Distributional Random Forests: Heterogeneity Adjustment and Multivariate Distributional Regression

Domagoj Ćevid   CEVID@STAT.MATH.ETHZ.CH

Loris Michel   MICHEL@STAT.MATH.ETHZ.CH

Jeffrey Näf   NAEF@STAT.MATH.ETHZ.CH

Peter Bühlmann   BUHLMANN@STAT.MATH.ETHZ.CH

Nicolai Meinshausen   MEINSHAUSEN@STAT.MATH.ETHZ.CH
Seminar für Statistik
ETH Zürich
8092 Zürich, Switzerland

Editor: Mladen Kolar

Abstract

Random Forest (Breiman, 2001) is a successful and widely used regression and classification algorithm. Part of its appeal and reason for its versatility is its (implicit) construction of a kernel-type weighting function on training data, which can also be used for targets other than the original mean estimation. We propose a novel forest construction for multivariate responses based on their joint conditional distribution, independent of the estimation target and the data model. It uses a new splitting criterion based on the MMD distributional metric, which is suitable for detecting heterogeneity in multivariate distributions. The induced weights define an estimate of the full conditional distribution, which in turn can be used for arbitrary and potentially complicated targets of interest. The method is very versatile and convenient to use, as we illustrate on a wide range of examples. The code is available as `Python` and `R` packages `drf`.

Keywords: causality, distributional regression, fairness, Maximal Mean Discrepancy, Random Forests, two-sample testing

1. Introduction

In practice, one often encounters heterogeneous data, whose distribution is not constant, but depends on certain covariates. For example, data can be collected from several different sources, its distribution might differ across certain subpopulations or it could even change with time, etc. Inferring valid conclusions about a certain target of interest from such data can be very challenging as many different aspects of the distribution could potentially change. As an example, in medical studies, the effectiveness of a certain treatment might not be constant throughout the population but depend on certain patient characteristics such as age, race, gender, or medical history. Another issue could be that different patient groups were not equally likely to receive the same treatment in the observed data.

Obviously, pooling all available data together can result in invalid conclusions. On the other hand, if for a given test point of interest one only considers similar training data points,

i.e. a small homogeneous subpopulation, one may end up with too few samples for accurate statistical estimation. In this paper, we propose a method based on the Random Forest algorithm (Breiman, 2001) which in a data-adaptive way determines for any given test point which training data points are relevant for it. This in turn can be used for drawing valid conclusions or for accurately estimating any quantity of interest.
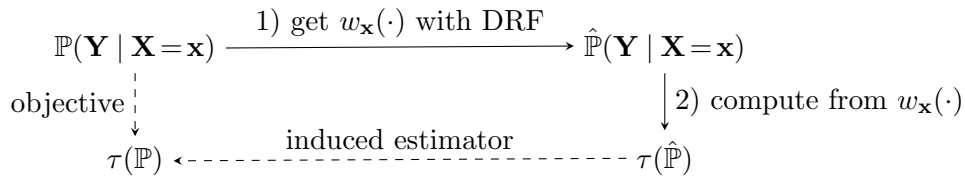
Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_d) \in \mathbb{R}^d$ be a multivariate random variable representing the data of interest, but whose joint distribution is heterogeneous and depends on some subset of a potentially large number of covariates $\mathbf{X} = (X_1, X_2, \ldots, X_p) \in \mathbb{R}^p$. Throughout the paper, vector quantities are denoted in bold. We aim to estimate a certain target object $\tau(\mathbf{x})$ that depends on the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}) = \mathbb{P}(\mathbf{Y} \mid X_1 = x_1, \ldots, X_p = x_p)$, where $\mathbf{x} = (x_1, \ldots, x_p)$ is an arbitrary point in $\mathbb{R}^p$. The estimation target $\tau(\mathbf{x})$ can range from simple quantities, such as the conditional expectations $\mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}]$ (Breiman, 2001) or quantiles $Q_\alpha[f(\mathbf{Y}) \mid \mathbf{X}]$ (Meinshausen, 2006) for some function $f : \mathbb{R}^d \to \mathbb{R}$, to some more complicated aspects of the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$, such as conditional copulas or conditional independence measures. Given the observed data $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, the most straightforward way of estimating $\tau(\mathbf{x})$ nonparametrically would be to consider only the data points in some neighborhood $\mathcal{N}_{\mathbf{x}}$ around $\mathbf{x}$, e.g. by considering the k nearest neighbors according to some metric. However, such methods typically suffer from the curse of dimensionality even when p is only moderately large: for a reasonably small neighborhood, such that the distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} \in \mathcal{N}_{\mathbf{x}})$ is close to the distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$, the number of training data points contained in $\mathcal{N}_{\mathbf{x}}$ will be very small, thus making the accurate estimation of the target $\tau(\mathbf{x})$ difficult. The same phenomenon occurs with other methods which locally weight the training observations such as kernel methods (Silverman, 1986), local MLE (Fan et al., 1998) or weighted regression (Cleveland, 1979) even for the relatively simple problem of estimating the conditional mean $\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ for fairly small p. For that reason, more importance should be given to the training data points $(\mathbf{x}_i, \mathbf{y}_i)$ for which the response distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}_i)$ at point $\mathbf{x}_i$ is similar to the target distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$, even if $\mathbf{x}_i$ is not necessarily close to $\mathbf{x}$ in every component.

In this paper, we propose the Distributional Random Forest (DRF) algorithm which estimates the multivariate conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ in a locally adaptive fashion. This is done by repeatedly dividing the data points in the spirit of the Random Forest algorithm (Breiman, 2001): at each step, we split the data points into two groups based on some feature X_j in such a way that the distribution of $\mathbf{Y}$ for which $X_j \leq l$, for some level l, differs the most compared to the distribution of $\mathbf{Y}$ when $X_j > l$, according to some distributional metric. One can use any multivariate two-sample test statistic, provided it can detect a wide variety of distributional changes. As the default choice, we propose a criterion based on the Maximal Mean Discrepancy (MMD) statistic (Gretton et al., 2007a) with many interesting properties. This splitting procedure partitions the data such that the distribution of the multivariate response $\mathbf{Y}$ in the resulting leaf nodes is as homogeneous as possible, thus defining neighborhoods of relevant training data points for every $\mathbf{x}$. Repeating this many times with randomization induces a weighting function $w_{\mathbf{x}}(\mathbf{x}_i)$ as in Lin and Jeon (2002, 2006), described in detail in Section 2, which quantifies the relevance of each training data point $\mathbf{x}_i$ for a given test point $\mathbf{x}$. The conditional distribution is then estimated by an empirical distribution determined by these weights (Meinshausen, 2006). This construction

is data-adaptive as it assigns more weight to the training points $\mathbf{x}_i$ that are closer to the test point $\mathbf{x}$ in the components which are more relevant for the distribution of $\mathbf{Y}$.

Our forest construction does not depend on the estimation target $\tau(\mathbf{x})$, but it rather estimates the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ directly and the induced forest weights can be used to estimate $\tau(\mathbf{x})$ in a second step. This approach has several advantages. First, only one DRF fit is required to obtain estimates of many different targets, which has a big computational advantage. Furthermore, since those estimates are obtained from the same forest fit, they are mutually compatible. For example, if the conditional correlation matrix $\{\mathrm{Cor}(Y_i, Y_j \mid \mathbf{X} = \mathbf{x})\}_{i,j=1}^d$ is estimated componentwise using some other method, the resulting matrix might not be positive semidefinite, and as another example, the CDF estimates $\hat{\mathbb{P}}(\mathbf{Y} \leq \mathbf{y} \mid \mathbf{X} = \mathbf{x})$ might not be monotone in $\mathbf{y}$, see Figure 6. Finally, it could be extremely difficult to tailor forest construction to some complex targets $\tau(\mathbf{x})$. The induced weighting function can thus be used not only for obtaining simple distributional aspects such as, for example, the conditional quantiles, conditional correlations, or joint conditional probability statements, but also to obtain more complex objectives, such as conditional independence tests (Zhang et al., 2011), heterogeneous regression (see also Section 4.4 for more details) (Künzel et al., 2019; Wager and Athey, 2018) or semiparametric estimation by fitting a parametric model for $\mathbf{Y}$, having nonparametrically adjusted for $\mathbf{X}$ (Bickel et al., 1993). Representation of the conditional distribution via the weighting function has a great potential for applications in causality such as causal effect estimation or as a way of implementing do-calculus (Pearl, 2009) for finite samples, as we discuss in Section 4.4.

Therefore, DRF is used in two steps: in the first step, we obtain the weighting function $w_{\mathbf{x}}(\cdot)$ describing the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ in a target- and model-free way, which is then used as an input for the second step. Even if the method used in the second step does not directly support weighting of the training data points, one can easily resample the data set with the sampling probabilities equal to $\{w_{\mathbf{x}}(\mathbf{x}_i)\}_{i=1}^n$. This two-step approach is visualized in the following diagram:

$$
\begin{array}{ccc}
\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}) & \xrightarrow{\text{1) get } w_{\mathbf{x}}(\cdot) \text{ with DRF}} & \hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}) \\
\text{objective} \downarrow & & \downarrow \text{2) compute from } w_{\mathbf{x}}(\cdot) \\
\tau(\mathbb{P}) & \xleftarrow{\quad\text{induced estimator}\quad} & \tau(\hat{\mathbb{P}})
\end{array}
$$

1.1 Related work and our contribution

Several adaptations of the Random Forest algorithm have been proposed for targets beyond the original one of the univariate conditional mean $\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$: for survival analysis (Hothorn et al., 2006), conditional quantiles (Meinshausen, 2006), density estimation (Pospisil and Lee, 2018), CDF estimation (Hothorn and Zeileis, 2021) or heterogeneous treatment effects (Wager and Athey, 2018). Almost all such methods use the weights induced by the forest, as described in Section 2, rather than averaging the estimates obtained per tree. This view of Random Forests as a powerful adaptive nearest neighbor method is well known and dates back to Lin and Jeon (2002, 2006). It was first used for targets beyond the conditional mean in Meinshausen (2006), where the original forest construction with univariate Y was

used (Breiman, 2001). However, the univariate response setting considered there severely restricts the number of interesting targets $\tau(\mathbf{x})$ and DRF can thus be viewed as an important generalization of this approach to the multivariate setting.

In order to be able to perform certain tasks or to achieve a better accuracy, many forest-based methods adapt the forest construction by using a custom splitting criterion tailored to their specific target, instead of relying on the standard CART criterion. In Zeileis et al. (2008) and Hothorn and Zeileis (2021), a parametric model for the response $\mathbf{Y} \mid \mathbf{X} = \mathbf{x} \sim f(\theta(\mathbf{x}), \cdot)$ is assumed and recursive splitting is performed based on a permutation test which uses the user-provided score functions. Similarly, Athey et al. (2019) estimate certain univariate targets for which there exist corresponding score functions defining the local estimating equations. The data is split so that the estimates of the target in resulting child nodes differ the most. This is different, though, to the target-free splitting criterion of DRF, which splits so that the distribution of $\mathbf{Y}$ in child nodes is as different as possible.

Since the splitting step is extensively used in the algorithm, its complexity is crucial for the overall computational efficiency of the method, and one often needs to resort to approximating the splitting criterion (Pospisil and Lee, 2018; Athey et al., 2019) to obtain good computational run time. We propose a splitting criterion based on a fast random approximation of the MMD statistic (Gretton et al., 2012a; Zhao and Meng, 2015), which is commonly used in practice for two-sample testing as it is able to detect any change in the multivariate distribution of $\mathbf{Y}$ with good power (Gretton et al., 2007a). DRF with the MMD splitting criterion also has interesting theoretical properties as shown in Section 3 below.

The multivariate response case has not received much attention in the Random Forest literature. Most of the existing forest-based methods focus on either a univariate response Y or on a certain univariate target $\tau(\mathbf{x})$. One interesting line of work considers density estimation (Pospisil and Lee, 2018) and uses aggregation of the CART criteria for different response transformations. Another approach (Kocev et al., 2007; Segal and Xiao, 2011; Ishwaran and Kogalur, 2022) is based on aggregating standard univariate CART splitting criteria for $Y_1, \ldots, Y_d$ and targets only the conditional mean of the responses, a task which could also be solved by separate regression fits for each Y_i. In order to capture any change in the distribution of the multivariate response $\mathbf{Y}$, one needs to not only consider the marginal distributions for each component Y_i, but also to determine whether their dependence structure changes, see e.g. Figure 8.

There is an increasing number of methods that nonparametrically estimate the joint multivariate conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ in the statistics and machine learning literature. In addition to a few simple classical methods such as k-nearest neighbors and kernel regression, there exist methods based on normalizing flows such as Inverse Autoregressive Flow (Kingma et al., 2016) or Masked Autoregressive Flow (Papamakarios et al., 2017) and also conditional variants of several popular generative models such as Conditional Generative Adversarial Networks (Mirza and Osindero, 2014) or Conditional Variational Autoencoder (Sohn et al., 2015). The focus of these methods is more on the settings with large response dimension d and small covariate dimension p, such as image or text generation. Another interesting and related line of research focuses on estimating the conditional mean embedding (CME), as described e.g., in Song et al. (2009); Muandet et al. (2017); Song et al. (2013); Park and Muandet (2020), rather than estimating the conditional distribution directly. CMEs generalize the concept of embedding (marginal) probability distributions

into a Reproducing Kernel Hilbert Space (RKHS) to the conditional case. Interestingly, DRF with the MMD-based splitting criterion can also be viewed as a method for estimating the CME, as discussed in Section 3 below. This viewpoint provides a natural connection between the Random Forest and kernel embedding literature. A comparison of DRF with the methods for distributional estimation listed above can be found in Section 4.1.

Our contribution, resulting in the proposal of the Distributional Random Forest (DRF), can be summarized as follows: First, we introduce the idea of forest construction based on sequential multivariate two-sample test statistics. It does not depend on a particular estimation target and is completely nonparametric, which makes its implementation and usage very simple and universal. Not only does it not require additional user input such as the log-likelihoods or score functions, but it can be used even for complicated targets for which there is no obvious forest construction. Furthermore, it has a computational advantage as only a single forest fit is needed for producing estimates of many different targets that are additionally compatible with each other. Second, we propose an MMD-based splitting criterion with good statistical and computational properties, for which we also derive interesting theoretical results in Section 3. It underpins our implementation, which we provide as R and `Python` packages `drf`. Finally, we show on a broad range of examples in Section 4 how many different statistical estimation problems, some of which not being easily tractable by existing forest-based methods, can be cast to our framework, thus illustrating the usefulness and versatility of DRF.

2. Method

In this section we describe the details of the Distributional Random Forest (DRF) algorithm. We closely follow the implementations of the `grf` (Athey et al., 2019) and `ranger` (Wright and Ziegler, 2017) R-packages. A detailed description of the method and its implementation and the corresponding pseudocode can be found in the Appendix A.

2.1 Forest Building

The trees are grown recursively in a model-free and target-free way as follows: For every parent node P, we determine how to best split it into two child nodes of the form $C_L = \{X_j \leq l\}$ and $C_R = \{X_j > l\}$, where the variable X_j is one of the randomly chosen splitting candidates and l denotes its level based on which we perform the splitting. The split is chosen such that we maximize a certain (multivariate) two-sample test statistic

$$\mathcal{D}\left(\{\mathbf{y}_i \mid \mathbf{x}_i \in C_L\}, \{\mathbf{y}_i \mid \mathbf{x}_i \in C_R\}\right), \tag{1}$$

which measures the difference of the empirical distributions of the data $\mathbf{Y}$ in the two resulting child nodes C_L and C_R. Therefore, in each step we select the candidate predictor X_j which seems to affect the distribution of $\mathbf{Y}$ the most, as measured by the metric $\mathcal{D}(\cdot, \cdot)$. Intuitively, in this way we ensure that the distribution of the data points in every leaf of the resulting tree is as homogeneous as possible, which helps mitigate the bias caused by pooling the heterogeneous data together. A related idea can be found in GRF (Athey et al., 2019), where one attempts to split the data so that the resulting estimates $\hat{\tau}_L$ and $\hat{\tau}_R$, obtained

respectively from data points in C_L and C_R, differ the most:

$$\frac{n_L n_R}{n_P^2}\left(\hat{\tau}_L - \hat{\tau}_R\right)^2, \tag{2}$$

where we write $n_P = |\{i \mid \mathbf{x}_i \in P\}|$ and n_L, n_R are defined analogously.

One could construct the forest using any metric $\mathcal{D}(\cdot, \cdot)$ for empirical distributions. However, in order to have a good accuracy of the overall method, the corresponding two-sample test using $\mathcal{D}(\cdot, \cdot)$ needs to have a good power for detecting any kind of change in distribution, which is a difficult task in general, especially for multivariate data (Bai and Saranadasa, 1996; Székely and Rizzo, 2004). Another very important aspect of the choice of distributional metric $\mathcal{D}(\cdot, \cdot)$ is the computational efficiency; one needs to be able to sequentially compute the values of $\mathcal{D}\left(\{\mathbf{y}_i \mid \mathbf{x}_i \in C_L\}, \{\mathbf{y}_i \mid \mathbf{x}_i \in C_R\}\right)$ for every possible split very fast for the overall algorithm to be computationally feasible, even for moderately large data sets. Below, we propose a splitting criterion based on the MMD two-sample test statistic (Gretton et al., 2007a) which has both good statistical and computational properties.

In contrast to other forest-based methods, we do not use any information about our estimation target τ in order to find the best split of the data, which comes with a certain trade-off. On one hand, it is sensible that tailoring the splitting criterion to the target should improve the estimation accuracy; for example, some predictors might affect the conditional distribution of $\mathbf{Y}$, but not necessarily the estimation target τ and splitting on such predictors unnecessarily reduces the number of training points used for estimating τ. On the other hand, our approach has multiple benefits: it is easier to use as it does not require any user input such as the likelihood or score functions and it can also be used for very complicated targets for which one could not easily adapt the splitting criterion. Furthermore, only one DRF fit is necessary for producing estimates of many different targets, which has both computational advantage and the practical advantage that the resulting estimates are mutually compatible (see e.g. Figure 5).

Interestingly, sometimes it could even be beneficial to split based on a predictor which does not affect the target of estimation, but which affects the conditional distribution. This is illustrated by the following toy example. Suppose that for a bivariate response (Y_1, Y_2) we are interested in estimating the slope of the linear regression of Y_2 on Y_1 conditionally on $p = 30$ predictors $\mathbf{X}$, i.e. our target is $\tau(\mathbf{x}) = \text{Cov}(Y_1, Y_2 \mid \mathbf{X}=\mathbf{x})/\text{Var}(Y_1 \mid \mathbf{X}=\mathbf{x})$. This is one of the main use cases for GRF and its variant which estimates this target is called Causal Forest (Wager and Athey, 2018; Athey et al., 2019). Let us assume that the data has the following distribution:

$$\mathbb{P}\left(\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \,\middle|\, \mathbf{X}=\mathbf{x}\right) \sim N\left(\begin{bmatrix} x_1 \\ x_1 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}\right) \qquad \mathbf{X} \sim N(\mathbf{0}, I_p), \tag{3}$$

i.e. X_1 affects only the mean of the responses, while the other $p-1$ predictors have no effect. In Figure 1 we illustrate the distribution of the data when $n = 300, p = 30, \sigma = 0.2$, together with the DRF and GRF splitting criteria. The true value of the target is $\tau(\mathbf{x}) = 0$, but when σ is not too big, the slope estimates $\hat{\tau}$ on pooled data will be closer to 1. Therefore, the difference of $\hat{\tau}_L$ and $\hat{\tau}_R$ between the induced slope estimates for a candidate split, which is used for splitting criterion (2) of GRF, might not be large enough for us to decide to split

on X_1, or the resulting split might be too unbalanced. This results in worse forest estimates for this toy example, see Figure 1.
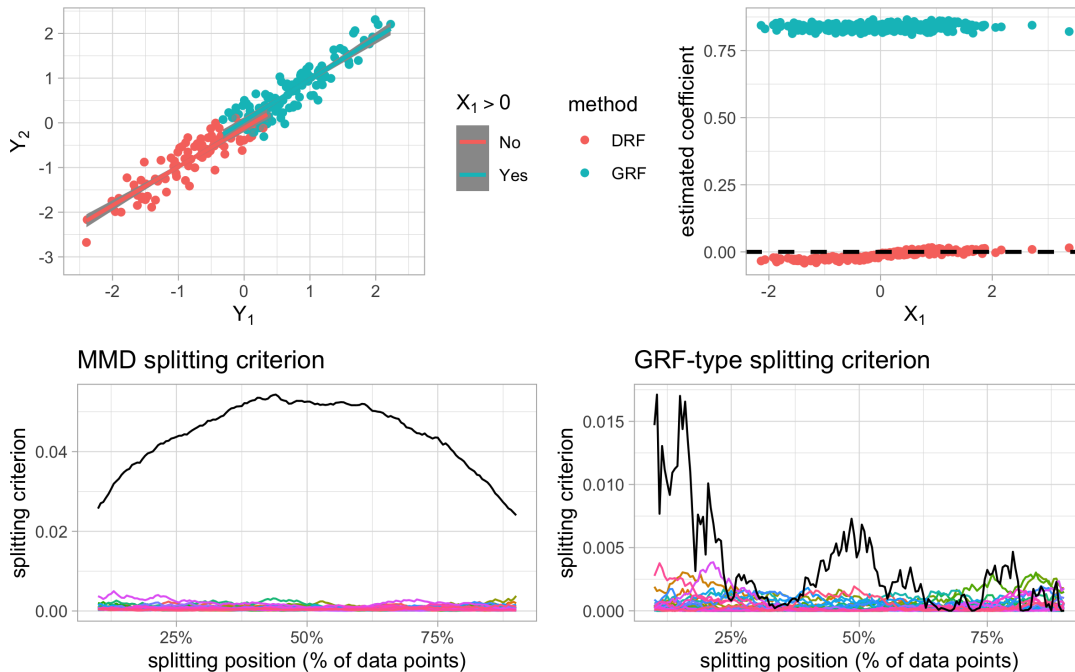

Figure 1: Top left: Illustration of data distribution for the toy example (3) when $n = 300$, $p = 30$. Bottom: The corresponding MMD (12) (left) and GRF (2) splitting criteria (right) at the root node. The curves of different colors correspond to different predictors, with X_1 denoted in black. Top right: Comparison of the estimates of DRF and Causal Forest (Athey et al., 2019) which respectively use those splitting criteria. Test points were randomly generated from the same distribution as the training data. Black dashed line indicates the correct value of the target quantity.

2.2 Weighting Function

Having constructed our forest, just as the standard Random Forest (Breiman, 2001) can be viewed as the weighted nearest neighbor method (Lin and Jeon, 2002), we can use the induced weighting function to estimate the conditional distribution at any given test point $\mathbf{x}$ and thus any other quantity of interest $\tau(\mathbf{x})$. This approach is commonly used in various forest-based methods for obtaining predictions, see e.g., Hothorn and Zeileis (2021); Pospisil and Lee (2018); Athey et al. (2019).

Suppose that we have built N trees $\mathcal{T}_1, \ldots, \mathcal{T}_N$. Let $\mathcal{L}_k(\mathbf{x})$ be the set of the training data points which end up in the same leaf as $\mathbf{x}$ in the tree $\mathcal{T}_k$. The weighting function $w_{\mathbf{x}}(\mathbf{x}_i)$ is defined as the average of the corresponding weighting functions per tree (Lin and Jeon,

2006):

$$w_{\mathbf{x}}(\mathbf{x}_i) = \frac{1}{N} \sum_{k=1}^{N} \frac{\mathbb{1}\left(\mathbf{x}_i \in \mathcal{L}_k(\mathbf{x})\right)}{|\mathcal{L}_k(\mathbf{x})|}. \tag{4}$$

The weights are positive and add up to 1: $\sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) = 1$. In the case of equally sized leaf nodes, the assigned weight to a training point $\mathbf{x}_i$ is proportional to the number of trees where the test point $\mathbf{x}$ and $\mathbf{x}_i$ end up in the same leaf node. This shows that forest-based methods can in general be viewed as adaptive nearest neighbor methods. The sets $\mathcal{L}_k(\mathbf{x})$ of DRF will contain data points $(\mathbf{x}_i, \mathbf{y}_i)$ such that $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}_i)$ is close to $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$, thus removing bias due to heterogeneity of $\mathbf{Y}$ caused by $\mathbf{X}$. On the other hand, since the trees are constructed randomly and are thus fairly independent (Breiman, 2001), the leaf sets $\mathcal{L}_k(\mathbf{x})$ will be different enough so that the induced weights $w_{\mathbf{x}}(\mathbf{x}_i)$ are not concentrated on a small set of data points, which would lead to high estimation variance. Such good bias-variance tradeoff properties of forest-based methods are also implied by their asymptotic properties (Biau, 2012; Wager, 2014), even though this is a still active area of research and not much can be shown rigorously.

One can estimate the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ from the weighting function by using the corresponding empirical distribution:

$$\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}) = \sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) \cdot \delta_{\mathbf{y}_i}, \tag{5}$$

where $\delta_{\mathbf{y}_i}$ is the point mass at $\mathbf{y}_i$.

Two-step approach using weights. The weighting function $w_{\mathbf{x}}(\mathbf{x}_i)$ can directly be used for any target $\tau(\mathbf{x})$ in a second step and not just for estimating the conditional distribution. For example, the estimated conditional joint CDF is given by

$$\hat{F}_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}}(\mathbf{t}) = \hat{\mathbb{P}}(Y_1 \leq t_1, \ldots, Y_d \leq t_d \mid \mathbf{X} = \mathbf{x}) = \sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) \mathbb{1}(\cap_{j=1}^{d}\{(\mathbf{y}_i)_j \leq t_j\}). \tag{6}$$

It is important to point out that using the induced weighting function for locally weighted estimation is different than the approach of averaging the noisy estimates obtained per tree (Wager and Athey, 2018), originally used in standard Random Forests (Breiman, 2001). Even though the two approaches are equivalent for conditional mean estimation, the former approach is often much more efficient for more complex targets (Athey et al., 2019), since the number of data points in a single leaf is very small, leading to large variance of the estimates.

For the univariate response, the idea of using the induced weights for estimating targets different than the original target of conditional mean considered in Breiman (2001) dates back to Quantile Regression Forests (QRF) (Meinshausen, 2006), where a lot of emphasis is put on the quantile estimation, as the number of interesting targets is quite limited in the univariate setting. In the multivariate case, on the other hand, many interesting quantities such as, for example, conditional quantiles, conditional correlations or various conditional probability statements can easily be directly estimated from the weights.

By using the weights as an input for some other method, we can accomplish some more complicated objectives, such as conditional independence testing, causal effect estimation,
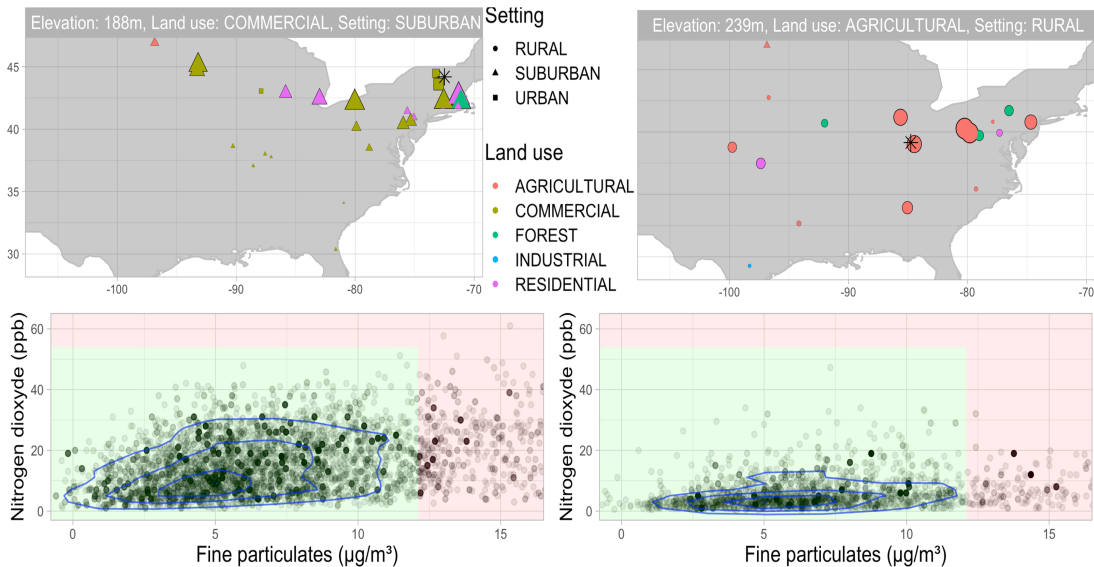
Figure 2: Top: the characteristics of the important training sites, for a fixed test site whose position is indicated by a black star and whose characteristics are indicated in the title. The total weight assigned corresponds to the symbol size. Bottom: estimated joint conditional distribution of two pollutants NO_2 and PM2.5, where the weights correspond to the transparency of the data points. Green area corresponds to 'Good' air quality category (AQI $\leq$ 50).

semiparametric learning, time series prediction or tail-index estimation in extreme value analysis. As an example, suppose that our data $\mathbf{Y}$ come from a certain parametric model, where the parameter θ is not constant, but depends on $\mathbf{X}$ instead, i.e. $\mathbf{Y} \mid \mathbf{X} = \mathbf{x} \sim f(\theta(\mathbf{x}), \cdot)$, see also Zeileis et al. (2008). One can then estimate the parameter $\theta(\mathbf{x})$ by using weighted maximum likelihood estimation:

$$\hat{\theta}(\mathbf{x}) = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) \log f(\theta, \mathbf{y}_i).$$

Another example is heterogeneous regression, where we are interested in the regression fit of an outcome $Y \in \mathbb{R}$ on certain predicting variables $\mathbf{W} \in \mathbb{R}^s$ conditionally on some event $\{\mathbf{X} = \mathbf{x}\}$. This can be achieved by weighted regression of Y on $\mathbf{W}$, where the weights $w_{\mathbf{x}}(\mathbf{x}_i)$ assigned to each data point $(\mathbf{w}_i, y_i)$ are obtained from DRF with the multivariate response $(Y, \mathbf{W}) \in \mathbb{R}^{s+1}$ and predictors $\mathbf{X} \in \mathbb{R}^p$, for an illustration see Section 4.4.

The weighting function of DRF is illustrated on the air quality data in Figure 2. Five years $(2015 - 2019)$ of air pollution measurements were obtained from the US Environmental Protection Agency (EPA) website. Six main air pollutants (nitrogen dioxide (NO_2), carbon monoxide (CO), sulphur dioxide (SO_2), ozone (O_3) and coarse and fine particulate matter (PM10 and PM2.5)) that form the air quality index (AQI) were measured at many different measuring sites in the US for which we know the longitude, latitude, elevation, location setting (rural, urban, suburban) and how the land is used within a 1/4 mile radius. Suppose

we would want to know the distribution of the pollutant measurements at some new, unobserved, measurement site. We train DRF with the measurements (intraday maximum) of the two pollutants PM2.5 and NO$_2$ as the responses, and the site longitude, latitude, elevation, land use and location settings as the predictors and choose two decommissioned measurement sites as test points. For each test point we obtain the weights to all training measurements. We further combine the weights for all measurements corresponding to the same site. The top row illustrates for a given test site, whose characteristics are indicated in the plot title, how much weight in total is assigned to the measurements from a specific training site. We see that the important sites share many characteristics with the test site and that DRF determines the relevance of each characteristic in a data-adaptive way. The bottom row shows the corresponding estimates of the joint conditional distribution of the pollutants (we choose 2 of them for visualization purposes), where the transparency of each training point reflects the assigned weight. One can clearly see how the estimated pollution levels are larger for the suburban site than for the rural site. The forest weights can be used, for example, for estimating the joint density (whose contours can be seen in the plot) or for estimating the probability that the AQI is below a certain value by summing the weights in the corresponding region of space.

2.3 Distributional Metric

In order to determine the best split of a parent node P, i.e. such that the distributions of the responses $\mathbf{Y}$ in the resulting child nodes C_L and C_R differ the most, one needs a good distributional metric $\mathcal{D}(\cdot, \cdot)$ (see Equation (1)) which can detect change in distribution of the response $\mathbf{Y}$ when additionally conditioning on an event $\{X_j > l\}$. Testing equality of distributions from the corresponding samples is an old problem in statistics, called two-sample testing problem. For univariate data, many good tests exist such as Wilcoxon rank test (Wilcoxon, 1946), Welch's t-test (Welch, 1947), Wasserstein two-sample testing (Ramdas et al., 2017), Kolmogorov-Smirnov test (Massey Jr, 1951) and many others, but obtaining an efficient test for multivariate distributions has proven to be quite challenging due to the curse of dimensionality (Friedman and Rafsky, 1979; Baringhaus and Franz, 2004).

Additional requirement for the choice of distributional metric $\mathcal{D}(\cdot, \cdot)$ used for data splitting is that it needs to be computationally very efficient as splitting is used extensively in the algorithm. If we construct N trees from n data points and in each node we consider mtry candidate variables for splitting, the complexity of the standard Random Forest algorithm (Breiman, 2001) in the univariate case is $\mathcal{O}(N \times \text{mtry} \times n \log n)$ provided our splits are balanced. It uses the CART splitting criterion, given by:

$$\frac{1}{n_P} \left(\sum_{\mathbf{x}_i \in C_L} (y_i - \overline{y}_L)^2 + \sum_{\mathbf{x}_i \in C_R} (y_i - \overline{y}_R)^2 \right), \tag{7}$$

where $\overline{y}_L = \frac{1}{n_L} \sum_{\mathbf{x}_i \in C_L} y_i$ and $\overline{y}_R$ is defined analogously. This criterion has an advantage that not only it can be computed in $\mathcal{O}(n_P)$ complexity, but this can be done for all possible splits $\{X_j \le l\}$ as cutoff level l varies, since updating the splitting criterion when moving a single training data point from one child node to the other requires only $\mathcal{O}(1)$ computational steps (most easily seen by rewriting the CART criterion as in (13)).

If the time complexity of evaluating the DRF splitting criterion (1) for a single splitting candidate X_j and all cutoffs l of interest (usually taken to range over all possible values) is at least n^c for some $c > 1$, say $\mathcal{O}(f(n))$ for some function $f : \mathbb{R} \to \mathbb{R}$, then by solving the recursive relation we obtain that the overall complexity of the method is given by $\mathcal{O}(N \times \text{mtry} \times f(n))$ (Akra and Bazzi, 1998), which can be unfeasible even for moderately large n if f grows too fast.

The problem of sequential two-sample testing is also central to the field of change-point detection (Wolfe and Schechtman, 1984; Brodsky and Darkhovsky, 2013), with the slight difference that in the change-point problems the distribution is assumed to change abruptly at certain points in time, whereas for our forest construction we only are interested in finding the best split of the form $\{X_j \leq l\}$ and the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \{\mathbf{X} \in P\} \cap \{X_j \leq l\})$ usually changes gradually with l. The testing power and the computational feasibility of the method play a big role in change-point detection as well. However, the state-of-the-art change-point detection algorithms (Li et al., 2019; Matteson and James, 2014) are often too slow for our purpose as sequential testing is done $\mathcal{O}(N \times \text{mtry} \times n)$ times for forest construction, much more frequently than in change-point problems.

2.3.1 MMD SPLITTING CRITERION

Even though DRF could in theory be constructed with any distributional metric $\mathcal{D}(\cdot, \cdot)$, as a default choice we propose splitting criterion based on the Maximum Mean Discrepancy (MMD) statistic (Gretton et al., 2007a). Let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be the RKHS of real-valued functions on $\mathbb{R}^d$ induced by some positive-definite kernel k, and let $\varphi : \mathbb{R}^d \to \mathcal{H}$ be the corresponding feature map satisfying that $k(\mathbf{u}, \mathbf{v}) = \langle \varphi(\mathbf{u}), \varphi(\mathbf{v}) \rangle_{\mathcal{H}}$.

The MMD statistic $\mathcal{D}_{\text{MMD}(k)}(U, V)$ for kernel k and two samples $U = \{\mathbf{u}_1, \ldots, \mathbf{u}_{|U|}\}$ and $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_{|V|}\}$ is given by:

$$\mathcal{D}_{\text{MMD}(k)}(U, V) = \frac{1}{|U|^2} \sum_{i,j=1}^{|U|} k(\mathbf{u}_i, \mathbf{u}_j) + \frac{1}{|V|^2} \sum_{i,j=1}^{|V|} k(\mathbf{v}_i, \mathbf{v}_j) - \frac{2}{|U||V|} \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} k(\mathbf{u}_i, \mathbf{v}_j). \quad (8)$$

MMD compares the similarities, described by the kernel k, within each sample with the similarities across samples and is commonly used in practice for two-sample testing. It is based on the idea that one can assign to each distribution $\mathcal{P}$ its embedding $\mu(\mathcal{P})$ into the RKHS $\mathcal{H}$, which is the unique element of $\mathcal{H}$ given by

$$\mu(\mathcal{P}) = \mathbb{E}_{\mathbf{Y} \sim \mathcal{P}}[\varphi(\mathbf{Y})]. \quad (9)$$

The MMD two-sample statistic (8) can then equivalently be written as the squared distance between the embeddings of the empirical distributions with respect to the RKHS norm $\|\cdot\|_{\mathcal{H}}$:

$$\mathcal{D}_{\text{MMD}(k)}(U, V) = \left\| \mu \left(\frac{1}{|U|} \sum_{i=1}^{|U|} \delta_{\mathbf{u}_i} \right) - \mu \left(\frac{1}{|V|} \sum_{i=1}^{|V|} \delta_{\mathbf{v}_i} \right) \right\|_{\mathcal{H}}^2, \quad (10)$$

recalling that $\delta_{\mathbf{y}}$ is the point mass at $\mathbf{y}$.

As the sample sizes $|U|$ and $|V|$ grow, the MMD statistic (10) converges to its population version, which is the squared RKHS distance between the corresponding embeddings of

the data-generating distributions of U and V. Since the embedding map μ is injective for a characteristic kernel k, we see that MMD is able to detect any difference in the distribution. Even though the power of the MMD two sample test also deteriorates as the data dimensionality grows, since the testing problem becomes intrinsically harder (Ramdas et al., 2015), it still has good empirical power compared to other multivariate two-sample tests for a wide range of k (Gretton et al., 2012a).

Fast random splitting criterion approximation.   The $\mathcal{O}((|U| + |V|)^2)$ complexity for computing $\mathcal{D}_{\mathrm{MMD}(k)}(U, V)$ from (8) is nevertheless too large for many applications. For that reason, several fast approximations of MMD have been suggested in the literature (Gretton et al., 2012a,b; Zaremba et al., 2013; Chwialkowski et al., 2015; Jitkrittum et al., 2016). As already mentioned, the complexity of the distributional metric $\mathcal{D}(\cdot, \cdot)$ used for DRF is crucial for the overall method to be computationally efficient, since the splitting step is used extensively in the forest construction. We therefore propose splitting based on an MMD statistic computed with an approximate kernel $\tilde{k}$, which is also a fast random approximation of the original MMD statistic (Zhao and Meng, 2015).

Bochner's theorem (see e.g. Wendland (2004, Theorem 6.6)) gives us that any bounded shift-invariant kernel can be written as

$$k(\mathbf{u}, \mathbf{v}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^T(\mathbf{u}-\mathbf{v})} d\nu(\boldsymbol{\omega}), \tag{11}$$

i.e. as a Fourier transform of some measure ν. Therefore, by randomly sampling the frequency vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_B$ from normalized ν, we can approximate our kernel k by another kernel $\tilde{k}$ (up to a scaling factor) as follows:

$$k(\mathbf{u}, \mathbf{v}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^T(\mathbf{u}-\mathbf{v})} d\nu(\boldsymbol{\omega}) \approx \frac{1}{B} \sum_{b=1}^{B} e^{i\boldsymbol{\omega}_b^T(\mathbf{u}-\mathbf{v})} = \tilde{k}(\mathbf{u}, \mathbf{v}),$$

where we define $\tilde{k}(\mathbf{u}, \mathbf{v}) = \langle \widetilde{\varphi}(\mathbf{u}), \widetilde{\varphi}(\mathbf{v}) \rangle_{\mathbb{C}^B}$ as the kernel function with the feature map given by

$$\widetilde{\varphi}(\mathbf{u}) = \frac{1}{\sqrt{B}} \left(\tilde{\varphi}_{\boldsymbol{\omega}_1}(\mathbf{u}), \ldots, \tilde{\varphi}_{\boldsymbol{\omega}_B}(\mathbf{u}) \right)^T = \frac{1}{\sqrt{B}} \left(e^{i\boldsymbol{\omega}_1^T\mathbf{u}}, \ldots, e^{i\boldsymbol{\omega}_B^T\mathbf{u}} \right)^T,$$

which is a random vector consisting of the Fourier features $\widetilde{\varphi}_{\boldsymbol{\omega}}(\mathbf{u}) = e^{i\boldsymbol{\omega}^T\mathbf{u}} \in \mathbb{C}$ (Rahimi and Recht, 2008). Such kernel approximations are frequently used in practice for computational efficiency (Rahimi and Recht, 2009; Le et al., 2013). As a default choice of k we take the Gaussian kernel with bandwidth σ, since in this case we have a convenient expression for the measure ν and we sample $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_B \sim N_d(\mathbf{0}, \sigma^{-2}I_d)$. The bandwidth σ is chosen as the median pairwise distance between all training responses $\{\mathbf{y}_i\}_{i=1}^n$, commonly referred to as the 'median heuristic' (Gretton et al., 2012c).

From the representation of MMD via the distribution embeddings (10), we can obtain that MMD two-sample test statistic $\mathcal{D}_{\mathrm{MMD}(\tilde{k})}$ using the approximate kernel $\tilde{k}$ is given by

$$\mathcal{D}_{\mathrm{MMD}(\tilde{k})} \left(\{\mathbf{u}_i\}_{i=1}^{|U|}, \{\mathbf{v}_i\}_{i=1}^{|V|} \right) = \frac{1}{B} \sum_{b=1}^{B} \left| \frac{1}{|U|} \sum_{i=1}^{|U|} \tilde{\varphi}_{\boldsymbol{\omega}_b}(\mathbf{u}_i) - \frac{1}{|V|} \sum_{i=1}^{|V|} \tilde{\varphi}_{\boldsymbol{\omega}_b}(\mathbf{v}_i) \right|^2.$$

Interestingly, $\mathcal{D}_{\mathrm{MMD}(\tilde{k})}$ is not only an MMD statistic on its own, but can also be viewed as a random approximation of the original MMD statistic $\mathcal{D}_{\mathrm{MMD}(k)}$ (8) using kernel k; by using the kernel representation (11), it can be written as

$$\mathcal{D}_{\mathrm{MMD}(k)} \left(\{\mathbf{u}_i\}_{i=1}^{|U|}, \{\mathbf{v}_i\}_{i=1}^{|V|} \right) = \int_{\mathbb{R}^d} \left| \frac{1}{|U|} \sum_{i=1}^{|U|} \tilde{\varphi}_{\boldsymbol{\omega}}(\mathbf{u}_i) - \frac{1}{|V|} \sum_{i=1}^{|V|} \tilde{\varphi}_{\boldsymbol{\omega}}(\mathbf{v}_i) \right|^2 d\nu(\boldsymbol{\omega}).$$

Finally, our DRF splitting criterion $\mathcal{D}(\cdot, \cdot)$ (1) is then taken to be the (scaled) MMD statistic $\frac{n_L n_R}{n_P^2} \mathcal{D}_{\mathrm{MMD}(\tilde{k})} \left(\{\mathbf{y}_i \mid \mathbf{x}_i \in C_L\}, \{\mathbf{y}_i \mid \mathbf{x}_i \in C_R\} \right)$ with the approximate random kernel $\tilde{k}$ used instead of k, which can thus be conveniently written as:

$$\frac{1}{B} \sum_{b=1}^{B} \frac{n_L n_R}{n_P^2} \left| \frac{1}{n_L} \sum_{\mathbf{x}_i \in C_L} \tilde{\varphi}_{\boldsymbol{\omega}_b}(\mathbf{y}_i) - \frac{1}{n_R} \sum_{\mathbf{x}_i \in C_R} \tilde{\varphi}_{\boldsymbol{\omega}_b}(\mathbf{y}_i) \right|^2, \tag{12}$$

where we recall that $n_P = |\{i \mid \mathbf{x}_i \in P\}|$ and n_L, n_R are defined analogously. The additional scaling factor $\frac{n_L n_R}{n_P^2}$ in (12) occurs naturally and compensates the increased variance of the test statistic for unbalanced splits; it also appears in the GRF (2) and CART (see representation (13)) splitting criteria.

The main advantage of the splitting criterion based on $\mathcal{D}_{\mathrm{MMD}(\tilde{k})}$ is that by using the representation (1) it can be easily computed for every possible splitting level l in $\mathcal{O}(B n_P)$ complexity, whereas the MMD statistic $\mathcal{D}_{\mathrm{MMD}(k)}$ using kernel k would require $\mathcal{O}(n_P^2)$ computational steps, which makes the overall complexity of the algorithm $\mathcal{O}(B \times N \times \mathrm{mtry} \times n \log n)$ instead of much slower $\mathcal{O}(N \times \mathrm{mtry} \times n^2)$.

We do not use the same approximate random kernel $\tilde{k}$ for different splits; for every parent node P we resample the frequency vectors $\{\omega_b\}_{b=1}^{B}$ defining the corresponding feature map $\tilde{\varphi}$. Using different $\tilde{k}$ at each node might help to better detect different distributional changes. Furthermore, having different random kernels for each node agrees well with the randomness of the Random Forests and helps making the trees more independent. Since the MMD statistic $\mathcal{D}_{\mathrm{MMD}(\tilde{k})}$ used for our splitting criterion is not only an approximation of $\mathcal{D}_{\mathrm{MMD}(k)}$, but is itself an MMD statistic, it inherits good power for detecting any difference in distribution of $\mathbf{Y}$ in the child nodes for moderately large data dimensionality d, even when B is reasonably small. One could even consider changing the number of random Fourier features B at different levels of the tree, as n_P varies, but for simplicity we take it to be fixed.

Relationship to CART. There is some similarity of our MMD-based splitting criterion (12) with the standard variance reduction CART splitting criterion (7) when $d = 1$, which can be rewritten as:

$$\frac{n_L n_R}{n_P^2} \left(\frac{1}{n_L} \sum_{\mathbf{x}_i \in C_L} y_i - \frac{1}{n_R} \sum_{\mathbf{x}_i \in C_R} y_i \right)^2. \tag{13}$$

The derivation can be found in Appendix B. From this representation, we see that the CART splitting criterion (7) is also equivalent to the GRF splitting criterion (2) when our target is the univariate conditional mean $\tau(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$ which is estimated for C_L and C_R

by the sample means $\hat{\tau}_L = \overline{y}_L$ and $\hat{\tau}_R = \overline{y}_R$. Therefore, as it compares the means of the univariate response Y in the child nodes, the CART criterion can only detect changes in the response mean well, which is sufficient for prediction of Y from $\mathbf{X}$, but might not be suitable for more complex targets. Similarly, for multivariate applications, aggregating the marginal CART criteria (Kocev et al., 2007; Segal and Xiao, 2011) across different components Y_i of the response can only detect changes in the means of their marginal distributions. However, it is possible in the multivariate case that the pairwise correlations or the variances of the responses change, while the marginal means stay (almost) constant. For an illustration on simulated data, see Figure 7. Additionally, aggregating the splitting criteria over d components of the response $\mathbf{Y}$ can reduce the signal size if only the distribution of a few components change. Our MMD-based splitting criterion (12) is able to avoid such difficulties as it implicitly inspects all aspects of the multivariate response distribution.

If one takes a trivial kernel $k_{\mathrm{id}}(y_i, y_j) = y_i y_j$ with the identity feature map $\varphi_{\mathrm{id}}(y) = y$, the distributional embedding (9) is given by $\mu(\mathcal{P}) = \mathbb{E}_{Y \sim \mathcal{P}}[Y]$ and thus the corresponding splitting criterion based on $\mathcal{D}_{\mathrm{MMD}(k_{\mathrm{id}})}$ (10) is exactly equal to the CART splitting criterion (7), which can be seen from its equivalent representation (13). Interestingly, Theorem 1 in Section 3 shows that the MMD splitting criterion with general kernel k can also be viewed as the abstract version of the CART criterion in the RKHS $\mathcal{H}$ corresponding to k (Fan et al., 2010), with the response variable being the feature map $\varphi(\mathbf{Y}) \in \mathcal{H}$. Therefore, DRF with the MMD splitting criterion can also be viewed as a forest-based method for estimation of the conditional embedding, which further justifies the proposed method. In Section 3 below, we use this relationship to derive interesting theoretical properties of DRF with the MMD splitting criterion.

3. Theoretical Results

In this section we first use the properties of the kernel mean embedding in order to relate DRF with the MMD splitting criterion to an abstract version of the standard Random Forest with the CART splitting criterion (Breiman, 2001), where the response is taking values in the corresponding RKHS. This representation reveals that DRF with the MMD splitting criterion can be viewed as a Random Forest estimator of the conditional mean embedding (CME) (Park and Muandet, 2020), similarly as the standard Random Forest estimates the conditional mean. This relationship is further exploited to adapt the existing theoretical results from the Random Forest literature to show that our estimate (5) of the conditional distribution of the response is consistent with respect to the MMD metric for probability measures and with a good rate. Finally, we show that this implies consistency of the induced DRF estimates for a range interesting targets $\tau(\mathbf{x})$, such as conditional CDFs or quantiles. The proofs of all results can be found in the Appendix B.

3.1 Casting DRF as a Random Forest in an RKHS

Recalling the notation from above, let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be the Reproducing kernel Hilbert space induced by the positive definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and let $\varphi : \mathbb{R}^d \to \mathcal{H}$ be its corresponding feature map. The kernel embedding function $\mu : \mathcal{M}_b(\mathbb{R}^d) \to \mathcal{H}$ maps any

bounded signed Borel measure $\mathcal{P}$ on $\mathbb{R}^d$ to an element $\mu(\mathcal{P}) \in \mathcal{H}$ defined by

$$\mu(\mathcal{P}) = \int_{\mathbb{R}^d} \varphi(\mathbf{y}) \, d\mathcal{P}(\mathbf{y}), \tag{14}$$

see also (9). Boundedness of k ensures that μ is indeed defined on all of $\mathcal{M}_b(\mathbb{R}^d)$, while continuity of k ensures that $\mathcal{H}$ is separable (Hsing and Eubank, 2015).

By considering the kernel embedding $\mu(\cdot)$ and using its linearity, the embedding of the distributional estimate $\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}))$ of DRF (5) can be written as the average of the embeddings of the empirical distributions of $\mathbf{Y}$ in the leaves containing $\mathbf{x}$ over all trees:

$$\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})) = \frac{1}{N} \sum_{k=1}^{N} \mu \left(\frac{1}{|\mathcal{L}_k(\mathbf{x})|} \sum_{\mathbf{x}_i \in \mathcal{L}_k(\mathbf{x})} \delta_{\mathbf{y}_i} \right) = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{|\mathcal{L}_k(\mathbf{x})|} \sum_{\mathbf{x}_i \in \mathcal{L}_k(\mathbf{x})} \mu(\delta_{\mathbf{y}_i}). \tag{15}$$

This is analogous to the prediction of the response for the standard univariate Random Forest, but where we average the embeddings $\mu(\delta_{\mathbf{y}_i}) = \varphi(\mathbf{y}_i) \in \mathcal{H}$ instead of the response values y_i themselves.

Furthermore, one can relate the MMD splitting criterion to the original CART criterion (7), which measures the mean squared prediction error for splitting a certain parent node P into children C_L and C_R. On one hand, from Equation (13) we see that the CART criterion also measures the squared distance between the response averages $\frac{1}{n_L} \sum_{\mathbf{x}_i \in C_L} y_i$ and $\frac{1}{n_R} \sum_{\mathbf{x}_i \in C_R} y_i$ in the child nodes, but on the other hand, Equation (10) shows that the MMD splitting criterion measures the RKHS distance between the embeddings of the empirical response distributions in C_L and C_R. This is summarized in the following theorem, which not only shows that the MMD splitting criterion can be viewed as the abstract CART criterion in the RKHS $\mathcal{H}$ (Fan et al., 2010), but also that DRF with the MMD splitting criterion can be viewed asymptotically as a greedy minimization of the average squared MMD distance between our estimate $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ and the truth $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$:

Theorem 1 *For any split of a parent node P into child nodes C_L and C_R, let $\hat{\mathbb{P}}_{split}(\mathbf{x}) = \sum_{j \in \{L,R\}} \mathbb{1}(\mathbf{x} \in C_j) \frac{1}{n_j} \sum_{\mathbf{x}_i \in C_j} \delta_{\mathbf{y}_i}$ denote the resulting estimate of the distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ when $\mathbf{x} \in P$. Then the MMD splitting criterion can be viewed as the version of the CART criterion (7) on $\mathcal{H}$:*

$$\arg\max_{split} \frac{n_L n_R}{n_P^2} \mathcal{D}_{MMD(k)} \left(\{\mathbf{y}_i \mid \mathbf{x}_i \in C_L\}, \{\mathbf{y}_i \mid \mathbf{x}_i \in C_R\} \right)$$

$$= \arg\min_{split} \frac{1}{n_P} \sum_{\mathbf{x}_i \in P} \left\| \mu(\delta_{\mathbf{y}_i}) - \mu(\hat{\mathbb{P}}_{split}(\mathbf{x}_i)) \right\|_{\mathcal{H}}^2.$$

Moreover, for any node P and any fixed distributional estimator $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$, we have:

$$\frac{1}{n_P} \sum_{\mathbf{x}_i \in P} \left\| \mu(\delta_{\mathbf{y}_i}) - \mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}_i)) \right\|_{\mathcal{H}}^2$$

$$= V_P + \mathbb{E} \left[\| \mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X})) - \mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X})) \|_{\mathcal{H}}^2 \mid \mathbf{X} \in P \right] + \mathcal{O}_p(n^{-1/2}),$$

where $V_P = \mathbb{E} \left[\| \mu(\delta_{\mathbf{Y}}) - \mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X})) \|_{\mathcal{H}}^2 \mid \mathbf{X} \in P \right]$ is a deterministic term not depending on the estimates $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$.

In conclusion, from the above results we see that by applying the kernel embedding (14), we can shift the perspective to the RKHS $\mathcal{H}$ and view DRF as the analogue of the original Random Forest for estimation of the CME $\mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})) = \mathbb{E}[\varphi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}]$ in an abstract Hilbert space $\mathcal{H}$. Like some traditional CME estimators (Song et al., 2009, 2013; Muandet et al., 2017; Park and Muandet, 2020) it is also of the form

$$\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})) = \sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) \cdot k(\mathbf{y}_i, \cdot). \tag{16}$$

It can be shown that, since DRF produces nonnegative weights that sum to one in (16), there is one-to-one correspondence between the resulting estimate in $\mathcal{H}$ and the empirical probability distribution in $\mathcal{M}_b(\mathbb{R}^d)$. Thus DRF can be seen as a CME estimator through (16), or directly as an estimator for the conditional distribution through (5). By contrast, other CME estimators of the form (16) have weights that are unconstrained and can be negative. Finding an appropriate distribution on $\mathbb{R}^d$ for a given mean-embedding (sometimes referred to as "distributional inverse image problem", see e.g. McCalman et al. (2013); Muandet et al. (2017)) is not straightforward in general. For certain tasks, such as sampling from the estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ or obtaining the plug-in estimates of some target functionals $\tau(\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}))$, this is crucial.

3.2 Convergence of Conditional Distribution Estimates

As we have seen, DRF can be viewed as the abstract version of the standard Random Forest when the response takes value in an RKHS. In principle, one could thus derive properties of DRF by adapting any existing theoretical result from the literature to the RKHS case. However, a lot of care is needed for making the results rigorous in this abstract setup, as many useful properties of $\mathbb{R}$ need not hold for infinite-dimensional $\mathcal{H}$. This section is inspired by the results from Wager and Athey (2018).

We suppose that the forest construction satisfies the following properties, which significantly facilitate the theoretical considerations of the method and ensure that our forest estimator is well behaved, as stated in Wager and Athey (2018):

(P1) (*Data sampling*) The bootstrap sampling with replacement, usually used in forest-based methods, is replaced by a subsampling step, where for each tree we choose a random subset of size s_n out of n training data points. We consider s_n going to infinity with n, with the rate specified below.

(P2) (*Honesty*) The data used for constructing each tree is split into two partOn the pitfalls of Gaussian scoring for causal discoverys; the first is used for determining the splits and the second for populating the leaves and thus for estimating the response.

(P3) (α-*regularity*) Each split leaves at least a fraction $0 < \alpha \leq 0.2$ of the available training sample on each side. Moreover, the trees are grown until every leaf contains between κ and $2\kappa - 1$ observations, for some fixed tuning parameter $\kappa \in \mathbb{N}$.

(P4) (*Symmetry*) The (randomized) output of a tree does not depend on the ordering of the training samples.

(P5) (*Random-split*) At every split point, the probability that the split occurs along the feature X_j is bounded below by π/p, for some $\pi > 0$ and for all $j = 1, \ldots, p$.

The validity of the above properties are easily ensured by the forest construction used.

From Equation (15), the prediction of DRF for a given test point $\mathbf{x}$ can be viewed as an element of $\mathcal{H}$. If we denote the i-th training observation by $\mathbf{Z}_i = (\mathbf{x}_i, \mu(\delta_{\mathbf{y}_i})) \in \mathbb{R}^p \times \mathcal{H}$, then by (15) we estimate the embedding of the true conditional distribution $\mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}))$ by the average of the corresponding estimates per tree:

$$\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})) = \frac{1}{N} \sum_{j=1}^{N} T(\mathbf{x}; \varepsilon_j, \mathcal{Z}_j),$$

where $\mathcal{Z}_k$ is a random subset of $\{\mathbf{Z}_i\}_{i=1}^n$ of size s_n chosen for constructing the j-th tree $\mathcal{T}_j$ and ε_j is a random variable capturing all randomness in growing $\mathcal{T}_j$, such as the choice of the splitting candidates. $T(\mathbf{x}; \varepsilon, \mathcal{Z})$ denotes the output of a single tree: i.e. the average of the terms $\mu(\delta_{\mathbf{Y}_i})$ over all data points $\mathbf{Z}_i$ contained in the leaf $\mathcal{L}(\mathbf{x})$ of the tree constructed from ε and $\mathcal{Z}$.

Since one can take the number of trees N to be arbitrarily large, we consider an "idealized" version of our estimator, as done in Wager and Athey (2017), which we denote as $\hat{\mu}_n(\mathbf{x})$:

$$\hat{\mu}_n(\mathbf{x}) = \binom{n}{s_n}^{-1} \sum_{i_1 < i_2 < \ldots < i_{s_n}} \mathbb{E}_\varepsilon \, T(\mathbf{x}; \varepsilon; \{\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}\}), \tag{17}$$

where the sum is taken over all $\binom{n}{s_n}$ possible subsets of $\{\mathbf{Z}_i\}_{i=1}^n$. We have that $\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})) \to \hat{\mu}_n(\mathbf{x})$ as $N \to \infty$, while keeping the other variables constant, and thus we assume for simplicity that those two quantities are the same.

Our main result shows that, under similar assumptions as in Wager and Athey (2017), the embedding of our conditional distribution estimator $\hat{\mu}_n(\mathbf{x}) = \mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}))$ consistently estimates $\mu(\mathbf{x}) := \mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}))$ with respect to the RKHS norm with a certain rate:

Theorem 2 *Suppose that our forest construction satisfies properties **(P1)**–**(P5)**. Assume additionally that k is a bounded and continuous kernel and that we have a random design with $\mathbf{X}_1, \ldots, \mathbf{X}_n$ independent and identically distributed on $[0,1]^p$ with a density bounded away from 0 and infinity. If the subsample size s_n is of order n^β for some $0 < \beta < 1$, the mapping*

$$\mathbf{x} \mapsto \mu(\mathbf{x}) = \mathbb{E}[\mu(\delta_{\mathbf{Y}}) \mid \mathbf{X} = \mathbf{x}] \in \mathcal{H},$$

is Lipschitz and $\sup_{\mathbf{x} \in [0,1]^p} \mathbb{E}[\|\mu(\delta_{\mathbf{Y}})\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}] < \infty$, we obtain the consistency w.r.t. the RKHS norm:

$$\|\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\|_{\mathcal{H}} = \mathcal{O}_p\left(n^{-\gamma}\right), \tag{18}$$

for $\gamma = \frac{1}{2} \min\left(1 - \beta, \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})} \frac{\pi}{p} \cdot \beta\right)$.

Remark 3 *The rate in (18) is analogous to the one from Wager and Athey (2018), who used it further to derive the asymptotic normality of the Random Forest estimator in $\mathbb{R}$. Unfortunately, this alone is not enough to establish asymptotic normality of $(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}))/\sigma_n$*

as an element of $\mathcal{H}$. To do so, one needs to prove a functional central limit theorem with a Gaussian limiting process in the Hilbert space $\mathcal{H}$. This then allows to deduce asymptotic normality of smooth real-valued functionals. We will provide the detailed derivations in future work.

3.3 Convergence of the Induced Estimates

The above result shows that DRF estimate $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ converges fast to the truth $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ in the MMD distance, i.e. the RKHS distance between the corresponding embeddings. Even though this is interesting on its own, ultimately we want to relate this result to estimation of certain distributional targets $\tau(\mathbf{x}) = \tau(\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}))$.

For any $f \in \mathcal{H}$, we have that the DRF estimate of the target $\tau(\mathbf{x}) = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}]$ equals the dot product $\langle f, \hat{\mu}_n(\mathbf{x}) \rangle_{\mathcal{H}}$ in the RKHS:

$$\langle f, \hat{\mu}_n(\mathbf{x}) \rangle_{\mathcal{H}} = \left\langle f, \int_{\mathbb{R}^d} \varphi(\mathbf{y}) d\hat{\mathbb{P}}(\mathbf{y} \mid \mathbf{X} = \mathbf{x}) \right\rangle_{\mathcal{H}} = \int_{\mathbb{R}^d} f(\mathbf{y}) \, d\hat{\mathbb{P}}(\mathbf{y} \mid \mathbf{X} = \mathbf{x}) = \sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) f(\mathbf{y}_i),$$

where we recall the weighting function $w_{\mathbf{x}}(\cdot)$ induced by the forest (4). Therefore, the consistency result (18) in Theorem 2 directly implies that

$$\sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) f(\mathbf{y}_i) = \langle f, \hat{\mu}_n(\mathbf{x}) \rangle_{\mathcal{H}} \xrightarrow{p} \langle f, \mu(\mathbf{x}) \rangle_{\mathcal{H}} = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}] \quad \text{for any } f \in \mathcal{H}, \quad (19)$$

i.e. that the DRF consistently estimates the targets of the form $\tau(\mathbf{x}) = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}]$, for $f \in \mathcal{H}$. From (18) we also obtain the rate of convergence when $s_n \asymp n^{\beta}$:

$$\left| \sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) f(\mathbf{y}_i) - \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}] \right| = \mathcal{O}_p \left(n^{-\gamma} \|f\|_{\mathcal{H}} \right),$$

for γ as in Theorem 2. When k is continuous, it is well known that all elements of $\mathcal{H}$ are continuous, see e.g. Hsing and Eubank (2015). Under certain assumptions on the kernel and its input space, holding for several popular kernels, (e.g. the Gaussian kernel) (Sriperumbudur, 2016), we can generalize the convergence result (19) to any bounded and continuous function $f : \mathbb{R}^d \to \mathbb{R}$, as the convergence of measures $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}) \to \mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ in the MMD metric will also imply their weak convergence, i.e. k metrizes weak convergence (Sriperumbudur, 2016; Simon-Gabriel and Schölkopf, 2018; Simon-Gabriel et al., 2020):

Corollary 4 *Assume that one of the following two sets of conditions holds:*

(a) The kernel k is bounded, (jointly) continuous and has

$$\int \int k(\mathbf{x}, \mathbf{y}) d\mathcal{P}(\mathbf{x}) d\mathcal{P}(\mathbf{y}) > 0 \quad \forall \mathcal{P} \in \mathcal{M}_b(\mathbb{R}^d) \setminus \{0\}. \quad (20)$$

  Moreover, $\mathbf{y} \mapsto k(\mathbf{y}_0, \mathbf{y})$ is vanishing at infinity, for all $\mathbf{y}_0 \in \mathbb{R}^d$.

(b) The kernel k is bounded, shift-invariant, (jointly) continuous and ν in the Bochner representation in (11) is supported on all of $\mathbb{R}^d$. Moreover, $\mathbf{Y}$ takes its values almost surely in a closed and bounded subset of $\mathbb{R}^d$.

18

Then, under the conditions of Theorem 2, we have for any bounded and continuous function $f : \mathbb{R}^d \to \mathbb{R}$ that DRF consistently estimates the target $\tau(\mathbf{x}) = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}]$ for any $\mathbf{x} \in [0,1]^p$:

$$\sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) f(\mathbf{y}_i) \xrightarrow{p} \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}].$$

Recalling the Portmanteau Lemma on separable metric spaces, see e.g. Dudley (2002, Chapter 11), this has several other interesting consequences, such as the consistency of CDF and quantile estimates; Let $F_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}}(\cdot)$ be the conditional CDF of $\mathbf{Y}$ and for any index $1 \le i \le d$, let $F_{Y_i \mid \mathbf{X} = \mathbf{x}}(\cdot)$ be the conditional CDF of Y_i and $F_{Y_i \mid \mathbf{X} = \mathbf{x}}^{-1}(\cdot)$ its generalized inverse, i.e. the quantile function. Let $\hat{F}_{Y_i \mid \mathbf{X} = \mathbf{x}}(\cdot)$ and $\hat{F}_{Y_i \mid \mathbf{X} = \mathbf{x}}^{-1}(\cdot)$ be the corresponding DRF estimates via weighting function (6). Then we have the following result:

Corollary 5 *Under the conditions of Corollary 4, for any $1 \le i \le d$, we have*

$$\hat{F}_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}}(\mathbf{t}) \xrightarrow{p} F_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}}(\mathbf{t})$$
$$\hat{F}_{Y_i \mid \mathbf{X} = \mathbf{x}}^{-1}(t) \xrightarrow{p} F_{Y_i \mid \mathbf{X} = \mathbf{x}}^{-1}(t),$$

for all points of continuity $\mathbf{t} \in \mathbb{R}^d$ and $t \in \mathbb{R}$ of $F_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}}(\cdot)$ and $F_{Y_i \mid \mathbf{X} = \mathbf{x}}^{-1}(\cdot)$ respectively.

4. Applications and Numerical Experiments

The goal of this section is to demonstrate the versatility and applicability of DRF for many practical problems. We show that DRF can be used not only as an estimator of the multivariate conditional distribution, but also as a two-step method to easily obtain out-of-the box estimators for various, and potentially complex, targets $\tau(\mathbf{x})$.

Our main focus lies on the more complicated targets which cannot be that straight-forwardly approached by conventional methods. However, we also illustrate the usage of DRF for certain applications for which there already exist several well-established methods. Whenever possible in such cases, we compare the performance of DRF with the specialized, task-specific methods to show that, despite its generality, there is at most a very small loss of precision. However, we should point out that for many targets such as, that can not be written in a form of a conditional mean or a conditional quantile, for example, conditional correlation, direct comparison of the accuracy is not possible for real data, since no suitable loss function exists and the ground truth is unknown. Finally, we show that, in addition to directly estimating certain targets, DRF can also be a very useful tool for many different applications, such as causality and fairness.

Detailed descriptions of all competing methods, data sets and the corresponding analyses can be found in Appendix C, and some additional simulations can be found in the Appendix D.

4.1 Estimation of Conditional Multivariate Distributions

In order to provide good estimates for any target $\tau(\mathbf{x}) = \tau(\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}))$, our method needs to estimate the conditional multivariate distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ well. Therefore, we
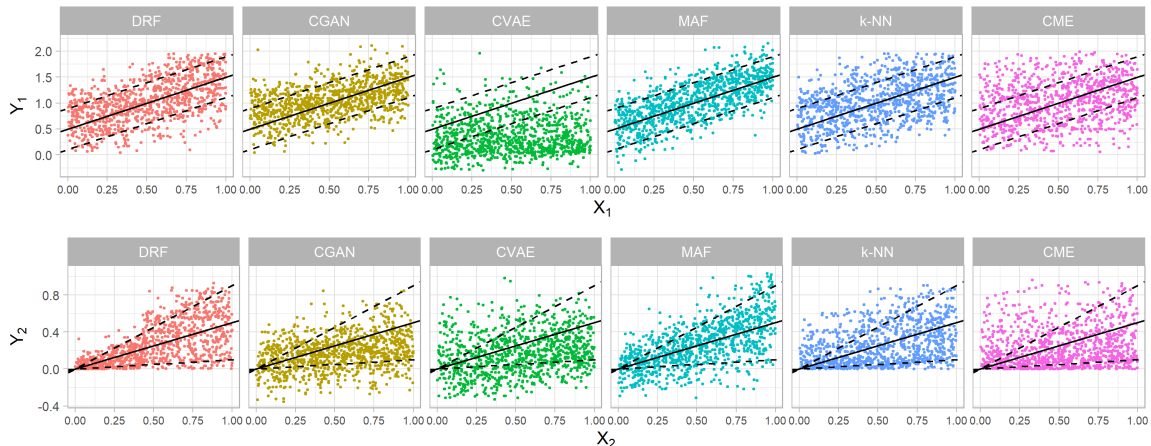
Figure 3: The illustration of the estimated joint conditional distribution obtained by different methods for the toy example (21). For 1000 randomly generated test points $\mathbf{X}_{\text{test}} \sim U(0,1)^p$ the top row shows the estimated distribution of the response component Y_1, whereas the bottom row shows the estimated distribution of Y_2. The 0.1 and 0.9 quantiles of the true conditional distribution are indicated by a dashed black line, whereas the conditional mean is shown as a black solid line.

first investigate here the accuracy of the DRF estimate (5) of the full conditional distribution and compare its performance with the performance of several existing methods.

In addition to a few simple methods such as the k-nearest neighbors or the kernel regression, which locally weight the training points, we also consider the CME estimator of Park and Muandet (2020) and several advanced machine learning methods such as the Conditional Generative Adversarial Network (CGAN) (Mirza and Osindero, 2014; Aggarwal et al., 2019), Conditional Variational Autoencoder (CVAE) (Sohn et al., 2015) and Masked Autoregressive Flow (Papamakarios et al., 2017). It is worth mentioning that the focus in the machine learning literature has been more on applications where d is very large (e.g. pixels of an image) and p is very small (such as image labels). Even though some methods do not provide the estimated conditional distribution in a form as simple as DRF, one is still able to sample from the estimated distribution and thus perform any subsequent analysis and make fair comparisons between the methods. For the CME estimator we simply set the negative weights to zero and renormalize, such that the weights are nonnegative and sum to one.

We first illustrate the estimated distributions by the above methods on a toy example where $n = 1000, p = 10, d = 2$ and

$$Y_1 \perp\!\!\!\perp Y_2 \mid \mathbf{X} = \mathbf{x}, \quad Y_1 \mid \mathbf{X} = \mathbf{x} \sim U(x_1, x_1 + 1), \quad Y_2 \mid \mathbf{X} = \mathbf{x} \sim U(0, x_2), \quad \mathbf{X} \sim U(0,1)^p. \tag{21}$$

In the above example X_1 affects the mean of Y_1, whereas X_2 affects the both mean and variance of Y_2, and $X_3, \ldots, X_p$ have no impact. The results can be seen in Figure 3. We see

	jura	slump	wq	enb	atp1d	atp7d	scpf	sf1	sf2	copula	wage	births1	births2	air
n	359	103	1K	768	337	296	143	323	1K	5K	10K	10K	10K	10K
p	15	7	16	8	370	370	8	21	22	10	73	23	24	15
d	3	3	14	2	6	6	3	3	6	2	2	2	4	6
DRF	3.9	**4.0**	**22.5**	2.1	7.3	**7.0**	**2.0**	-24.2	**-24.3**	**2.8**	**2.8**	2.5	**4.2**	**8.5**
CGAN	10.8	5.3	27.3	3.5	10.4	363	4.8	9.8	21.1	5.8	360	**2.4**	>1K	11.8
CVAE	4.8	37.8	36.8	2.6	>1K	>1K	108.8	8.6	>1K	2.9	>1K	>1K	49.7	9.6
MAF	4.6	4.5	23.9	3.0	8.0	8.1	2.6	4.7	3.8	2.9	3.0	2.5	>1K	8.5
k-NN	4.5	5.0	23.4	2.4	8.8	8.6	4.1	-22.4	-19.7	2.9	**2.8**	2.7	4.4	8.8
kernel	4.1	4.2	23.0	**2.0**	**6.6**	7.1	2.9	-23.0	-20.6	2.8	2.9	2.6	4.3	8.4
RF	7.1	12.1	35.2	5.7	12.7	13.3	16.7	3.9	2.2	5.8	6.1	5.0	8.3	13.9
XGBoost	11.4	38.3	25.9	3.0	>1K	>1K	>1K	0.3	1.6	3.5	2.9	>1K	>1K	12.8
DNN	4.0	4.2	23.3	2.6	8.6	8.7	2.6	2.3	2.2	2.9	3.0	2.6	5.4	8.6
CME	**3.2**	4.9	23.2	2.9	8.5	8.4	2.5	**-24.4**	**-24.3**	2.8	3.5	3.8	15.2	8.8

Table 1: NLPD loss computed on out-of-sample observations for the estimated conditional distributions obtained by several different methods (corresponding to rows) for many real data sets (corresponding to columns). The best method is indicated in bold.

that, unlike some other methods, DRF is able to balance the importance of the predictors X_1 and X_2 and thus to estimate the distributions of Y_1 and Y_2 well.

One can do a more extensive comparison on a collection of real data sets. We use the benchmark data sets from the multi-target regression literature (Tsoumakas et al., 2011) together with some additional ones created from the data sets described throughout this paper. The performance of DRF is compared with the performance of other existing methods for nonparametric estimation of multivariate distributions by using the Negative Log Predictive Density (NLPD) loss, which evaluates the logarithm of the induced multivariate density estimate (Quinonero-Candela et al., 2005). As the number of test points grows to infinity, NLPD loss becomes equivalent to the average KL divergence between the estimated and the true conditional distribution and is thus able to capture how well one estimates the whole distribution, instead of only its mean.

In addition to the methods mentioned above, we also include some methods that are intended only for mean prediction, by assuming that the distribution of the response around its mean is homogeneous, i.e. that the conditional distribution $\mathbb{P}(\mathbf{Y} - \mathbb{E}[\mathbf{Y} \mid \mathbf{X}] \mid \mathbf{X} = \mathbf{x})$ does not depend on $\mathbf{x}$. This is fitted by regressing each component of $\mathbf{Y}$ separately on $\mathbf{X}$ and using the pooled residuals. We consider the standard nonparametric regression methods such as Random Forest (Breiman, 2001), XGBoost (Chen and Guestrin, 2016), and Deep Neural Networks (Goodfellow et al., 2016).

The results are shown in Table 1. We see that DRF performs well for a wide range of sample size and problem dimensionality, especially in problems where p is large and

d is moderately big. It does so without the need for any tuning or involved numerical optimization.

4.2 Estimation of Statistical Functionals

Because DRF represents the estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}) = \sum_i w_{\mathbf{x}}(\mathbf{x}_i) \cdot \delta_{\mathbf{y}_i}$ in a convenient form by using weights $w_{\mathbf{x}}(\mathbf{x}_i)$, a plug-in estimator $\tau(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}))$ of many common real-valued statistical functionals $\tau(\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})) \in \mathbb{R}$ can be easily constructed from $w_{\mathbf{x}}(\cdot)$.
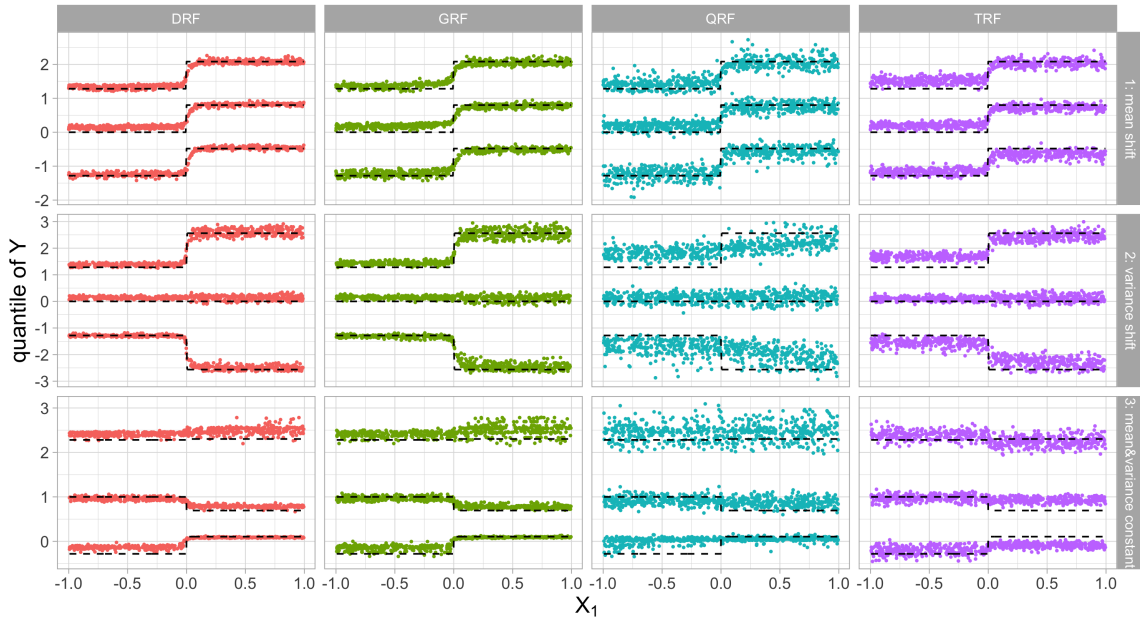
Figure 4: Scatter plot of predictions of the $0.1, 0.5$ and 0.9 quantiles against X_1 for randomly generated 500 test data points $\mathbf{X}_{\text{test}} \sim U(-1, 1)^p$. The true values of the quantiles are displayed by black dashed lines. The columns corresponds to different methods DRF (red), GRF (green), QRF (blue), TRF (purple). The rows correspond to different simulation scenarios. The first two are taken from Athey et al. (2019).

We first investigate the performance for the classical problem of univariate quantile estimation on simulated data. We consider the following three data generating mechanisms with $p = 40, n = 2000$ and $\mathbf{X}_i \overset{i.i.d.}{\sim} U(-1, 1)^p$:

- Scenario 1: $Y \sim N(0.8 \cdot \mathbb{1}(X_1 > 0), 1)$ (mean shift based on X_1)

- Scenario 2: $Y \sim N(0, (1 + \mathbb{1}(X_1 > 0))^2)$ (variance shift based on X_1)

- Scenario 3: $Y \sim \mathbb{1}(X_1 \leq 0) \cdot N(1, 1) + \mathbb{1}(X_1 > 0) \cdot \text{Exp}(1)$ (distribution shift based on X_1, constant mean and variance)

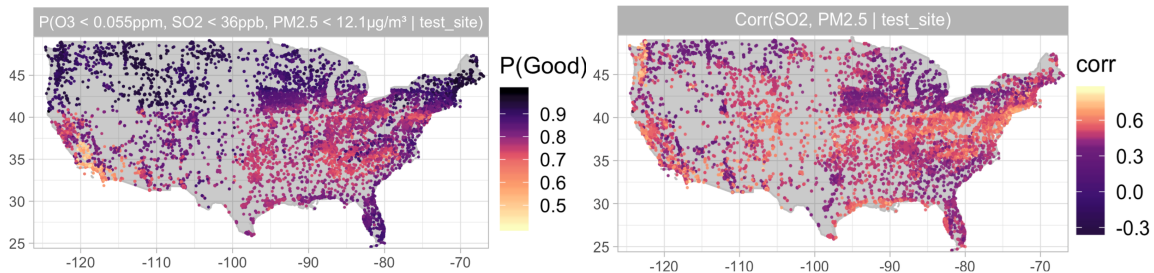The first two scenarios correspond exactly to the examples given in Athey et al. (2019).

Figure 5: Estimates of the probability $\mathbb{P}(AQI \leq 50 \,|\, \text{test site})$ (left) and the conditional correlation (right) derived from the DRF estimate of the multivariate conditional distribution.

In Figure 4 we can see the corresponding estimates of the conditional quantiles for DRF, Quantile Regression Forest (QRF) (Meinshausen, 2006), which uses the same forest construction with CART splitting criterion as the original Random Forest (Breiman, 2001) but estimates the quantiles from the induced weighting function, Generalized Random Forests (GRF) (Athey et al., 2019) with a splitting criterion specifically designed for quantile estimation and Transformation Forests (TRF) (Hothorn and Zeileis, 2021). We see that DRF is performing very well even compared to methods that are specifically tailored to quantile estimation.

The multivariate setting is however more interesting, as one can use DRF to compute much more interesting statistical functionals $\tau(\mathbf{x})$. We illustrate this in Figure 5 for the air quality data set, described in Section 2.2. The left plot shows one value of the estimated multivariate CDF, specifically the estimated probability of the event that the air quality index (AQI) is at most 50 at a given test site. This corresponds to the "Good" category and means that the amount of every air pollutant is below a certain threshold determined by the EPA. Such probability estimates can be easily obtained by summing the weights of the training points belonging to the event of interest. For both plots in Figures 2 and 5, we train the single DRF with the same set of predictor variables and take the three pollutants O_3, SO_2 and PM2.5 as the responses. In this way we still have training data from many different sites.

In order to investigate the accuracy of the conditional CDF obtained by DRF, we compare the estimated probabilities with estimates of the standard univariate classification forest (Breiman, 2001) with the response $\mathbb{1}(AQI \leq 50)$. In the left plot of Figure 6, we can see that the DRF estimates of the $\mathbb{P}(AQI \leq 50 \,|\, \mathbf{X} = \mathbf{x})$ (also visualized in Figure 5) are quite similar to the estimates of the classification forest predicting the outcome $\mathbb{1}(AQI \leq 50)$. Furthermore, the cross-entropy loss evaluated on the held-out measurements equals 0.4671 and 0.4663 respectively, showing almost no loss of precision. In general, estimating the simple functionals from the weights provided by DRF comes usually at a small to no loss compared to the classical methods specifically designed for this task.

In addition to the classical functionals $\tau(\mathbf{x})$ in the form of an expectation $\mathbb{E}(f(\mathbf{Y}) \,|\, \mathbf{X} = \mathbf{x})$ or a quantile $Q_\alpha(f(\mathbf{Y}) \,|\, \mathbf{X} = \mathbf{x})$ for some function $f : \mathbb{R}^d \to \mathbb{R}$, which can also be computed by solving the corresponding one-dimensional problems, additional interesting statistical
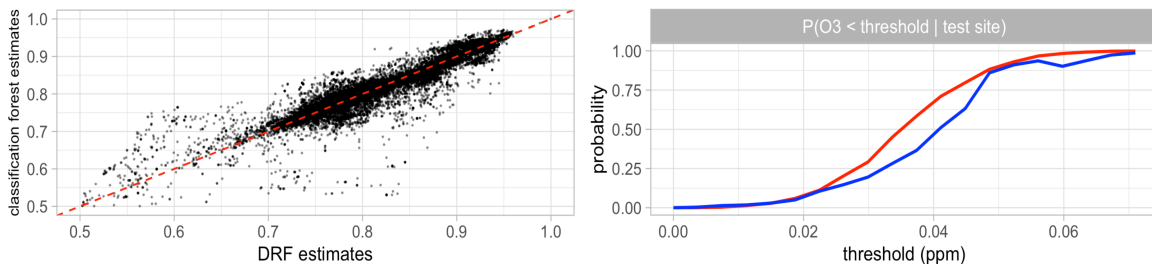
Figure 6: Left: Comparison of the CDF estimates obtained by DRF (displayed also in the left plot of Figure 5) and by the classification forest. Right: Example how the CDF estimated by using the classification forest (blue) need not be monotone, whereas the DRF estimates (red) are well-behaved.

functionals with intrinsically multivariate nature that are not that simple to estimate directly are accessible by DRF, such as, for example, the conditional correlations $\mathrm{Cor}(Y_i, Y_j \mid \mathbf{X} = \mathbf{x})$. As an illustration, the estimated correlation of the sulfur dioxide (SO_2) and fine particulate matter (PM2.5) is shown in the right plot of Figure 5. The plot reveals also that the local correlation in many big cities is slightly larger than in its surroundings, which can be explained by the fact that the industrial production directly affects the levels of both pollutants.

A big advantage of the target-free forest construction of DRF is that all subsequent targets are computed from same the weighting function $w_{\mathbf{x}}$ obtained from a single forest fit. First, this is computationally more efficient, since we do not need for every target of interest to fit the method specifically tailored to it. For example, estimating the CDF with classification forests requires fitting one forest for each function value. Secondly and even more importantly, since all statistical functionals are plug-in estimates computed from the same weighting function, the obtained estimates are mathematically well-behaved and mutually compatible. For example, if we estimate $\mathrm{Cor}(Y_i, Y_j \mid \mathbf{X} = \mathbf{x})$ by separately estimating the terms $\mathrm{Cov}(Y_i, Y_j \mid \mathbf{X} = \mathbf{x})$, $\mathrm{Var}(Y_i \mid \mathbf{X} = \mathbf{x})$, and $\mathrm{Var}(Y_j \mid \mathbf{X} = \mathbf{x})$, one can not in general guarantee the estimate to be in the range $[-1, 1]$, but this is possible with DRF. Alternatively, the correlation or covariance matrices that are estimated entrywise are guaranteed to be positive semi-definite if one uses DRF. As an additional illustration, Figure 6 shows that the estimated (univariate) CDF using the classification forest need not be monotone due to random errors in each predicted value, which can not happen with the DRF estimates.

4.3 Conditional Copulas and Conditional Independence Testing

One can use the weighting function not only to estimate certain functionals, but also to obtain more complex objects, such as, for example, the conditional copulas. The well-known Sklar's theorem (Sklar, 1959) implies that at a point $\mathbf{x} \in \mathbb{R}^p$, the conditional CDF $\mathbb{P}(\mathbf{Y} \leq \mathbf{y} \mid \mathbf{X} = \mathbf{x}) = \mathbb{P}(Y_1 \leq y_1, \ldots, Y_d \leq y_d \mid \mathbf{X} = \mathbf{x})$ can be represented by a CDF $C_{\mathbf{x}}$ on $[0, 1]^d$, the conditional copula at $\mathbf{x}$, and d conditional marginal CDFs $F_{Y_i \mid \mathbf{X} = \mathbf{x}}(y) = \mathbb{P}(Y_i \leq y \mid \mathbf{X} = \mathbf{x})$
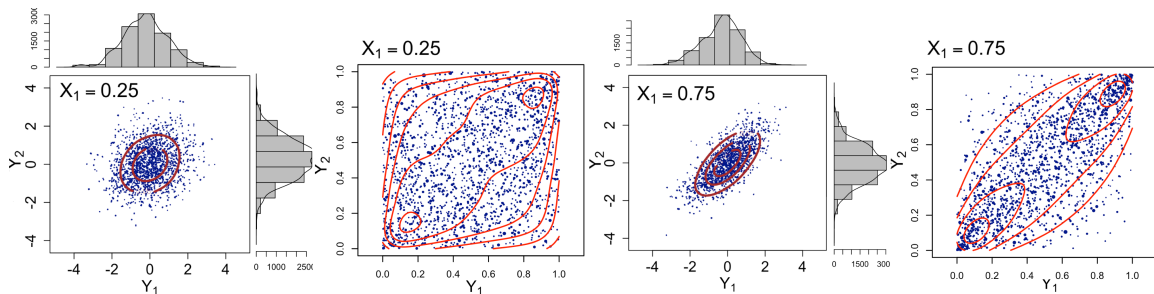
Figure 7: Estimated conditional joint distribution of (Y_1, Y_2) and conditional copulas obtained by DRF at different test points $\mathbf{x}$, where x_1 equals 0.25 and 0.75 respectively. The red lines are the contours of the true multivariate density function.

for $1 \leq i \leq d$, as follows:

$$\mathbb{P}(\mathbf{Y} \leq \mathbf{y} \mid \mathbf{X} = \mathbf{x}) = C_{\mathbf{x}} \left(F_{Y_1 \mid \mathbf{X} = \mathbf{x}}(y_1), \ldots, F_{Y_d \mid \mathbf{X} = \mathbf{x}}(y_d) \right). \tag{22}$$

Copulas capture the dependence of the components Y_i by the joint distribution of the corresponding quantile levels of the marginal distributions: $F_{Y_i \mid \mathbf{X} = \mathbf{x}}(Y_i) \in [0, 1]$. Decomposing the full multivariate distribution to marginal distributions and the copula is a very useful technique used in many fields such as risk analysis or finance (Cherubini et al., 2004). Using DRF enables us to estimate copulas conditionally, either by fitting certain parametric model or nonparametrically, directly from the weights.

To illustrate this, consider an example where the 5-dimensional $\mathbf{Y}$ is generated from the equicorrelated Gaussian copula $\mathbf{Y} = (Y_1, \ldots, Y_5) \mid \mathbf{X} = \mathbf{x} \sim C_{\rho(\mathbf{x})}^{\text{Gauss}}$ conditionally on the covariates $\mathbf{X}$ with distribution $\mathbf{X}_i \overset{i.i.d.}{\sim} U(0, 1)^p$, where $p = 30$ and $n = 5000$. All Y_i have a $N(0, 1)$ distribution marginally, but their conditional correlation for $i \neq j$ is given by $\text{Cor}(Y_i, Y_j) = \rho(\mathbf{x}) = x_1$. Figure 7 shows that DRF estimates the full conditional distribution at different test points $\mathbf{x}$ quite accurately and thus we can obtain a good nonparametric estimate of the conditional copula as follows. First, for each component Y_i, we compute the corresponding marginal CDF estimate $\hat{F}_{Y_i \mid \mathbf{X} = \mathbf{x}}(\cdot)$ from the weights. Second, we map each response $\mathbf{y}_i \rightarrow \mathbf{u}_i := \left(\hat{F}_{Y_1 \mid \mathbf{X} = \mathbf{x}}((\mathbf{y}_i)_1), \ldots, \hat{F}_{Y_d \mid \mathbf{X} = \mathbf{x}}((\mathbf{y}_i)_d) \right)$. The copula estimate is finally obtained from the weighted distribution $\sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) \delta_{\mathbf{u}_i}$, from which we sample the points in Figure 7 in order to visualize the copula.

If we want to instead estimate the copula parametrically, we need to find the choice of parameters for a given model family which best matches the estimated conditional distribution, e.g. by weighted maximum likelihood estimation (MLE). For the above example, the correlation parameter of the Gaussian copula can be estimated by computing the weighted correlation with weights $\{w_{\mathbf{x}}(\mathbf{x}_i)\}_{i=1}^{n}$. The left plot in Figure 8 shows the resulting estimates of the conditional correlation $\text{Cor}(Y_1, Y_2 \mid \mathbf{X} = \mathbf{x})$ obtained from DRF_{MMD}, which uses the MMD splitting criterion (12) described in Section 2.3.1, and DRF_{CART}, which aggregates the marginal CART criteria (Kocev et al., 2007; Segal and Xiao, 2011). We see that DRF_{MMD} is able to detect the distributional heterogeneity and provide good estimates
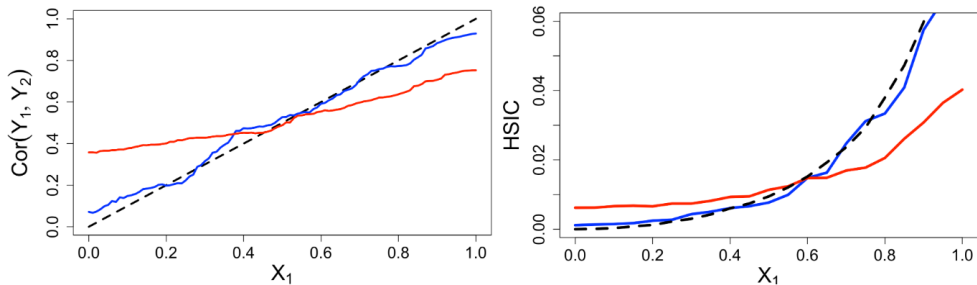
25

Figure 8: Estimated conditional correlation of Y_1 and Y_2 (left) and estimated conditional dependence quantified by HSIC statistic (right), obtained by $\mathrm{DRF}_{\mathrm{MMD}}$ (blue) and $\mathrm{DRF}_{\mathrm{CART}}$ (red) respectively. For every test point, we set $X_j = 0.5, j \neq 1$. Black dashed curve indicates the population values.

of the conditional correlation. On the other hand, $\mathrm{DRF}_{\mathrm{CART}}$ cannot detect the change in distribution of $\mathbf{Y}$ caused by X_1 that well. The distributional heterogeneity can not only occur in marginal distribution of the responses (a case extensively studied in the literature), but also in their interdependence structure described by the conditional copula $C_{\mathbf{x}}$, as one can see from decomposition (22). Since $\mathrm{DRF}_{\mathrm{MMD}}$ relies on a distributional metric for its splitting criterion, it is capable of detecting any change in distribution (Gretton et al., 2007a), whereas aggregating marginal CART criteria for $Y_1, \ldots, Y_d$ in $\mathrm{DRF}_{\mathrm{CART}}$ only captures the changes in the marginal means.

This is further illustrated for a related application of conditional independence testing, where we compute some dependence measure from the obtained weights. For example, we can test the independence $Y_1 \perp\!\!\!\perp Y_2$ conditionally on the event $\mathbf{X} = \mathbf{x}$ by using the Hilbert Schmidt Independence Criterion (HSIC) (Gretton et al., 2007b), which measures the difference between the joint distribution and the product of the marginal distributions. The right plot of Figure 8 shows that the $\mathrm{DRF}_{\mathrm{MMD}}$ estimates are quite close to the population value of the HSIC, unlike the ones obtained by $\mathrm{DRF}_{\mathrm{CART}}$.

4.4 Heterogeneous Regression and Causal Effect Estimation

In this and the following section, we illustrate that, in addition to direct estimation of certain targets, DRF can also be a useful tool for complex statistical problems and applications, such as causality.

Suppose we would like to investigate the relationship between some (univariate) quantity of interest Y and certain predictors $\mathbf{W}$ from heterogeneous data, where the change in distribution of $(\mathbf{W}, Y)$ can be explained by some other covariates $\mathbf{X}$. Very often in causality applications, $\mathbf{W}$ is a (multivariate) treatment variable, Y is the outcome, which is commonly, but not necessarily, binary, and $\mathbf{X}$ is a set of observed confounding variables for which we need to adjust if we are interested in the causal effect of $\mathbf{W}$ on Y. This is illustrated by the following causal graph:
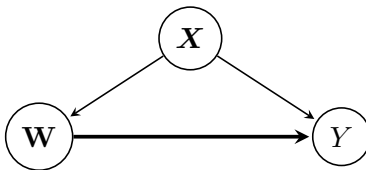
The problem of nonparametric confounding adjustment is hard; not only can the marginal distributions of Y and $\mathbf{W}$ be affected by $\mathbf{X}$, thus inducing spurious associations due to confounding, but the way how $\mathbf{W}$ affects Y can itself depend on $\mathbf{X}$, i.e. the treatment effect might be heterogeneous. The total causal effect can be computed by using the adjustment formula (Pearl, 2009):

$$\mathbb{E}[Y \mid do(\mathbf{W}=\mathbf{w})] = \int \mathbb{E}[Y \mid do(\mathbf{W}=\mathbf{w}), \mathbf{X}=\mathbf{x}] \, \mathbb{P}(\mathbf{X}=\mathbf{x} \mid do(\mathbf{W}=\mathbf{w}))d\mathbf{x}$$

$$= \int \mathbb{E}[Y \mid \mathbf{W}=\mathbf{w}, \mathbf{X}=\mathbf{x}] \, \mathbb{P}(\mathbf{X}=\mathbf{x})d\mathbf{x}. \tag{23}$$

In general, implementing do-calculus for finite samples and potentially non-discrete data might not be straightforward and comes with certain difficulties. In this case, the standard approach would be to estimate the conditional mean $\mathbb{E}[Y \mid \mathbf{W}=\mathbf{w}, \mathbf{X}=\mathbf{x}]$ nonparametrically by regressing Y on $(\mathbf{X}, \mathbf{W})$ with some method of choice and to average out the estimates over different $\mathbf{x}$ sampled from the observed distribution of $\mathbf{X}$. Using DRF for this approach is not necessary, but has an advantage that one can easily estimate the full interventional distribution $\mathbb{P}(Y \mid do(\mathbf{W}=\mathbf{w}))$ and not only the interventional mean $\mathbb{E}[Y \mid do(\mathbf{W}=\mathbf{w})]$.

Another way of computing the causal effect, which allows to add more structure to the problem, is explained in the following: We use DRF to first fit the forest with the multivariate response $(\mathbf{W}, Y)$ and the predictors $\mathbf{X}$. In this way, one can for any point of interest $\mathbf{x}$ obtain the joint distribution of $(\mathbf{W}, Y)$ conditionally on the event $\mathbf{X}=\mathbf{x}$ and then the weights $\{w_{\mathbf{x}}(\mathbf{x}_i)\}_{i=1}^{n}$ can be used as an input for some regression method for regressing Y on $\mathbf{W}$ in the second step. This conditional regression fit might be of an independent interest, but it can also be used for estimating the causal effect $\mathbb{E}[Y \mid do(\mathbf{W}=\mathbf{w})]$ from (23), by averaging the estimates $\mathbb{E}[Y \mid \mathbf{W}=\mathbf{w}, \mathbf{X}=\mathbf{x}]$ over $\mathbf{x}$, where $\mathbf{x}$ is sampled from the empirical observation of $\mathbf{X}$. In this way one can efficiently exploit and incorporate any prior knowledge of the relationship between $\mathbf{W}$ and Y, such as, for example, monotonicity, smoothness or that it satisfies a certain parametric regression model, without imposing any assumptions on the effect of $\mathbf{X}$ on $(\mathbf{W}, Y)$. Furthermore, one might be able to better extrapolate to the regions of space where $\mathbb{P}(\mathbf{W}=\mathbf{w}, \mathbf{X}=\mathbf{x})$ is small, compared to the standard approach which computes $\mathbb{E}[Y \mid \mathbf{W}=\mathbf{w}, \mathbf{X}=\mathbf{x}]$ directly, by regressing Y on $(\mathbf{W}, \mathbf{X})$. Extrapolation is crucial for causal applications, since for computing $\mathbb{E}[Y \mid do(\mathbf{W}=\mathbf{w})]$ we are interested in what would happen with Y when our treatment variable $\mathbf{W}$ is set to be $\mathbf{w}$, regardless of the value achieved by $\mathbf{X}$. However, it can easily happen that for this specific combination of $\mathbf{X}$ and $\mathbf{W}$ there are very few observed data points, thus making the estimation of the causal effect hard (Pearl, 2009).

As an illustration, we consider the following synthetic data example, with continuous outcome Y, continuous univariate treatment W, $n = 5000$ and $p = 20$:

$$\mathbf{X} \sim U(0,5)^p, \quad W \mid \mathbf{X} \sim N(X_2, 1), \quad Y \mid \mathbf{X}, W \sim N(X_2 + X_1 \sin(W), 1). \tag{24}$$
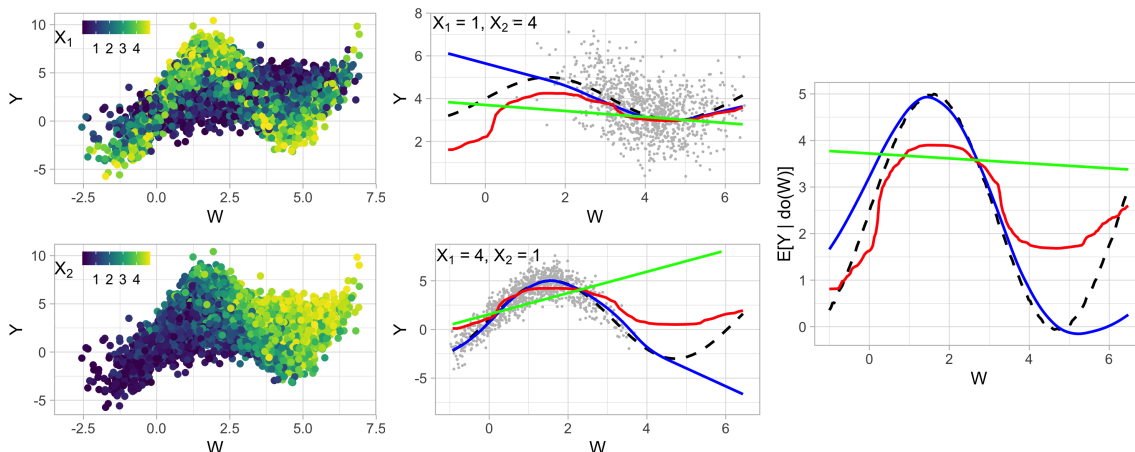
Figure 9: Left: Visualization of heterogeneous synthetic example (24). Middle: Gray points depict joint distribution of (W, Y) conditionally on $\mathbf{X} = \mathbf{x}$, for some choices of $\mathbf{x}$ indicated in the top left corner. Black curve indicates the true conditional mean $\mathbb{E}[Y \mid W = w, \mathbf{X} = \mathbf{x}]$, the blue curve represents the estimate obtained by DRF with response (W, Y) and predictors $\mathbf{X}$ in combination with smoothing splines regression, the red curve represents the estimate obtained by standard Random Forest, whereas the green line shows the estimate of the Causal Forest (Athey et al., 2019) which makes the linearity assumption and is thus misspecified. Right: The corresponding estimates for all the methods of the causal effect $\mathbb{E}[Y \mid do(W = w)]$ computed from (23). The true causal effect is denoted by a black dashed curve.

A visualization of the data can be seen on the left side of Figure 9; treatment W affects Y nonlinearly, X_2 is a confounding variable that affects the marginal distributions of Y and W and X_1 makes the treatment effect heterogeneous. The middle part of Figure 9 shows the conditional regression fits, i.e. the estimates of $\mathbb{E}[Y \mid W = w, \mathbf{X} = \mathbf{x}]$ as w varies and $\mathbf{x}$ is fixed. In general, the conditional regression fit is related to the concept of the conditional average treatment effect (CATE) as it quantifies the effect of $\mathbf{W}$ on Y for the subpopulation for which $\mathbf{X} = \mathbf{x}$. We see that combination of DRF with response (Y, W) and predictors $\mathbf{X}$ with the smoothing splines regression of Y on W (blue curve) is more accurate than the estimates obtained by standard Random Forest (Breiman, 2001) with response Y and predictors $(W, \mathbf{X})$ (red curve). Furthermore, we see that the former approach can extrapolate better to regions with small number of data points, which enables us to better estimate the causal effect $\mathbb{E}[Y \mid do(W = w)]$ from (23), by averaging the corresponding estimates of $\mathbb{E}[Y \mid W = w, \mathbf{X} = \mathbf{x}]$ over observed $\mathbf{x}$, as shown in the right plot of Figure 9.

There exist many successful methods in the literature for estimating the causal effects and the (conditional) average treatment effects for a wide range of settings (Abadie and Imbens, 2006; Chernozhukov et al., 2018; Wager and Athey, 2018; Künzel et al., 2019). However, some methods are not designed for the most general case and make certain modeling assumptions or are designed specifically for the (very common) case where the treatment variable is univariate or even binary. Due to its versatility, DRF can easily be used when the underlying

assumptions of conventional methods are violated, when some additional structure is given in the problem or for the general, nonparametric, settings (Imbens, 2004; Ernest and Bühlmann, 2015; Kennedy et al., 2017). Appendix D contains additional comparisons with some existing methods for causal effect estimation.

4.4.1 Births data

We further illustrate the applicability of DRF for causality-related problems on the natality data obtained from the Centers for Disease Control and Prevention (CDC) website, where we have information about all recorded births in the USA in 2018. We investigate the relationship between the pregnancy length and the birthweight, an important indicator of baby's health. Not only is this relationship complex, but it also depends on many different factors, such as parents' race, baby's gender, birth multiplicity (single, twins, triplets...) etc. In the left two plots of Figure 10 one can see the estimated joint distribution of birthweight and pregnancy length conditionally on many different covariates, as indicated in the plot. The black curves denote the subsequent regression fit, based on smoothing splines. In addition to the estimate of the mean, indicated by the solid curve, we also include the estimates of the conditional 0.1- and 0.9-quantiles, indicated by dashed curves, which is very useful in practice for determining whether a baby is large or small for its gestational age. Notice how DRF assigns less importance to the mother's race when the point of interest is a twin (middle plot), as in this case more weight is given to twin births, regardless of the race of the parents.

Suppose now we would like to understand how a twin birth T causally affects the birthweight B, but ignoring the obvious indirect effect due to shorter pregnancy length L. For example, sharing of resources between the babies might have some effect on their birthweight. We additionally need to be careful to adjust for other confounding variables $\mathbf{X}$, such as, for example, the parents' race, which can affect B, T and L. We assume that this is represented by the following causal graph:
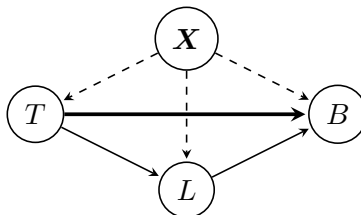
In order to answer the above question, we investigate the causal quantity $\mathbb{P}(B \mid do(T{=}t, L{=}l))$. Even though one cannot make such do-intervention in practice, this quantity describes the total causal effect if the birth multiplicity and the length of the pregnancy could be manipulated and thus for a fixed pregnancy length l, we can see the difference in birthweight due to T. We compute this quantity as above, by using DRF with subsequent regression fits, which has the advantage of better extrapolating to regions with small probability, such as long twin pregnancies (see the middle plot of Figure 10). In the right plot of Figure 10 we show the mean and quantiles of the estimated interventional distribution and we see that, as one might expect, a twin birth causes smaller birthweight on average, with the difference increasing with the length of the pregnancy.
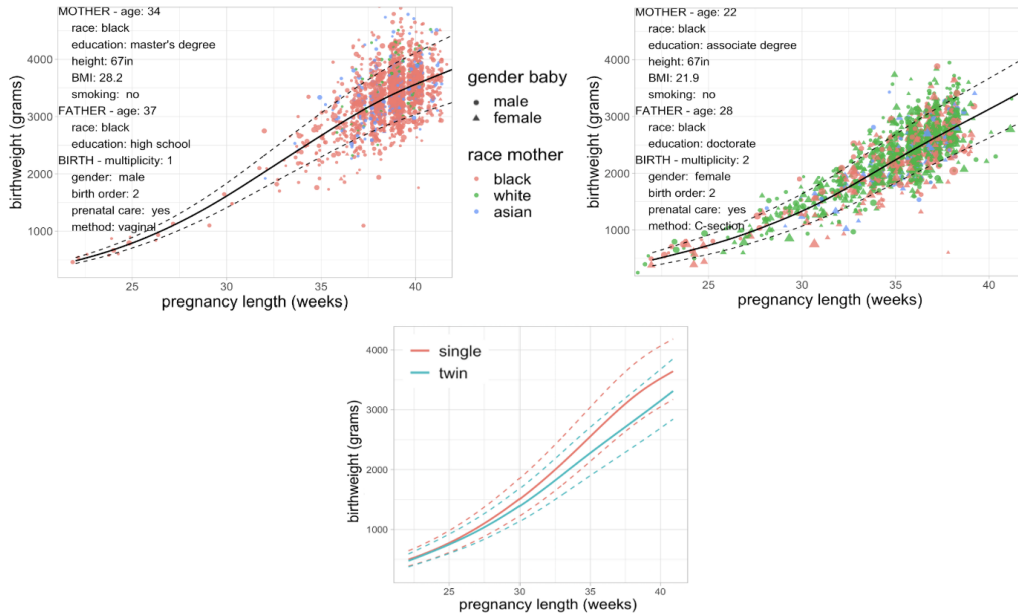
Figure 10: Above: estimated relationship of pregnancy length and birthweight, conditionally on the criteria indicated in the upper left corner. Below: estimated interventional effect of twin birth on the birthweight for a fixed pregnancy length. In all plots the solid curves denote the estimated conditional mean and the dashed denote the estimated 0.1 and 0.9 quantiles.

4.5 Fairness

Being able to compute different causal quantities with DRF could prove useful in a range of applications, including fairness (Kusner et al., 2017). We investigate the data on approximately 1 million full-time employees from the 2018 American Community Survey by the US Census Bureau from which we have extracted the salary information and all covariates that might be relevant for salaries. In the bottom left plot of Figure 11 one can see the distribution of hourly salary of men and women (on the logarithmic scale). The overall salary was scaled with working hours to account for working part-time and for the fact that certain jobs have different working hours. We can see that men are paid more in general, especially for the very high salaries. The difference between the median hourly salaries, a commonly used statistic in practice, amounts 17% for this data set.

We would like to answer whether the observed gender pay gap in the data is indeed unfair, i.e. only due to the gender, or whether it can at least in part be explained by some other factors, such as age, job type, number of children, geography, race, attained education level and many others. Hypothetically, it could be, for example, that women have a preference for jobs that are paid less, thus causing the gender pay gap.

In order to answer this question, we assume that the data is obtained from the following causal graph, where G denotes the gender, W the hourly wage and all other factors are denoted by $\mathbf{X}$:
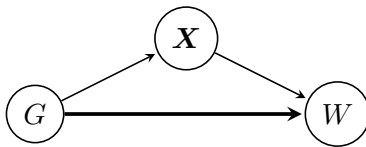
i.e. G is a source node and W is a sink node in the graph. In order to determine the direct effect of the gender on wage that is not mediated by other factors, we would like to compute the distribution of the nested counterfactual $W(\text{male}, \mathbf{X}(\text{female}))$, which is interpreted as the women's wage had they been treated in same way as men by their employers for determining the salary, but without changing their propensities for other characteristics, such as the choice of occupation (Chernozhukov et al., 2013). Therefore, it can be obtained from the observed distribution as follows:

$$\mathbb{P}\left(W(\text{male}, \mathbf{X}(\text{female}))\right) = \int \mathbb{P}\left(W(G\!=\!\text{male}, \mathbf{X}\!=\!\mathbf{x})\right) \mathbb{P}(\mathbf{X}\!=\!\mathbf{x} \mid G\!=\!\text{female}) d\mathbf{x}$$

$$= \int \mathbb{P}\left(W \mid G\!=\!\text{male}, \mathbf{X}\!=\!\mathbf{x}\right) \mathbb{P}(\mathbf{X}\!=\!\mathbf{x} \mid G\!=\!\text{female}) d\mathbf{x}, \qquad (25)$$

Put in the language of the fairness literature, it quantifies the unfairness when all variables $\mathbf{X}$ are assumed to be resolving (Kilbertus et al., 2017), meaning that any difference in salaries directly due to factors $\mathbf{X}$ is not viewed as gender discrimination. For example, one does not consider unfair if people with low education level get lower salaries, even if the gender distribution in this group is not balanced.

There are several ways how one can compute the distribution of $W(\text{male}, \mathbf{X}(\text{female}))$ from (25) with DRF. The most straightforward option is to take W as the response and $(G, \mathbf{X})$ as predictors in order to compute the conditional distribution $\mathbb{P}\left(W \mid G\!=\!\text{male}, \mathbf{X}\!=\!\mathbf{x}\right)$. However, with this approach it could happen that for predicting $\mathbb{P}\left(W \mid G\!=\!\text{male}, \mathbf{X}\!=\!\mathbf{x}\right)$ we also assign weight to training data points for which $G\!=\!\text{female}$. This happens if in some trees we did not split on variable G, which is likely, for example, if $\mathbb{P}(G = \text{male} \mid \mathbf{X}\!=\!\mathbf{x})$ is low. Using salaries of both genders to estimate the distribution of men's salaries might be an issue if our goal is to objectively compare how women and men are paid.

Another approach is to take (W, G) as a multivariate response and $\mathbf{X}$ as the predictors for DRF and thus obtain joint distribution of (W, G) conditionally on the event $\mathbf{X}\!=\!\mathbf{x}$. In this way we can also quantify the gender discrimination of a single individual with characteristics $\mathbf{x}$ by comparing his/her salary to the corresponding quantile of the salary distribution of people of the opposite gender with the same characteristics $\mathbf{x}$ (Plečko and Meinshausen, 2020). This is interesting because the distribution of salaries, and thus also the gender discrimination, can be quite different depending on other factors such as the industry sector or job type, as illustrated for a few choices of $\mathbf{x}$ in the top row of Figure 11.

Finally, by averaging the DRF estimates of $\mathbb{P}\left(W \mid \mathbf{X}\!=\!\mathbf{x}, G\!=\!\text{male}\right)$, conveniently represented via the weights, over different $\mathbf{x}$ sampled from the distribution $\mathbb{P}(\mathbf{X} \mid G\!=\!\text{female})$, we can compute the distribution of the nested counterfactual $W(\text{male}, \mathbf{X}(\text{female}))$ (Chernozhukov et al., 2013). In the middle panel in the bottom row of Figure 11 a noticeable difference in the means, also called natural direct effect in the causality literature (Pearl, 2009), is still visible between the observed distribution of women's salaries and the hypothetical distribution of their salaries had they been treated as men, despite adjusting for
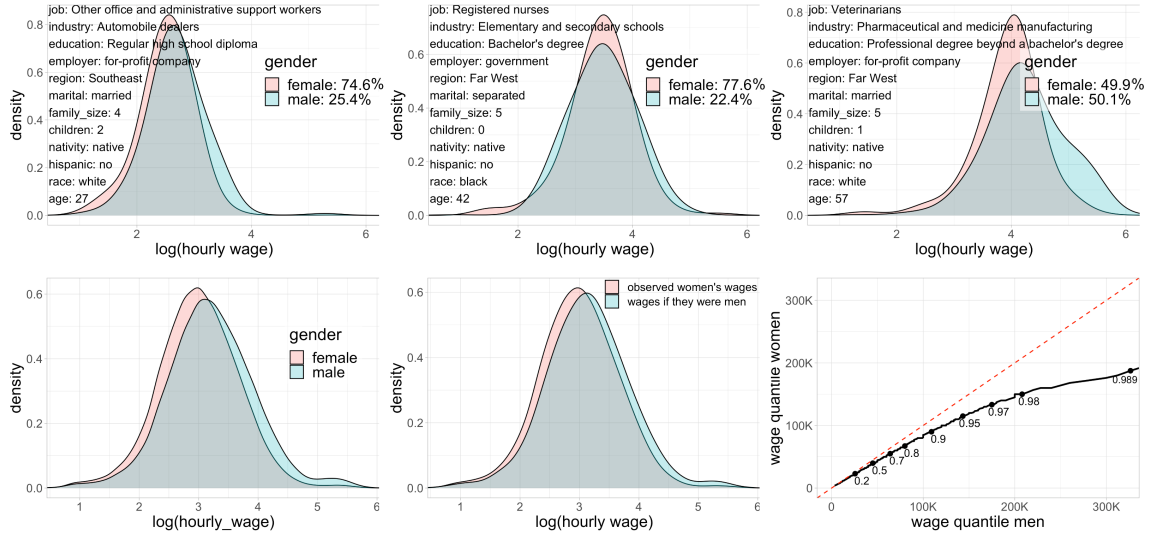
Figure 11: Top row: Estimated joint distribution of wage and gender for some fixed values of other covariates **X** indicated in the top left part of each plot. Bottom row: observed overall distribution of salaries (left), estimated counterfactual distribution $\mathbb{P}\left(W(\text{male}, \mathbf{X}(\text{female}))\right)$ of women's salaries (middle) and the quantile comparison of the counterfactual distribution of women's salaries and the observed distribution of men's salaries (right).

indirect effects of the gender via covariates **X**. By further matching the quantiles of the counterfactual distribution $\mathbb{P}\left(W(\text{male}, \mathbf{X}(\text{female}))\right)$ with the corresponding quantiles of the observed distribution of men's salaries in the bottom right panel of Figure 11, we can also see that the adjusted gender pay gap even increases for larger salaries. Median hourly wage for women is still 11% lower than the median wage for the hypothetical population of men with exactly the same characteristics **X** as women, indicating that only a minor proportion of the actually observed hourly wage difference of 17% can be explained by other demographic factors.

5. Conclusion

We have shown that DRF is a flexible, general and powerful tool, which exploits the well-known properties of the Random Forest as an adaptive nearest neighbor method via the induced weighting function. Not only does it estimate multivariate conditional distributions well, but it constructs the forest in a model- and target-free way and is thus an easy to use out-of-the-box algorithm for many, potentially complex, learning problems in a wide range of applications, including also causality and fairness, with competitive performance even for problems with existing tailored methods.

Appendix A. Implementation Details

Here we present the implementation of the Distributional Random Forests (DRF) in detail. The code is available in the R-package `drf` and the Python package `drf`. The implementation is based on the implementations of the R-packages `grf` (Athey et al., 2019) and `ranger` (Wright and Ziegler, 2017). The largest difference is in the splitting criterion itself and the provided user interface. Algorithm 1 gives the pseudocode for the forest construction and computation of the weighting function $w_{\mathbf{x}}(\cdot)$.

- Every tree is constructed based on a random subset of size s (taken to be 50% of the size of the training set by default) of the training data set, similar to Wager and Athey (2018). This differs from the original Random Forest algorithm (Breiman, 2001), where the bootstrap subsampling is done by drawing from the original sample with replacement.

- The principle of honesty (Biau, 2012; Denil et al., 2014; Wager and Athey, 2018) is used for building the trees (line 4), whereby for each tree one first performs the splitting based on one random set of data points $\mathcal{S}_{\text{build}}$, and then populates the leaves with a disjoint random set $\mathcal{S}_{\text{populate}}$ of data points for determining the weighting function $w_{\mathbf{x}}(\cdot)$. This prevents overfitting, since we do not assign weight to the data points which we used to built the tree.

- We borrow the method for selecting the number of candidate splitting variables from the `grf` package (Athey et al., 2019). This number is randomly generated as $\min(\max(\text{Poisson}(\text{mtry}), 1), p)$, where mtry is a tuning parameter. This differs from the original Random Forests algorithm, where the number of splitting candidates is fixed to be mtry.

- The number of trees built is $N = 2000$ by default.

- The factor variables in both the responses and the predictors are encoded by using the one-hot encoding, where we add an additional indicator variable for each level l of some factor variable X_k. This implies that in the building step, if we split on this indicator variable, we divide the current set of data points in the sets where $X_k = l$ and $X_k \neq l$. This works well if the number of levels is not too big, since otherwise one makes very uneven splits and the dimensionality of the problem increases significantly. Handling of categorical problems is a general challenge for the forest based methods and is an area of active research (Johannemann et al., 2019). We will leave improving on this approach for the future development.

- We try to enforce splits where each child has at least a fixed percentage (chosen to be 10% as the default value) of the current number of data points. In this way we achieve balanced splits and reduce the computational time. However, we cannot enforce this if we are trying to split on the variable X_i with only a few unique values, e.g. indicator variable for a level of some factor variable.

- All components of the response Y are scaled for the building step (but not when we populate the leaves). This ensures that each component of the response contributes

---

Algorithm 1 Pseudocode for Distributional Random Forest

1: **procedure** BUILDFOREST(set of samples $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, number of trees N)
2:      **for** $i = 1, \ldots, N$ **do**
3:          $\mathcal{S}_{\text{subsample}} = \text{SUBSAMPLE}(\mathcal{S})$
4:          $\mathcal{S}_{\text{build}}, \mathcal{S}_{\text{populate}} \leftarrow \text{SPLITSAMPLES}(\mathcal{S}_{\text{subsample}})$ ▷ Honesty principle, see above
5:          $\mathcal{T}_i \leftarrow \text{CREATENEWTREE}(\mathcal{S}_{\text{build}})$ ▷ Samples $\mathcal{S}_{\text{build}}$ used for building the tree
6:          $\text{BUILDTREE}(\text{ROOTNODE}(\mathcal{T}_i))$ ▷ Start recursion from the root node
7:          $\text{POPULATELEAVES}(\mathcal{T}_i, \mathcal{S}_{\text{populate}})$ ▷ Samples $\mathcal{S}_{\text{populate}}$ used for computing $w_{\mathbf{x}}(\cdot)$
8:      **end for**
9:      **return** $\mathcal{F} = \{\mathcal{T}_1, \ldots, \mathcal{T}_N\}$
10: **end procedure**

11: **procedure** BUILDTREE(current node $\mathcal{N}$)   ▷ Recursively constructs the trees
12:      **if** STOPPINGCRITERION($\mathcal{N}$) **then**   ▷ E.g. if only a few samples left
13:          **return**
14:      **end if**
15:      $\mathcal{S} \leftarrow \text{GETSAMPLES}(\mathcal{N})$
16:      $\mathcal{I} \leftarrow \text{GETSPLITVARIABLES}()$ ▷ Random set of candidate variables
17:      $\mathcal{C} \leftarrow \text{INITIALIZESPLITS}()$ ▷ Here we store info about candidate splits
18:      **for** $\text{idx} \in \mathcal{I}$, level l **do**   ▷ l iterates over all values of variable X_{idx}
19:          $\mathcal{S}_L, \mathcal{S}_R \leftarrow \text{CHILDSAMPLES}(\mathcal{S}, \text{idx}, l)$ ▷ Split samples based on $(\mathbf{x}_i)_{\text{idx}} \leq l$
20:          test statistic $v = \text{SPLITTINGCRITERION}(\mathcal{S}_L, \mathcal{S}_R)$ ▷ Two-sample test of choice
21:          $\text{ADDNEWSPLITCANDIDATE}(\mathcal{C}, v, \mathcal{S}_L, \mathcal{S}_R, \text{idx}, l)$
22:      **end for**
23:      $\mathcal{S}_L, \mathcal{S}_R, \text{idx}, l \leftarrow \text{FINDBESTSPLIT}(\mathcal{C})$
24:      $\mathcal{N}_L \leftarrow \text{CREATENODE}(\mathcal{S}_L)$ ▷ Create new node with set of samples $\mathcal{S}_L$
25:      $\mathcal{N}_R \leftarrow \text{CREATENODE}(\mathcal{S}_R)$ ▷ Create new node with set of samples $\mathcal{S}_R$
26:      $\text{BUILDTREE}(\mathcal{N}_L)$, $\text{BUILDTREE}(\mathcal{N}_R)$ ▷ Proceed building recursively
27:      $\text{CHILDREN}(\mathcal{N}) \leftarrow \mathcal{N}_L, \mathcal{N}_R$
28:      $\text{SPLIT}(\mathcal{N}) \leftarrow \text{idx}, l$ ▷ Store the split
29:      **return**
30: **end procedure**

31: **procedure** GETWEIGHTS(forest $\mathcal{F}$, test point $\mathbf{x}$)   ▷ Computes the weighting function
32:      vector of weights $w = \text{ZEROS}(n)$ ▷ n is the training set size
33:      **for** $i = 1, \ldots, |\mathcal{F}|$ **do**
34:          $\mathcal{L} = \text{GETLEAFSAMPLES}(\mathcal{T}_i, \mathbf{x})$ ▷ indices of training samples in same leaf as $\mathbf{x}$
35:          **for** $\text{idx} \in \mathcal{L}$ **do**
36:              $w[\text{idx}] = w[\text{idx}] + 1/(|\mathcal{L}| \cdot |\mathcal{F}|)$
37:          **end for**
38:      **end for**
39:      **return** w
40: **end procedure**

---

equally to the kernel values, and consequently to the MMD two-sample test statistic. Plain usage of the MMD two-sample test would scale the components of Y at each node. However, this approach favors always splitting on the same variables, even though their effect will diminish significantly after having split several times.

- By default, in step 20 of the Algorithm 1, we use the MMD-based splitting criterion given by

$$\frac{1}{B}\sum_{k=1}^{B}\frac{|\mathcal{S}_L||\mathcal{S}_L|}{(|\mathcal{S}_L|+|\mathcal{S}_R|)^2}\left|\frac{1}{|\mathcal{S}_L|}\sum_{(\mathbf{x}_i,\mathbf{y}_i)\in\mathcal{S}_L}\varphi_{\boldsymbol{\omega}_k}(\mathbf{y}_i)-\frac{1}{|\mathcal{S}_R|}\sum_{(\mathbf{x}_i,\mathbf{y}_i)\in\mathcal{S}_R}\varphi_{\boldsymbol{\omega}_k}(\mathbf{y}_i)\right|^2.$$

The Gaussian kernel $k(\mathbf{x},\mathbf{y})=\frac{1}{(\sqrt{2\pi}\sigma)^d}e^{\frac{-\|x-y\|_2^2}{2\sigma^2}}$ is used as the default choice, with the bandwidth σ chosen as the median pairwise distance between all training responses $\{\mathbf{y}_i\}_{i=1}^n$, commonly referred to as the 'median heuristic' (Gretton et al., 2012c). However the algorithm can be used with any choice of kernel, or in fact with any two-sample test.

- The number B of random Fourier features is fixed and taken to be 20 by default. The performance of the trees empirically shows stability for large range of B. Smaller values of B help making the trees more independent which could improve the performance. One could even use an adaptive strategy of choosing B, possibly increasing B as the depth of the tree increases, but we decided to keep B fixed for simplicity.

- We compute variable importance similarly as for the original Random Forest algorithm (Breiman, 2001; Wright and Ziegler, 2017), by sequentially permuting each variable and investigating the decrease in the performance. However, since we target the full conditional distribution of the multivariate response, as our performance measure we use for every test point $(\mathbf{x},\mathbf{y})$ the MMD distance between the estimated joint distribution $\hat{\mathbb{P}}(\mathbf{Y}\mid\mathbf{X}=\mathbf{x})$, described by the DRF weights, and the point mass $\delta_{\mathbf{y}}$.

Appendix B. Derivations and Proofs

In this section we present proofs and further details for the results in Sections 2.3.1 and 3, in the order in which they appear. Sections B.1 and B.3 derive the splitting criterion presented in Equation (12) in the main text, with Section B.3 showing that the CART criterion can be written in an analogous way. Section B.4 provides background and proves to the statements in Section 3.

B.1 Expressing MMD test statistic as an integral in the feature space

The biased MMD two-sample statistic is given as

$$
\mathcal{D}_{\mathrm{MMD}}\left(\{\mathbf{u}_i\}_{i=1}^m, \{\mathbf{v}_i\}_{i=1}^n\right)
$$
$$
= \frac{1}{m^2}\sum_{i,j} k(\mathbf{u}_i, \mathbf{u}_j) + \frac{1}{n^2}\sum_{i,j} k(\mathbf{v}_i, \mathbf{v}_j) - \frac{2}{mn}\sum_i \sum_j k(\mathbf{u}_i, \mathbf{v}_j)
$$
$$
= \frac{1}{m^2}\sum_{i,j} k(\mathbf{u}_i, \mathbf{u}_j) + \frac{1}{n^2}\sum_{i,j} k(\mathbf{v}_i, \mathbf{v}_j) - \frac{1}{mn}\sum_i \sum_j k(\mathbf{u}_i, \mathbf{v}_j) - \frac{1}{mn}\sum_i \sum_j k(\mathbf{v}_j, \mathbf{u}_i).
$$

Assume that the kernel k is bounded and shift-invariant, then by Bochner's theorem there exist a measure ν such that k can be written as $k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})} d\nu(\boldsymbol{\omega})$.

Let us write $\varphi_{\boldsymbol{\omega}}^U = \frac{1}{m}\sum_i e^{i\boldsymbol{\omega}^T \mathbf{u}_i}$ and $\varphi_{\boldsymbol{\omega}}^V = \frac{1}{n}\sum_i e^{i\boldsymbol{\omega}^T \mathbf{v}_i}$. We can now write $\mathcal{D}_{\mathrm{MMD}}$ as

$$
\mathcal{D}_{\mathrm{MMD}}\left(\{\mathbf{u}_i\}_{i=1}^m, \{\mathbf{v}_i\}_{i=1}^n\right) = \int_{\mathbb{R}^d} \left(\varphi_{\boldsymbol{\omega}}^U \overline{\varphi_{\boldsymbol{\omega}}^U} + \varphi_{\boldsymbol{\omega}}^V \overline{\varphi_{\boldsymbol{\omega}}^V} - \varphi_{\boldsymbol{\omega}}^U \overline{\varphi_{\boldsymbol{\omega}}^V} - \varphi_{\boldsymbol{\omega}}^V \overline{\varphi_{\boldsymbol{\omega}}^U}\right) d\nu(\boldsymbol{\omega})
$$
$$
= \int_{\mathbb{R}^d} \left|\varphi_{\boldsymbol{\omega}}^U - \varphi_{\boldsymbol{\omega}}^V\right|^2 d\nu(\boldsymbol{\omega})
$$
$$
= \int_{\mathbb{R}^d} \left|\frac{1}{m}\sum_{i=1}^m \varphi_{\boldsymbol{\omega}}(\mathbf{u}_i) - \frac{1}{n}\sum_{i=1}^n \varphi_{\boldsymbol{\omega}}(\mathbf{v}_i)\right|^2 d\nu(\boldsymbol{\omega}),
$$

where $\varphi_{\boldsymbol{\omega}}(\mathbf{y}) = e^{i\boldsymbol{\omega}^T \mathbf{y}} \in \mathbb{C}$ are the corresponding Fourier features, which is what we wanted to show.

B.2 Approximate kernel and its MMD

When the kernel k is bounded and shift invariant, we have seen that it can be written as $k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})} d\nu(\boldsymbol{\omega})$. This integral can be approximated by sampling from ν: Let $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_B \sim \nu$ be a random sample from the measure ν. Then we can write

$$
k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})} d\nu(\boldsymbol{\omega}) \approx \frac{1}{B}\sum_{b=1}^B e^{i\boldsymbol{\omega}_b^T(\mathbf{x}-\mathbf{y})} = \frac{1}{B}\langle \widetilde{\varphi}(\mathbf{u}), \widetilde{\varphi}(\mathbf{v})\rangle_{\mathbb{C}^B} := \tilde{k}(\mathbf{u}, \mathbf{v}),
$$

where $\widetilde{\varphi}(\mathbf{u}) = (\varphi_{\boldsymbol{\omega}_1}(\mathbf{u}), \ldots, \varphi_{\boldsymbol{\omega}_B}(\mathbf{u}))^T$ is a random complex vector consisting of the Fourier features $\varphi_{\boldsymbol{\omega}}(\mathbf{u}) = e^{i\boldsymbol{\omega}^T \mathbf{u}} \in \mathbb{C}$. The kernel $\tilde{k}$ is analogous to the kernel k, but where the measure ν is replaced by the empirical measure $\tilde{\nu} = \frac{1}{B}\sum_{b=1}^B \delta_{\boldsymbol{\omega}_b}$:

$$
\tilde{k}(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})} d\tilde{\nu}(\boldsymbol{\omega}).
$$

Analogously as in the section B.1, we can now write the MMD for the kernel $\tilde{k}$ as:

$$
\mathcal{D}_{\mathrm{MMD}(\tilde{k})} = \int_{\mathbb{R}^d} \left|\frac{1}{m}\sum_{i=1}^m \varphi_{\boldsymbol{\omega}}(\mathbf{u}_i) - \frac{1}{n}\sum_{i=1}^n \varphi_{\boldsymbol{\omega}}(\mathbf{v}_i)\right|^2 d\tilde{\nu}(\boldsymbol{\omega})
$$
$$
= \frac{1}{B}\sum_{b=1}^B \left|\frac{1}{m}\sum_{i=1}^m \varphi_{\boldsymbol{\omega}_b}(\mathbf{u}_i) - \frac{1}{n}\sum_{i=1}^n \varphi_{\boldsymbol{\omega}_b}(\mathbf{v}_i)\right|^2,
$$

which can also additionally be interpreted as the approximation of $\mathcal{D}_{\text{MMD}}$. Therefore, our splitting criterion is obtained as the MMD of the random approximate kernel $\tilde{k}$:

$$\frac{1}{B} \sum_{b=1}^{B} \frac{n_L n_R}{n_P^2} \left| \frac{1}{n_L} \sum_{\mathbf{x}_i \in C_L} \varphi_{\boldsymbol{\omega}_b}(\mathbf{y}_i) - \frac{1}{n_R} \sum_{\mathbf{x}_i \in C_R} \varphi_{\boldsymbol{\omega}_b}(\mathbf{y}_i) \right|^2 .$$

The scaling factor $\frac{n_L n_R}{n_P^2}$ occurs naturally and penalizes the increased variance of the sample MMD statistic when n_L or n_R are small: it appears when we rewrite the CART criterion in the related form, see section B.3.

This representation of the MMD is the key why we use the approximate kernel $\tilde{k}$ instead of k. This splitting criterion can be computed in $\mathcal{O}(B n_P)$ complexity, by updating the sums $\sum_{\mathbf{x}_i \in C_L} \varphi_{\boldsymbol{\omega}_k}(\mathbf{y}_i)$ and $\sum_{\mathbf{x}_i \in C_R} \varphi_{\boldsymbol{\omega}_k}(\mathbf{y}_i)$ in $\mathcal{O}(1)$ computations, whereas this is not possible for $\mathcal{D}_{\text{MMD}}$.

B.3 CART criterion rewritten

Standard CART criterion used in Random Forests (Breiman, 2001) is the following: we repeatedly choose to split the parent node P of size n_P in two children C_L and C_R, of sizes n_L and n_R respectively, such that the expression

$$\frac{1}{n_P} \left(\sum_{i \in C_L} (Y_i - \overline{Y}_L)^2 + \sum_{i \in C_R} (Y_i - \overline{Y}_R)^2 \right) \tag{26}$$

is minimized, where $\overline{Y}_L = \frac{1}{n_L} \sum_{i \in C_L} Y_i$ and Y_R is defined similarly.

We now have $\overline{Y} = \frac{1}{n_P} \sum_{i \in P} Y_i = \frac{n_L}{n_P} \overline{Y}_L + \frac{n_R}{n_P} \overline{Y}_R$, which gives $\overline{Y} - \overline{Y}_L = \frac{n_R}{n_P} (\overline{Y}_R - \overline{Y}_L)$, so we can write

$$\sum_{i \in C_L} (Y_i - \overline{Y}_L)^2 = \sum_{i \in C_L} (Y_i - \overline{Y} + \overline{Y} - \overline{Y}_L)^2 = \sum_{i \in C_L} (Y_i - \overline{Y} + \frac{n_R}{n_P}(\overline{Y}_R - \overline{Y}_L))^2$$

$$= \sum_{i \in C_L} (Y_i - \overline{Y})^2 + 2 \frac{n_R}{n_P}(\overline{Y}_R - \overline{Y}_L) \sum_{i \in C_L} (Y_i - \overline{Y}) + \frac{n_L n_R^2}{n_P^2}(\overline{Y}_R - \overline{Y}_L)^2$$

$$= \sum_{i \in C_L} (Y_i - \overline{Y})^2 + 2 \frac{n_R}{n_P}(\overline{Y}_R - \overline{Y}_L) \cdot n_L(\overline{Y}_L - \overline{Y}) + \frac{n_L n_R^2}{n_P^2}(\overline{Y}_R - \overline{Y}_L)^2$$

$$= \sum_{i \in C_L} (Y_i - \overline{Y})^2 + 2 \frac{n_R n_L}{n_P}(\overline{Y}_R - \overline{Y}_L) \cdot \frac{n_R}{n_P}(\overline{Y}_L - \overline{Y}_R) + \frac{n_L n_R^2}{n_P^2}(\overline{Y}_R - \overline{Y}_L)^2$$

$$= \sum_{i \in C_L} (Y_i - \overline{Y})^2 - \frac{n_L n_R^2}{n_P^2}(\overline{Y}_R - \overline{Y}_L)^2.$$

Similarly we obtain

$$\sum_{i \in C_R} (Y_i - \overline{Y}_R)^2 = \sum_{i \in C_R} (Y_i - \overline{Y} + \overline{Y} - \overline{Y}_R)^2 = \sum_{i \in C_R} (Y_i - \overline{Y})^2 - \frac{n_L^2 n_R}{n_P^2}(\overline{Y}_R - \overline{Y}_L)^2,$$

which gives us that the CART criterion (26) can be written as

$$\frac{1}{n_P} \left(\sum_{i \in C_L} (Y_i - \overline{Y})^2 - \frac{n_L n_R^2}{n_P^2} (\overline{Y}_R - \overline{Y}_L)^2 + \sum_{i \in C_R} (Y_i - \overline{Y})^2 - \frac{n_L^2 n_R}{n_P^2} (\overline{Y}_R - \overline{Y}_L)^2 \right)$$

$$= \frac{1}{n_P} \sum_{i \in P} (Y_i - \overline{Y})^2 - \frac{n_L n_R}{n_P^2} (\overline{Y}_R - \overline{Y}_L)^2,$$

since $n_L + n_R = n_P$. Since the first term depends only on the parent node and not on the chosen split, we conclude that minimizing the CART criterion (26) is equivalent to maximizing the following expression

$$\frac{n_L n_R}{n_P^2} (\overline{Y}_L - \overline{Y}_R)^2. \tag{27}$$

This equivalent criterion can be interpreted as comparing the difference in the means of the resulting child nodes, i.e. we will choose the split such that the means in the child nodes are as heterogeneous as possible. The scaling factor $\frac{n_L n_R}{n_P^2}$ appears naturally, penalizing uneven splits due to the increased variance of $\overline{Y}_L$ or $\overline{Y}_R$.

B.4 Proofs for Section 3

Preliminaries. We first set notation and define basic probabilistic concepts on the separable Hilbert space $(\mathcal{H}, \langle, \cdot, \rangle_{\mathcal{H}})$. We thereby mostly refer to Hsing and Eubank (2015) and Pisier (2016). The initial results derived here parallel some of the results derived in Park and Muandet (2020), but where derived independently. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be the underlying probability space. Let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be the Hilbert space induced by the kernel k and $\mu : \mathcal{M}_b(\mathbb{R}^d) \to \mathcal{H}$ be the embedding function of, $\mu(P) \in \mathcal{H}$ for all bounded signed Borel measures P on $\mathbb{R}^d$. Throughout we assume that k is *bounded* and *continuous* in its two arguments. Boundedness of k ensures that μ is indeed defined on all of $\mathcal{M}_b(\mathbb{R}^d)$, while continuity of $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ ensures $\mathcal{H}$ is *separable*. Thus measurability issues can be avoided, in particular, a map $\xi : (\Omega, \mathcal{A}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ is measurable iff $\langle \xi, f \rangle_{\mathcal{H}}$ is measurable for all $f \in \mathcal{H}$. Moreover, a quick check reveals that $\mu(P)$ is linear on $\mathcal{M}_b(\mathbb{R}^d)$. If $\mathbb{E}[\|\xi\|_{\mathcal{H}}] < \infty$, we define

$$\mathbb{E}[\xi] := \int_{\mathcal{H}} \xi d\mathbb{P} \in \mathcal{H},$$

where the integral is meant in a Bochner sense. Separability and $\mathbb{E}[\|\xi\|_{\mathcal{H}}] < \infty$ mean this integral is well-defined and moreover

$$F(\mathbb{E}[\xi]) = \mathbb{E}[F(\xi)],$$

for any continuous linear function $F : \mathcal{H} \to \mathbb{R}$.[1] In particular, $\mathbb{E}[\langle \xi, f \rangle_{\mathcal{H}}] = \langle \mathbb{E}[\xi], f \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. Define moreover for $q \geq 1$, and $\xi, \xi_1, \xi_2 \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H})$,

$$\mathcal{L}^q(\Omega, \mathcal{A}, \mathcal{H}) = \{\xi : (\Omega, \mathcal{F}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H})) \text{ measurable, with } \mathbb{E}[\|\xi\|^q] < \infty]\}$$

$$\mathbb{L}^q(\Omega, \mathcal{A}, \mathcal{H}) = \text{Set of equivalence classes in } \mathcal{L}^q(\Omega, \mathcal{A}, \mathcal{H})$$

$$\mathrm{Var}(\xi) := \mathbb{E}[\|\xi - \mathbb{E}[\xi]\|^2] = \mathbb{E}[\|\xi\|^2] - \|\mathbb{E}[\xi]\|^2, \quad \xi \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H})$$

$$\mathrm{Cov}(\xi_1, \xi_2) = \mathbb{E}[\langle \xi_1 - \mathbb{E}[\xi_1], \xi_2 - \mathbb{E}[\xi_2] \rangle_{\mathcal{H}}] = \mathbb{E}[\langle \xi_1, \xi_2 \rangle_{\mathcal{H}}] - \langle \mathbb{E}[\xi_1], \mathbb{E}[\xi_2] \rangle_{\mathcal{H}}.$$

---

1. Here and later $F(\xi)$ is meant to mean $F(\xi(\omega))$ for all $\omega \in \Omega$.

It is well-known, that $(\mathbb{L}^q, \| \cdot \|_{\mathbb{L}^q(\mathcal{H})})$ is a Banach space, with

$$\|\xi\|_{\mathbb{L}^q(\mathcal{H})} = \mathbb{E}[\|\xi\|_{\mathcal{H}}^q]^{1/q}.$$

We can then also define *conditional* expectation. For a sub $\sigma-$ algebra $\mathcal{F} \subset \mathcal{A}$, $\xi \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H})$, $\mathbb{E}[\xi \,|\, \mathcal{F}]$ is the (a.s.) unique element such that

(C1) $\mathbb{E}[\xi \,|\, \mathcal{F}] : (\Omega, \mathcal{F}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ is measurable and $\mathbb{E}[\xi \,|\, \mathcal{F}] \in \mathbb{L}^1(\Omega, \mathcal{F}, \mathcal{H})$,

(C2) $\mathbb{E}[\xi \mathbb{1}_F] = \mathbb{E}[\mathbb{E}[\xi \,|\, \mathcal{F}] \mathbb{1}_F]$ for all $F \in \mathcal{F}$.

See e.g. Umegaki and Bharucha-Reid (1970) or Pisier (2016, Chapter 1). (C2) in particular means that $\mathbb{E}[\mathbb{E}[\xi \,|\, \mathcal{F}]] = \mathbb{E}[\mathbb{E}[\xi \,|\, \mathcal{F}] \mathbb{1}_\Omega] = \mathbb{E}[\xi]$, since $\Omega \in \mathcal{F}$ for any σ-algebra. It can also be shown that $F(\mathbb{E}[\xi \,|\, \mathcal{F}]) = \mathbb{E}[F(\xi) \,|\, \mathcal{F}]$ for all linear and continuous $F : \mathcal{H} \to \mathbb{R}$ and that $\|\mathbb{E}[\xi \,|\, \mathcal{F}]\|_{\mathcal{H}} \leq \mathbb{E}[\|\xi\|_{\mathcal{H}} \,|\, \mathcal{F}]$ (Pisier, 2016, Chapter 1). Moreover,

(C3) For $\xi \in \mathbb{L}^2(\Omega, \mathcal{A}, \mathcal{H})$, $\mathbb{E}[\xi \,|\, \mathcal{F}]$ is the orthogonal projection into $\mathbb{L}^2(\Omega, \mathcal{F}, \mathcal{H})$,

again we refer to (Umegaki and Bharucha-Reid, 1970). We note that, as with conditional expectation on $\mathbb{R}$, $\mathbb{E}[\xi \,|\, \mathcal{F}]$ is only defined uniquely a.s. As such all (in)equalitie statements hold only a.s. However, we will often not explicitly write this going forward.

We then define $\mathbb{E}[\xi \,|\, \mathbf{X}] = \mathbb{E}[\xi \,|\, \sigma(\mathbf{X})]$. The following Proposition shows that this notion is well-defined and some further properties of Hilbert space-valued conditional expectation, in addition to (C1) – (C3):

Proposition 6 *Let $(\mathcal{H}_1, \langle \cdot, \cdot \rangle_1)$, $(\mathcal{H}_2, \langle \cdot, \cdot \rangle_2)$ be two separable Hilbert spaces, $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2 \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H}_1)$ and $\xi_1, \xi_2, \xi \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H}_2)$.*[2]

(C4) There exists a measurable function $h : (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1)) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$, such that $\mathbb{E}[\xi \,|\, \sigma(\mathbf{X})] = h(\mathbf{X}) = \mathbb{E}[\xi \,|\, \mathbf{X}]$,

(C5) If $\xi_1 \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}_1)$, $\xi_2 \in \mathcal{L}^2(\Omega, \sigma(\mathbf{X}), \mathcal{H}_1)$, then $\mathbb{E}[\langle \xi_1, \xi_2 \rangle_{\mathcal{H}_1} \,|\, \mathbf{X}] = \langle \mathbb{E}[\xi_1 \,|\, \mathbf{X}], \xi_2 \rangle_{\mathcal{H}_1}$,

(C6) If $\mathbf{X}_2$ and $(\xi, \mathbf{X}_1)$ are independent, then $\mathbb{E}[\xi \,|\, \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[\xi \,|\, \mathbf{X}_1]$,

(C7) $\mathbb{E}[\mathbb{E}[\xi \,|\, \mathbf{X}_1, \mathbf{X}_2] \,|\, \mathbf{X}_1] = \mathbb{E}[\mathbb{E}[\xi \,|\, \mathbf{X}_1] \,|\, \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[\xi \,|\, \mathbf{X}_1]$.

Proof We will use the following fact in the proof: Under the assumption of separability, all relevant notions of measurability are the same, see e.g. Hsing and Eubank (2015, Chapter 2, 7). In particular, $\xi \in \mathbb{L}^q(\Omega, \mathcal{F}, \mathcal{H}_2)$, for any $q \geq 1$, means that $\xi : (\Omega, \mathcal{F}) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ is measurable, which in turns means there exists a sequence of simple functions

$$f_n = \sum_{k=1}^{m_n} g_k \mathbb{1}_{A_k}, \tag{28}$$

with $g_k \in \mathcal{H}_2$ and $A_k \in \mathcal{F}$ for all k and such that $f_n \to \xi$ a.s. on $\mathcal{H}_2$ and even $\|f_n - \xi\|_{\mathbb{L}^q(\mathcal{H})} \to 0$, see e.g. Pisier (2016, Proposition 1.2).

---

2. We again note that all equalities technically only hold a.s.

For (C4), we note that by (C1), $\mathbb{E}[\xi \mid \sigma(\mathbf{X})] \in \mathbb{L}^1(\Omega, \sigma(\mathbf{X}), \mathcal{H}_2)$ and thus there exists a sequence of functions $f_n : (\Omega, \sigma(\mathbf{X})) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ of the form (28), such that $f_n \to \mathbb{E}[\xi \mid \sigma(\mathbf{X})]$ a.s. on $\mathcal{H}_2$. Since $A_k \in \sigma(\mathbf{X})$, $A_k = \{\omega : \mathbf{X}(\omega) \in B_k\}$ for some $B_k \in \mathcal{B}(\mathcal{H}_1)$, we may transform f_n from a function on Ω to a function on $\mathcal{H}_1$ into $\mathcal{H}_2$:

$$f_n(\omega) = \sum_{k=1}^{m_n} g_k \mathbb{1}_{A_k}(\omega) = \sum_{k=1}^{m_n} g_k \mathbb{1}_{B_k}(\mathbf{X}(\omega)) := h_n(\mathbf{X}(\omega)).$$

This defines a sequence of measurable functions $h_n : (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1)) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ with $h(\mathbf{X}) = \lim_n h_n(\mathbf{X}) = \mathbb{E}[\xi \mid \sigma(\mathbf{X})]$ a.s., proving the result.

We first show (C5) for simple functions and then extend this to $\mathcal{L}^2(\Omega, \sigma(\mathbf{X}), \mathcal{H}_1)$, using the fact at the beginning of the proof. Let thus f_n be of the form (28). Then for all $F \in \sigma(\mathbf{X})$,

$$
\begin{aligned}
\mathbb{E}[\langle \mathbb{E}[f_n \mid \mathbf{X}], \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F] &= \sum_{k=1}^{m_n} \mathbb{E}[\langle \mathbb{E}[\mathbb{1}_{A_k} \mid \mathbf{X}] g_k, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F] \\
&= \sum_{k=1}^{m_n} \mathbb{E}[\mathbb{E}[\mathbb{1}_{A_k} \mid \mathbf{X}] \langle g_k, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F] \\
&= \sum_{k=1}^{m_n} \mathbb{E}[\mathbb{E}[\mathbb{1}_{A_k} \langle g_k, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F \mid \mathbf{X}]] \\
&= \mathbb{E}[\langle f_n, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F],
\end{aligned}
$$

from the properties of real-valued conditional expectation. As additionally $\langle \mathbb{E}[f_n \mid \mathbf{X}], \xi_2 \rangle_{\mathcal{H}}$ is clearly $\sigma(\mathbf{X})$ measurable, (C1) and (C2) are met for this candidate. Since conditional expectation is (a.s.) uniquely defined by (C1) and (C2), (C5) holds true for the special case of simple functions. For general $\xi_1 \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}_1)$, let f_n have $\|f_n - \xi_1\|_{\mathbb{L}^2(\mathcal{H})} \to 0$. The goal is to show that

$$|\mathbb{E}[\langle \mathbb{E}[f_n \mid \mathbf{X}], \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F] - \mathbb{E}[\langle \mathbb{E}[\xi_1 \mid \mathbf{X}], \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F]| \to 0, \tag{29}$$

$$|\mathbb{E}[\langle f_n, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F] - \mathbb{E}[\langle \xi_1, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F]| \to 0. \tag{30}$$

We can bound both terms by the same quantity, using Cauchy–Schwarz:

$$
\begin{aligned}
|\mathbb{E}[\langle \mathbb{E}[f_n \mid \mathbf{X}], \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F] - \mathbb{E}[\langle \mathbb{E}[\xi_1 \mid \mathbf{X}], \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F]| &= |\mathbb{E}[\langle \mathbb{E}[f_n - \xi_1 \mid \mathbf{X}], \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F]| \\
&\leq \mathbb{E}[\mathbb{E}[\|f_n - \xi_1\|_{\mathcal{H}} \mathbb{1}_F \|\xi_2\|_{\mathcal{H}} \mid \mathbf{X}]] \\
&= \mathbb{E}[\|f_n - \xi_1\|_{\mathcal{H}} \mathbb{1}_F \|\xi_2\|_{\mathcal{H}}],
\end{aligned}
$$

as the random variable $\mathbb{1}_F \|\xi_2\|_{\mathcal{H}}$ is $\sigma(\mathbf{X})$ measurable by assumption and

$$
\begin{aligned}
|\mathbb{E}[\langle f_n, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F] - \mathbb{E}[\langle \xi_1, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F]| &= |\mathbb{E}[\langle f_n - \xi_1, \xi_2 \rangle_{\mathcal{H}} \mathbb{1}_F]| \\
&\leq \mathbb{E}[\|f_n - \xi_1\|_{\mathcal{H}} \|\xi_2\|_{\mathcal{H}} \mathbb{1}_F].
\end{aligned}
$$

The result thus follows from the Hölder inequality,

$$\mathbb{E}[\|f_n - \xi_1\|_{\mathcal{H}} \|\xi_2\|_{\mathcal{H}} \mathbb{1}_F] \leq \|f_n - \xi_1\|_{\mathbb{L}^2(\mathcal{H})} \cdot \|\xi_2\|_{\mathbb{L}^2(\mathcal{H})} \to 0.$$

Finally (C6) and (C7) is easily proven with the technique of "scalarization" (Pisier, 2016, Chapter 1). We do the full argument for (C6), (C7) can be shown analogously. That is, we show that for any $F : \mathcal{H}_2 \to \mathbb{R}$ linear and continuous,

$$F(\mathbb{E}[\xi \mid \mathbf{X}_1, \mathbf{X}_2](\omega)) = F(\mathbb{E}[\xi \mid \mathbf{X}_1](\omega)). \tag{31}$$

for almost all ω and some representation of $\mathbb{E}[\xi \mid \mathbf{X}_1, \mathbf{X}_2]$. This then immediately implies the result. Now, using the property of real-valued conditional expectations

$$F(\mathbb{E}[\xi \mid \mathbf{X}_1, \mathbf{X}_2]) = \mathbb{E}[F(\xi) \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[F(\xi) \mid \mathbf{X}_1],$$

since $F(\xi)$ is real-valued and independent of $\mathbf{X}_2$. Since $\mathbb{E}[F(\xi) \mid \mathbf{X}_1] = F(\mathbb{E}[\xi \mid \mathbf{X}_1])$, we obtain (31). ∎

(C4) in particular allows to see $\mathbb{E}[\xi \mid \sigma(\mathbf{X})]$ as a function in $\mathbf{X}$ and thus justifies the notation $\mathbb{E}[\xi \mid \mathbf{X}]$ and all the subsequent derivations. We may also define *conditional independence* through conditional expectation: With the notation of Proposition 6, ξ and $\mathbf{X}_1$ are conditionally independent given $\mathbf{X}_2$, if $\mathbb{E}[f(\xi) \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[f(\xi) \mid \mathbf{X}_1]$ for all $f : (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ bounded and measurable, see e.g., Constantinou and Dawid (2017, Proposition 2.3). This leads to two further important properties:

Proposition 7 *Let* $(\mathcal{H}_1, \langle \cdot, \cdot \rangle_1)$, $(\mathcal{H}_2, \langle \cdot, \cdot \rangle_2)$ *be two separable Hilbert spaces,* $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2 \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H}_1)$ *and* $\xi_1, \xi_2, \xi \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H}_2)$.

(C8) If ξ and $\mathbf{X}_2$ are conditionally independent given $\mathbf{X}_1$, then $\mathbb{E}[\xi \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[\xi \mid \mathbf{X}_1]$,

(C9) If ξ_1, ξ_2 are conditionally independent given $\mathbf{X}$, $\mathbb{E}[\langle \xi_1, \xi_2 \rangle \mid \mathbf{X}] = \langle \mathbb{E}[\xi_1 \mid \mathbf{X}], \mathbb{E}[\xi_2 \mid \mathbf{X}] \rangle$.

Proof We prove (C8) again using the scalarization trick: For any $F : \mathcal{H}_2 \to \mathbb{R}$ continuous and linear, it holds that

$$F_n(f) := F(f)\mathbb{1}\{|F(f)| \leq n\} \quad \forall f \in \mathcal{H}_2,$$

is a bounded and measurable function. Thus by assumption,

$$\mathbb{E}[F_n(\xi) \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[F_n(\xi) \mid \mathbf{X}_1].$$

We now show that $\mathbb{E}[F_n(\xi) \mid \mathbf{X}_1, \mathbf{X}_2] \to \mathbb{E}[F(\xi) \mid \mathbf{X}_1, \mathbf{X}_2]$ and $\mathbb{E}[F_n(\xi) \mid \mathbf{X}_1] \to \mathbb{E}[F(\xi) \mid \mathbf{X}_1]$ a.s. Let $\mathcal{F}$ stand for either $\sigma(\mathbf{X}_1, \mathbf{X}_2)$ or $\sigma(\mathbf{X}_1)$. Then, as $F(f) = F_n(f) + F(f)\mathbb{1}\{|F(f)| > n\}$,

$$|\mathbb{E}[F_n(\xi) \mid \mathcal{F}] - \mathbb{E}[F(\xi) \mid \mathcal{F}]| \leq \mathbb{E}[|F_n(\xi) - F(\xi)| \mid \mathcal{F}]$$
$$= \mathbb{E}[|F(f)|\mathbb{1}\{|F(f)| > n\} \mid \mathcal{F}].$$

Now, since for all n, $|F(f)|\mathbb{1}\{|F(f)| > n\} \geq 0$ and $\mathbb{E}[|F(f)|\mathbb{1}\{|F(f)| > n\} \mid \mathcal{F}] \leq \mathbb{E}[|F(f)| \mid \mathcal{F}] < \infty$ a.s. an application of Fatou's Lemma for (real-valued) conditional expectation (see e.g., Dudley (2002, Problem 10.7)) to $|F(f)| - |F(f)|\mathbb{1}\{|F(f)| > n\}$ implies

$$\limsup_n |\mathbb{E}[F_n(\xi) \mid \mathcal{F}] - \mathbb{E}[F(\xi) \mid \mathcal{F}]| \leq \mathbb{E}[\limsup_n |F(f)|\mathbb{1}\{|F(f)| > n\} \mid \mathcal{F}] = 0.$$

Thus, we have shown that for all $F : \mathcal{H}_2 \to \mathbb{R}$ continuous and linear, $F(\mathbb{E}[\xi \mid \mathbf{X}_1, \mathbf{X}_2]) = F(\mathbb{E}[\xi \mid \mathbf{X}_1])$, proving the claim.

Finally, combining (C5), (C7) and (C8), we obtain for ξ_1, ξ_2 conditionally independent given $\mathbf{X}$

$$
\begin{aligned}
\mathbb{E}[\langle \xi_1, \xi_2 \rangle \mid \mathbf{X}] &= \mathbb{E}[\mathbb{E}[\langle \xi_1, \xi_2 \rangle \mid \mathbf{X}, \xi_2] \mid \mathbf{X}] \\
&= \mathbb{E}[\langle \mathbb{E}[\xi_1 \mid \mathbf{X}, \xi_2], \xi_2 \rangle \mid \mathbf{X}] \\
&= \mathbb{E}[\langle \mathbb{E}[\xi_1 \mid \mathbf{X}], \xi_2 \rangle \mid \mathbf{X}] \\
&= \langle \mathbb{E}[\xi_1 \mid \mathbf{X}], \mathbb{E}[\xi_2 \mid \mathbf{X}] \rangle.
\end{aligned}
$$

$\blacksquare$

Let for $\mathbf{x} \in \mathcal{X}$, $P_{\mathbf{x}}$ be the conditional distribution of Y given $\mathbf{x}$ on $\mathbb{R}^d$ and similarly with $P_{\mathbf{X}}$ (i.e. the regular conditional probability measure). Note that $P_{\mathbf{x}} \in \mathcal{H}$, while $P_{\mathbf{X}}$ is a random element mapping into $\mathcal{H}$.

As in Wager and Athey (2018), we define for f, g two functions, with $\liminf_{s \to \infty} g(s) > 0$, $f(s) = \mathcal{O}(g(s))$ if

$$
\limsup_{s \to \infty} \frac{|f(s)|}{g(s)} \leq C,
$$

for some $C > 0$. If $C = 1$, then we write $f(s) \precsim g(s)$. For a sequence of random variables $X_n : \Omega \to \mathbb{R}$, and $a_n \in (0, +\infty)$, $n \in \mathbb{N}$, we write as usual $X_n = \mathcal{O}_p(a_n)$, if

$$
\lim_{M \to \infty} \sup_n \mathbb{P}(a_n^{-1}|X_n| > M) = 0,
$$

i.e. if X_n is bounded in probability. We write $X_n = o_p(a_n)$, if $a_n^{-1} X_n$ converges in probability to zero. Similarly, for (S, d) a separable metric space, $\mathbf{X}_n : (\Omega, \mathcal{A}) \to (S, \mathcal{B}(S))$, $n \in \mathbb{N}$ and $\mathbf{X} : (\Omega, \mathcal{A}) \to (S, \mathcal{B}(S))$ measurable, we write $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$, if $d(\mathbf{X}_n, \mathbf{X}) = o_p(1)$.

Finally let $\mathbf{X} \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}_1)$, $\xi \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}_2)$ and assume that $A \subset \Omega$ depends on $\mathbf{X}$, $A = A(\mathbf{X})$. Thus for $\mathbf{X}$ fixed to a certain value, A is a fixed set. If $\mathbb{P}(A \mid \mathbf{X}) > 0$ almost everywhere, we define

$$
\mathbb{E}[\xi \mid A] = \mathbb{E}[\xi \mid \mathbf{X}, A] := \frac{\mathbb{E}[\xi \mathbb{1}_A \mid \mathbf{X}]}{\mathbb{P}(A \mid \mathbf{X})} \in \mathcal{L}^2(\Omega, \sigma(\mathbf{X}), \mathcal{H}_2).
$$

It then holds by construction that

$$
\mathbb{E}[\xi \mathbb{1}_A \mid \mathbf{X}] = \mathbb{E}[\xi \mid \mathbf{X}, A] \cdot \mathbb{P}(A \mid \mathbf{X}). \tag{32}
$$

Let again $\mu(\mathbf{x}) := \mu(P_{\mathbf{x}})$ be the embedding of the true conditional distribution into $\mathcal{H}$. We first state 3 preliminary results:

Lemma 8 *It holds that* $\mathbb{E}[\mu(\delta_{\mathbf{Y}}) \mid \mathbf{X} = \mathbf{x}] = \mu(P_{\mathbf{x}})$.

Proof We first note that $P_{\mathbf{x}}$ exists and is a probability measure on $\mathbb{R}^d$. Since k is bounded $\mu(P_{\mathbf{x}}) \in \mathcal{H}$ exists and is uniquely defined by the relation

$$
\langle f, \mu(P_{\mathbf{x}}) \rangle_{\mathcal{H}} = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}] \quad \forall f \in \mathcal{H}.
$$

On the other hand, by the above $\mathbb{E}[\mu(\delta_\mathbf{Y}) \mid \mathbf{X}] \in \mathbb{L}^2(\Omega, \mathcal{A}, \mathcal{H})$ exists and for all $f \in \mathcal{H}$,

$$\langle \mathbb{E}[\mu(\delta_\mathbf{Y}) \mid \mathbf{X}], f \rangle_\mathcal{H} = \mathbb{E}[\langle \mu(\delta_\mathbf{Y}), f \rangle_\mathcal{H} \mid \mathbf{X}] = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}],$$

since $F : H \to \mathbb{R}$, $F(g) = \langle g, f \rangle_\mathcal{H}$ defines a continuous linear function. In particular, for all $f \in \mathcal{H}$,

$$\langle \mathbb{E}[\mu(\delta_\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}], f \rangle_\mathcal{H} = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}],$$

or $\mathbb{E}[\mu(\delta_\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}] = \mu(P_\mathbf{x})$. ∎

Since $\mu(\delta_\mathbf{Y}) = k(\mathbf{Y}, \cdot)$, Lemma 8 in fact corresponds to Lemma 3.2 in Park and Muandet (2020).

For a more compact notation in the following Lemma, let $N = \{1, \ldots, n\}$ and let for $A \subset N$ and $k \leq |A|$, $C_k(A)$ be the set of all subsets of size k drawn from A without replacement, with $C_0 := \emptyset$.

Lemma 9 (H-Decomposition of a Hilbert-space valued Kernel) *Let* $(\mathcal{H}_1, \langle \cdot, \cdot \rangle_1)$, $(\mathcal{H}_2, \langle \cdot, \cdot \rangle_2)$ *be two separable Hilbert spaces,* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *be i.i.d. copies of a random element* $\mathbf{X} : (\Omega, \mathcal{A}) \to (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1))$. *Write* $\mathcal{X}_n = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ *and let* $T : (\mathcal{H}_1^n, \mathcal{B}(\mathcal{H}_1^n)) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ *measurable with* $\mathbb{E}[\|T(\mathcal{X}_n)\|_{\mathcal{H}_2}^2] < \infty$. *If* T *is symmetric, there exists functions* T_j, $j = 1, \ldots, n$, *such that*

$$T(\mathcal{X}_n) = \mathbb{E}[T(\mathbf{X})] + \sum_{i=1}^n T_1(\mathbf{X}_i) + \sum_{i_1 < i_2} T_2(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}) + \ldots T_n(\mathcal{X}_n), \quad (33)$$

and it holds that

$$\mathrm{Var}(T(\mathcal{X}_n)) = \sum_{i=1}^n \binom{n}{i} \mathrm{Var}(T_i(\mathbf{X}_1, \ldots, \mathbf{X}_i)), \quad (34)$$

and

$$T_1(\mathbf{X}_i) = \mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_i] - \mathbb{E}[T(\mathcal{X}_n)].$$

Proof Composition (33) was proven in a range of different ways for real-valued T, see e.g. Hoeffding (1961), Dynkin and Mandelbaum (1983), Efron and Stein (1981), or Vitale (1992). We consider and slightly extend the elegant proof of Dynkin and Mandelbaum (1983) to also prove (34). See also Vitale (1992). Let

$$T_1(\mathbf{X}_i) = \mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_i] - \mathbb{E}[T(\mathcal{X}_n)]$$
$$T_2(\mathbf{X}_i, \mathbf{X}_j) = \mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_i, \mathbf{X}_j] - \mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_i] - \mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_j] + \mathbb{E}[T(\mathcal{X}_n)]$$
$$\vdots$$
$$T_\ell(\pi_{NA_\ell}(\mathcal{X}_n)) = \sum_{k=0}^{\ell-1} (-1)^{\ell-k} \sum_{B \in C_k(A_\ell)} \mathbb{E}[T(\mathcal{X}_n) \mid \pi_{NB}(\mathcal{X}_n)].$$

We note that T_ℓ does not depend on the exact indices in A_ℓ thanks to the assumed symmetry of T. Since $\mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_1, \ldots, \mathbf{X}_n] = T(\mathcal{X}_n)$ is part of $T_n(\mathcal{X}_n)$, this leads to a telescoping sum, already proving (33).

Adapting the approach of Dynkin and Mandelbaum (1983), consider now $\mathcal{L}^2(\Omega, \sigma(\mathcal{X}_n), \mathcal{H})$ and let Q_i be the projection operator into $\mathcal{L}^2(\Omega, \sigma(\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n), \mathcal{H})$. That is

$$(Q_i T)(\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n) = \mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n],$$

by (C3). Now as in Vitale (1992),

(I) the Q_i commute,

(II) For $A_\ell = \{i_1, \ldots, i_\ell\} \subset N$,

$$(Q_{i_1} \cdots Q_{i_\ell} T)(\pi_{N A_\ell^c}(\mathcal{X}_n)) = \mathbb{E}[T(\mathcal{X}_n) \mid \pi_{N A_\ell^c}(\mathcal{X}_n)].$$

In particular,

$$(Q_1 \cdots Q_\ell T)(\mathbf{X}_{\ell+1}, \ldots, \mathbf{X}_n) = \mathbb{E}[T(\mathcal{X}_n) \mid \mathbf{X}_{\ell+1}, \ldots, \mathbf{X}_n].$$

Moreover, it holds that

$$T_\ell(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_\ell}) = ([I - Q_{i_1}][I - Q_{i_2}] \ldots [I - Q_{i_\ell}] Q_{i_{\ell+1}} \cdots Q_{i_n} T)(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_\ell}).$$

Expanding the identity

$$T(\mathcal{X}_n) = (I^n T)(\mathcal{X}_n) = ([[(I - Q_1) + Q_1][(I - Q_2) + Q_2] \cdots [(I - Q_n) + Q_n] T)(\mathcal{X}_n),$$

as in Vitale (1992) and using $Q_i(1 - Q_i) = 0$, proves (33). Furthermore the above implies that for any subset $A_l \subset N$ that intersects with $A_\ell = \{i_1, \ldots, i_\ell\}$, we must have $\mathbb{E}[T_\ell(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_\ell}) \mid \pi_{N A_l}(\mathcal{X}_n)] = 0$. Indeed assume $A_l \cap A_\ell = \{i_2, \ldots, i_\ell\}$, then since the elements of $\mathcal{X}_n$ are independent, by (C6),

$$\mathbb{E}[T_\ell(\pi_{N A_\ell}(\mathcal{X}_n)) \mid \pi_{N A_l}(\mathcal{X}_n)] = \mathbb{E}[T_\ell(\pi_{N A_\ell}(\mathcal{X}_n)) \mid \pi_{N(A_l \cap A_\ell)}(\mathcal{X}_n)],$$

i.e. all elements outside the intersection are irrelevant. Moreover,

$$\begin{aligned}
&\mathbb{E}[T_\ell(\pi_{N A_\ell}(\mathcal{X}_n)) \mid \pi_{N(A_l \cap A_\ell)}(\mathcal{X}_n)] \\
&= (Q_{i_1}[I - Q_{i_1}][I - Q_{i_2}] \ldots [I - Q_{i_\ell}] Q_{i_{\ell+1}} \cdots Q_{i_n} T)(\pi_{N(A_l \cap A_\ell)}(\mathcal{X}_n))
\end{aligned}$$

and clearly this projection can only be 0. The same argument can be made for any other intersection set $A_l \cap A_\ell$, even if it is the empty set.

Thus combining the above with (C5), for $A_\ell \neq A_l$, it holds that

$$\begin{aligned}
\mathbb{E}[\langle T_\ell(\pi_{N A_\ell}(\mathcal{X}_n)), T_l(\pi_{N A_l}(\mathcal{X}_n)) \rangle_{\mathcal{H}}] &= \mathbb{E}[\mathbb{E}[\langle T_\ell(\pi_{N A_\ell}(\mathcal{X}_n)), T_l(\pi_{N A_l}(\mathcal{X}_n)) \rangle_{\mathcal{H}} \mid \pi_{N A_l}(\mathcal{X}_n)]] \\
&= \mathbb{E}[\langle \mathbb{E}[T_\ell(\pi_{N A_\ell}(\mathcal{X}_n)) \mid \pi_{N A_l}(\mathcal{X}_n)], T_l(\pi_{N A_l}(\mathcal{X}_n)) \rangle_{\mathcal{H}}] \\
&= 0.
\end{aligned}$$

In other words, the covariance between any two elements in the decomposition of $T(\mathcal{X}_n) - \mathbb{E}[T(\mathbf{X})]$ in (33) is uncorrelated. Finally then

$$\mathrm{Var}(T(\mathbf{X}_1, \ldots, \mathbf{X}_n)) = \mathbb{E}[\langle T(\mathcal{X}_n) - \mathbb{E}[T(\mathcal{X}_n)], T(\mathcal{X}_n) - \mathbb{E}[T(\mathcal{X}_n)]\rangle]$$
$$= \sum_{i=1}^n \binom{n}{i} \mathrm{Var}(T_i(\mathbf{X}_1, \ldots, \mathbf{X}_i)).$$

■

In order to proceed, we first prove Theorem 1, which is somewhat separate from the remainder of this section:

Theorem 1 *For any split of a parent node P into child nodes C_L and C_R, let $\hat{\mathbb{P}}_{split}(\mathbf{x}) = \sum_{j \in \{L,R\}} \mathbb{1}(\mathbf{x} \in C_j) \frac{1}{n_j} \sum_{\mathbf{x}_i \in C_j} \delta_{\mathbf{y}_i}$ denote the resulting estimate of the distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ when $\mathbf{x} \in P$. Then the MMD splitting criterion can be viewed as the version of the CART criterion (7) on $\mathcal{H}$:*

$$\underset{split}{\arg\max} \frac{n_L n_R}{n_P^2} \mathcal{D}_{MMD(k)}\left(\{\mathbf{y}_i \mid \mathbf{x}_i \in C_L\}, \{\mathbf{y}_i \mid \mathbf{x}_i \in C_R\}\right)$$
$$= \underset{split}{\arg\min} \frac{1}{n_P} \sum_{\mathbf{x}_i \in P} \left\| \mu(\delta_{\mathbf{y}_i}) - \mu(\hat{\mathbb{P}}_{split}(\mathbf{x}_i)) \right\|_{\mathcal{H}}^2.$$

Moreover, for any node P and any fixed distributional estimator $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$, we have:

$$\frac{1}{n_P} \sum_{\mathbf{x}_i \in P} \left\| \mu(\delta_{\mathbf{y}_i}) - \mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}_i)) \right\|_{\mathcal{H}}^2$$
$$= V_P + \mathbb{E}\left[\|\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X})) - \mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X}))\|_{\mathcal{H}}^2 \mid \mathbf{X} \in P \right] + \mathcal{O}_p(n^{-1/2}),$$

where $V_P = \mathbb{E}\left[\|\mu(\delta_{\mathbf{Y}}) - \mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X}))\|_{\mathcal{H}}^2 \mid \mathbf{X} \in P \right]$ is a deterministic term not depending on the estimates $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$.

Proof The first part of the Theorem is shown analogously to the proof in Section B.3 of the appendix, but where we replace the standard dot product in $\mathbb{R}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ associated with $(\mathcal{H}, k)$ and use the induced RKHS norm $\|\cdot\|_{\mathcal{H}}$. Also, since k is bounded, the embedding $\mu(\mathcal{D})$ into RKHS $\mathcal{H}$ exists for any distribution $\mathcal{D}$, so everything is well-defined.

For the second statement of the Theorem 1, note that $n_P \sim \mathrm{Binomial}(\pi, n)$, where $\pi := \mathbb{P}(\mathbf{X} \in P) > 0$. Let $\hat{P}_{\mathbf{x}} = \hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ be a fixed conditional distribution estimator and recall that $P_{\mathbf{x}} = \mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$. We now write

$$\sum_{\mathbf{x}_i \in P} \left\| \mu(\delta_{\mathbf{y}_i}) - \mu(\hat{P}_{\mathbf{x}_i}) \right\|_{\mathcal{H}}^2 = \sum_{i=1}^n \left\| \mu(\delta_{\mathbf{y}_i}) - \mu(\hat{P}_{\mathbf{x}_i}) \right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{x}_i \in P\}.$$

Then it holds that

$$\mathbb{E}\left[\left\| \mu(\delta_{\mathbf{Y}_i}) - \mu(\hat{P}_{\mathbf{X}_i}) \right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{X}_i \in P\} \right] = \mathbb{E}\left[\mathbb{E}\left[\left\| \mu(\delta_{\mathbf{Y}}) - \mu(\hat{P}_{\mathbf{X}}) \right\|_{\mathcal{H}}^2 \mid \mathbf{X} \right] \mathbb{1}\{\mathbf{X} \in P\} \right],$$

and

$$\mathbb{E}\left[\left\|\mu(\delta_{\boldsymbol{Y}}) - \mu(\hat{P}_{\mathbf{X}})\right\|_{\mathcal{H}}^2 \mid \mathbf{X}\right] = \mathbb{E}\left[\mathbb{E}\|\mu(\delta_{\boldsymbol{Y}}) - \mu(P_{\mathbf{X}})\|_{\mathcal{H}}^2 + \mathbb{E}\left\|\mu(P_{\mathbf{X}}) - \mu(\hat{P}_{\mathbf{X}})\right\|_{\mathcal{H}}^2 \mid \mathbf{X}\right]$$
$$+ 2\mathbb{E}\left[\langle\mu(\delta_{\boldsymbol{Y}}) - \mu(P_{\mathbf{X}}),\, \mu(P_{\mathbf{X}}) - \mu(\hat{P}_{\mathbf{X}})\rangle_{\mathcal{H}} \mid \mathbf{X}\right].$$

It follows with Lemma 8 and (C5) that,

$$\mathbb{E}\left[\langle\mu(\delta_{\boldsymbol{Y}}) - \mu(P_{\mathbf{X}}),\, \mu(P_{\mathbf{X}}) - \mu(\hat{P}_{\mathbf{X}})\rangle_{\mathcal{H}} \mid \mathbf{X}\right] =$$
$$\langle\mathbb{E}\left[\mu(\delta_{\boldsymbol{Y}}) - \mu(P_{\mathbf{X}}) \mid \mathbf{X}\right],\, \mu(P_{\mathbf{X}}) - \mu(\hat{P}_{\mathbf{X}})\rangle_{\mathcal{H}} = 0.$$

Combining the three equations, this means

$$\mathbb{E}\left[\left\|\mu(\delta_{\boldsymbol{Y}_i}) - \mu(\hat{P}_{\mathbf{X}_i})\right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{X}_i \in P\}\right]$$
$$= \mathbb{E}\left[\|\mu(\delta_{\boldsymbol{Y}}) - \mu(P_{\mathbf{X}})\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{X} \in P\} + \left\|\mu(P_{\mathbf{X}}) - \mu(\hat{P}_{\mathbf{X}})\right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{X} \in P\}\right]$$
$$= \pi\left(V_P + \mathbb{E}\left[\left\|\mu(P_{\mathbf{X}}) - \mu(\hat{P}_{\mathbf{X}})\right\|_{\mathcal{H}}^2 \mid \mathbf{X} \in P\right]\right), \tag{35}$$

using that $\mathbb{E}[g(\mathbf{X}) \mid \mathbf{X} \in P] = \mathbb{E}[g(\mathbf{X})\mathbb{1}\{\mathbf{X} \in P\}]/\mathbb{P}(\mathbf{X} \in P)$. We now show that the difference between $\frac{1}{n_P}\sum_{\mathbf{x}_i \in P}\left\|\mu(\delta_{\boldsymbol{y}_i}) - \mu(\hat{\mathbb{P}}(\boldsymbol{Y} \mid \mathbf{X}=\mathbf{x}_i))\right\|_{\mathcal{H}}^2$ and the expectation on the left of Equation (35) is $\mathcal{O}_p(n^{-1/2})$, using standard CLT arguments. Define $K = \sup_{z,z'}|k(z,z')| < \infty$, as we have assumed that k is bounded. For any two distributions $\mathcal{D}_1, \mathcal{D}_2$ we now obtain

$$\|\mu(\mathcal{D}_1) - \mu(\mathcal{D}_2)\|_{\mathcal{H}}^2 = \mathbb{E}[k(\mathbf{Z}_1, \mathbf{Z}_1')] - 2\mathbb{E}[k(\mathbf{Z}_1, \mathbf{Z}_2)] + \mathbb{E}[k(\mathbf{Z}_2, \mathbf{Z}_2')] \le 4K,$$

where $\mathbf{Z}_1, \mathbf{Z}_1' \sim \mathcal{D}_1$ and $\mathbf{Z}_2, \mathbf{Z}_2' \sim \mathcal{D}_2$ are independent random variables. Thus,

$$\mathbb{E}\left[\left\|\mu(\delta_{\boldsymbol{Y}}) - \mu(\hat{P}_{\mathbf{X}})\right\|_{\mathcal{H}}^2\right] \le 4K, \quad \mathbb{E}\left[\left\|\mu(\delta_{\boldsymbol{Y}}) - \mu(\hat{P}_{\mathbf{X}})\right\|_{\mathcal{H}}^4\right] \le 16K^2,$$

implying that both first and second moments of the random variable $\left\|\mu(\delta_{\boldsymbol{Y}}) - \mu(\hat{P}_{\mathbf{X}})\right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{X} \in P\}$ are finite. Moreover, since $\left\|\mu(\delta_{\boldsymbol{y}_i}) - \mu(\hat{P}_{\mathbf{x}_i})\right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{x}_i \in P\}$ for $i = 1, \ldots, n$ are its i.i.d. realizations, it follows directly from the CLT that:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n\left\|\mu(\delta_{\boldsymbol{y}_i}) - \mu(\hat{P}_{\mathbf{x}_i})\right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{x}_i \in P\} - \mathbb{E}\left[\left\|\mu(\delta_{\boldsymbol{Y}_i}) - \mu(\hat{P}_{\mathbf{X}_i})\right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{X}_i \in P\}\right]\right)$$
$$= \mathcal{O}_p(1).$$

By multiplying the above equation with $n/n_P = (1/\pi + o_p(1)) = \mathcal{O}_p(1)$, it also holds that

$$\sqrt{n}\left(\frac{1}{n_P}\sum_{\mathbf{x}_i \in P}\left\|\mu(\delta_{\boldsymbol{y}_i}) - \mu(\hat{P}_{\mathbf{x}_i})\right\|_{\mathcal{H}}^2 - \frac{n}{n_P}\mathbb{E}\left[\left\|\mu(\delta_{\boldsymbol{Y}_i}) - \mu(\hat{P}_{\mathbf{X}_i})\right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{X}_i \in P\}\right]\right) = \mathcal{O}_p(1).$$
$$\tag{36}$$

Thus,

$$\sqrt{n}\left(\frac{1}{n_P}\sum_{\mathbf{x}_i \in P}\left\|\mu(\delta_{\mathbf{y}_i}) - \mu(\hat{P}_{\mathbf{x}_i})\right\|_{\mathcal{H}}^2 - \frac{1}{\pi}\mathbb{E}\left[\left\|\mu(\delta_{\mathbf{Y}_i}) - \mu(\hat{P}_{\mathbf{X}_i})\right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{X}_i \in P\}\right]\right)$$

$$= \sqrt{n}\left(\frac{1}{n_P}\sum_{\mathbf{x}_i \in P}\left\|\mu(\delta_{\mathbf{y}_i}) - \mu(\hat{P}_{\mathbf{x}_i})\right\|_{\mathcal{H}}^2 - \frac{n}{n_P}\mathbb{E}\left[\left\|\mu(\delta_{\mathbf{Y}_i}) - \mu(\hat{P}_{\mathbf{X}_i})\right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{X}_i \in P\}\right]\right)$$

$$- \sqrt{n}\left(1 - \frac{n\pi}{n_P}\right)\frac{1}{\pi}\mathbb{E}\left[\left\|\mu(\delta_{\mathbf{Y}_i}) - \mu(\hat{P}_{\mathbf{X}_i})\right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathbf{X}_i \in P\}\right]. \tag{37}$$

Now both terms in (37) are $\mathcal{O}_p(1)$: For the first term this follows from (36). For the second term it holds, since $\mathbb{E}\left[\left\|\mu(\delta_{\mathbf{Y}}) - \mu(\hat{P}_{\mathbf{X}})\right\|_{\mathcal{H}}^2\right] \le 4K$ and

$$\sqrt{n}\left(1 - \frac{n\pi}{n_P}\right) = \mathcal{O}_p(1),$$

which in turn is true by another application of the CLT on random variables $\mathbb{1}\{\mathbf{X}_i \in P\}$ and the fact that $n/n_P = \mathcal{O}_p(1)$:

$$\sqrt{n}\left(1 - \frac{n\pi}{n_P}\right) = \sqrt{n}\frac{n_P - n\pi}{n_P} = \frac{n}{n_P}\frac{n_P - n\pi}{\sqrt{n}} = \mathcal{O}_p(1).$$

Combining the fact that the expression in (37) is $\mathcal{O}_p(1)$ with (35) gives the result. ∎

Let $\hat{\mu}_n(\mathbf{x})$ be defined as in (17):

$$\hat{\mu}_n(\mathbf{x}) = \binom{n}{s_n}^{-1}\sum_{i_1 < i_2 < \ldots < i_{s_n}}\mathbb{E}_\varepsilon\left[T(\mathbf{x}, \varepsilon; \mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}})\right], \tag{38}$$

where the sum is taken over all $\binom{n}{s_n}$ possible subsamples $\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}$ of $\mathbf{Z}_1, \ldots \mathbf{Z}_n$ and $s_n \to \infty$ with n. Moreover, a single tree is given as

$$T(\mathbf{x}; \epsilon_k, \mathbf{Z}_{k_1}, \ldots \mathbf{Z}_{k_{s_n}}) = \sum_{j=1}^{s_n}\frac{\mathbb{1}(\mathbf{X}_{k_j} \in \mathcal{L}_k(\mathbf{x}))}{|\mathcal{L}_k(\mathbf{x})|}\mu(\delta_{\mathbf{Y}_{k_j}}). \tag{39}$$

Given these preliminary results, the proofs we use are for the most part analogous to the ones in Wager and Athey (2018). Thus the proofs are given mostly for completeness and sometimes omitted altogether. We introduce the following additional notation, similar to Section 3: Let $\mathcal{Z}_s = (\mathbf{Z}_1, \ldots, \mathbf{Z}_s)$ collect s i.i.d. copies of $\mathbf{Z}$ and define for $j = 1, \ldots, s_n$,

$$\text{Var}(T) = \text{Var}(T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n}))$$
$$\text{Var}(T_j) = \text{Var}(\mathbb{E}[T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})|\mathbf{Z}_1, \ldots, \mathbf{Z}_j]).$$

The index s will take the role of s_n or n, depending on the situation. We note that, due to i.i.d. sampling, it doesn't matter for variance or expectation what kind of subset $\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_s}$ we are considering. In particular, we might just take $\mathcal{Z}_s$ each time.

Before going on to the main proofs, we repeat here the assumed properties of the trees for completeness:

(P1) (*Data sampling*) The bootstrap sampling with replacement, usually used in forest-based methods, is replaced by a subsampling step, where for each tree we choose a random subset of size s_n out of n training data points. We consider s_n going to infinity with n, with the rate specified below.

(P2) (*Honesty*) An observation $\mathbf{Z} = (\mathbf{X}, \mu(\delta_{\mathbf{Y}}))$ is either used to place the splits in a tree or to estimate the response, but never both.

(P3) (α-*regularity*) Each split leaves at least a fraction $\alpha \leq 0.2$ of the available training sample on each side. Moreover, the trees are grown until every leaf contains between κ and $2\kappa - 1$ observations, for some fixed tuning parameter $\kappa \in \mathbb{N}$.

(P4) (*Symmetry*) The (randomized) output of a tree does not depend on the ordering of the training samples.

(P5) (*Random-split*) At every split point, the probability that the split occurs along the feature X_j is bounded below by π/p, for some $\pi > 0$ and for all $j = 1, \ldots, p$.

Note that assumption **(P2)** differentiates from assumption **(P2)** in the main text. We refer to it as

(P2') (*Double Sampling*) The data used for constructing each tree is split into two parts; the first is used for determining the splits and the second for populating the leaves and thus for estimating the response.

(P2) will allow us to assume that all s_n observations of the tree are used to estimate the response, as in (39) and the trees are built with some auxiliary data. This is done for simplicity of exposition, the results can be extended to hold in case of **(P2')** as well. In fact, since the (random) division into the two data sets can be seen as part of ε_k, the adaptation simply involves changing s_n to $s_n/2$.

Now, we may directly apply Lemma (9) to the U-statistics $\hat{\mu}_n(\mathbf{x})$:

Lemma 10 *Let $\hat{\mu}_n(\mathbf{x})$ be as in* (38) *and assume T satisfies* **(P4)** *and*

$$\mathrm{Var}(T) < \infty.$$

Then

$$\mathrm{Var}(\hat{\mu}_n(\mathbf{x})) \leq \frac{s_n^2}{n}\mathrm{Var}(T_1) + \frac{s_n^2}{n^2}\mathrm{Var}(T)$$

$$\leq \left(\frac{s_n}{n} + \frac{s_n^2}{n^2}\right)\mathrm{Var}(T).$$

Proof Using the composition in (33) on $T(\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_s}) := \mathbb{E}_\varepsilon\left[T(\mathbf{x}, \varepsilon; \mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_s})\right]$, we have for $A \subset N = \{1, \ldots, n\}$, $|A| = s_n$,

$$T(\pi_{NA}(\mathcal{Z}_n)) = \mathbb{E}[T(\pi_{NA}(\mathcal{Z}_n))] + \sum_{\ell=1}^{s} \sum_{B \in C_\ell(A)} T_\ell(\pi_{NB}(\mathcal{Z}_n)), \tag{40}$$

Moreover, it holds by symmetry and i.i.d. sampling, that for $A_1 = A_2 \subset N$, $|A_1| = |A_2| = \ell$, $T_\ell(\pi_{A_1}(\mathcal{Z}_n)) = T_\ell(\pi_{A_2}(\mathcal{Z}_n))$. Thus we obtain,

$$\hat{\mu}_n(\mathbf{x}) = \mathbb{E}[T(\mathcal{Z}_{s_n})] + \binom{n}{s_n}^{-1} \left(\binom{n-1}{s_n-1} \sum_{i=1}^{n} T_1(\mathbf{Z}_i) + \binom{n-2}{s_n-2} \sum_{i_1 < i_2} T_2(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}) \right.$$

$$\left. + \ldots + \sum_{i_1 < i_2 < \ldots < i_{s_n}} T_s(\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}) \right).$$

Now

$$\binom{n}{s_n}^{-1} \binom{n-j}{s_n-j} = \frac{s_n!}{n!} \frac{(n-j)!}{(s_n-j)!}$$

$$= \frac{s_n \cdot (s_n-1) \cdots (s_n-j+1)}{n \cdot (n-1) \cdots (n-j+1)}$$

$$= \frac{(s_n)_j}{(n)_j},$$

where $(s_n)_j = s_n(s_n-1) \cdot (s_n - (j-1)) = s_n!/(s_n-j)!$. In particular

$$\binom{n}{s_n}^{-1} \binom{n-1}{s_n-1} = \frac{s_n}{n}.$$

Consequently,

$$\hat{\mu}_n(\mathbf{x}) = \frac{s_n}{n} \sum_{i=1}^{n} T_1(\mathbf{Z}_i) + \frac{(s_n)_2}{(n)_2} \sum_{i_1 < i_2} T_2(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}) + \frac{(s_n)_3}{(n)_3} \sum_{i_1 < i_2 < i_3} T_3(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \mathbf{Z}_{i_3}) +$$

$$\ldots + \frac{(s_n)_{s_n}}{(n)_{s_n}} \sum_{i_1 < i_2 < \ldots < i_{s_n}} T_{s_n}(\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}),$$

with covariances between terms equal to 0, as in Lemma 9. Thus

$$\text{Var}(\hat{\mu}_n(\mathbf{x})) = \frac{s_n^2}{n^2} n \text{Var}(T_1) + \sum_{i=2}^{s_n} \left(\frac{(s_n)_i}{(n)_i} \right)^2 \binom{n}{i} \text{Var}(T_i)$$

$$= \frac{s_n^2}{n} \text{Var}(T_1) + \sum_{i=2}^{s_n} \left(\frac{(s_n)_i}{(n)_i} \right) \binom{s_n}{i} \text{Var}(T_i)$$

$$\leq \frac{s_n^2}{n} \text{Var}(T_1) + \frac{(s_n)_2}{(n)_2} \sum_{i=2}^{s_n} \binom{s_n}{i} \text{Var}(T_i)$$

$$\leq \frac{s_n^2}{n} \text{Var}(T_1) + \frac{s_n^2}{n^2} \text{Var}(T),$$

where the last step followed from (34). This proves the first inequality. On the other hand, we have from Lemma 9 that

$$\text{Var}(T) = \sum_{i=1}^{s_n} \binom{s_n}{i} \text{Var}(T_i) \geq s_n \text{Var}(T_1),$$

leading to the second inequality. ∎

Lemma 11 *[Lemma 2 from Wager and Athey (2018)] Let T is a tree satisfying **(P3)**, **(P5)** trained on $\mathcal{Z}_s = (\xi_1, \mathbf{X}_1), \ldots (\xi_s, \mathbf{X}_s)$ and let $L(x, \mathcal{Z}_s)$ be the leaf containing $\mathbf{x}$. Suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_s$ are i.i.d. on $[0,1]^p$ independently with a density f bounded away from 0 and infinity. Then,*

$$\mathbb{P}\left(diam(L(\mathbf{x}, \mathcal{Z}_s)) \geq \sqrt{d}\left(\frac{s}{2k-1}\right)^{-0.51 \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})} \frac{\pi}{p}}\right) \leq d\left(\frac{s}{2k-1}\right)^{-1/2 \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})} \frac{\pi}{p}}.$$

(41)

Lemma 12 *Let T be a tree satisfying **(P1)**, **(P2)** and $L(\mathbf{x}, \mathcal{Z}_s)$ be the leaf containing $\mathbf{x}$. Then*

$$\mathbb{E}[T(\mathcal{Z}_s)] = \mathbb{E}[\mathbb{E}[\xi_1 \mid \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)]],$$

(42)

and

$$\mathrm{Var}(T(\mathcal{Z}_s)) \leq \sup_{\mathbf{x} \in [0,1]^p} \mathbb{E}[\|\xi_1\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}].$$

(43)

Proof We want to prove

$$\mathbb{E}[T(\mathcal{Z}_s)] = \mathbb{E}[\mathbb{E}[T(\mathcal{Z}_s) \mid L(\mathbf{x}, \mathcal{Z}_s)]] = \mathbb{E}[\mathbb{E}[\xi_1 \mid \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s), L(\mathbf{x}, \mathcal{Z}_s)]].$$

(44)

Let for the following $N_{\mathbf{x}} = \sum_{i=1}^s \mathbb{1}\{\mathbf{X}_i \in L(\mathbf{x}, \mathcal{Z}_s)\}$. Then due to i.i.d. sampling:

$$\mathbb{E}[\mathbb{E}[T(\mathcal{Z}_s) \mid L(\mathbf{x}, \mathcal{Z}_s)]] = \mathbb{E}[\mathbb{E}[\sum_{i=1}^s S_i \xi_i \mid L(\mathbf{x}, \mathcal{Z}_s)]] = s\mathbb{E}[S_1 \xi_1 \mid L(\mathbf{x}, \mathcal{Z}_s)].$$

The last expression can be broken into

$$s\mathbb{E}[\mathbb{E}[S_1 \xi_1 \mid L(\mathbf{x}, \mathcal{Z}_s)]] = s\mathbb{E}\left[\mathbb{E}[\mathbb{E}[S_1 \xi_1 \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s)] \mid L(\mathbf{x}, \mathcal{Z}_s)]\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{s}{N_{\mathbf{x}}}\mathbb{E}[\mathbb{1}\{\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)\}\xi_1 \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s)] \mid L(\mathbf{x}, \mathcal{Z}_s)\right]\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{s}{N_{\mathbf{x}}}\mathbb{E}[\xi_1 \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)]\right.\right.$$

$$\left.\left.\mathbb{P}\left(\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s) \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s)\right) \mid L(\mathbf{x}, \mathcal{Z}_s)\right]\right].$$

(45)

Now, by honesty, given the knowledge that $\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)$, ξ_1 is independent of $N_{\mathbf{x}}$, thus:

$$s\mathbb{E}[S_1 \xi_1 \mid L(\mathbf{x}, \mathcal{Z}_s)]$$

$$= \mathbb{E}\left[\mathbb{E}[\xi_1 \mid L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)]\mathbb{E}\left[\frac{s}{N_{\mathbf{x}}}\mathbb{P}\left(\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s) \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s)\right) \mid L(\mathbf{x}, \mathcal{Z}_s)\right]\right]$$

$$= \mathbb{E}\left[\mathbb{E}[\xi_1 \mid L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)]s\mathbb{E}[S_1 \mid L(\mathbf{x}, \mathcal{Z}_s)]\right].$$

Now it holds by i.i.d. sampling that,

$$s\mathbb{E}\left[S_1 \mid L(\mathbf{x}, \mathcal{Z}_s)\right] = \sum_{i=1}^{s} \mathbb{E}\left[S_i \mid L(\mathbf{x}, \mathcal{Z}_s)\right] = \mathbb{E}\left[\sum_{i=1}^{s} S_i \mid L(\mathbf{x}, \mathcal{Z}_s)\right] = 1,$$

as $\sum_{i=1}^{s} S_i = 1$ by definition.

For (43), we write

$$\mathrm{Var}(T(\mathcal{Z}_s)) \leq \mathbb{E}\left[\left\|\sum_{i=1}^{s} S_i \xi_i\right\|_{\mathcal{H}}^2\right] = \mathbb{E}\left[\frac{1}{N_{\mathbf{x}}^2}\sum_{i=1}^{s} \mathbb{1}\{\mathbf{X}_i \in L(\mathbf{x}, \mathcal{Z}_s)\}\|\xi_i\|_{\mathcal{H}}^2\right] +$$

$$\mathbb{E}\left[\frac{1}{N_{\mathbf{x}}^2}\sum_{i=1}^{s}\sum_{j\neq i} \mathbb{1}\{\mathbf{X}_j \in L(\mathbf{x}, \mathcal{Z}_s)\}\mathbb{1}\{\mathbf{X}_i \in L(\mathbf{x}, \mathcal{Z}_s)\}\langle\xi_i, \xi_j\rangle_{\mathcal{H}}\right].$$

We focus on the second term. For the first, the bound follows by analogous arguments. Similar as before,

$$\mathbb{E}\left[\frac{1}{N_{\mathbf{x}}^2}\sum_{i=1}^{s}\sum_{j\neq i} \mathbb{1}\{\mathbf{X}_j \in L(\mathbf{x}, \mathcal{Z}_s)\}\mathbb{1}\{\mathbf{X}_i \in L(\mathbf{x}, \mathcal{Z}_s)\}\langle\xi_i, \xi_j\rangle_{\mathcal{H}}\right]$$

$$= s(s-1)\mathbb{E}\left[\frac{1}{N_{\mathbf{x}}^2}\mathbb{1}\{\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)\}\mathbb{1}\{\mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s)\}\langle\xi_1, \xi_2\rangle_{\mathcal{H}}\right]$$

$$= s(s-1)\mathbb{E}\left[\frac{1}{N_{\mathbf{x}}^2}\mathbb{E}\left[\mathbb{1}\{\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)\}\mathbb{1}\{\mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s)\}\langle\xi_1, \xi_2\rangle_{\mathcal{H}} \mid N_{\mathbf{x}}\right]\right]. \qquad (46)$$

By the same argument as above

$$\mathbb{P}(\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s) \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s))$$

$$= \frac{1}{s(s-1)}\mathbb{E}\left[\sum_{i=1}^{s}\sum_{j\neq i} \mathbb{1}\{\mathbf{X}_i \in L(\mathbf{x}, \mathcal{Z}_s)\}\mathbb{1}\{\mathbf{X}_j \in L(\mathbf{x}, \mathcal{Z}_s)\} \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s)\right]$$

$$= \frac{1}{s(s-1)}\mathbb{E}\left[N_{\mathbf{x}}(N_{\mathbf{x}} - 1) \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s)\right]$$

$$= \frac{N_{\mathbf{x}}(N_{\mathbf{x}} - 1)}{s(s-1)},$$

and

$$\mathbb{E}\left[\mathbb{1}\{\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)\}\mathbb{1}\{\mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s)\}\langle\xi_1, \xi_2\rangle_{\mathcal{H}} \mid N_{\mathbf{x}}\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s)\}\mathbb{1}\{\mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s)\}\langle\xi_1, \xi_2\rangle_{\mathcal{H}} \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s)\right] \mid N_{\mathbf{x}}\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\langle\xi_1, \xi_2\rangle_{\mathcal{H}} \mid N_{\mathbf{x}}, A, \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s)\right]\right.$$

$$\mathbb{P}(\mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s) \mid N_{\mathbf{x}}, L(\mathbf{x}, \mathcal{Z}_s)) \mid N_{\mathbf{x}}]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\langle\xi_1, \xi_2\rangle_{\mathcal{H}} \mid \mathbf{X}_1 \in L(\mathbf{x}, \mathcal{Z}_s), \mathbf{X}_2 \in L(\mathbf{x}, \mathcal{Z}_s), L(\mathbf{x}, \mathcal{Z}_s)\right] \mid N_{\mathbf{x}}\right]\frac{N_{\mathbf{x}}(N_{\mathbf{x}} - 1)}{s(s-1)}. \qquad (47)$$

Thus combining (46) and (47),

$$
\mathbb{E}\left[\frac{1}{N_{\mathbf{x}}^2}\sum_{i=1}^{s}\sum_{j\neq i}\mathbb{1}\{\mathbf{X}_j \in L(\mathbf{x},\mathcal{Z}_s)\}\mathbb{1}\{\mathbf{X}_i \in L(\mathbf{x},\mathcal{Z}_s)\}\langle\xi_i,\xi_j\rangle_{\mathcal{H}}\right]
$$

$$
= \mathbb{E}\left[\frac{N_{\mathbf{x}}(N_{\mathbf{x}}-1)}{N_{\mathbf{x}}^2}\mathbb{E}\left[\mathbb{E}\left[\langle\xi_1,\xi_2\rangle_{\mathcal{H}} \mid \mathbf{X}_1 \in L(\mathbf{x},\mathcal{Z}_s), \mathbf{X}_2 \in L(\mathbf{x},\mathcal{Z}_s), L(\mathbf{x},\mathcal{Z}_s)\right] \mid N_{\mathbf{x}}\right]\right]
$$

$$
\leq \mathbb{E}\left[\mathbb{E}\left[\langle\xi_1,\xi_2\rangle_{\mathcal{H}} \mid \mathbf{X}_1 \in L(\mathbf{x},\mathcal{Z}_s), \mathbf{X}_2 \in L(\mathbf{x},\mathcal{Z}_s), L(\mathbf{x},\mathcal{Z}_s)\right]\right]
$$

$$
= \mathbb{E}\left[\langle\mathbb{E}[\xi_1 \mid \mathbf{X}_1 \in L(\mathbf{x},\mathcal{Z}_s)], \mathbb{E}[\xi_2 \mid \mathbf{X}_2 \in L(\mathbf{x},\mathcal{Z}_s)]\rangle_{\mathcal{H}}\right],
$$

where in the last step we used independence of $(\xi_1,\mathbb{1}\{\mathbf{X}_1 \in L(\mathbf{x},\mathcal{Z}_s)\}), (\xi_2,\mathbb{1}\{\mathbf{X}_2 \in L(\mathbf{x},\mathcal{Z}_s)\})$ conditionally on $L(\mathbf{x},\mathcal{Z}_s)$ and (C9). Finally,

$$
\mathbb{E}\left[\langle\mathbb{E}[\xi_1 \mid \mathbf{X}_1 \in L(\mathbf{x},\mathcal{Z}_s)], \mathbb{E}[\xi_2 \mid \mathbf{X}_2 \in L(\mathbf{x},\mathcal{Z}_s)]\rangle_{\mathcal{H}}\right] \leq \sup_{\mathbf{x}\in[0,1]^p}\|\mathbb{E}[\xi_1 \mid \mathbf{X}=\mathbf{x}]\|_{\mathcal{H}}^2
$$

$$
\leq \sup_{\mathbf{x}\in[0,1]^p}\mathbb{E}[\|\xi_1\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}],
$$

proving the claim.

∎

Corollary 13 *Under the conditions of Lemma 11, assume*

$$
\mathbf{x} \mapsto \mu(\mathbf{x}) = \mathbb{E}[\xi \mid \mathbf{X}=\mathbf{x}] \in \mathcal{H},
$$

is Lipschitz and that the trees T in the forest satisfy (P2) and (P3). Then

$$
\|\mathbb{E}[\hat{\mu}_n(\mathbf{x})] - \mu(\mathbf{x})\|_{\mathcal{H}} = \mathcal{O}\left(s^{-1/2\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}\right), \tag{48}
$$

and

$$
\|\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x},\mathcal{Z}_s)]\|_{\mathcal{H}} \xrightarrow{p} \|\mathbb{E}[\xi \mid \mathbf{X}=\mathbf{x}]\|_{\mathcal{H}}. \tag{49}
$$

If moreover,

$$
\mathbf{x} \mapsto \mathbb{E}[\|\xi\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}] \in \mathbb{R},
$$

is Lipschitz, then:

$$
\mathbb{E}[\|\xi\|_{\mathcal{H}}^2 \mid \mathbf{X} \in L(\mathbf{x},\mathcal{Z}_s)] \xrightarrow{p} \mathbb{E}[\|\xi\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}]. \tag{50}
$$

Proof By (42), it holds as in Wager and Athey (2018)

$$
\|\mathbb{E}[T(x,\mathbf{Z})] - \mathbb{E}[\xi \mid \mathbf{X}=\mathbf{x}]\|_{\mathcal{H}} = \|\mathbb{E}[\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x},\mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X}=\mathbf{x}]]\|_{\mathcal{H}}.
$$

Let

$$
s_1^* = \sqrt{d}\left(\frac{s}{2k-1}\right)^{-0.51\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}, s_2^* = d\left(\frac{s}{2k-1}\right)^{-0.5\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}
$$

Then it follows from Lemma 11 that

$$\mathbb{P}(L(\mathbf{x}, \mathcal{Z}_s) \geq s_1^*) \leq s_2^*,$$

while by Lipschitz continuity: $\mathbb{E}[\|\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}}] \leq L\mathbb{E}[\text{diam}(L(\mathbf{x}, \mathcal{Z}_s))]$, where L is the Lipschitz constant. Thus,

$$\begin{aligned}
&\|\mathbb{E}[\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]]\|_{\mathcal{H}} \\
&\leq \|\mathbb{E}[(\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]) \mathbb{1}\{\text{diam}(L(\mathbf{x}, \mathcal{Z}_s)) \geq s_1^*\}]\|_{\mathcal{H}} \\
&\quad + \|\mathbb{E}[(\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]) \mathbb{1}\{\text{diam}(L(\mathbf{x}, \mathcal{Z}_s)) < s_1^*\}]\|_{\mathcal{H}} \\
&\leq \mathbb{E}[\|\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}} \mathbb{1}\{\text{diam}(L(\mathbf{x}, \mathcal{Z}_s)) \geq s_1^*\}] \\
&\quad + \mathbb{E}[\|\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}} \mathbb{1}\{\text{diam}(L(\mathbf{x}, \mathcal{Z}_s)) < s_1^*\}] \\
&\leq \left(\sup_{\mathbf{x}_1, \mathbf{x}_2 \in [0,1]^p} \|\mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}_1] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}_2]\|_{\mathcal{H}} \right) \mathbb{P}(\text{diam}(L(\mathbf{x}, \mathcal{Z}_s)) \geq s_1^*) + L s_1^* \\
&\leq \left(\sup_{\mathbf{x}_1, \mathbf{x}_2 \in [0,1]^p} \|\mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}_1] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}_2]\|_{\mathcal{H}} \right) s_2^* + L s_1^* \\
&\precsim \left(\sup_{\mathbf{x}_1, \mathbf{x}_2 \in [0,1]^p} \|\mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}_1] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}_2]\|_{\mathcal{H}} \right) s_2^*,
\end{aligned}$$

since $s_1^*/s_2^* \to 0$. Due to the Lipschitz condition

$$\sup_{\mathbf{x}_1, \mathbf{x}_2 \in [0,1]^p} \|\mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}_1] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}_2]\|_{\mathcal{H}} \leq L \sup_{\mathbf{x}_1, \mathbf{x}_2 \in [0,1]^p} \|\mathbf{x}_1 - \mathbf{x}_2\|_{\mathbb{R}^d} = \mathcal{O}(1).$$

Finally by the reverse triangle inequality

$$\begin{aligned}
\left| \|\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)]\|_{\mathcal{H}} - \|\mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}} \right| &\leq \|\mathbb{E}[\xi \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] - \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}} \\
&\leq L \text{diam}(L(\mathbf{x}, \mathcal{Z}_s)),
\end{aligned}$$

and if $\mathbf{x} \mapsto \mathbb{E}[\|\xi\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}]$ is Lipschitz as well, also

$$|\mathbb{E}[\|\xi\|_{\mathcal{H}}^2 \mid \mathbf{X} \in L(\mathbf{x}, \mathcal{Z}_s)] - \mathbb{E}[\|\xi\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}]| \leq C \text{diam}(L(\mathbf{x}, \mathcal{Z}_s)).$$

Since $\text{diam}(L(\mathbf{x}, \mathcal{Z}_s)) \xrightarrow{p} 0$, as $s \to \infty$, (49), respectively (50) hold true.

As the expectation of the forest is the same as that of one tree:

$$\mathbb{E}[\hat{\mu}_n(\mathbf{x})] = \mathbb{E}\left[T(\mathbf{x}, \varepsilon; \mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}})\right],$$

the result follows. ∎

This leads us to the proof of of Theorem 2 in the main text.

Theorem 2 *Suppose that our forest construction satisfies properties (P1)–(P5). Assume additionally that k is a bounded and continuous kernel and that we have a random design with $\mathbf{X}_1, \ldots, \mathbf{X}_n$ independent and identically distributed on $[0,1]^p$ with a density bounded*

away from 0 *and infinity. If the subsample size* s_n *is of order* n^β *for some* $0 < \beta < 1$, *the mapping*

$$\mathbf{x} \mapsto \mu(\mathbf{x}) = \mathbb{E}[\mu(\delta_{\mathbf{Y}}) \mid \mathbf{X} = \mathbf{x}] \in \mathcal{H},$$

is Lipschitz and $\sup_{\mathbf{x} \in [0,1]^p} \mathbb{E}[\|\mu(\delta_{\mathbf{Y}})\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}] < \infty$, *we obtain the consistency w.r.t. the RKHS norm:*

$$\|\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\|_{\mathcal{H}} = \mathcal{O}_p\left(n^{-\gamma}\right), \tag{18}$$

for $\gamma = \frac{1}{2}\min\left(1 - \beta, \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})} \frac{\pi}{p} \cdot \beta\right)$.

Proof We first note that $\sup_{\mathbf{x} \in [0,1]^p} \mathbb{E}[\|\mu(\delta_{\mathbf{Y}})\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}] < \infty$ together with (43) implies $\mathrm{Var}(T) < \infty$. Thus, from Markov's inequality and Lemma 10,

$$\mathbb{P}\left(n^\gamma \|\hat{\mu}_n(\mathbf{x}) - \mathbb{E}[\hat{\mu}_n(\mathbf{x})]\|_{\mathcal{H}} > \varepsilon\right) \leq \frac{n^{2\gamma}}{\varepsilon^2}(s/n + s^2/n^2)\mathrm{Var}(T) = \frac{1}{\varepsilon^2}\mathcal{O}(n^{2\gamma + \beta - 1}).$$

Thus

$$n^\gamma \|\hat{\mu}_n(\mathbf{x}) - \mathbb{E}[\hat{\mu}_n(\mathbf{x})]\|_{\mathcal{H}} = \mathcal{O}_p(1),$$

for $\gamma \leq (1-\beta)/2$. In particular, it goes to zero for any $\varepsilon > 0$, if $\gamma < (1-\beta)/2$. Since,

$$n^\gamma \|\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\|_{\mathcal{H}} \leq n^\gamma \|\hat{\mu}_n(\mathbf{x}) - \mathbb{E}[\hat{\mu}(\mathbf{x})]\| + n^\gamma \|\mathbb{E}[\hat{\mu}(\mathbf{x})] - \mu(\mathbf{x})\|_{\mathcal{H}},$$

the result follows as soon as the second expression goes to zero. Now from Theorem 13, with $C_\alpha = \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}$,

$$\|\mathbb{E}[\hat{\mu}_n(\mathbf{x})] - \mu(\mathbf{x})\|_{\mathcal{H}} = \mathcal{O}\left(s_n^{-1/2C_\alpha \frac{\pi}{p}}\right) = \mathcal{O}\left(n^{-1/2\beta C_\alpha \frac{\pi}{p}}\right).$$

This goes to zero provided that,

$$1/2\beta C_\alpha \frac{\pi}{p} > \gamma.$$

∎

To prove Corollaries 4 and 5, we first need another auxiliary result:

Lemma 14 *Let* (S, d) *be a separable metric space and* $\mathbf{X}_n : (\Omega, \mathcal{A}) \to (S, \mathcal{B}(S))$, $n \in \mathbb{N}$ *and* $\mathbf{X} : (\Omega, \mathcal{A}) \to (S, \mathcal{B}(S))$ *be measurable. Then* $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ *if and only if for every subsequence* $n(k)$ *there exists a further subsequence* $n(k(l))$ *such that*

$$\mathbf{X}_{n(k(l))} \to \mathbf{X} \quad a.s. \tag{51}$$

Proof If $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$, then so does any subsequence $\mathbf{X}_{n(k)}$. By well-known results, see e.g. van der Vaart (1998, Chapter 2), this implies that there exists a further subsequence $\mathbf{X}_{n(k(l))}$ such that a.s. convergence holds.

As is well known, there exists a metric ρ on $\mathcal{P}(S)$ such that $\rho(\mathbf{X}_n, \mathbf{X}) \to 0$ iff $\mathbf{X}_n \xrightarrow{p} X$, see e.g., Dudley (2002, Chapter 11). Now assume that for any subsequence we can find

a further subsequence such that (51) holds, but the overall sequence does not converge in probability. Then we can build a subsequence such that for some $\varepsilon > 0$,

$$\rho(\mathbf{X}_{n(k)}, \mathbf{X}) \geq \varepsilon$$

for all elements of that subsequence. Thus any further subsequence will also not convergence in probability and consequently cannot converge a.s. This proves the claim. ∎

We note that the set A with $P(A) = 1$ on which (51) holds is allowed to depend on the subsequence. Corollary 4 and 5 are finally proven jointly in the following Corollary. The proof is motivated by the tools used in Berti et al. (2006).

Corollary 15 *Assume that one of the following two sets of conditions holds:*

(a) The kernel k is bounded, (jointly) continuous and has

$$\int \int k(\mathbf{x}, \mathbf{y}) d\mathcal{P}(\mathbf{x}) d\mathcal{P}(\mathbf{y}) > 0 \quad \forall \mathcal{P} \in \mathcal{M}_b(\mathbb{R}^d) \setminus \{0\}. \tag{52}$$

Moreover, $\mathbf{y} \mapsto k(\mathbf{y}_0, \mathbf{y})$ is vanishing at infinity, for all $\mathbf{y}_0 \in \mathbb{R}^d$.

(b) The kernel k is bounded, shift-invariant, (jointly) continuous and ν in the Bochner representation in (11) is supported on all of $\mathbb{R}^d$. Moreover, $\mathbf{Y}$ takes its values almost surely in a closed and bounded subset of $\mathbb{R}^d$.

Then, under the conditions of Theorem 2, we have for any bounded and continuous function $f : \mathbb{R}^d \to \mathbb{R}$ that DRF consistently estimates the target $\tau(\mathbf{x}) = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}]$ for any $\mathbf{x} \in [0, 1]^p$:

$$\sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) f(\mathbf{y}_i) \xrightarrow{p} \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}].$$

Moreover,

$$\hat{F}_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}}(\mathbf{t}) \xrightarrow{p} F_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}}(\mathbf{t})$$
$$\hat{F}_{Y_i \mid \mathbf{X} = \mathbf{x}}^{-1}(t) \xrightarrow{p} F_{Y_i \mid \mathbf{X} = \mathbf{x}}^{-1}(t),$$

for all points of continuity $\mathbf{t} \in \mathbb{R}^d$ and $t \in \mathbb{R}$ of $F_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}}(\cdot)$ and $F_{Y_i \mid \mathbf{X} = \mathbf{x}}^{-1}(\cdot)$ respectively.

Proof As shown in (Sriperumbudur, 2016, Theorem 3.2), (a) implies that k metrizes weak convergence. Similarly, from Theorem 9 in Sriperumbudur et al. (2010), it follows that k is characteristic on the compact subspace of $\mathbb{R}^d$ in which Y takes its value almost surely. Thus, ignoring the Null set, Theorem 23 in Sriperumbudur et al. (2010) implies that k metrizes the weak convergence in this case as well. Thus in both cases $\|\hat{\mu}_{n(k(l))}(\mathbf{x}) - \mu(\mathbf{x})\|_{\mathcal{H}} \to 0$ implies weak convergence of $\hat{\mu}_{n(k(l))}(\mathbf{x})$ to $\mu(\mathbf{x})$.

From Theorem 14, for any subsequence, we can choose a further subsequence, such that

$$\|\hat{\mu}_{n(k(l))}(\mathbf{x}) - \mu(\mathbf{x})\|_{\mathcal{H}} \to 0 \text{ , a.s.}$$

and since it is assumed that $\| \cdot \|_{\mathcal{H}}$ metrizes weak convergence, $\hat{\mu}_{n(k(l))}(\mathbf{x})$ converges weakly to $\mu(\mathbf{x})$ on a set A, depending on the subsequence, with $\mathbb{P}(A) = 1$. Let $C_b(\mathbb{R}^d)$ denote the space of all bounded continuous functions on $\mathbb{R}^d$. By the Portmanteau theorem (see e.g. Dudley (2002)), this implies that on A

(I) $\int f d\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}) \to \int f d\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ for all $f \in C_b(\mathbb{R}^d)$

(II) $\hat{F}_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}}(\mathbf{t}) \to F_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}}(\mathbf{t})$ for all continuity points $\mathbf{t} \in \mathbb{R}^d$ of $F_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}}(\cdot)$,

where we omitted the dependence on the subsequence. But, since the subsequence $n(k)$ was arbitrary, this immediately implies

(I') $\int f d\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}) \xrightarrow{p} \int f d\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ for all $f \in C_b(\mathbb{R}^d)$

(II') $\hat{F}_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}}(t) \xrightarrow{p} F_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}}(t)$ for all continuity points t of $F_{\mathbf{Y} \mid \mathbf{X} = \mathbf{x}}(\cdot)$,

for the overall sequence.

On the other hand, (II) implies that on A, for the given subsequence $n(k(l))$, $\hat{F}_{Y_i \mid \mathbf{X} = \mathbf{x}}(t) \to F_{Y_i \mid \mathbf{X} = \mathbf{x}}(t)$ for all $t \in \mathbb{R}$ at which $F_{Y_i \mid \mathbf{X} = \mathbf{x}}(\cdot)$ is continuous. Using for each $\omega \in A$ the arguments in van der Vaart (1998, Chapter 21), this implies that

$$\hat{F}^{-1}_{Y_i \mid \mathbf{X} = \mathbf{x}}(t) \to F^{-1}_{Y_i \mid \mathbf{X} = \mathbf{x}}(t), \text{ for all continuity points } t \text{ of } F^{-1}_{Y_i \mid \mathbf{X} = \mathbf{x}}(t)$$

on A for the given subsequence. Again, as the subsequence $n(k)$ was arbitrary, this implies the result. ∎

Appendix C. Simulation Details

In this section we describe in detail all our simulations shown in the main paper, together with the data used in the analysis. The data sets are available in the R-package `drf` as well.

C.1 Air quality data

Data. This data is obtained from the website of the Environmental Protection Agency website (`https://aqs.epa.gov/aqsweb/airdata/download_files.html`). We have daily measurements for 5 years of data (2015-2019) for 6 'criteria' pollutants that form the Air Quality Index (AQI):

- O_3 - ground ozone (8 hours' average, expressed in pieces per million (ppm))

- SO_2 - sulfur dioxide (1 hour average, expressed in pieces per billion (ppb))

- CO - carbon monoxide (8 hours' average, expressed in pieces per million (ppm))

- NO_2 - nitrogen dioxide (1 hour average, expressed in pieces per billion (ppb))

- PM2.5 - fine particulate matter smaller than 2.5 micrometers (24 hours' average, expressed in $\nu g/m^3$)

- PM10 - large particulate matter, smaller than 10 micrometers (24 hours' average, expressed in $\mu g/m^3$)

For the above quantities, we have the maximal and mean value within the same day. In our analysis we have used only the maximal intraday values.

The pollutants are measured at different measurement sites. For each site we have information about

- site address (street, city, county, state, zip code)

- site coordinates (longitude and latitude)

- site elevation

- location setting (rural, urban, suburban)

- how the land is used within a 1/4 mile radius (agricultural, forest, desert, industrial, commercial, residential, blighted area, military reservation, mobile)

- date when the measurement site was put in operation

- date when the measurement site was decommissioned (NA if the site is still operational)

We have information about $19'739$ sites, much more than the number of $2'419$ sites from which we have measurements in years 2015-2019, since many sites were only operating in the past and are decommissioned.

In total there is $5'305'859$ pollutant measurements. Many pollutants are measured at the same site, but it is important to note that not every site measures every pollutant, so there is a lot of 'missing' measurements. It can also occur that there are several measuring devices for the same pollutant at the same site, in which case we just average the measurements across the devices and do not report those measurements separately.

Analysis. Since we have a lot of missing data, we use only the data points (identified by the measurement date and the measurement site) for which we have measurements of all the pollutants chosen as the responses. For that reason we also do not train DRF with all 6 pollutants as the responses, but only those that we are interested in, since only 64 sites measure all pollutants. For computational feasibility, we only use $50'000$ of the available measurements for the training step. We also omit the states Alaska and Hawaii and the US territories for plotting purposes.

To obtain the results displayed in Figure 2, we train the DRF with the measurements (intraday maximum) of the two pollutants PM2.5 and NO_2 as the responses, and the site longitude, latitude, elevation, land use and location settings as the predictors. We manually choose two decommissioned measurement sites (for which we have no measurements in years 2015-2019) as the test points. For each test point we obtain the weights to all training measurements. We further combine the weights for all measurements corresponding to the same site, which is represented by the symbol size in the top row. The bottom row shows the estimated distribution of the response, where the transparency (alpha) each training point corresponds to the assigned weight. We also add some estimated contours.

For all plots in Figures 2 and 5, we train the single DRF with the same set of predictor variables and take the three pollutants O_3, SO_2 and PM2.5 as the responses. In this way we still have training data from many different sites (see the above discussion on missing data) and moreover, those are the 3 pollutants that most likely cross the threshold for the "Good" AQI category set by the EPA. Carbon monoxide (CO), for example, almost never crosses this threshold.

In left plot of Figure 6, we compare the estimated CDF value with the standard classification forest which has the indicator $\mathbb{1}(O_3 < 0.055\text{ppm}, SO_2 < 36\text{ppb}, \text{PM2.5} < 12.1\mu g/m^3)$ as the univariate response. In the right plot, we obtain the estimated CDF by fitting for each threshold a separate classification forest with an indicator $\mathbb{1}(O_3 \leq \text{threshold})$. We pick a test point such that the classification performs bad, just to illustrate that its estimated CDF need not be monotone, which cannot happen with DRF. In most of the cases, the estimated CDFs are very similar, as can also be seen from the left plot in Figure 6.

C.2 Benchmark Analysis

In this part we compare the resulting distributional estimates of DRF with several benchmark methods on a number of data sets. Because our target of estimation is now the whole conditional distribution one needs to use a distributional loss, and there appears to be no well-established choice in the literature. Furthermore, for any test point $\mathbf{x}_i$ we only have one observation $\mathbf{y}_i$ from $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$, which makes performance evaluation of our estimator $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ hard. We thus use the following performance measure:

- **(NLPD loss)** For a fixed conditional distribution estimator $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}_i)$, we sample a set of $m = 500$ observations from which we estimate the conditional density via a Gaussian kernel estimator, using the L_2 loss with scale components and the median heuristic for the choice of the bandwidth parameter. We then evaluate the negative log-likelihood of the test observation $\mathbf{y}_i$ implied by the kernel estimate of the distribution and average over the test set (consisting of multiple pairs $(\mathbf{x}_i, \mathbf{y}_i)$). To reduce the dependence of these results on single large values of the log-likelihood, we use an $0.05-$trimmed mean to average the losses over the training set.

This loss definition provides a fair way to compare the ability to estimate the conditional distribution since most of the candidate methods only allow for sampling from the estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}_i)$.

C.2.1 Competing methods

We compare DRF that uses the MMD splitting criterion with many existing methods that can be used for estimation of the conditional distribution.

- **Nearest Neighbor (k-NN)**: The standard k-nearest neighbors algorithm with the Euclidean metric. An estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ at a test point $\mathbf{x}$ is defined by a uniform distribution over the k nearest observations in the training set. k is chosen to be the square root of the training set size.

- **Gaussian kernel (kernel)**: The estimate of the conditional distribution $\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ at a test point $\mathbf{x}$ is obtained by assigning to each training observation $(\mathbf{x}_i, \mathbf{y}_i)$ the weight

proportional to the Gaussian kernel $k(\mathbf{x}, \mathbf{x}_i)$, analogously to usual kernel estimation methods. Median heuristic is used for bandwidth selection.

- **Homogeneous distribution models**: This method makes the homogeneity assumption that the residuals have constant distribution and only the conditional mean changes. The estimate of the conditional distribution $\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ is obtained by first fitting a regression method of choice, computing the residuals, assigning the same weight to every residual and then adding those residuals to the predicted mean. We chose three different methods for the mean estimation:

  1. **Random Forests (RF)**, a classical univariate regression forest is fitted independently for each response component;

  2. **Extreme Gradient Boosting (XGBoost)**, a tree gradient boosting model (as described in Chen and Guestrin (2016)) is fitted independently for each response;

  3. **Deep Neural Network (DNN)**, a single deep neural network is fitted to predict the conditional mean of each response.

- **Conditional Generative Adversarial Neural Network (CGAN)**: The estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ is obtained through sampling from the discriminator with conditional feature $\mathbf{x}$. The implementation of the CGAN is taken from Aggarwal et al. (2019). The architecture of the neural networks was taken to be the best one for the considered data sets among a set of candidates.

- **Conditional Variational Auto-Encoder (CVAE)**: The estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ is obtained through sampling from the decoder of the CVAE with conditional feature $\mathbf{x}$. The implementation of the CVAE follows the one in Sohn et al. (2015). The architecture of the neural networks was taken to be the best one for the considered data sets among a set of candidates.

- **Masked Autoregressive Flow (MAF)**: The estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ is obtained through sampling from the normalizing flow model with conditional feature $\mathbf{X}$. The implementation of the model follows the one presented in Papamakarios et al. (2017). The number of layers is chosen to be the best value from a set $\{5, 10\}$ for the considered data set.

- **Conditional Mean Embedding (CME)**: The CME is calculated as in (16) with the weights given as in e.g., Song et al. (2009); Muandet et al. (2017); Park and Muandet (2020). We choose both kernels to be Gaussian kernels with $\sigma = 0.01$, as in Park and Muandet (2020) and set $\lambda = 0.01$. The estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ is obtained through sampling from the obtained weights, renormalized such that they lie in $[0, 1]$ and sum up to one.

C.2.2 Benchmark data sets

Many benchmark data sets used come from the multiple target regression literature, where only the conditional means of the multivariate response is considered. We have used the data sets: **jura**, **slump**, **wq**, **enb**, **atp1d**, **atp7d**, **scpf**, **sf1** and **sf2** collected in the

`Mulan` (Tsoumakas et al., 2011) library. Description about the dimensionality of the data sets, together with the descriptions of the outcomes and the regressors can be found in Tsoumakas et al. (2011) with links to the relevant papers introducing these data sets. In each data set categorical variables have been represented by the one-hot dummy encoding, the observations with missing data were removed together with constant regressors.

We additionally added 5 data sets obtained from the data sets used in the main paper:

- **copula**: Simulated Gaussian copula example where the response Y is bivariate and whose marginal distribution is $N(0, 1)$, but the correlation between Y_1 and Y_2 depends on X_1.

- **birth1**: This data set is created from the CDC natality data and contains many covariates as predictors and the pregnancy length and birthweights as the responses.

- **birth2**: This data set is similar as the above one, but we take pregnancy length as the predictor and add 3 more measures of baby's health as the response: APGAR score measured 5 minutes after birth and indicators whether there were any abnormal conditions and congenital anomalies.

- **wage**: This data set is created from the 2018 American Community Survey. We take the logarithmic hourly wage and gender as the response, as it was done in the fairness example in the main paper.

- **air**: This data set is obtained from the EPA air quality data. All six pollutants were taken as the response and we add both the information about the measuring site (location, which setting it is in, etc.), as well as the temporal information when the measurement has taken place (month, day of the week).

C.3 Births data

Data. This data set is obtained from the CDC Vital Statistics Data Online Portal (`https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm`) and contains the information about the ≈ 3.8 million births in 2018. However, as we do not need this many data points, we subsample $300'000$ of them. Even though the original data contains a lot of variables, we have taken only the following variables from the source data:

- mother's age, height, weight before the pregnancy and BMI before pregnancy

- mother's race (black, white, asian, NHOPI, AIAN or mixed), marital status (married or unmarried) and the level of education (in total 8 levels)

- father's age, race and education level

- month and year of birth

- plurality of the birth (how many babies were born at once)

- whether and when the prenatal care started
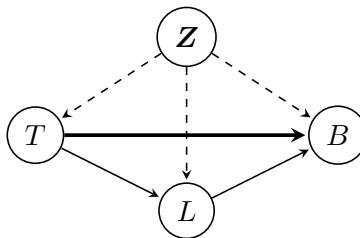
- length of the pregnancy

- delivery method (vaginal or C-section)

- birth order - the total number of babies born by the same mother (including the current one)

- birth interval - number of months passed since last birth (NA if this is the first child)

- number of cigarettes smoked per day on average during the pregnancy

- birthweight (in grams) and gender of the baby

- APGAR score (taken after 5min and 10min)

- indicators whether baby had any abnormal condition or some congenital anomalies

Analysis. After removing the data points with any missing entries and taking only the data points where the race of both parents is either black, white or Asian (for nicer plotting), we are left with $183'881$ data points. We use randomly chosen $100'000$ data points for training the DRF. We take the birthweight and the pregnancy length as the bivariate response and for the predictors we take: mother's age, race, education, marital status, height, BMI; father's age, race and education level; birth plurality, birth order, delivery method, baby's gender, number of cigarettes and indicator whether prenatal care took place.

For arbitrary test points from the data we can get the estimated weights by the fitted DRF, thus estimating the joint distribution of birthweight and pregnancy length conditional on all other variables mentioned above. Two such distributions are shown in Figure 10. In addition we use the weights to fit a parametric model for the mean and 0.1 and 0.9 quantiles. This is done as follows:

- We slightly upweight the data points where the pregnancy length is significantly above or below the usual range. This is to avoid the bulk of the data points to dominate the fit obtained for very long or short pregnancies.

- We apply the transformation $f(\cdot) = \log(\log(\cdot))$ on both the pregnancy length and the birthweight since then the scatterplots look much nicer.

- We estimate the mean with smoothing splines with a small manually chosen number of degrees of freedom.

- The fitted mean is subtracted from the response (birthweight). The residuals seem well behaved with maybe slight, seemingly linear trend in standard deviation.

- We fit the 0.1 and 0.9 quantiles as the best linear functions that minimize the sum of quantile losses, by using the `quantreg` package (Koenker et al., 2012).

- The data is transformed back on the original scale by using the function $f^{-1}(\cdot) = \exp(\exp(\cdot))$.

For the right plot in Figure 10, we have the following causal graph, as mentioned in the main paper:

We want to determine the direct effect (indicated in bold) of the twin pregnancy T on the birthweight B that is due to sharing of resources by the babies (space, food etc.) and is not due to the fact that twin pregnancy causes shorter pregnancy length L, which in turn causes the smaller birthweight. Another big issue is that we have confounding factors $\mathbf{Z}$ which can directly affect B, L and T. For example, the number of twin pregnancies significantly depends on the parents' race, but so do the pregnancy length and the birthweight, e.g. black people have more twins, shorter pregnancies and smaller babies. We take all other variables as the potential confounders $\mathbf{Z}$ and adjust for all of them (mother's age, race, education, marital status, height, BMI; father's age, race and education level; birth plurality, birth order, baby's gender, number of cigarettes and indicator whether prenatal care took place). In order to do it, we fit the same DRF as before, where $\mathbf{Z}$ and T are the predictors and (B, L) is the bivariate response for which we can fit the parametric model described above. We compute then the interventional distribution $\mathbb{P}(B \mid do(T = t, L = l))$ for all values of t and l, by using the do-calculus to adjust the confounding $\mathbf{Z}$ via the backdoor criterion (Pearl, 2009), where we also use the obtained parametric regression fit. In this way we can generalize the fit well, which is important when doing the do-calculus, since we are interested in some hypothetical combinations of covariates which might not occur frequently in the observed data, such as very long twin pregnancies.

C.4 Wage data

Data. The PUMS (Public Use Microdata Area) data from the 2018 1-Year American Community Survey is obtained from the US Census Bureau API (`https://www.census.gov/content/dam/Census/data/developers/api-user-guide/api-guide.pdf`). The survey is sent to ≈ 3.5 million people annually and aims to give more up to date data than the official census that is carried out every decade. The 2018 data set has $3'214'539$ anonymized data points for the 51 states and District of Columbia. Even though the original survey contains many questions, we have retrieved only the subset of variables that might be relevant for the salaries:

- person's gender, age, race (AIAN, black, white, asian, mix, NHOPI, other), indicator of hispanic origin, state of residence, US citizenship indicator (5 ordered levels), indicator whether the person is foreign-born

- person's marital status, number of own children in the same household and the number of family members in the same household

- person's education level (24 ordered levels) and level of English knowledge (5 ordered levels)

- person's employment status (employed, not at work, not in workforce, unemployed)

- for employed people we have annual salary earnings, number of weeks worked in a year and average number of hours worked per week

- for employed people we have employer type (government, non-profit company, for-profit company, self-employed), occupation (530 levels), industry where the person works (271 levels) and the geographical unit where the person works (59 levels)

- statistical weight determined by the US Census Bureau which aims to correct sampling bias

For our purposes, since we want to analyze the unfairness of the gender pay gap, we consider only employed people that are at least 17 years of age, have worked full-time (at least 48 weeks in a year) and have worked at least 16 hours a week on average. We also omit the self-employed persons, since they often report zero annual salary and the pay gap there, if exists, cannot be called unfair as the salary is not determined by any employer. Since there are no missing data which would need to be omitted, we finally end up with $1'071'866$ data points.

Analysis. We scale the salary with the amount of time spent working (determined from the number of weeks worked and average hours worked per week) to compute the logarithm of the hourly wages. The scaling with the time spent working is necessary, since full-time employed men spend on average 11% more time working than women. The logarithmic transformation is used since the salaries are very skewed (positively) and logarithmic wages show nice behavior.

We also reduce the large number of levels of some of the categorical variables: for the occupation we use the group of 530 jobs into 20 categories provided in the SOC system (`https://www.bls.gov/soc/`); for the industry information we group the 271 possibilities in 23 categories as is done in the NIACS classification (`https://www.bls.gov/bls/naics.htm`); for the work place we group the 59 US states and foreign territories into 9 economic regions (including the "abroad" category), as determined by the Bureau of the Economic Analysis (`https://apps.bea.gov/regional/docs/regions.cfm`).

We want to investigate how the logarithmic hourly wage W is affected by the gender G, depending on the other factors $\mathbf{Z}$: age, race, hispanic origin, citizenship, being foreign-born, marital status, family size, number of children, education level, knowledge of English, occupation, industry type and place of work. To do this, we train DRF with bivariate response (W, G) and predictors $\mathbf{Z}$ on a subsample of $300'000$ data points. With it we can answer the following: For fixed values of the covariates $\mathbf{Z} = \mathbf{z}$, what are the distributions of salaries of men and women. In addition, we can determine the "propensities", i.e. the proportion of men and women corresponding to $\mathbf{Z} = \mathbf{z}$. This information is displayed in the top row of Figure 11 for a combination of covariates corresponding to some person in the left-out data. It illustrates how the distribution of salaries and their relationship can vary with different covariates $\mathbf{Z}$.

We do not only want to determine how different covariates $\mathbf{Z}$ affect the salary distribution, but we want to quantify the overall fairness of the pay, after appropriate adjustments. In Figure 11, we can see that the observed salaries of men and women differ noticeably, and this difference in the logarithmic wages means that an average woman has 17% smaller salary than an average men. However, the question is how much of this difference is "fair". For example, the effect of the gender on the salary can be mediated through some variables such as, for example, the occupation, workplace or the level of education and we are only interested in the direct effect. This is illustrated in the following causal graph:
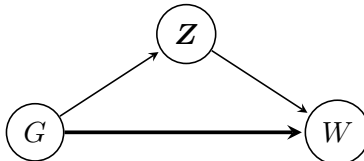
If we assume that people have the freedom to choose such variables themselves, the pay gap which arises from such different choices for men and women is fair and those variables are resolving variables (Kilbertus et al., 2017). Another way that the pay gap can be explained is that some of the variables are not statistically independent of the gender in the population of full-time employed people (e.g. the race or the age), but they themselves have an effect on the salary.

In order to address those issues, we compute the distribution of the nested counterfactual $W(\text{male}, \mathbf{Z}(\text{female}))$, corresponding to the wages of a person that has characteristics $\mathbf{Z}$ as a woman, but which was treated as a man for obtaining the salary. Such distribution can be computed from the DRF, as described in the main paper: we randomly draw a female person and for its characteristics $\mathbf{z}$ we obtain the conditional distribution of wages of men with those characteristics $\mathbb{P}(W \mid G = \text{male}, \mathbf{Z} = \mathbf{z})$ via the weights. Those distributions are averaged over random draw of $1'000$ women (that were not used in the training step of the DRF). In case that the difference in salary is fair, the distribution of the counterfactual salary $W(\text{male}, \mathbf{Z}(\text{female}))$ should be exactly the same as the observed distribution of women's wages. However, we can see that this is not the case and that the median salaries of the two distributions differ by 11%. Even though this is smaller than the 17% we obtain by comparing only the observational distributions, it still shows that women are paid less compared to men.

Appendix D. Additional Synthetic Examples

D.1 Univariate distributional regression.

The univariate response (case $d = 1$) is by far the most studied case in the regression literature. However, at the level of the whole conditional distribution and compared to the multivariate case, the range of practically interesting targets $\tau(\mathbf{x})$ is quite reduced, e.g. conditional mean of some functional $\mathbb{E}(f(Y) \mid \mathbf{X})$ or conditional quantiles $Q_\alpha(Y \mid \mathbf{X})$. In Figure 4, we have compared the performance of DRF (which uses the MMD splitting criterion) with 3 different tree-based univariate methods that can estimate the conditional quantiles in the univariate case:

- **QRF**: the quantile regression forest introduced in Meinshausen (2006), which is equivalent to DRF$_{\text{CART}}$ in the univariate case and uses the standard forest construction (Breiman, 2001) to get the weights.

- **GRF**: the quantile forest proposed in Athey et al. (2019) based on the generalized random forest algorithm.

- **TRF**: the transformation forest, a model-based recursive partitioning approach, introduced in Hothorn and Zeileis (2021).

Additionally to the visual inspection of the performance given in Figure 4, we present here a formal performance comparison for the three simulation scenarios also described in the main paper. The first two scenarios correspond exactly to the examples given in Athey et al. (2019) for the quantile version of the GRF, which serve to illustrate its advantage compared to the conventional quantile regression forest (QRF) (Meinshausen, 2006). Scenario 3, in addition, aims at assessing the ability to detect a change of distribution that does not relate to a change in the first two moments.

method	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
DRF	**0.180**	**0.353**	**0.402**	**0.349**	**0.177**	**0.267**	**0.518**	**0.589**	**0.514**	**0.264**	0.140	**0.298**	0.371	**0.351**	**0.198**
QRF	0.182	0.357	0.482	0.351	0.179	0.285	0.526	0.592	0.521	0.281	0.144	0.299	0.376	0.357	0.204
GRF	0.183	0.359	0.409	0.354	0.180	0.278	0.522	0.590	0.517	0.274	**0.139**	0.299	**0.371**	**0.351**	0.200
TRF	0.183	0.358	0.408	0.353	0.180	0.272	0.519	0.590	0.516	0.268	0.145	0.300	0.373	**0.351**	0.200
5-NN	0.232	0.402	0.452	0.404	0.239	0.354	0.587	0.657	0.584	0.340	0.187	0.348	.424	0.406	0.260
20-NN	0.192	0.368	0.418	0.366	0.192	0.290	0.535	0.606	0.533	0.283	0.146	0.310	0.382	0.365	0.211
40-NN	0.187	0.364	0.413	0.360	0.185	0.283	0.528	0.596	0.522	0.273	0.141	0.303	0.376	0.357	0.204

Table 2: Average quantile losses for scenarios 1 (left), 2 (middle), 3 (right) over the repeated out-of-sample validations.

The performance of each method is evaluated as follows: We consider the quantile (pinball) loss for the resulting quantile estimates provided by each candidate method for the different percentiles $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The losses are presented and computed based on repeated (10 times) out-of-sample validation (with a $70 - 30\%$ ratio between the training and testing sets sizes). The results are presented respectively for each scenario in Table 2. We additionally include the estimates obtained by k-nearest neighbor algorithm for several different values of k.

Furthermore, Table 3 shows non-inferiority of DRF compared to the standard Random Forest for the classical task of estimating the conditional mean. We observe that DRF has a good relative performance that makes it on par with existing algorithms, some of which specially designed for the problem of estimating conditional quantiles. Furthermore, it seems that the MMD splitting criterion improves the CART criterion for distributional regression in a general heterogeneous case (see e.g. scenarios 2 and 3), since the CART criterion is suitable only for detecting the change in the conditional mean, unlike MMD.

Table 3: Average mean squared errors (MSE) for the three scenarios described above over 10 repeated out-of-sample validations for estimating the conditional mean.

method	SC1	SC2	SC3
RF	1.0545	2.4940	0.9624
DRF	**1.0412**	**2.4561**	**0.9340**

Dependence of the estimated quantiles on X_1 for each method (except the k-nearest neighbors) is displayed in the main paper in Figure 4. In addition, the estimates of 2-Wasserstein distance to the true conditional distribution, quantifying the difference in the estimated CDFs, are shown in Figure 12.
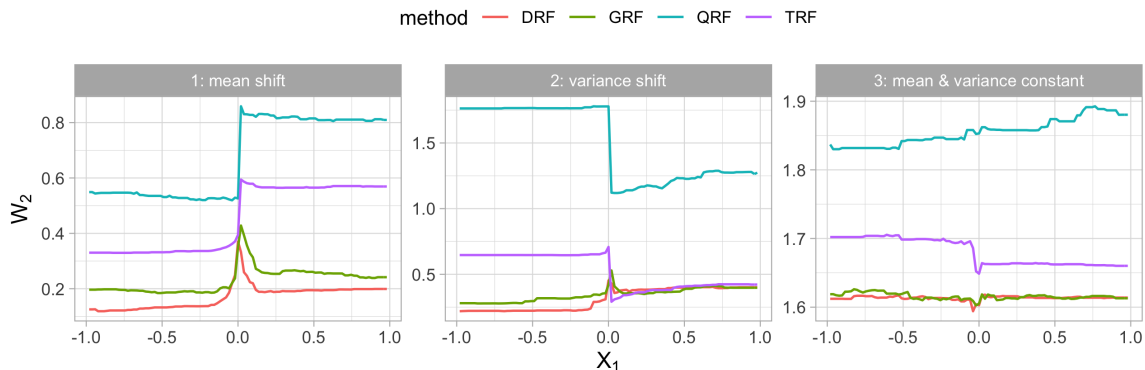

Figure 12: Scatter plot of discrete estimates of the 2-Wasserstein distance between the estimated and true conditional distribution against X_1 for a grid of test points of the form $(x_1, 0, \ldots, 0)$. The 2-Wasserstein distance is estimated over a grid of 100 quantiles with levels equally spaced on $[0, 1]$. Different colors corresponds to different methods: DRF (red), GRF (green), QRF (blue), TRF (purple).

D.2 Heterogeneous regression and causal effects

We explore here the performance of DRF on the synthetic data for the setup of heterogeneous regression, where we want to obtain the regression fit of Y on the explanatory (or treatment) variables $\mathbf{W}$, but where this fit might change depending on values of $\mathbf{X}$. This can be done by DRF by using $\mathbf{X}$ as predictors and $(\mathbf{W}, Y)$ as the response and then using some standard regression method for regressing Y on $\mathbf{W}$ in the second step, having already obtained the weights that describe the conditional distribution $\mathbb{P}((\mathbf{W}, Y) \mid \mathbf{X} = \mathbf{x})$.

The most important such setup is when the data come from the following causal graph:
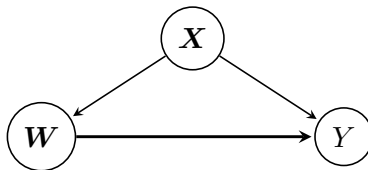
In the case of such a causal graph, $\mathbf{X}$ are confounding variables, which we need to adjust for to understand the causal effect of $\mathbf{W}$ on Y. Not only can the marginal distributions of Y and $\mathbf{W}$ be affected by $\mathbf{X}$, but also the regression fit (e.g. the regression coefficients).

D.2.1 CATE AND ATE

One special case of this setup that is intensively studied in the causal literature is when W is a (univariate) binary treatment variable. In this case we are interested in the distribution of the potential outcomes $Y(W = 0)$ and $Y(W = 1)$ and especially in their difference. It is commonly measured by using the Conditional Average Treatment Effect

$$\mathrm{CATE}(\mathbf{x}) = \mathbb{E}[Y(W = 1) - Y(W = 0) \mid \mathbf{X} = \mathbf{x}]$$

and the Average Treatment Effect

$$\mathrm{ATE} = \mathbb{E}[Y(W = 1) - Y(W = 0)] = \mathbb{E}[\mathrm{CATE}(\mathbf{X})].$$

Competing methods. We will compare the performance of the DRF with the following methods, specially designed for estimation of the CATE (or ATE)

- Double Machine Learning (DML) of Chernozhukov et al. (2018), which assumes the model $Y = m(\mathbf{X}) + W\theta + \epsilon$ with constant treatment effect and can thus only be used for estimating ATE and not CATE.

- X-learner (XL) introduced in Künzel et al. (2019) (the version with RF learners)

- Causal Forest (CF) introduced in Wager and Athey (2018); Athey et al. (2019) (we use the GRF version (Athey et al., 2019) with local centering that substantially improves on the version in Wager and Athey (2018))

In order to make the comparison fair, we use the local centering approach for DRF as well.

Data. We will use the following data models for our simulations, where the first three are taken directly from Athey et al. (2019):

1. In this model X_3 is a confounder affecting both W and Y:

$$\mathbf{X} \sim U(0,1)^p, \quad W \mid \mathbf{X} \sim \mathrm{Bernoulli}\left(\frac{1}{4}(1 + \beta_{2,4}(X_3))\right),$$

$$Y \mid \mathbf{X}, W \sim 2\left(X_3 - \frac{1}{2}\right) + N(0,1),$$

where $\beta_{a}, b(x)$ is the density of the beta random variable with parameters a and b.

2. In this model the treatment effect is heterogeneous, i.e. how W affects Y changes with X_1 and X_2:

$$\mathbf{X} \sim U(0,1)^p, \quad W \mid \mathbf{X} \sim \text{Bernoulli}(0.5),$$

$$Y \mid \mathbf{X}, W \sim \left(W - \frac{1}{2}\right)\eta(X_1)\eta(X_2) + N(0,1),$$

where $\eta(x) = 1 + \left(1 + e^{-20(x - \frac{1}{3})}\right)^{-1}$.

3. This model is a combination of the previous two, so the treatment effect is heterogeneous and we have confounding:

$$\mathbf{X} \sim U(0,1)^p, \quad W \mid \mathbf{X} \sim \text{Bernoulli}\left(\frac{1}{4}(1 + \beta_{2,4}(X_3))\right),$$

$$Y \mid \mathbf{X}, W \sim 2\left(X_3 - \frac{1}{2}\right) + \left(W - \frac{1}{2}\right)\eta(X_1)\eta(X_2) + N(0,1).$$

4. The following model is similar to above, with slightly different structure, where X_2 induces the confounding effects and X_1 makes the treatment heterogeneous:

$$\mathbf{X} \sim U(0,1)^p, \quad W \mid \mathbf{X} \sim \text{Bernoulli}\left(\text{expit}(4X_2 - 2)\right),$$

$$Y \mid \mathbf{X}, W \sim 100X_2^2 + \left(W - \frac{1}{2}\right)\sin(3X_1) + N(0,1).$$

Results. For every model we generate n data points $(X_1, \ldots, X_p, W, Y)_{i=1,\ldots,n}$. We run all methods and compute the root mean squared error of the obtained CATE estimate on a randomly generated test set $\mathbf{X}_{\text{test}}$ containing 1000 data points. CATE corresponds to the coefficient of W in the data generating mechanism of Y. We repeat the same procedure 100 times and report the average result. For methods other than the DML, we estimate ATE by averaging the CATE estimates over the randomly generated test set. The results can be seen in Table 4 and Figure 13. Even though DRF is performing less well in general compared to the methods that are specially designed for the task of estimating CATE, we can still see that its estimates are fairly good.

D.2.2 Continuous W, linear treatment effect

When the treatment variable W is continuous, many methods designed for binary treatments, such as the X-learner (Künzel et al., 2019) cannot be used. However, many important real-word examples fall within this framework. As an example, we might be interested in how the amount of medicine W affects some biological parameter of interest Y (conditionally on $\mathbf{X}$). When W affects Y linearly conditionally on $\mathbf{X}$, one can still use the Causal Forest (CF) (Athey et al., 2019; Wager and Athey, 2018) method, which makes the splits based on the slope of the conditional linear fit $Y \sim W$. Due to its generality and versatility, DRF can trivially be used in such setting as well.

To illustrate this, we consider the Model 3, as described in the previous section, which is also taken from Athey et al. (2019), but where we change the distribution of the binary

Table 4: RMSE for the CATE, averaged over 1000 test points and 100 overall repetitions.

model	n	p	DRF	CF	XL
1	800	10	0.140	**0.109**	0.149
1	1600	10	0.119	**0.085**	0.122
1	800	20	0.125	**0.094**	0.128
1	1600	20	0.105	**0.076**	0.107
2	800	10	0.452	0.319	**0.288**
2	1600	10	0.285	0.234	**0.228**
2	800	20	0.568	0.336	**0.306**
2	1600	20	0.341	0.254	**0.241**
3	800	10	0.621	0.328	**0.319**
3	1600	10	0.453	0.243	**0.237**
3	800	20	0.708	**0.343**	0.346
3	1600	20	0.533	0.257	**0.256**
4	800	10	0.320	**0.273**	0.682
4	1600	10	0.285	**0.228**	0.389
4	800	20	0.316	**0.289**	0.722
4	1600	20	0.291	**0.248**	0.412

Table 5: RMSE for the ATE, averaged over 100 repetitions.

| model	n	p	DRF	CF	XL	DML
1	800	10	0.0841	**0.0806**	0.0858	0.0843
1	1600	10	0.0526	**0.0517**	0.0523	0.0538
1	800	20	0.0786	**0.0762**	0.0802	0.0785
1	1600	20	**0.0585**	0.0588	0.0609	0.0625
2	800	10	**0.0844**	0.0880	0.0877	0.0891
2	1600	10	**0.0567**	0.0587	0.0593	0.0584
2	800	20	0.0783	**0.0767**	0.0788	0.0872
2	1600	20	**0.0645**	0.0657	0.0665	0.0656
3	800	10	**0.0914**	0.0916	0.0932	0.1116
3	1600	10	**0.0573**	0.0581	0.0599	0.0778
3	800	20	**0.0858**	0.0917	0.0922	0.1114
3	1600	20	**0.0623**	0.0673	0.0599	0.0925
4	800	10	**0.1061**	0.1075	0.2554	0.9734
4	1600	10	0.0677	0.0665	0.1028	0.5542
4	800	20	**0.1008**	0.1046	0.2308	2.7805
4	1600	20	**0.0655**	0.0660	0.0947	1.834
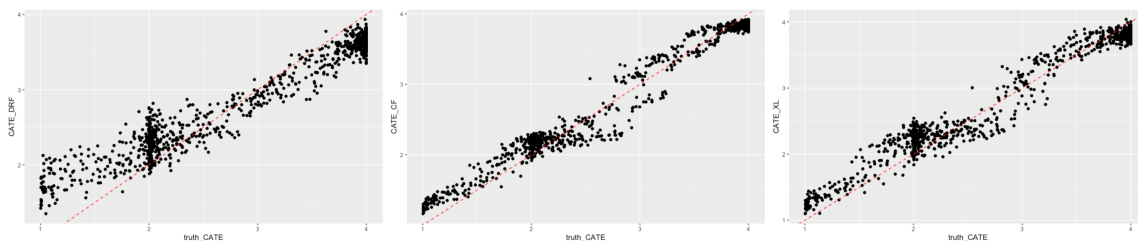
Figure 13: Estimates of the CATE for DRF (left), Causal Forest (middle), and X-learner (right) plotted against the true cate on the x-axis for Model 3 with $n = 1600$, $p = 20$.

treatment variable so that it is continuous and it has a normal distribution with the same mean and variance, which depend on **X**. In this model W affects Y linearly, which is a
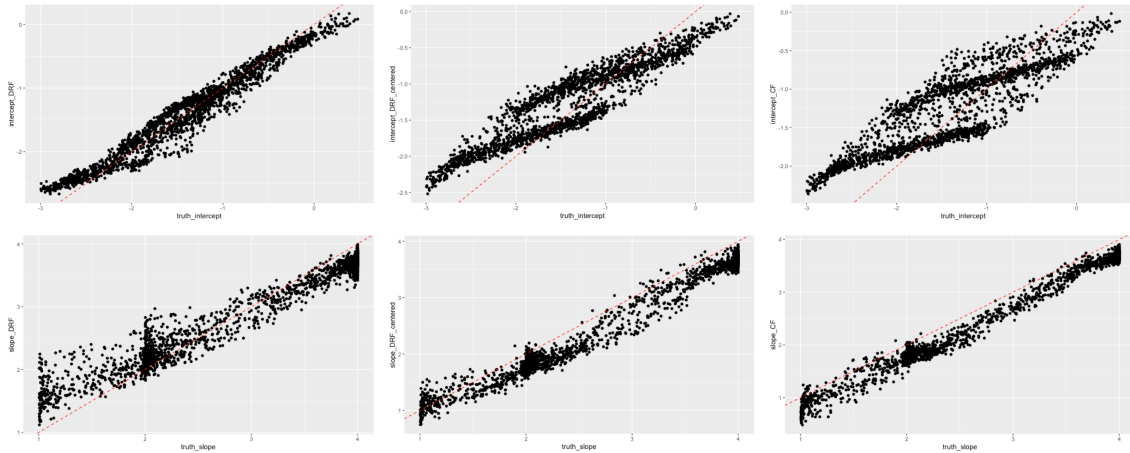
Figure 14: The estimates of the intercept (top row) and the slope (bottom row) for the linear fit conditional on $\mathbf{X}$, against the true values on the x-axis, obtained for the plain DRF (left), DRF with local centering as in Athey et al. (2019) (middle) and CF (right), which uses the local centering approach. The data is generated from Model 3 described in the previous section, with continuous treatment variable W with the same mean and variance conditionally on $\mathbf{X}$.

crucial assumption for the CF approach to work. We take $n = 10000$ and $p = 10$. The concept of CATE does not exist in this form in such setup and therefore we consider how the forest obtained by each method estimates both the intercept and the slope of the fit $Y \sim W$, conditionally on $\mathbf{X}$. The results can be seen in Figure 14. We see that the estimate of the slope for DRF is slightly worse than for the CF, whose forest construction is specially designed for estimating the conditional slope. However, DRF estimates the intercept significantly better than the CF, especially in combination the local centering approach, which uses the centered data $Y - \widehat{Y}(\mathbf{X})$ and $W - \widehat{W}(\mathbf{X})$ instead. In this example the slope depends only on X_2, whereas the intercept depends on both X_1 and X_2. Since CF targets only the slope for forest construction, it will split mostly on X_2 and not on X_1, which leads to poor estimate of the intercept term. On the other hand, DRF splits both on X_1 and X_2, depending on the size of their effect on the joint distribution of (W, Y). For many applications, especially in causality (see also the example in next section), it is essential to know the whole conditional distribution $\mathbb{P}(Y \mid W, \mathbf{X})$ so the DRF approach might be more beneficial than the CF.

D.2.3 NONLINEAR TREATMENT EFFECT

There are very few methods that can estimate the treatment effect when the treatment variable W is continuous and affects Y nonlinearly, as it is commonly the case in real world settings. Here we demonstrate that DRF can be easily used in such setup as well, as opposed to the CF, which assumes a linear, though heterogeneous, treatment effect of W on Y. We further show how the obtained regression fits $Y \sim W$ conditional on $\mathbf{X}$ can be used to

estimate the causal effect $\mathbb{E}[Y \mid do(W = w)]$, as it is done in the main paper in the birth data example:

$$\mathbb{E}[Y \mid do(W = w)] = \int \mathbb{E}[Y \mid W = w, \mathbf{X} = \mathbf{x}]\mathbb{P}(\mathbf{X} = \mathbf{x})d\mathbf{x}.$$

We compare the performance of DRF with the straightforward and commonly used approach, where we first regress Y on $(W, \mathbf{X})$ and use this regression fit which estimates $\mathbb{E}[Y \mid W = w, \mathbf{X} = \mathbf{x}]$ together with the above formula to estimate the causal effect.

Data We consider the following example, similar to the previous examples:

$$\mathbf{X} \sim U(0, 1)^p, \quad W \mid \mathbf{X} \sim \frac{1}{2}\left|1 + 4X_3 + N(0, 1)\right|,$$

$$Y \mid \mathbf{X}, W \sim 3\left(X_3 - \frac{1}{2}\right) + 3X_1 \sin(3W) + X_2 N(0, 1).$$

Therefore, W affects Y highly nonlinearly through a sine function. X_3 is a confounding variable that affects the marginal distributions of Y and W. X_2 regulates the error level for Y, whereas X_1 makes the treatment effect heterogeneous. This is illustrated in the following plot:
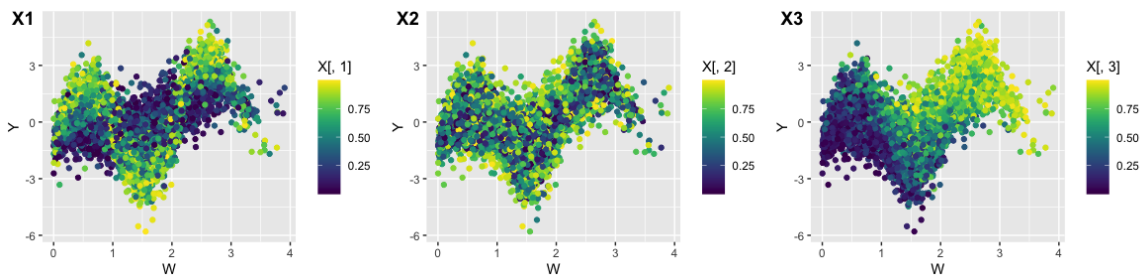

Figure 15: Visualization how X_1, X_2, X_3 affect the conditional nonlinear regression fit $Y \sim W$. X_1 changes the effect size, X_2 changes the noise level, whereas X_3 is a confounding variable which affects the means of W and Y.

Results In Figure 16 we can see the estimated joint distribution of (Y, W) conditionally on $\mathbf{X} = \mathbf{x}$, where the values of X_1 and X_3 vary, while the rest are fixed (even though X_2 also affects the conditional distribution, the effect is much weaker than for X_1 of X_3, see Figure 16). We see that the estimated distribution matches the true regression line, denoted in red, very well. The estimated distribution induced by the DRF weights enables us to fit some specialised regression method for regressing Y on W for every fixed value of $\mathbf{X}$. The blue line indicates the fit obtained by using smoothing splines. Compared to the green line, which shows the predicted values for regression $Y \sim (W, \mathbf{X})$, it is nicer looking and also is able to extrapolate much better to values of W which have low probability conditionally on X.

This extrapolation is crucial for causal applications, since for computing $\mathbb{E}[Y \mid do(W = w)]$ we are interested in what would happen with Y when our treatment variable W is fixed to
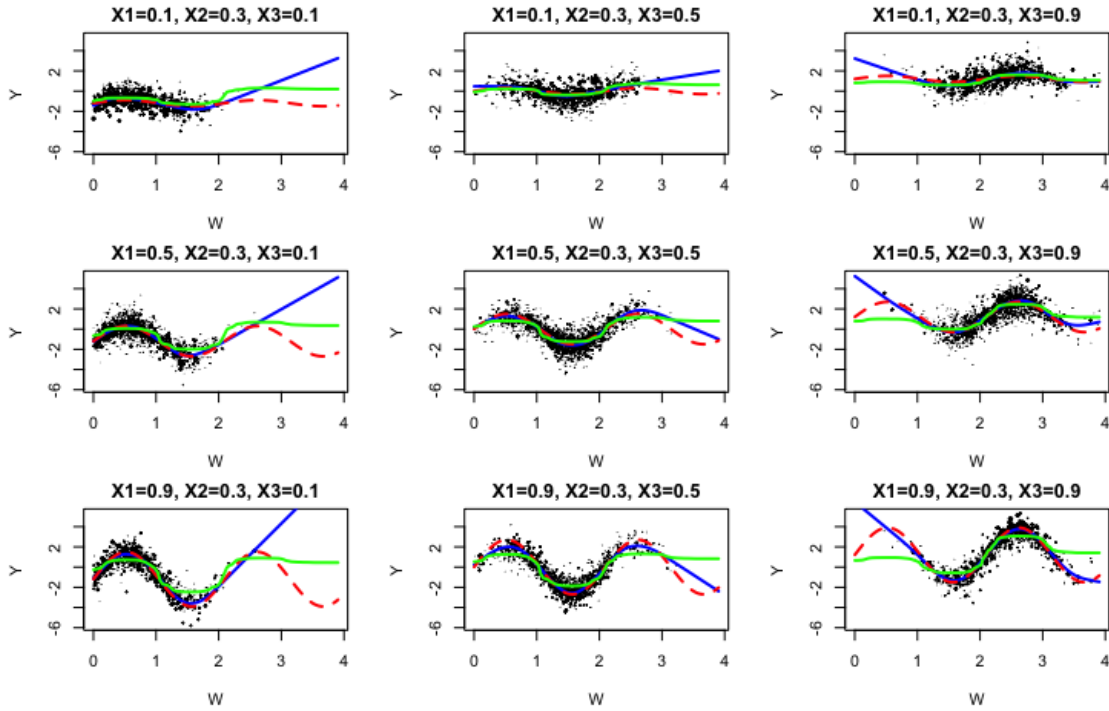
Figure 16: For a grid of values for the test point $\mathbf{x}$, the scatterplot illustrates the estimated joint distribution (Y, W) by DRF. The subsequent regression fit using smoothing splines is denoted in blue, whereas the true conditional mean $\mathbb{E}[Y \mid W = w, \mathbf{X} = \mathbf{x}]$ is denoted with red dashed line. Green line shows the estimate of the conditional mean $\mathbb{E}[Y \mid W = w, \mathbf{X} = \mathbf{x}]$ with plain random forest.

be w, regardless of which values are achieved by $\mathbf{X}$. However, it can easily happen that for this specific combination of $\mathbf{X}$ and W there are very few observed data points, which makes the estimation hard (Pearl, 2009). In this example, W tends to be small for small values of X_3 and vice-versa and thus is hard to say what would happen with Y when X_3 is large and W is set to a small value by an outside intervention.

In Figure 17, we indeed see that the estimates of the causal effect $\mathbb{E}[Y \mid do(W = w)]$ by DRF are much better. One can still see that the error increases for the border values of W, which have small probability for some values of X, since the estimation there is much harder, but this error is much less pronounced for DRF than for the standard regression approach.
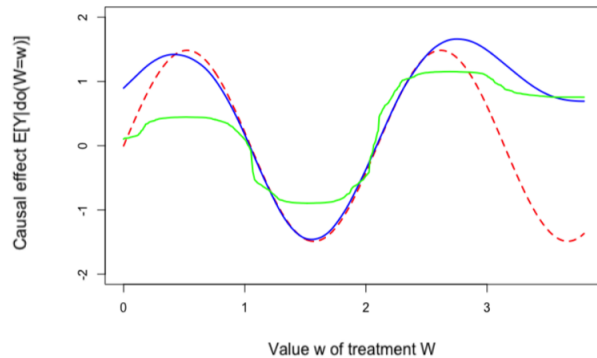
Figure 17: The estimated causal effect $\mathbb{E}[Y \mid do(W = w)]$ with DRF (blue) and with conventional method which regresses Y on $(W, \mathbf{X})$ using plain Random Forest (green). The true value is denoted by a red dashed line.

References

Alberto Abadie and Guido W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

Karan Aggarwal, Matthieu Kirchmeyer, Pranjul Yadav, Sathiya S. Keerthi, and Patrick Gallinari. Benchmarking regression methods: A comparison with CGAN. *arXiv preprint arXiv:1905.12868*, 2019.

Mohamad Akra and Louay Bazzi. On the solution of linear recurrence equations. *Computational Optimization and Applications*, 10(2):195–210, 1998.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178, 2019.

Zhidong Bai and Hewa Saranadasa. Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, pages 311–329, 1996.

Ludwig Baringhaus and Carsten Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

Patrizia Berti, Luca Pratelli, and Pietro Rigo. Almost sure weak convergence of random probability measures. *Stochastics*, 78(2):91–97, 2006.

Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1): 1063–1095, 2012.

Peter J. Bickel, Chris AJ Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Boris E. Brodsky and Boris S. Darkhovsky. *Nonparametric Methods in Change Point Problems*, volume 243. Springer Science & Business Media, 2013.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 16, pages 785–794, New York, NY, USA, 2016.

Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Umberto Cherubini, Elisa Luciano, and Walter Vecchiato. *Copula Methods in Finance*. John Wiley & Sons, 2004.

Kacper P. Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, volume 28, pages 1981–1989. 2015.

William S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

Panayiota Constantinou and Alexander P. Dawid. Extended conditional independence and applications in causal inference. *The Annals of Statistics*, 45(6):2618 – 2653, 2017.

Misha Denil, David Matheson, and Nando De Freitas. Narrowing the gap: Random forests in theory and in practice. In *31st International Conference on Machine Learning (ICML)*, volume 32 of *ICML'14*, pages 665–673, 2014.

Richard M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002.

Eugene B. Dynkin and Avishai Mandelbaum. Symmetric statistics, Poisson point processes, and multiple Wiener integrals. *The Annals of Statistics*, 11(3):739 – 745, 1983.

Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596, 05 1981.

Jan Ernest and Peter Bühlmann. Marginal integration for nonparametric causal inference. *Electronic Journal of Statistics*, 9(2):3155 – 3194, 2015.

Guangzhe Fan, Zhou Wang, and Jiheng Wang. Cw-ssim kernel based random forest for image classification. In *Visual Communications and Image Processing 2010*, volume 7744, page 774425. International Society for Optics and Photonics, 2010.

Jianqing Fan, Mark Farmen, and Irene Gijbels. Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):591–608, 1998.

Jerome H. Friedman and Lawrence C. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, pages 697–717, 1979.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep Learning*, volume 1. MIT press Cambridge, 2016.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, volume 19, pages 513–520, 2007a.

Arthur Gretton, Kenji Fukumizu, Choon H. Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20, page 585–592, 2007b.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012a.

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, volume 25, pages 1205–1213. 2012b.

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, volume 25, pages 1205–1213, 2012c.

Wassily Hoeffding. The strong law of large numbers for U-statistics. Mimeograph Series 302, Institute of Statistics, University of Noth Carolina, 1961.

Torsten Hothorn and Achim Zeileis. Predictive distribution modeling using transformation forests. *Journal of Computational and Graphical Statistics*, 30(4):1181–1196, 2021.

Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J. Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley, 2015.

Guido W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

Hemant Ishwaran and Udaya B. Kogalur. Package 'randomForestSRC', 2022.

Wittawat Jitkrittum, Zoltán Szabó, Kacper P. Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems*, volume 29, pages 181–189. 2016.

Jonathan Johannemann, Vitor Hadad, Susan Athey, and Stefan Wager. Sufficient representations for categorical variables. *arXiv preprint arXiv:1908.09874*, 2019.

Edward H. Kennedy, Zongming Ma, Matthew D. McHugh, and Dylan S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79(4):1229, 2017.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, volume 31, pages 656–666, 2017.

Durk P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, volume 29, page 4743–4751, 2016.

Dragi Kocev, Celine Vens, Jan Struyf, and Sašo Džeroski. Ensembles of multi-objective decision trees. In *European Conference on Machine Learning*, pages 624–631, 2007.

Roger Koenker, Stephen Portnoy, Pin T. Ng, Achim Zeileis, Philip Grosjean, and Brian D. Ripley. Package 'quantreg', 2012.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076, 2017.

Quoc Le, Tamás Sarlós, and Alex Smola. Fastfood: Approximating kernel expansions in loglinear time. In *30th International Conference on Machine Learning (ICML)*, volume 28 of *ICML'13*, page 244–252, 2013.

Shuang Li, Yao Xie, Hanjun Dai, and Le Song. Scan B-statistic for kernel change-point detection. *Sequential Analysis*, 38(4):503–544, 2019.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors (technical report no. 1055). *University of Wisconsin*, 2002.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

Frank J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

David S. Matteson and Nicholas A. James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345, 2014.

Lachlan McCalman, Simon T. O'Callaghan, and Fabio Ramos. Multi-modal estimation with kernel embeddings for learning motion models. In *IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, May 6-10, 2013*, pages 2845–2852, 2013.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(35): 983–999, 2006.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Krikamol Muandet, Kenji Fukumizu, Bharath K. Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, volume 30, page 2335–2344, 2017.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Gilles Pisier. *Martingales in Banach Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2016.

Drago Plečko and Nicolai Meinshausen. Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44, 2020.

Taylor Pospisil and Ann B. Lee. Rfcde: Random forests for conditional density estimation. *arXiv preprint arXiv:1804.05753*, 2018.

Joaquin Quinonero-Candela, Carl Edward Rasmussen, Fabian Sinz, Olivier Bousquet, and Bernhard Schölkopf. Evaluating predictive uncertainty challenge. In *Machine Learning Challenges Workshop*, pages 1–27. Springer, 2005.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pages 1177–1184, 2008.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems*, volume 21, pages 1313–1320, 2009.

Aaditya Ramdas, Sashank J. Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. AAAI'15, page 3571–3577, 2015.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2), 2017.

Mark Segal and Yuanyuan Xiao. Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):80–87, 2011.

Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26. CRC press, 1986.

Carl-Johann Simon-Gabriel and Bernhard Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *The Journal of Machine Learning Research*, 19(1):1708–1736, 2018.

Carl-Johann Simon-Gabriel, Alessandro Barp, and Lester Mackey. Metrizing weak convergence with maximum mean discrepancies. *arXiv preprint arXiv:2006.09268*, 2020.

Abe Sklar. *Fonctions de Répartition À N Dimensions Et Leurs Marges*. Université Paris 8, 1959.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, volume 28, pages 3483–3491, 2015.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *26th Annual International Conference on Machine Learning (ICML)*, ICML '09, page 961–968, 2009.

Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

Bharath K. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839 – 1893, 2016.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Gábor J. Székely and Maria L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5 (16.10):1249–1272, 2004.

Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. MULAN: A Java Library for Multi-Label Learning. *Journal of Machine Learning Research*, 12(71):2411–2414, 2011.

Hisaharu Umegaki and Albert T. Bharucha-Reid. Banach space-valued random variables and tensor products of Banach spaces. *Journal of Mathematical Analysis and Applications*, 31(1):49–67, 1970.

Aad van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Richard A. Vitale. Covariances of symmetric statistics. *Journal of Multivariate Analysis*, 41(1):14–26, 1992.

Stefan Wager. Asymptotic theory for random forests. *arXiv preprint arXiv:1405.0352*, 2014.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342*, 2017.

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

Bernard L. Welch. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

Holger Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004.

Frank Wilcoxon. Individual comparisons of grouped data by ranking methods. *Journal of Economic Entomology*, 39(2):269–270, 1946.

Douglas A. Wolfe and Edna Schechtman. Nonparametric statistical procedures for the changepoint problem. *Journal of Statistical Planning and Inference*, 9(3):389–396, 1984.

Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017.

Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 26, pages 755–763, 2013.

Achim Zeileis, Torsten Hothorn, and Kurt Hornik. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514, 2008.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence*, page 804–813, 2011.

Ji Zhao and Deyu Meng. Fastmmd: Ensemble of circular discrepancy for efficient two-sample test. *Neural computation*, 27(6):1345–1372, 2015.