Wassertein Random Forests and Applications in Heterogeneous Treatment Effects: Supplementary Materials

1 Additional simulation study for univariate case

In order to compare with the other conditional density estimation methods such as RFCDE (Pospisil and Lee, 2019) and take into account the influence of the propensity score function, we consider a slightly modified model:

- $X \sim \text{Unif}\left([0,1]^d\right)$ with $d = 50$;
- $Y(0) \sim \mathcal{N}(m_0(X), \sigma_0^2(X))$ and $Y(1) \sim \frac{1}{2}\mathcal{N}(-1,1) + \frac{1}{2}\mathcal{N}\left(m_1(X), \sigma_1^2(X)\right)$;
- $T \sim \text{Bernoulli}\left(\frac{1}{2}\sin\left(2X^{(1)}X^{(2)} + 6X^{(3)}\right) + \frac{1}{2}\right)$,

with

- $m_0(x) = 10x^{(2)}x^{(4)} + x^{(3)} + \exp\left\{x^{(4)} - 2x^{(1)}\right\}$;
- $\sigma_0^2(x) = \left\{-x^{(1)}x^{(2)} + 4\left(x^{(3)}\right)^2\right\} \vee \frac{1}{5}$;
- $m_1(x) = 2m_0(x) + 1 - 5x^{(2)}x^{(5)}$;
- $\sigma_1^2(x) = 3x^{(2)} + x^{(3)}x^{(4)} + x^{(6)}$.

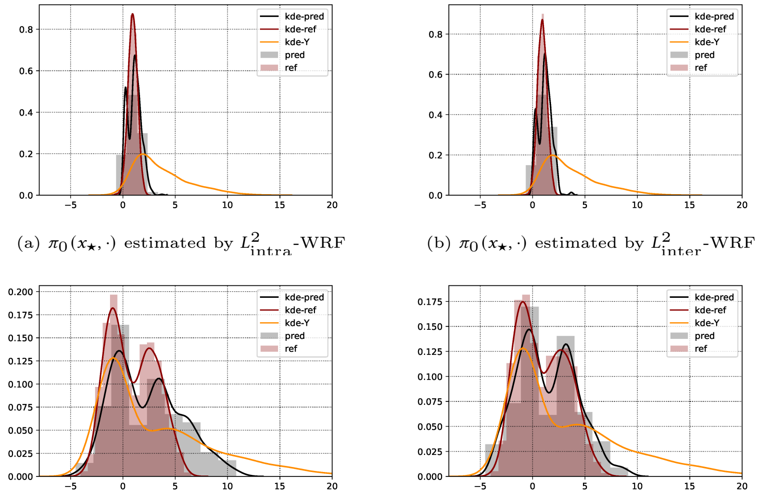
Basically, the distributions of X and $Y(0)$ remain the same, while the conditional distribution of $Y(1)$ given X is replaced by a mixture of two Gaussians, which admits a density w.r.t. Lebesgue measure on $\mathbb{R}^d$. The propensity score function is also modified in order to model the complexity of observational studies.

First, to illustrate the good quality of the estimation provided by WRF, we randomly select an individual x_* such that the associated CATE function is 0 (i.e., $x_*^{(2)}x_*^{(5)} = 0$), for which a CATE-based inference cannot provide sufficient insight in the causality. The visualization can be found in Figure 1. Note that we add a standard kernel smoothing since conditional density is assumed to exist in this case. It is clear that both L^2_{intra}-WRF and L^2_{inter}-WRF can provide a good approximation of both $\pi_0(x_*, \cdot)$ and $\pi_1(x_*, \cdot)$. A more detailed benchmark can be found in Table 1. The setting of the experiment (for all considered forests) remains the same as in the main text: The dataset is of size 1000 and the associated parameters for the forests are $a_n = 500$, $M = 200$ and **nodesize** $= 2$.

Table 1: Estimation of π_0 (i.e., $\mathcal{L}(Y(0) \mid X = x)$) and π_1 (i.e., $\mathcal{L}(Y(0) \mid X = x)$)

Methods	π_0-$\overline{\mathcal{W}}_1(1000)$	π_0-$\overline{\mathcal{W}}_2(1000)$	π_1-$\overline{\mathcal{W}}_1(1000)$	π_1-$\overline{\mathcal{W}}_2(1000)$
L^2_{intra}-WRF	0.6967	0.8523	1.5406	2.2493
L^2_{inter}-WRF	**0.6869**	0.8403	**1.3844**	**1.9881**
L^1_{inter}-WRF	0.6915	**0.8397**	1.4210	2.0428
MF	2.0110	2.0321	2.3958	2.8991
ERT	0.7025	0.8961	1.6490	2.4223
RFCDE	0.7979	3.1471	0.9503	3.3630

It is clear that L^2_{inter}-WRF provides the overall most accurate prediction for this synthetic dataset. The difference between intra-class and inter-class WRF are more noticeable in the estimation of π_1, which provides more evidence that inter-class variants of WRF are better suited for more complex situation (multimodality or large variance). The fact that L^2_{inter}-WRF outperforms L^1_{inter}-WRF may be due to the existence of conditional density functions. This case can be regarded as more "smooth" than the case considered in the main text, where conditional density does not exist for π_1.



(a) $\pi_0(x_\star, \cdot)$ estimated by L^2_{intra}-WRF

(b) $\pi_0(x_\star, \cdot)$ estimated by L^2_{inter}-WRF

(c) $\pi_1(x_\star, \cdot)$ estimated by L^2_{intra}-WRF

(d) $\pi_1(x_\star, \cdot)$ estimated by L^2_{inter}-WRF

Figure 1: An illustration of estimated conditional distributions provided by different variants of WRF with the same parameters: $a_n = 500$ (with repetition), $M = 200$, **mtry** $= 50$, **nodesize** $= 2$. In the legend, `pred` and `ref` denote respectively the prediction given by WRF and reference values sampled directly from the true conditional distribution with sample size fixed to be 2000. The acronyms `kde-pred` and `kde-ref` stand for the outputs of the `kdeplot` function of `seaborn` package (Waskom et al., 2020), which provides a standard kernel smoothing. Finally, `kde-Y` denotes the `kdeplot` of the Y-population, i.e., all the $Y_i(1)$ or $Y_i(0)$ in the training dataset according to the treatment/control group.

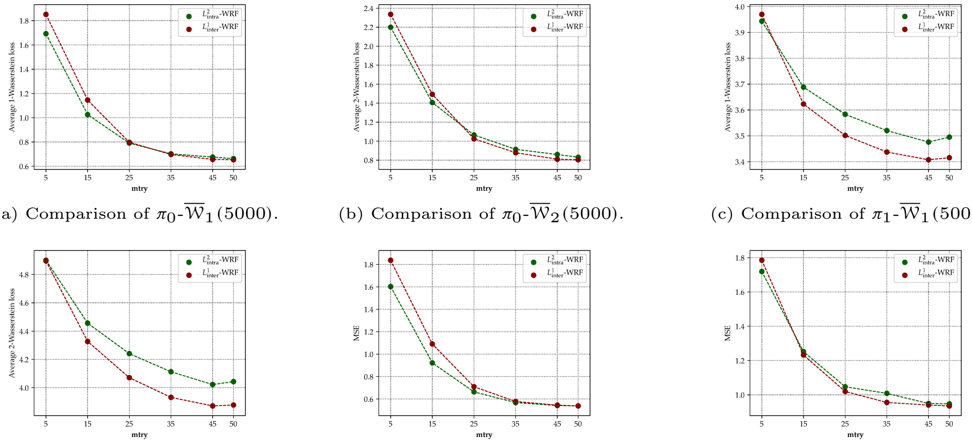
2 On the parameter tuning of WRF

We discuss in this section the influence of the choice of parameters (i.e., **mtry**, a_n and **nodesize**) of the WRF and try to provide some suggestions on the algorithm tuning. We stick to the model provided in Section 3.2 of the main text and compare the π_t-$\overline{\mathcal{W}}_p(5000)$ respectively for $t \in \{0, 1\}$ and $p \in \{1, 2\}$ to illustrate the performance of our method in unimodal and multimodal situations. Unlike the conditional expectation estimation, the cross validation-based tuning strategy is not straightforward to implement for conditional distribution estimation. Indeed, we have only a single sample at each point X_i, and it does not provide enough information for the conditional distribution. Therefore, we also track the performance of the associated conditional expectation estimations in terms of Mean Squared Error (MSE). The conditional expectation functions given $X = x$ of $Y(0)$ and $Y(1)$ are denoted respectively by $\mu_0(x)$ and $\mu_1(x)$. Our goal is to illustrate whether the tuning for the conditional expectation can be exploited to guide the tuning for the conditional distribution estimation problem. We also note that since each tree is constructed using only part of the data, the *out-of-bag* errors for the forest can thus be obtained by averaging the empirical error of each tree on the unused sub-dataset (see, e.g., Biau and Scornet, 2015, Section 2.4) in the case where an independent test dataset is not available.

First, it is well-known that in the classical RF context the number of trees M should be taken as large as possible, according to the available computing budget, in order to reduce the variance of the forest. Although the goal in the WRF framework is changed to the conditional distribution estimation, it is still suggested to use a large M if possible.

Second, let us investigate the number of directions to be explored at each cell **mtry**. The result is illustrated in Figure 2 ((a)-(d) for average Wasserstein loss and (e)-(f) for MSE of conditional expectation estimation). Roughly speaking, the value of **mtry** reflects the strength of greedy optimization at each cell during the construction of decision trees. A conservative approach is to choose **mtry** as large as possible according to the available computing resources.

Then, let us see the influence brought by the change of **nodesize**. The illustration can be found in Figure 3 ((a)-(d) for average Wasserstein loss and (e)-(f) for MSE of conditional expectation estimation). In the classical RF context, the motivation of the choice **nodesize** > 2 can be interpreted as introducing some local averaging procedure at each cell in order to deal with the variance or noise of the sample. Here, as discussed in the main text, we are interested in the conditional distribution estimation in the HTE context, where the variance



(a) Comparison of π_0-$\overline{\mathcal{W}}_1(5000)$.

(b) Comparison of π_0-$\overline{\mathcal{W}}_2(5000)$.

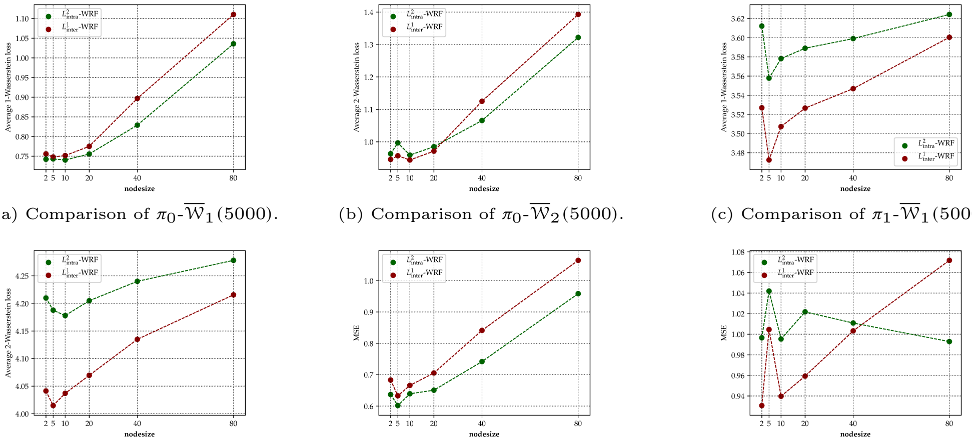
(c) Comparison of π_1-$\overline{\mathcal{W}}_1(5000)$.

(d) Comparison of π_1-$\overline{\mathcal{W}}_2(5000)$.

(e) Comparison of the estimation of μ_0.

(f) Comparison of the estimation of μ_1.

Figure 2: An illustration of the performance of different variants of WRF (namely, L^2_{intra}-WRF and L^1_{inter}-WRF) with **mtry** varying in $\{5, 15, 25, 35, 45, 50\}$, $a_n = 500$ (with repetition), $M = 300$ and **nodesize** $= 3$.



(a) Comparison of π_0-$\overline{\mathcal{W}}_1(5000)$.

(b) Comparison of π_0-$\overline{\mathcal{W}}_2(5000)$.

(c) Comparison of π_1-$\overline{\mathcal{W}}_1(5000)$.

(d) Comparison of π_1-$\overline{\mathcal{W}}_2(5000)$.

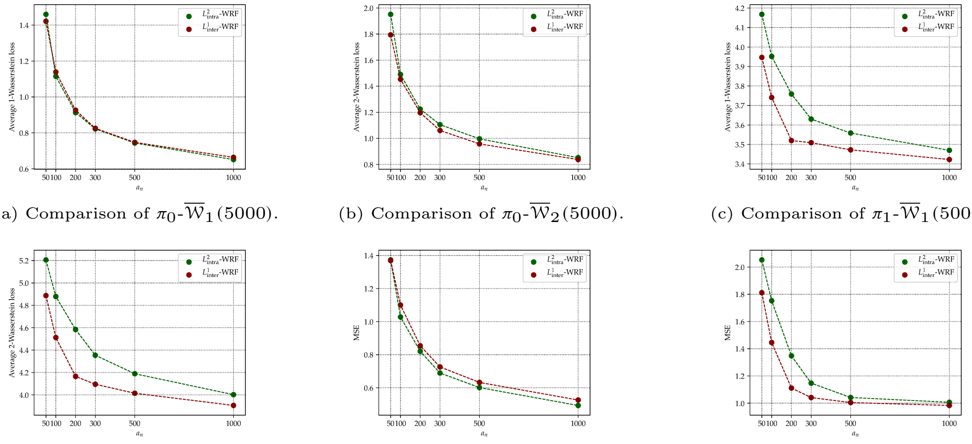
(e) Comparison of the estimation of μ_0.

(f) Comparison of the estimation of μ_1.

Figure 3: An illustration of the performance of different variants of WRF (namely, L^2_{intra}-WRF and L^1_{inter}-WRF) with **nodesize** varying in $\{2, 5, 10, 20, 40, 80\}$, $a_n = 500$ (with repetition), $M = 300$ and **mtry** $= 30$.

or other fluctuation of the conditional distribution is part of the information to be estimated. Hence, the interpretation of the choice **nodesize** > 2 should be adapted accordingly, as the minimum sample size that is used to describe the conditional distribution at each cell. This interpretation is better suited when it comes to the estimation of multimodal conditional distributions. As shown in Figure 3 (a)-(d), there are some optimal choices of **nodesize** between 2 and a_n. In the simple cases, such as the estimation of π_0 (unimodal), the MSE of the associated conditional expectation (Figure 3 (e)) can be used, accordingly, to tune the algorithm for conditional distribution estimation. However, in the more complex case such as the estimation of π_1 (bi-modal), the MSE of the conditional expectation estimation is no as stable (Figure 3 (f)). Nevertheless, it is also recommended to use small **nodesize** in this situation as a conservative choice.

Finally, we discuss the size a_n of the sub-dataset used to construct each decision tree. Note that the choice of a_n is still not well-understood even in the classical RF context (see, e.g., Biau and Scornet, 2015; Scornet et al., 2015). When the computing budget allows to implement $a_n = n$ (with replacement, which corresponds to the classical Bootstrap), we recommend to use this choice. Otherwise, we recommend to fix the a_n from one fifth to one third of the whole data size in order to maintain a reasonably good performance without heavy computations.



(a) Comparison of π_0-$\overline{\mathcal{W}}_1(5000)$.

(b) Comparison of π_0-$\overline{\mathcal{W}}_2(5000)$.

(c) Comparison of π_1-$\overline{\mathcal{W}}_1(5000)$.

(d) Comparison of π_1-$\overline{\mathcal{W}}_2(5000)$.

(e) Comparison of the estimation of μ_0.

(f) Comparison of the estimation of μ_1.

Figure 4: An illustration of the performance of different variants of WRF (namely, L^2_{intra}-WRF and L^1_{inter}-WRF) with a_n varies in $\{50, 100, 200, 300, 500, 1000\}$ (with repetition), **nodesize** $= 5$, $M = 300$ and **mtry** $= 30$.

Suggestions on the parameter tuning The take-home message for the parameter tuning of WRF is simple: We recommend to use large M and **mtry** according to the available computing resources. The parameter **nodesize** can be tuned via a cross validation-based strategy using the MSE of the associated conditional expectation estimation. In addition, we suggest to choose smaller **nodesize** when there is abnormal fluctuation of the MSE score. It is also proposed to use classical bootstrap (i.e., $a_n = n$ with replacement) when possible. Otherwise, we suggest to fix a smaller a_n according to the computing budget. Finally, although there is no theoretical guarantee, we advocate to use L^1_{inter}-WRF or L^2_{inter}-WRF for univariate objective, since it has a better overall accuracy with a reasonable additional computational cost.

3 On the propensity score function

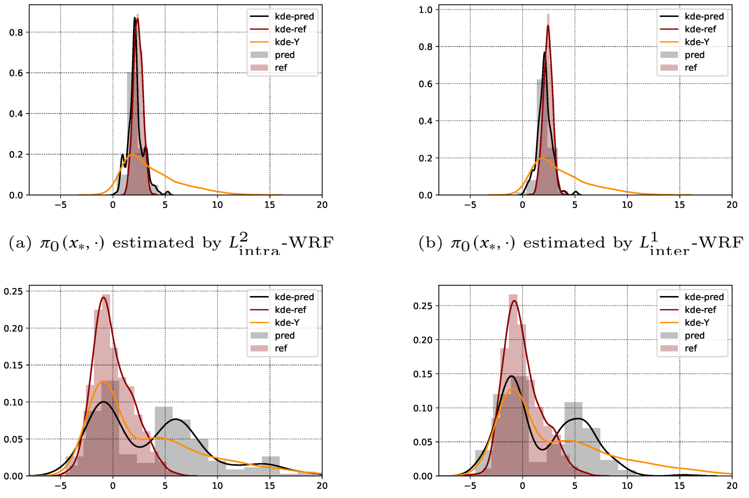
The propensity score function $e(\cdot)$ measures the probability that the treatment is assigned to a certain individual, which basically determines the distribution of the available dataset for the estimation of π_0 and π_1 in the population. More precisely, imagine that x is an individual such that in the neighbourhood of x, the value of $e(\cdot)$ is close to 0. Then, it is expected that only very few training data for the estimation of $\pi_1(x, \cdot)$ can be collected during the observational study. As a consequence, it is expected that the estimation $\hat{\pi}_1$ at such point is of reasonably bad quality. For example, the propensity score function is

$$e(x) = \frac{1}{2}\sin(2x^{(1)}x^{(2)} + 6x^{(3)}) + \frac{1}{2}.$$

Denote by $x_\star$ an individual such that $x_\star^{(1)} = \frac{\pi}{4}$, $x_\star^{(2)} = 1$, and $x_\star^{(6)} = \frac{\pi}{6}$. It is readily checked that $e(x_\star) = 0$. As shown in Figure 5, the estimation of $\pi_0(x_\star, \cdot)$ is very accurate (see Figure 5 (a)-(b)), while the estimation of $\pi_1(x_\star, \cdot)$ is of poor quality (see Figure 5 (c)-(d)).

From a theoretical perspective, one may suppose that the propensity score function is bounded away from 0 and 1 uniformly for all $x \in \mathbb{R}^d$ (see, e.g., Künzel et al., 2019; Nie and Wager, 2017). However, it is, unfortunately, not possible to control the propensity score during an observational study. As a consequence, it is usually very difficult to verify such an assumption in practice. Therefore, a more meaningful question can be how to detect if our estimation is reliable or not for a certain individual. A straightforward strategy is to estimate the propensity score function independently, as done for example in (Athey and Wager, 2019), and to test whether the value of this score is close to 0 and 1. Another approach is to exploit the information encoded in the splits/weights of the forest to detect whether enough data is collected for the prediction at target individual. The details are left for future research.

Finally, let us mention that if the goal is to estimate the function $\Lambda_p(\cdot)$ defined in Section 3.1 of the main text, we expect that more dedicated variants of WRF can be constructed, in the same spirit of Causal Forests introduced in (Athey and Wager, 2019).



(a) $\pi_0(x_*, \cdot)$ estimated by L^2_{intra}-WRF

(b) $\pi_0(x_*, \cdot)$ estimated by L^1_{inter}-WRF

(c) $\pi_1(x_*, \cdot)$ estimated by L^2_{intra}-WRF

(d) $\pi_1(x_*, \cdot)$ estimated by L^1_{inter}-WRF

Figure 5: An illustration of estimated conditional distributions provided by different variants of WRF with the same parameters: $a_n = 500$ (with repetition), $M = 200$, **mtry** $= 50$, **nodesize** $= 2$. In the legend, `pred` and `ref` denote respectively the prediction given by WRF and reference values sampled directly from the true conditional distribution with sample size fixed to be 2000. The acronyms `kde-pred` and `kde-ref` stand for the outputs of the `kdeplot` function of `seaborn` package (Waskom et al., 2020), which provides a standard kernel smoothing. Finally, `kde-Y` denotes the `kdeplot` of the Y-population, i.e., all the $Y_i(1)$ or $Y_i(0)$ in the training dataset according to the treatment/control group.

4 Possible extensions

In this section, we discuss two natural extensions of WRF that we did not investigate in details.

First, inspired by the Random Rotation Ensembles introduced in (Blaser and Fryzlewicz, 2016), it is natural to consider the implementation of oblique splits, i.e., the splits are not necessarily axis-aligned. More precisely, for each tree, by sampling a uniformly distributed rotation matrix (e.g. Blaser and Fryzlewicz, 2016, Section 3), we are able to construct the decision tree by using the rotated sub-dataset (or equivalently, one can also implement randomly rotated cuts in the tree's construction). Intuitively speaking, the rotation variants of WRF will be more consistent when it comes to performance, while the additional computing resources are required for both training and prediction.

Another direction is to replace the Dirac mass in the empirical measures by some kernel $K(x, dy)$, as proposed in (Pospisil and Lee, 2019). For instance, the L^p_{inter}-WRF can be modified by using the following splitting criteria:

$$\begin{aligned}\tilde{L}^p_{\text{inter}}(A_L, A_R) :=& \frac{N_L}{N_A}\mathcal{W}_p^p\left(\frac{1}{N_L}\sum_{X_i\in A_L}K(Y_i, dy), \frac{1}{N_A}\sum_{X_i\in A}K(Y_i, dy)\right)\\ &+ \frac{N_R}{N_A}\mathcal{W}_p^p\left(\frac{1}{N_R}\sum_{X_i\in A_R}K(Y_i, dy), \frac{1}{N_A}\sum_{X_i\in A}K(Y_i, dy)\right),\end{aligned}$$

where the kernel $K(\cdot, \cdot)$ is chosen according to prior knowledge of the problem. At the same time, the final prediction will be replaced by

$$\tilde{\pi}_{M,n}(x, dy; \Theta_{[M]}, \mathcal{D}_n) = \sum_{i=1}^{n} \alpha_i(x) K(Y_i, dy),$$

where $\alpha_i(\cdot)$ remains the same as defined in Section 2.1 of the main text. When the associated $\mathcal{W}_p$-distance is easy to compute, we expect that this extension will be more accurate for small datasets. Nevertheless, the performances of these natural extensions are still not clear. The details are therefore left for future research.

References

Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational Studies*, 5.

Biau, G. and Scornet, E. (2015). A random forest guided tour. *TEST*, 25:197–227.

Blaser, R. and Fryzlewicz, P. (2016). Random rotation ensembles. *Journal of Machine Learning Research*, 17:1–26.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116:4156–4165.

Nie, X. and Wager, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects. *arXiv*, 1712.04912.

Pospisil, T. and Lee, A. B. (2019). RFCDE: Random Forests for Conditional Density Estimation and functional data. *arXiv*, 1906.07177.

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

Waskom, M., Botvinnik, O., Ostblom, J., Gelbart, M., Lukauskas, S., Hobson, P., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Swain, C., Miles, A., Brunner, T., O'Kane, D., Yarkoni, T., Williams, M. L., Evans, C., Fitzgerald, C., and Brian (2020). mwaskom/seaborn: v0.10.1 (april 2020).