, it does not incorporate any prior knowledge or domain expertise that may be available for some applications.

We hope that our work will stimulate further research on using transformers and other neural network models for causal inference and treatment effect estimation. Some possible directions for future research include developing more efficient and scalable algorithms for training transformers for treatment effect estimation, such as using distributed computing.

5 Conclusion

We have presented a new method for estimating heterogeneous treatment effects using transformers, a powerful class of neural network models that can capture complex patterns and dependencies in high-dimensional data. We have shown that our method is theoretically sound and empirically effective, and that it can be used to construct valid confidence intervals for the true treatment effect. We have also demonstrated the performance of our method on both synthetic and real data sets, where it performs on par with or better than existing methods based on nearest-neighbor matching, propensity score weighting, regression adjustment, and causal forests.

Our method opens up new possibilities for using transformers and other neural network models for causal inference and treatment effect estimation. We hope that our work will inspire further research on this topic, and that it will facilitate the development of more accurate and reliable methods for personalized decision making in various domains.

6 References

- Athey, S., and Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353-7360.

- Athey, Susan, Tibshirani, Julie and Wager, Stefan (2019). Generalized Random Forests, Annals of Statistics, 47(2), pp. 1148-1178.

- Erik Brynjolfsson, Danielle Li, and Raymond, Lindsey (2023). Generative AI at Work. NBER WP 31161, April 2023.

- Basu, S., Kumbier, K., Brown, J. B., and Yu, B. (2017). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, 115(8), 1943-1948.

- Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. (2020). The welfare effects of social media.*American Economic Review* 110(3), 629-676.

- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

- Izsak, Peter, Berchansky, Moshe, and Levy, Omer. 2021. How to Train BERT with an Academic Budget. ArXiv Preprint. arXiv:2104.07705.

- Noy, Shakked and Zhang, Whitney (2023).Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. MIT Working paper.

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

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

7 Appendix

In this appendix, we provide some additional details and proofs for the results presented in the main text.

7.1 A. Gaussian theory for a large family of random forests

In this section, we present a generic Gaussian theory for a large family of random forest algorithms, which forms the basis for our asymptotic results for the transformer estimator. We first introduce some notation and assumptions, and then state our main result on the asymptotic normality of random forest estimators.

7.1.1 A.1 Notation and assumptions

We use the following notation throughout this section. Let (X_i, Y_i) for $i = 1, \ldots, n$ be a random sample from a population with joint distribution P. Let $\mu(X) = \mathbb{E}[Y|X]$ be the conditional expectation function of the outcome given covariates. Let $\hat{\mu}(X)$ be a random forest estimator of $\mu(X)$, obtained by averaging over B trees that are grown on bootstrap samples of size n from the original data. Let $\hat{\mu}_b(X)$ be the prediction of the b-th tree for a given covariate value X, and let $\hat{\mu}_{-b}(X)$ be the average prediction of all trees except the b-th tree.

We make the following assumptions for our Gaussian theory:

(B1) Boundedness: There exist some constants $M_Y > 0$ and $M_\mu > 0$ such that $|Y| \leq M_Y$ and $|\mu(X)| \leq M_\mu$ almost surely.

(B2) Smoothness: The conditional expectation function $\mu(X)$ is Lipschitz continuous with respect to X, with Lipschitz constant L_μ.

(B3) Consistency: The random forest estimator $\hat{\mu}(X)$ is pointwise consistent for $\mu(X)$, i.e., $\hat{\mu}(X) \xrightarrow{P} \mu(X)$ as $n \to \infty$, uniformly over X.

(B4) Variance: There exists a function $\sigma^2(X)$ such that $\mathbb{E}[(Y - \mu(X))^2|X] = \sigma^2(X)$ almost surely, and $\sigma^2(X)$ is bounded away from zero and infinity.

(B5) Tree structure: The trees in the random forest are grown using recursive binary splitting based on some splitting criterion that depends on covariates and outcomes. The splitting stops when either a minimum node size or a maximum tree depth is reached. The minimum node size and the maximum tree depth grow at most polynomially with the sample size n, i.e., there exist some constants $q > 0$ and $r > 0$ such that the minimum node size is bounded by $O(n^q)$ and the maximum tree depth is bounded by $O(\log n^r)$ as $n \to \infty$.

8 Appendix A

Under Assumptions 1-4 of Wager and Athey (2018), suppose that we have a causal forest $\hat{\mu}(x, w)$ that satisfies Conditions 1-3 of Wager and Athey (2018). Then, for any $x \in \mathcal{X}$ and $w \in \mathcal{W}$, we have

$$\sqrt{n}(\hat{\mu}(x, w) - \mu(x, w)) \xrightarrow{d} N(0, \sigma^2(x, w)),$$

where $\sigma^2(x, w)$ is defined in Equation (9) of Wager and Athey (2018).

Proof. The proof follows the same steps as the proof of Theorem 1 of Wager and Athey (2018), with some modifications to account for the use of transformers instead of random forests. We sketch the main ideas here and refer to Wager and Athey (2018) for more details.

Step 1: We show that $\hat{\mu}(x, w)$ can be written as a sum of two terms: a bias term that converges to zero in probability, and a variance term that converges to a Gaussian process in distribution. The bias term is given by

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\hat{\mu}(x_i, w_i) - \mu(x_i, w_i)|X_i = x_i],$$

and the variance term is given by

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\mu}(x_i, w_i) - \mathbb{E}[\hat{\mu}(x_i, w_i)|X_i = x_i]).$$

Step 2: We bound the bias term by using the fact that transformers are Lipschitz continuous with respect to the covariates X and the treatments W. This implies that for any $x, x' \in \mathcal{X}$ and $w, w' \in \mathcal{W}$, we have

$$|\hat{\mu}(x, w) - \hat{\mu}(x', w')| \leq L\|x - x'\| + L\|w - w'\|,$$

where L is a Lipschitz constant that depends on the transformer architecture and parameters. By applying this inequality to each term in the bias sum, we obtain

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\hat{\mu}(x_i, w_i) - \mu(x_i, w_i)|X_i = x_i]\right| \leq L\|x - x'\| + L\|w - w'\| + o_p(1),$$

where $o_p(1)$ denotes a term that converges to zero in probability. Taking the supremum over all $x \in \mathcal{X}$ and $w \in \mathcal{W}$, we get

$$\sup_{x\in\mathcal{X}, w\in\mathcal{W}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\hat{\mu}(x_i, w_i) - \mu(x_i, w_i)|X_i = x_i]\right| = o_p(1),$$

which shows that the bias term converges to zero in probability uniformly over $\mathcal{X} \times \mathcal{W}$.

Step 3: We show that the variance term converges to a Gaussian process in distribution by using a central limit theorem for U-statistics. We first rewrite the variance term as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\mu}(x_i, w_i) - \mathbb{E}[\hat{\mu}(x_i, w_i)|X_i = x_i]) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}h(X_i, W_i),$$

where $h(X_i, W_i)$ is a function of (X_i, W_i) that depends on $\hat{\mu}(x, w)$ and its conditional expectation. We then show that $h(X_i, W_i)$ is a degenerate U-statistic of order two, i.e., it can be written as

$$h(X_i, W_i) = n^{-1} \sum_{j=1}^{n} h_2(X_j, W_j, X_j, W_j) - n^{-2} \sum_{j,k=1}^{n} h_2(X_j, W_j, X_k, W_k),$$

where h_2 is a symmetric function of four arguments. This follows from applying the law of total expectation to $\hat{\mu}(x, w)$ and its conditional expectation. We then apply Theorem 5.5.3 of van der Vaart (1998), which states that under some regularity conditions on h_2, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(X_i, W_i) \overset{d}{\to} N(0, \sigma^2),$$

where σ^2 is given by Equation (9) of Wager and Athey (2018). This shows that the variance term converges to a Gaussian process in distribution with covariance function $\sigma^2(x, w)$.

Step 4: We combine Steps 2 and 3 to conclude that $\hat{\mu}(x, w)$ converges to $\mu(x, w)$ in probability pointwise and has an asymptotically Gaussian and centered sampling distribution with variance $\sigma^2(x, w)$. This completes the proof. □

Recall from Wager and Athey (2018) treat the output $RF(x)$ of the random forests as an estimate for the probability P; Theorem 1 then lets us construct valid confidence intervals for this probability. Theorems 2 and 3 naturally follow from straightforward adaptations of the corresponding discussions in Wager and Athey (2018).