# Wasserstein Random Forests at First Glance

Qiming Du

February 17, 2020

## Contents

# 1 Overview

## 1.1 Motivation

The classical setting of supervised learning focus on the estimation of the conditional expectation $\mathbf{E}[Y \mid X = x]$ for some underlying 1-dimensional objective $Y$ and a multidimensional covariate $X$. In many real-world applications, it is also important to track the additional information encoded in the conditional distribution $\mathcal{L}(Y \mid X = x)$. For example, in Heterogeneous Treatment Effect (HTE), the traditional object of interest is the *Conditional Average Treatment Effect* (CATE) function, defined by

$$\tau(x) = \mathbf{E}[Y(1) - Y(0) \mid X = x],$$

where $Y(1)$ (resp. $Y(0)$) denotes the *potential outcomes* (see, e.g., [Rub74] and [IR15]) of treatment (resp. no treatment). The data is usually of form $(Y_i(W_i), W_i, X_i; 1 \le i \le N)$, where $W_i$ denotes the treatment assignment indicators. Recently, many approaches based on modern statistical learning are investigated on the estimation of the CATE function (see, e.g., [KSBY17, AW19, NW17]).

However, the information provided by conditional expectation

$$\mu_0 := \mathbf{E}[Y(0) \mid X = x] \quad \text{and} \quad \mu_1 := \mathbf{E}[Y(1) \mid X = x]$$

is limited. For example. when the conditional distribution of $Y(1)$ (resp. $Y(0)$) is multimodal, the conditional expectation can not provide insights on the causal inference. A simple inference based on the CATE function may also be dangerous without any information of the conditional distribution of $Y(0)$ and $Y(1)$.

In an ideal case, one is interested to estimate the joint conditional distribution

$$\mathcal{L}((Y(0), Y(1)) \mid X = x).$$

Unfortunately, by the essential of the problem, there is not a single sample of $(Y_i(0), Y_i(1))$ at point $X_i$ that can be collected in the observational study. Unlike the conditional expectation, the dependence between $Y(0)$ and $Y(1)$ is much more complex. It is in general difficult to design a strategy to estimate the covariance of $Y(0)$ and $Y(1)$ based on a similar strategy used to obtain the estimation of the conditional expectation. Therefore, it is extremely difficult to estimate the joint distribution $\mathcal{L}((Y(0), Y(1)) \mid X = x)$ in practice without any further assumptions on the dependence of $Y(0)$ and $Y(1)$. As a consequence, we concentrate on a subproblem: instead of estimating $\mathcal{L}((Y(0), Y(1)) \mid X = x)$, we are interested in the estimation of the conditional marginal distributions $\mathcal{L}(Y(0) \mid X = x)$ and $\mathcal{L}(Y(1) \mid X = x)$. By considering the two subgroupes

$$\{(X_i, Y(W_i)) \mid W_i = 0, 1 \le i \le N\} \quad \text{and} \quad \{(X_i, Y(W_i)) \mid W_i = 1, 1 \le i \le N\},$$

the subproblem thus enters in a classical supervised learning context (cf. T-learners [KSBY17]), while the objective is replaced by estimating the conditional distribution.

## 1.2 Fundamental idea

The success of Random Forests (RF) [Bre01] has be proven in many real-world applications. If we look at the final prediction at each point $x$ provided by the RF algorithm, it can be regarded as a weighted average of some $(Y_i; 1 \le i \le N)$ in the training dataset, where the weights are some random variables depending on the covariates $(X_i; 1 \le i \le N)$. A very natural idea is to use this weighted empirical measure to approximate the conditional distribution. This is also the fundamental idea in the construction of forests quantile regression [Mei06, AW19] and other literatures that combine the kernel density estimations and Random Forests[cite lot of papers]. To achieve this, the splitting criterion should be modified, in order to be compatible with conditional distribution estimation.

A natural metric to measure the distance of two measures is Wasserstein distance. More precisely, the Wasserstein distance $\mathcal{W}_p$ between two measures $\mu, \nu$ on $\mathbf{R}^d$ is defined by

$$\mathcal{W}_p(\mu,\nu) := \left(\inf_{\gamma\in\Gamma(\mu,\nu)}\int |x-y|^p\,\gamma(dx,dy)\right)^{\frac{1}{p}},$$

where $|\cdot|$ is the Euclidean norm in $\mathbf{R}^d$ and $\Gamma(\mu,\nu)$ denotes the set of all couplings of $\mu$ and $\nu$, namely,

$$\int_{y\in\mathbf{R}^q}\gamma(dx,dy)=\mu(dx)\quad\text{and}\quad\int_{x\in\mathbf{R}^q}\gamma(dx,dy)=\nu(dy).$$

When $\mu$ and $\nu$ are the probability measures on $\mathbf{R}$, it is easily checked that

$$\mathcal{W}_p(\mu,\nu) := \left(\int_0^1\left|F_\mu^{-1}(u)-F_\nu^{-1}(u)\right|^p du\right)^{\frac{1}{p}}, \tag{1}$$

where $F_\mu^{-1}(u)$ (resp. $F_\nu^{-1}(u)$) is the generalized inverse distribution function defined by

$$F_\mu^{-1}(u) := \inf\left\{x\in\mathbf{R}\;\middle|\;F_\mu(x)\geq u\right\},$$

with $F_\mu(x)$ the cumulative distribution function of $\mu$. Denote by $E$ the support of $X$ and denote by $\pi_x(dy)$ the probability measure associated to $\mathcal{L}(Y\mid X=x)$. Now, let us provide some possible regularity assumptions in the following.

**Assumption 1** (Lipschitz). *For any $(x,y)\in E\times E$, we assume that there exists a constant $C_\pi<+\infty$ such that*

$$\mathcal{W}_2(\pi_x,\pi_y)\leq C_\pi\,|x-y|\,.$$

**Assumption 2** (Hölder). *For any $(x,y)\in E\times E$, we assume that there exists $\alpha>0$ and a constant $C_\pi^\alpha<+\infty$ such that*

$$\mathcal{W}_2(\pi_x,\pi_y)\leq C_\pi^\alpha\,|x-y|^\alpha\,.$$

The ultimate goal is to construct an estimator $\pi_x^N$ of $\pi_x$ such that for all $x\in E$, we have

$$W_2(\pi_x^N,\pi_x)\xrightarrow[N\to\infty]{a.s.\text{ or }\mathbf{P}\text{ or }\mathbb{L}^p}0.$$

This is inspired by the classical consistence analysis of Random Forests, where the conditional expectation is often assumed to be Lipschitz or Hölder w.r.t. covariate.

### 1.3 Splitting criterion

We present an intuitive idea on how to perform the splitting which is compatible with the Wasserstein distance. Roughly speaking, we try to maximize the $\mathcal{W}_2$-distance between the empirical measures w.r.t. $Y_i$ of the newly created left node and right node, which obeys the same philosophy of the classical Random Forests: the splitting is performed greedily in order to maximize the homogeneity between the subgroups. This criterion also works when the dimension of $Y$ is greater than 1. However, in this case, the calculation of $\mathcal{W}_2$-distance may becomes intractable (with $\mathcal{O}(N^3)$ time complexity where $N$ represents the sum of the number of data in the two associated empirical measures). Some regularization strategies (such as sinkhorn distance, i.e., Optimal Transport with entropic constraints) may be applied. In Breiman's original Random Forests, the splitting is performed to maximize the inter-group variance, which is compatible with the conditional expectation estimation as the conditional expectation minimize the $\mathbb{L}^2$-error. The underlying philosophy is to maximize the heterogeneity between the two subgroups. Our idea is similar: we use Wasserstein distance to measure the distance between distributions, and the splitting is done by maximizing the Wasserstein distance between the associated empirical measures in the two subgroups.

**Toy example** Before proceeding further, we illustrate the splitting mechanism in a 2-dimensional toy example. Let $X = (X^{(1)}, X^{(2)})$ be the canonical random variable of covariate, and denote by $Y$ the associate objective. Considering a dataset $(X_i, Y_i; 1 \leq i \leq 8)$ with $n = 8$, we explain the first splitting in the direction (1) (see Figure 1). Fixing a splitting point $z$, we denote respectively $A_L$ and $A_R$ the data points $X_i$ in the left and right sides of $z$. We denote $\pi_L$ and $\pi_R$ the empirical measures associated respectively to $A_L$ and $A_R$, i.e.,

$$\pi_L = \frac{1}{\#(A_L)} \sum_{i=1}^{n} \delta_{Y_i} \mathbf{1}_{\{X_i \in A_L\}},$$

and

$$\pi_R = \frac{1}{\#(A_R)} \sum_{i=1}^{n} \delta_{Y_i} \mathbf{1}_{\{X_i \in A_R\}}.$$

The best splitting point $z^*$ is determined by

$$z^* = \arg\max_z \mathcal{W}_p(\pi_L, \pi_R).$$

Note that the splitting $z$ is always performed in the middle of two consecutive data points along the direction (1) in order to remove the possible ties in the arg max.

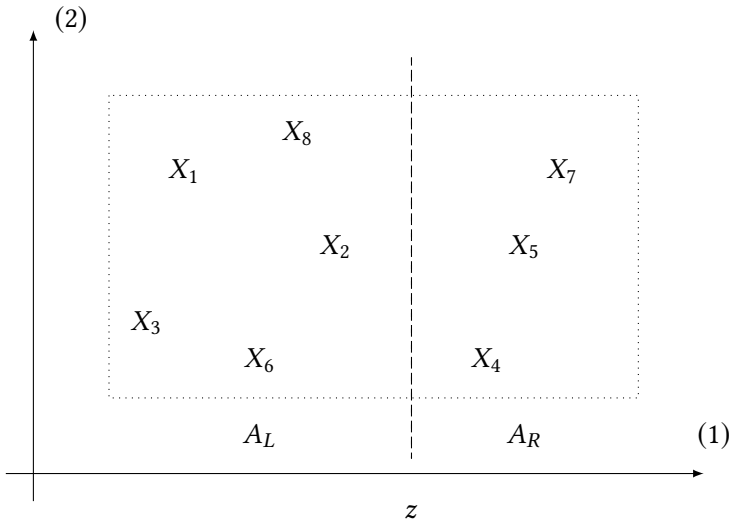


**Figure 1:** An illustration of splitting.

More concretely, in Figure 1 the Wasserstein distance to be evaluated is

$$\mathcal{W}_p\left(\frac{1}{5}(\delta_{Y_1} + \delta_{Y_2} + \delta_{Y_3} + \delta_{Y_6} + \delta_{Y_8}), \frac{1}{3}(\delta_{Y_4} + \delta_{Y_5} + \delta_{Y_7})\right).$$

The computation can be done by the equation given in (1), with $\mathcal{O}(N \log(N))$ time complexity, where $N$ is the number of the data points.

## 2 Mechanism

We use the notation compatible with [BS15]:

- $\forall n \in \mathbf{N}^*$, $[n] := \{1, 2, \ldots, n\}$;
- $\mathcal{D}_n := ((X_1, Y_1), \ldots, (X_n, Y_n))$;

- $\Theta$: generic random variable to describe the randomness of tree's construction;
- $\Theta_j$: random variable associated with $j$-th tree;
- $\mathcal{D}_n^*(\Theta_j)$: dataset used to construct the $j$-th tree;
- $A_n(x;\Theta_j,\mathcal{D}_n)$: the cell in the $j$-th tree that contains the point $x$;
- $N_n(x;\Theta_j,\mathcal{D}_n)$: the number of points in $\mathcal{D}_n^*(\Theta_j)$ that fall into $A_n(x;\Theta_j,\mathcal{D}_n)$;
- $m_n(x;\Theta_j,\mathcal{D}_n)$: prediction of conditional expectation given by $j$-th tree, defined by

$$m_n(x;\Theta_j,\mathcal{D}_n) = \sum_{i\in\mathcal{D}_n^*(\Theta_j)} \frac{Y_i \mathbf{1}_{\{X_i\in A_n(x;\Theta_j,\mathcal{D}_n)\}}}{N_n(x;\Theta_j,\mathcal{D}_n)};$$

- $\mu_n(x;\Theta_j,\mathcal{D}_n)$: prediction of conditional distribution given by $j$-th tree, defined by

$$\mu_n(x;\Theta_j,\mathcal{D}_n) = \sum_{i\in\mathcal{D}_n^*(\Theta_j)} \frac{\mathbf{1}_{\{X_i\in A_n(x;\Theta_j,\mathcal{D}_n)\}}}{N_n(x;\Theta_j,\mathcal{D}_n)}\delta_{Y_i};$$

- $m_{M,n}(x;\Theta_j,\mathcal{D}_n)$: prediction of conditional expectation given by a Random Forest that contains $M$ trees, that is

$$\begin{aligned} m_{M,n}(x;\Theta_j,\mathcal{D}_n) =& \frac{1}{M}\sum_{j=1}^{M} m_n(x;\Theta_j,\mathcal{D}_n) \\ =& \frac{1}{M}\sum_{j=1}^{M}\sum_{i\in\mathcal{D}_n^*(\Theta_j)} \frac{Y_i \mathbf{1}_{\{X_i\in A_n(x;\Theta_j,\mathcal{D}_n)\}}}{N_n(x;\Theta_j,\mathcal{D}_n)} \\ & (\text{when } \bigcup_{j=1}^{M}\mathcal{D}_n^*(\Theta_j) = [n]) \\ =& \sum_{i=1}^{n} \underbrace{\sum_{j=1}^{M} \frac{\mathbf{1}_{\{X_i\in A_n(x;\Theta_j,\mathcal{D}_n)\}}}{MN_n(x;\Theta_j,\mathcal{D}_n)}}_{\alpha_i(x)} Y_i; \end{aligned}$$

- $\mu_{M,n}(x;\Theta_j,\mathcal{D}_n)$: prediction of conditional distribution given by a Random Forest that contains $M$ trees, that is

$$\begin{aligned} \mu_{M,n}(x;\Theta_j,\mathcal{D}_n) =& \frac{1}{M}\sum_{j=1}^{M} \mu_n(x;\Theta_j,\mathcal{D}_n) \\ & (\text{when } \bigcup_{j=1}^{M}\mathcal{D}_n^*(\Theta_j) = [n]) \\ =& \sum_{i=1}^{n} \alpha_i(x)\delta_{Y_i}. \end{aligned}$$

Roughly speaking, Wasserstein Random Forests (WRF) are constructed by the same stochastic mechanism of Breiman's Random Forests. The main differences are splitting criterion and output of the prediction. As a consequence, for tuning the algorithm, we consider the following parameters that are identical to Breiman's Random Forests:

(i) $a_n \in [n]$: the number of data points sampled in each tree;

(ii) **mtry**$\in [p]$: the number of possible splitting directions to be explored at each node of each tree;

(iii) **nodesize** $\in [a_n]$: the number of points below which the splitting is forced to be rejected.

Now, let us provide the mechanism of the Wasserstein Random Forests:

# 3 Applications

## 3.1 Uncertainty control of CATE function

A by-product is to provide an uncertainty control of CATE function, namely, we are interested in estimating the probability of the following form:

$$\mathbf{P}\left(|(Y(1)-Y(0))-\tau(x)|>a \mid X=x\right). \tag{2}$$

Before proceeding further, we introduce some additional notation to simplify the writing. We denote

$$\mu_{00}(x) := \mathbf{E}\left[Y(0)^2 \,\middle|\, X=x\right],$$

and

$$\mu_{11}(x) := \mathbf{E}\left[Y(1)^2 \,\middle|\, X=x\right],$$

as well as

$$\mu_{01}(x) := \mathbf{E}\left[Y(0)Y(1) \mid X=x\right].$$

The conditional variances are denoted respectively by

$$\begin{aligned}\sigma_{00}(x) :=& \mathbf{E}\left[Y(0)^2 \,\middle|\, X=x\right] - \mathbf{E}\left[Y(0) \mid X=x\right]^2,\\ =& \mu_{00}(x) - \mu_0(x)^2,\end{aligned}$$

and

$$\begin{aligned}\sigma_{11}(x) :=& \mathbf{E}\left[Y(1)^2 \,\middle|\, X=x\right] - \mathbf{E}\left[Y(1) \mid X=x\right]^2,\\ =& \mu_{11}(x) - \mu_1(x)^2.\end{aligned}$$

Thanks to Chebyshev's inequality, we have

$$\begin{aligned}&\mathbf{P}\left(|(Y(1)-Y(0))-\tau(x)|>a \mid X=x\right)\\ \leq& \frac{\mathrm{Var}\left[Y(1)-Y(0) \mid X=x\right]}{a^2}\\ =& \frac{\sigma_{00}(x)+\sigma_{11}(x)-2\mu_{01}(x)+2\mu_0(x)\mu_1(x)}{a^2}.\end{aligned} \tag{3}$$

Moreover, Cauchy-Schwarz inequality gives

$$|\mu_{01}(x)| \leq \sqrt{\mu_{00}(x)\mu_{11}(x)},$$

whence

$$\begin{aligned}&\mathbf{P}\left(|(Y(1)-Y(0))-\tau(x)|>a \mid X=x\right)\\ \leq& \frac{\sigma_{00}(x)+\sigma_{11}(x)+2\mu_0(x)\mu_1(x)+2\sqrt{\mu_{00}(x)\mu_{11}(x)}}{a^2}.\end{aligned} \tag{4}$$

The interests of the upper bound given in (4) lies in the fact that all the terms at the r.h.s. can be approximated with supervised learning algorithm, under certain regularity assumptions. In practice, this allows a finer analysis on the causal inference of the treatment. In particular, all this terms can be derived by providing an estimation of the conditional distribution.

## References


[AW19] Susan Athey and Stefan Wager. Estimating Treatment Effects with Causal Forests: An Application. *arXiv e-prints*, page arXiv:1902.07409, Feb 2019.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[BS15] Gérard Biau and Erwan Scornet. A random forest guided tour. *TEST*, 25:197–227, 2015.

[IR15] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015.

[KSBY17] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning. *arXiv e-prints*, page arXiv:1706.03461, Jun 2017.

[Mei06] Nicolai Meinshausen. Quantile regression forests. *J. Mach. Learn. Res.*, 7:983–999, 2006.

[NW17] Xinkun Nie and Stefan Wager. Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *arXiv e-prints*, page arXiv:1712.04912, Dec 2017.

[Rub74] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.