A Unified Framework for Random Forest Prediction Error Estimation

Benjamin Lu
Department of Statistics
University of California, Berkeley
Evans Hall
Berkeley, CA 94720, USA

Johanna Hardin
Department of Mathematics
Pomona College
610 N. College Ave.
Claremont, CA 91711, USA

Abstract
We introduce a unified framework for random forest prediction error estimation based on a novel estimator of the conditional prediction error distribution function. Our framework enables immediate estimation of key parameters often of interest, including conditional mean squared prediction errors, conditional biases, and conditional quantiles, by a straightforward plug-in routine. Our approach is particularly well-adapted for prediction interval estimation, which has received less attention in the random forest literature despite its practical utility; we show via simulations that our proposed prediction intervals are competitive with, and in some settings outperform, existing methods. To establish theoretical grounding for our framework, we prove pointwise uniform consistency of a more stringent version of our estimator of the conditional prediction error distribution. In addition to providing a suite of measures of prediction uncertainty, our general framework is applicable to many variants of the random forest algorithm. The estimators introduced here are implemented in the R package forestError.

Keywords: Prediction error, bias, prediction intervals, bagging

1. Introduction
Random forests and other tree-based methods are often used by researchers for regression—that is, to relate a real-valued response $Y$ to covariates $X$ (Criminisi et al., 2010; Grimm et al., 2008; Wei et al., 2010). The objective in many of these applications is to predict the unknown response of observations given their covariates; we denote these predictions by $\hat{Y}(X)$. For example, researchers in precision medicine seek to predict the health outcomes of individual patients under some treatment regime given patient, clinical, and environmental characteristics, with the ultimate goal of developing individualized therapies for patients (Fang et al., 2018). Other researchers seek to predict bird migration patterns to reduce collisions with airplanes, wind turbines, and buildings (Van Doren and Horton, 2018).

When using any regression method for prediction, quantifying the uncertainty associated with the predictions can enhance their practical value. One central function for quantifying uncertainty is the conditional prediction error distribution, which, letting $E := Y - \hat{Y}(X)$...
denote the error of a prediction, is given by

\[ F_E(e \mid x) := \Pr (E \leq e \mid X = x) = \Pr (Y - \hat{\phi}(X) \leq e \mid X = x). \]

The conditional error distribution can be mapped to a number of useful parameters. For example, the conditional mean squared prediction error,

\[ \text{MSPE}(x) := \mathbb{E} [(Y - \hat{\phi}(X))^2 \mid X = x] = \int e^2 f_E(e \mid x) \, de, \]

can provide an informative summary of how erroneous a given point prediction is expected to be. Additionally, the conditional bias,

\[ \text{Bias}(x) := \mathbb{E} [Y - \hat{\phi}(X) \mid X = x] = \int e f_E(e \mid x) \, de, \]

provides one measure of systematic over- or under-prediction of responses for units with a given set of covariates. Finally, the conditional error distribution can be mapped to \( \alpha \)-quantiles of the prediction error,

\[ Q^\alpha_E(x) := \inf \{ e : F_E(e \mid x) \geq \alpha \}, \]

from which conditional prediction intervals containing the unknown response with a specified probability can be constructed.

This paper proposes a method of estimating the conditional prediction error distribution \( F_E(e \mid x) \) of random forests. With this estimate, conditional mean squared prediction error, conditional bias, conditional quantiles, and other parameters of the distribution can all be estimated with ease. By contrast, the current literature on characterizing random forest prediction uncertainty has been piecemeal, with different proposed methods, if any, for estimating each of the parameters mentioned above. Thus, the central contribution of this paper is a unified framework for assessing random forest prediction uncertainty, with a suite of estimators that empirically are competitive with, and in some cases outperform, existing methods, particularly for the tasks of prediction interval estimation and quantile regression.

In addition to providing a unified framework, our method is general in the sense that it can be implemented for many variants of the random forest algorithm. For example, it is compatible with a wide range of decision tree algorithms that partition the covariate space based on different criteria, as well as various resampling and subsampling regimes that have been proposed in recent literature. Biau et al. (2008) provide an excellent review and analysis of some of these variants.

The remainder of this manuscript is organized as follows. Section 2 reviews the literature on estimating parameters of \( F_E(e \mid x) \) that are commonly of interest. We establish the setting and relevant notation for our problem in Section 3 before introducing in Section 4 our proposed estimator of \( F_E(e \mid x) \) and showing how it enables straightforward plug-in estimation of certain parameters of \( F_E(e \mid x) \). In Section 5, we assess the empirical performance of some of these resulting plug-in estimators. In Section 6, we propose and prove uniform consistency of an estimator of \( F_E(e \mid x) \) that is similar to but more stringently constructed than the one proposed in Section 4. Section 7 concludes.
2. Related Work

To our knowledge, we are the first to propose a method of estimating the conditional prediction error distribution of random forests. However, random forest mean squared prediction error, bias, and prediction intervals have each been studied individually in previous works. We briefly review the literature on each in turn.

The most, and perhaps only, widely adopted summary metric for random forest prediction error is the unconditional mean squared prediction error,

\[ \text{MSPE} := \mathbb{E} \left[ (Y - \hat{\varphi}(X))^2 \right], \]

which is usually estimated by an out-of-bag procedure (Breiman, 1996; Liaw and Wiener, 2002). We build on this work by enabling straightforward estimation of the conditional mean squared prediction error \( \text{MSPE}(x) \), which, as we illustrate in Section 4.2, is often a more informative metric.

The literature on random forest bias has been more active recently. Wager and Athey (2018) show that random forests are biased and provide a bound on the magnitude of the bias under certain assumptions about the tree-growing mechanism and the underlying data distribution. Ghosal and Hooker (2018) leverage this work to investigate a method of bias correction, initially proposed by Breiman (1999), that entails fitting a random forest on the out-of-bag prediction errors. This boosting approach, which is similar to gradient boosting (Friedman, 2001), is also studied by Zhang and Lu (2012), who propose additional model-based bias corrections for random forests. Hooker and Mentch (2018) propose a different method of bias correction in which a decision tree is fit to a bootstrap sample of the training set with responses modified by a residual bootstrap procedure. We contribute to this literature by proposing a new bias correction procedure and comparing it to the boosting method examined by Ghosal and Hooker (2018) and Zhang and Lu (2012).

The literature on conditional prediction interval estimation for random forests has also grown in recent years, beginning with the development by Meinshausen (2006) of quantile regression forests, a random forest-based algorithm that enables consistent estimation of conditional prediction intervals; our approach hews closely to this work. Since then, Athey et al. (2019) have proposed generalized random forests, a method of estimating quantities identified by local moment conditions that grows trees specifically designed to express heterogeneity in the quantity of interest; they show that their algorithm can be used for quantile regression. Additionally, recent developments have been made in conformal inference, a generic method of prediction interval estimation that can be applied to virtually any estimator of the regression function, including random forests (Lei and Wasserman, 2014; Lei et al., 2018; Johansson et al., 2014). We build on this literature by enabling an alternative approach to conditional prediction interval estimation and assessing the relative strengths and weaknesses of each procedure through simulation.

Finally, we emphasize here that our work is separate from the rich literature investigating the performance of random forest-based algorithms for conditional mean estimation and inference (Sexton and Laake, 2009; Wager et al., 2014; Mentch and Hooker, 2016; Wager and Athey, 2018). Although, for many regression methods, the point estimator for an individual response is equivalent to the point estimator for the conditional mean, the statistical challenges of conditional mean estimation are different from those of prediction
error estimation. For example, many methods of conditional mean estimation and inference invoke some type of central limit theorem to characterize their estimators’ behavior; such approaches are generally less applicable to prediction error estimation, which concerns individual responses rather than their expected value.

3. Setup and Notation

Consider an observed training sample \( D_n := \{(X_i, Y_i)\}_{i=1}^n \), where \((X_i, Y_i) \overset{i.i.d.}{\sim} \mathbb{P} \) for some distribution \( \mathbb{P} \), \( X_i \in \mathcal{X} \) is a \( p \)-dimensional covariate with support \( \mathcal{X} \), and \( Y_i \in \mathbb{R} \) is a real-valued response with a continuous conditional distribution function. For convenience, we let \( Z_i := (X_i, Y_i) \). A standard implementation of random forests fits a tree on each of \( B \) bootstrap samples of the training set, \( D_{n,1}, \ldots, D_{n,B} \), using the classification and regression tree (CART) algorithm, with the \( b \)th tree’s construction governed by a random parameter \( \theta_b \) drawn i.i.d. from some distribution independently of \( D_n \) (Breiman, 2001). Included in \( \theta_b \), for example, is the randomization of eligible covariates for each split. Each tree splits the bootstrap training sample \( D_{n,b} \) into terminal nodes; this splitting corresponds to a partitioning of the predictor space \( \mathcal{X} \) into rectangular subspaces. For the \( b \)th tree, let \( \ell(x, \theta_b) \) index the terminal node corresponding to the subspace containing \( x \) and \( R_{\ell(x, \theta_b)} \) denote the subspace itself. With this notation, we introduce the following terminology; to our knowledge, no settled term for observations satisfying Definition 1 exists in the current literature.

**Definition 1.** A training observation \( Z_i \) is a **cohabitant** of \( x \) in tree \( b \) if and only if \( \ell(X_i, \theta_b) = \ell(x, \theta_b) \).

When predicting the response of a test observation with realized covariate values \( x \), each tree in the random forest employs a weighted average of the in-bag training responses, with weights corresponding to cohabitation. In particular, the in-bag weight given to the \( i \)th observation in the \( b \)th tree is

\[
w_i(x, \theta_b) := \frac{\#\{Z_i \in D^*_n \} \mathbb{1}(X_i \in R_{\ell(x, \theta_b)})}{\sum_{j=1}^n \#\{Z_j \in D^*_n \} \mathbb{1}(X_j \in R_{\ell(x, \theta_b)})},
\]

where \( \#\{Z_i \in D^*_n \} \) denotes the number of times the \( i \)th observation is in \( D^*_n \) and \( \mathbb{1}(\cdot) \) is the indicator function. The random forest prediction of the response of units with covariate value \( x \) is the average of the tree predictions:

\[
\hat{\varphi}(x) := \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^n w_i(x, \theta_b) Y_i.
\]

It is well-known that, with a sufficiently large number of trees grown on bootstrap samples of \( D_n \), each training observation will be out of bag—that is, not included in the bootstrap sample—for at least \( B/4 \) of the trees with high probability. Thus, we can define out-of-bag analogues to \( w_i(x, \theta_b) \) and \( \hat{\varphi}(x) \). The out-of-bag weight given to the \( i \)th training observation is the proportion of times the \( i \)th training observation is an out-of-bag cohabitant.
of \( x \), relative to all training observations:

\[
v_i(x) := \frac{\sum_{b=1}^{B} 1(Z_i \notin D_{n,b}^* \text{ and } X_i \in R_{t(x,\theta_b)})}{\sum_{j=1}^{n} \sum_{b=1}^{B} 1(Z_j \notin D_{n,b}^* \text{ and } X_j \in R_{t(x,\theta_b)})}.
\]

Notice that, unlike \( w_i(x, \theta_b) \), \( v_i(x) \) is defined over all trees because \( x \) is guaranteed an in-bag cohabitant in each tree but not an out-of-bag cohabitant. The out-of-bag prediction of the \( i^{th} \) training unit is the average prediction of the unit’s response among the trees for which the unit is out of bag:

\[
\hat{\varphi}^{(i)}(X_i) := \frac{1}{\sum_{b=1}^{B} 1(Z_i \notin D_{n,b}^*) \sum_{j=1}^{n} w_j(X_i, \theta_b) Y_j}.
\]

4. A Unified Framework for Assessing Prediction Uncertainty

In this section, we present a practical implementation of our proposed method of estimating the conditional prediction error distribution \( F_E(e \mid x) \) and show how it can easily facilitate estimation of conditional mean squared prediction errors, conditional biases, and conditional prediction intervals. This practical implementation is similar in spirit to but less stringent in its construction than our more rigorous method of estimating \( F_E(e \mid x) \), which we detail and prove is uniformly consistent in Section 6. Nonetheless, we present the practical version first to build intuition, demonstrate its viability in empirical applications (see Section 5), and suggest potential areas of future research; similar simplifications have been made in other recent work on random forests (Meinshausen, 2006). The estimators discussed in this section are implemented in the R package forestError.

4.1 Estimating the Conditional Prediction Error Distribution

The practical implementation of our proposed method estimates \( F_E(e \mid x) \) by out-of-bag weighting of the out-of-bag prediction errors:

\[
\hat{F}_E(e \mid x) := \sum_{i=1}^{n} v_i(x) 1 \left( Y_i - \hat{\varphi}^{(i)}(X_i) \leq e \right).
\]

This approach is grounded in the principle that, because training observations are not used in the construction of trees for which they are out of bag, the relationship between a training observation and the subset of trees for which it is out of bag is analogous to the relationship between the test observation and the random forest when the number of training observations and trees is large. In particular, not only are the out-of-bag prediction errors a reasonable proxy for the error of future test predictions in general, but also the out-of-bag prediction errors of training observations that are more frequently out-of-bag cohabitants of a given test observation make better proxies for the prediction error of that specific test observation than the out-of-bag prediction errors of training observations that are out-of-bag cohabitants less often. Broadly speaking, this notion of similarity also motivates the “proximity” measure commonly included in random forest implementations, but proximity
takes on a different form from $v_i(x)$ and has traditionally been measured between training observations to identify structures in the data (Liaw and Wiener, 2002; Breiman, 2002).

One minor caveat for the analogy between out-of-bag observations and test observations is that fewer trees are used to generate the out-of-bag weights and out-of-bag predictions. In this respect, the out-of-bag errors more closely resemble test errors from a fraction of the random forest’s trees, chosen randomly. However, as the following proposition shows, the distribution of prediction errors $E^*$ from a fraction $\pi \in (0, 1]$ of the $B$ trees in a random forest becomes arbitrarily similar to the distribution of prediction errors $E$ from the full random forest as $B$ increases; the proof is provided in the appendix.

**Proposition 1.** For every $x \in \mathcal{X}$,

$$\lim_{B \to \infty} \sup_{e \in \mathbb{R}} |F_{E^*}(e \mid x) - F_E(e \mid x)| = 0.$$ 

Other issues, primarily concerning the dependence relations induced by the construction of the random forest and $\hat{F}_E(e \mid x)$, prevent us from proving consistency of $\hat{F}_E(e \mid x)$. These issues, which touch on recent areas of research, are discussed further in Section 6, where we prove uniform consistency of a similar but more stringently constructed estimator of $F_E(e \mid x)$ (Theorem 1). However, we believe, based on simulations presented in Section 5, that they are minor in practice.

**4.2 Extensions**

Estimators for conditional mean squared prediction errors, conditional biases, and conditional prediction intervals follow immediately by plugging in $\hat{F}_E(e \mid x)$. Each is described in turn.

**Conditional Mean Squared Prediction Error**

We propose a plug-in estimator for the conditional mean squared prediction error $\text{MSPE}(x)$ obtained by averaging the squared out-of-bag prediction errors over $\hat{F}_E(e \mid x)$:

$$\hat{\text{MSPE}}(x) := \int e^2 \hat{f}_E(e \mid x) \, de = \sum_{i=1}^n v_i(x) \left( Y_i - \hat{\varphi}^{(t)}(X_i) \right)^2.$$ 

To our knowledge, no other method of estimating $\text{MSPE}(x)$ has been proposed; current implementations of random forests, such as the R package `randomForest` (Liaw and Wiener, 2002), generally estimate MSPE instead, using an unweighted average of the out-of-bag prediction errors. While MSPE can be an informative summary of the predictive performance of the random forest overall, $\text{MSPE}(x)$ is usually more appropriate for assessing the reliability of any individual prediction.

Figure 1 illustrates the distinction between $\hat{\text{MSPE}}$ and $\hat{\text{MSPE}}(x)$. To create this figure, we repeatedly drew 1,000 training observations $X \overset{i.i.d.}{\sim} \text{Unif}[-1, 1]^{10}$ with response $Y \overset{i.i.d.}{\sim} \mathcal{N}(10 \cdot 1(X_1 > 0), 1 + 2 \cdot 1(X_1 > 0))$. Note that, throughout this paper, we drop the subscript $i$ when discussing simulations for notational simplicity. For each draw, we fit a random forest to the training observations and predicted 500 test observations whose covariate...
values were fixed across the simulation repetitions but whose response values were randomly sampled from the same distribution as the training data. Figure 1 plots the average MSPE, the average $\hat{\text{MSPE}}(x)$ of each test point, and the actual average squared prediction error of each test point against the first covariate $X_1$. As expected, the true squared prediction error is greater for test observations with $X_1 > 0$. Our estimator $\hat{\text{MSPE}}(x)$ reflects this difference in prediction uncertainty, whereas $\hat{\text{MSPE}}$, while descriptive of overall prediction error, does not accurately assess the error one would expect from any individual prediction.

**Conditional Bias**

Our proposed plug-in estimator for the conditional bias is the average of the out-of-bag prediction errors over $\hat{F}_E(e \mid x)$:

$$\hat{\text{Bias}}(x) := \int e\hat{f}_E(e \mid x)\,de = \sum_{i=1}^n v_i(x) \left( Y_i - \hat{\varphi}^{(i)}(X_i) \right).$$

Thus, our bias-corrected random forest prediction at $x$ is given by

$$\hat{\varphi}^{\text{BC}}(x) := \hat{\varphi}(x) + \hat{\text{Bias}}(x).$$

We compare the empirical performance of $\hat{\varphi}^{\text{BC}}(x)$ to that of the boosting method investigated by Zhang and Lu (2012) and Ghosal and Hooker (2018) in Section 5.1.
Conditional Prediction Intervals and Response Quantiles

For a given type-I error rate $\alpha \in (0, 1)$, a conditional $\alpha$-level prediction interval $\text{PI}_\alpha(x)$ for the response at $x$ satisfies the inequality

$$\Pr(Y \in \text{PI}_\alpha(X) \mid X = x) \geq 1 - \alpha.$$  

We propose estimating a prediction interval $\text{PI}_\alpha(x)$ by adding the $\alpha/2$ and $1 - \alpha/2$ quantiles of $\hat{F}_E(e \mid x)$ to the random forest prediction at $x$:

$$\hat{\text{PI}}_\alpha(x) := \left[ \hat{\varphi}(x) + \hat{Q}^{\alpha/2}_E(x), \hat{\varphi}(x) + \hat{Q}^{1-\alpha/2}_E(x) \right],$$

where $\hat{Q}^\alpha_E(x) := \inf\{e : \hat{F}_E(e \mid x) \geq \alpha\}$. The bounds of $\hat{\text{PI}}_\alpha(x)$ correspond to plug-in estimates of the $\alpha/2$ and $1 - \alpha/2$ quantiles of the conditional response distribution at $x$; more generally, we propose estimating the $\alpha$-quantile of the conditional response distribution $Q^\alpha_Y(x) := \inf\{y : F_Y(y \mid x) \geq \alpha\}$ by the plug-in estimator

$$\hat{Q}^\alpha_Y(x) := \hat{\varphi}(x) + \hat{Q}^\alpha_E(x).$$

We compare the empirical performance of $\hat{\text{PI}}_\alpha(x)$ to those of prediction intervals obtained by quantile regression forests, generalized random forests, and conformal inference in Section 5.2.

Conditional Misclassification Rate for Categorical Outcomes

While this paper focuses on settings in which the response is continuous, it is worth noting that our proposed method of conditional mean squared prediction error estimation can be extended to random forest classification of categorical outcomes as well. In this setting, one common measure of predictive accuracy is the misclassification rate $\text{MCR} := \Pr(\hat{\varphi}(X) \neq Y)$. This is commonly estimated by

$$\widehat{\text{MCR}} := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(\hat{\varphi}^{(i)}(X_i) \neq Y_i).$$

By analogy to the continuous case discussed earlier, the conditional misclassification rate $\text{MCR}(x) := \Pr(\hat{\varphi}(X) \neq Y \mid X = x)$ is often a more informative metric, but, to our knowledge, no estimator for it has been introduced in the literature. We propose estimating this by

$$\widehat{\text{MCR}}(x) := \sum_{i=1}^{n} v_i(x) \mathbb{1}(\hat{\varphi}^{(i)}(X_i) \neq Y_i).$$

However, a detailed examination of this estimator and its behavior is beyond the scope of this paper.

5. Simulation Studies

In this section, we compare the empirical performance of our proposed bias correction and prediction intervals to existing methods reviewed in Section 2 across a variety of synthetic and benchmark datasets.
5.1 Conditional Bias Estimation

We compare our bias-corrected random forest $\hat{\phi}_{BC}(x)$ to the boosting-based bias correction discussed by Zhang and Lu (2012), who refer to it as “BC3,” and Ghosal and Hooker (2018). One metric for comparison is the mean squared bias,

$$\text{MSB} := \mathbb{E} [\text{Bias}(X)^2] ,$$

where the outer expectation is taken over the distribution of covariates; a lower value of MSB indicates a lower level of bias overall. Since correcting the bias of a prediction may increase the prediction variance, a second metric for comparison is the mean squared prediction error MSPE, which measures overall predictive accuracy, accounting for both bias and variance.

We tested each method on five synthetic datasets, in which the conditional means are known by design. In each dataset, the covariates were sampled as $X \overset{i.i.d.}\sim \text{Unif}[0,1]^{10}$. The responses were sampled as follows.

**Baseline:** $Y \overset{i.i.d.}\sim \mathcal{N}(0,1)$.

**Linear:** $Y \overset{\text{ind.}}{\sim} \mathcal{N}(X_1, 1)$.

**Step:** $Y \overset{\text{ind.}}{\sim} \mathcal{N}(10 \cdot \mathbb{1}(X_1 > 1/2), 1)$.

**Exponential:** $Y = \exp\{X_1 \epsilon\}$, where $\epsilon \overset{i.i.d.}\sim \mathcal{N}(0,1)$.

**Friedman:** $Y \overset{\text{ind.}}{\sim} \mathcal{N}\left(10 \sin(\pi X_1 X_2) + 20 (X_3 - 1/2)^2 + 10 X_4 + 5 X_5, 1\right)$ (Friedman, 1991).

In each repetition of the synthetic dataset simulations, we drew 200 training observations, following the simulation setup of Zhang and Lu (2012), and 2,000 test observations. We fit an uncorrected random forest and each bias-corrected predictor to the training set, and predicted the responses of the sampled test observations using each predictor. We then averaged the squared prediction errors; doing this repeatedly allowed us to estimate MSPE. In each repetition, we also predicted the responses of a held-out set of 2,000 observations whose covariate values were fixed over all repetitions. Averaging these predictions over the repetitions enabled us to estimate the mean prediction of the uncorrected random forest and each bias-corrected predictor at each of the fixed 2,000 points; we then combined this with the true conditional mean at each point, which we knew by design, to estimate MSB. We ran 1,000 repetitions for each synthetic dataset. We also assessed the MSPE of each estimator on the Boston Housing, Abalone, and Servo benchmark datasets via the above procedure, using the same train-test ratios as the simulations in Zhang and Lu (2012). These benchmark datasets were obtained through the UCI Machine Learning Repository and the MASS and mlbench packages in R (Dua and Graff, 2019; Leisch and Dimitriadou, 2010; Venables and Ripley, 2002).

Table 1 reports the results, and Figure 2 plots the conditional bias of each method against the signaling covariate(s). Overall, our bias-corrected predictor $\hat{\phi}_{BC}(x)$ appears to be more conservative but also more robust than the boosting approach. With respect to both MSB and MSPE, our bias correction generally improved upon but, at a minimum, did not much worse than the uncorrected random forest predictor. By comparison, the boosting approach...
sometimes improved bias more than $\hat{\varphi}^\text{BC}(x)$ did, but in other instances it had worse bias than even the uncorrected random forest predictor. Moreover, it sometimes reduced bias at the expense of greater variance, as reflected by the MSPE, less efficiently than $\hat{\varphi}^\text{BC}(x)$.

| Dataset     | MSB |         |         | MSPE |         |         |
|-------------|-----|---------|---------|------|---------|---------|
|             | RF  | Boost   | Error   | RF   | Boost   | Error   |
| Baseline    | 0.000 | 0.000 | 0.000   | 1.065 | 1.171 | 1.087   |
| Linear      | 0.008 | 0.001 | 0.003   | 1.075 | 1.174 | 1.097   |
| Step        | 0.837 | 0.190 | 0.240   | 2.015 | 1.512 | 1.460   |
| Exponential | 0.022 | 0.032 | 0.010   | 0.992 | 1.126 | 0.998   |
| Friedman    | 5.177 | 1.880 | 2.804   | 7.058 | 3.987 | 4.968   |
| Boston      | -    | -     | -       | 10.568 | 8.564 | 8.860   |
| Abalone     | -    | -     | -       | 4.566 | 4.812 | 4.628   |
| Servo       | -    | -     | -       | 50.946 | 27.022 | 34.714 |

Table 1: Mean squared bias and mean squared prediction error of the standard random forest predictor, the bias-corrected random forest predictor based on boosting, and our bias-corrected random forest predictor for each dataset.

### 5.2 Conditional Prediction Interval Estimation

Next, we compare our prediction interval estimator $\hat{\Pi}_\alpha(x)$ to the estimators obtained by quantile regression forests (Meinshausen, 2006), generalized random forests (Athey et al., 2019), and conformal inference (Lei and Wasserman, 2014; Lei et al., 2018; Johansson et al., 2014). For the conformal inference estimator, we specifically used the locally weighted split conformal inference procedure proposed by Lei et al. (2018), which is a special case of the standard conformal inference approach that attempts to account for residual heterogeneity across the predictor space by using standardized residuals for the conformity scores. Via simulation, we evaluate these methods with respect to three metrics: coverage rate, interval width, and qualitative behavior. In each simulation, we randomly sampled 1,000 training units and 1,000 test units, and applied each method to construct prediction intervals for the test units. We repeated this procedure 1,000 times for each of the following datasets.

**Linear:** $X \overset{\text{i.i.d.}}{\sim} \text{Unif}[-1, 1]$, and $Y \overset{\text{ind.}}{\sim} \mathcal{N}(X_1, 4)$.

**Clustered:** $X \in [0, 1]^{10}$ is drawn i.i.d. from a population consisting of five distinct clusters, with no overlap between clusters. $Y$ is independently drawn from a normal distribution with mean and variance determined by the cluster to which $X$ belongs. The response means and variances within the clusters are $\{(0, 1), (40, 4), (80, 9), (120, 16), (160, 25)\}$. See Maitra and Melnykov (2010) for details. The data were generated using the `MixSim` package (Melnykov et al., 2012).

**Step:** With probability 0.05, $X \overset{\text{i.i.d.}}{\sim} \text{Unif}([-1, 0] \times [-1, 1]^{9})$; else, $X \overset{\text{i.i.d.}}{\sim} \text{Unif}([0, 1] \times [-1, 1]^{9})$. $Y \overset{\text{ind.}}{\sim} \mathcal{N}(20 \cdot 1(X_1 > 0), 4)$.  


Figure 2: Conditional bias of each method for the Step, Exponential, Linear, and Friedman datasets (clockwise from top left).
Friedman: \(X \overset{\text{i.i.d.}}{\sim} \text{Unif}[-1, 1]^{10}\), and \(Y \overset{\text{ind.}}{\sim} \mathcal{N}\left(10 \sin(\pi X_1 X_2) + 20 (X_3 - 1/2)^2 + 10X_4 + 5X_5, 1\right)\).

Parabola: With probability 0.05, \(X \overset{\text{i.i.d.}}{\sim} \text{Unif}([-1, -1/3] \times [-1, 1]^9)\); with probability 0.9, \(X \overset{\text{i.i.d.}}{\sim} \text{Unif}([-1/3, 1/3] \times [-1, 1]^9)\); and with probability 0.05, \(X \overset{\text{i.i.d.}}{\sim} \text{Unif}([1/3, 1] \times [-1, 1]^9)\). \(Y \overset{\text{ind.}}{\sim} \mathcal{N}(0, X_4^2)\).

In addition to conducting simulations on the above synthetic datasets, we also randomly partitioned each of the Boston, Abalone, and Servo benchmark datasets into training and test sets using the same train-test ratios as in Section 5.1 and estimated prediction intervals for the test points; we repeated this 1,000 times for each of the three benchmark datasets.

Table 2 shows the average coverage rate of each method in each simulation, with average interval widths shown in parentheses. Overall, all four methods performed fairly well with respect to these two metrics. However, it is notable that, in the Clustered dataset, generalized random forests generated extremely wide prediction intervals, and the prediction intervals of quantile regression forests were wider than \(\hat{\Pi}_\alpha(x)\) despite having a lower coverage rate. Additionally, all of the methods other than conformal inference over-covered in the Friedman, Parabola, Boston, and Servo datasets; unlike quantile regression forests and generalized random forests, however, our \(\hat{\Pi}_\alpha(x)\) intervals were still narrower than the conformal inference intervals despite over-covering in the Friedman, Boston, and Servo datasets. On the other hand, the \(\hat{\Pi}_\alpha(x)\) intervals were wider than generalized random forests despite having a lower coverage rate in the Parabola dataset (see Figure 3).

| Dataset | \(1 - \alpha\) | QRF | GRF | Split | Error |
|---------|---------------|-----|-----|-------|-------|
| Linear  | 0.8           | 0.805 (5.28) | 0.806 (5.32) | 0.801 (5.34) | 0.799 (5.21) |
| Clustered | 0.95       | 0.930 (15.45) | 0.966 (46.07) | 0.950 (15.28) | 0.943 (13.66) |
| Step    | 0.8           | 0.798 (5.49) | 0.812 (6.41) | 0.801 (5.78) | 0.797 (5.33) |
| Friedman | 0.8          | 0.927 (22.50) | 0.939 (28.94) | 0.799 (16.49) | 0.857 (16.13) |
| Parabola | 0.8         | 0.830 (0.20) | 0.854 (0.17) | 0.800 (0.18) | 0.845 (0.22) |
| Boston  | 0.8           | 0.908 (9.37) | 0.950 (13.74) | 0.801 (7.59) | 0.814 (6.38) |
| Abalone | 0.8           | 0.883 (4.81) | 0.906 (5.38) | 0.801 (4.78) | 0.796 (4.70) |
| Servo   | 0.8           | 0.887 (16.04) | 0.909 (24.39) | 0.806 (15.39) | 0.868 (12.30) |

Table 2: Average coverage rate and width of prediction intervals constructed by quantile regression forests, generalized random forests, split conformal inference, and \(\hat{\Pi}_\alpha(x)\).

The second column denotes the desired type-I error rate.

Table 2, however, does not tell the full story. Figure 3 plots the estimated conditional response quantiles against the true conditional quantiles for the Linear, Clustered, Step, and Parabola datasets. Overall, \(\hat{\Pi}_\alpha(x)\) clearly captured the nuances in the structure of the data better than the other estimators; in all four datasets, \(\hat{\Pi}_\alpha(x)\) best tracked the changes in the conditional quantiles across the predictor space, with conformal inference performing equally well in the Clustered dataset. Only \(\hat{\Pi}_\alpha(x)\) correctly estimated the upper quantile when \(X_1 < 0\) in the Step dataset. Moreover, conformal inference and generalized random forests did not capture the strong curvature of the quantiles in the Parabola dataset, and
Figure 3: Average upper and lower bounds of prediction intervals generated by our method, generalized random forests, quantile regression forests, and split conformal inference for the Linear, Step, Parabola, and Clustered datasets (clockwise from top left) over 1,000 simulation repetitions. The true target conditional response quantiles are shown in black.

Quantile regression forests and generalized random forests produced erratic intervals in the Clustered dataset. Additionally, while all methods performed fairly well in the Linear dataset, ours exhibited the least bias at the boundaries of the covariate space.

We believe that at least some of the undesirable behaviors exhibited by generalized random forests, quantile regression forests, and conformal inference in these simulations can be attributed to the methods’ direct use of the empirical distribution of covariates or responses in the training set. For example, the quantile regression variant of generalized random forests partitions the covariate space based on the observed quantiles of the training responses, so they are less able to detect changes in the conditional response quantiles in low-density regions of the covariate space; this can be seen in the Step and Parabola datasets. Similarly, conformal inference directly uses the empirical distribution of conformity scores,
weighting each score equally rather than giving greater weight to scores from training units that more closely resemble the test point of interest; this can also be seen in the Step and Parabola datasets. Finally, quantile regression forests directly use training responses to impute the conditional response distribution, so their prediction intervals are more sensitive to sharp discontinuities in the conditional response distribution that are not detected well by the CART algorithm; this can be seen in the Step and Clustered datasets.

Our method of prediction interval estimation takes a different approach. Rather than directly using the responses of training units, as quantile regression forests do, our method uses their out-of-bag prediction errors, thus more fully leveraging the predictive power of the random forest. Additionally, our method weights these errors by how closely the training units resemble the test point of interest rather than directly using an empirical distribution, as generalized random forests and conformal inference do.

Of course, this is not to say that our method is uniformly better than the others. For example, although they may not properly estimate the conditional response quantiles in some settings, conformal inference prediction intervals are guaranteed to achieve the desired coverage rate even in finite samples. Additionally, Athey et al. (2019) show that generalized random forests outperform quantile regression forests when the response variance follows a step function across the predictor space but the mean response is constant because quantile regression forests grow trees using the CART algorithm, which is sensitive only to mean shifts. Since our method as discussed in this paper also employs the CART algorithm, it also inherits this limitation; it is worth noting, however, that we believe that our method can be generalized to other tree construction algorithms that do not suffer from this phenomenon.

6. Theoretical Result

In this section, we propose a similar but more stringent estimator of $F_E(e \mid x)$ and prove that it is uniformly consistent. In addition to regularity assumptions on the tree construction procedure that have become somewhat standard in recent literature, the most notable difference between $\hat{F}_E(e \mid x)$ as defined in (1) and the estimator discussed here is that the latter uses two random forests fit on disjoint subsets of the training data; one random forest produces weights, and the other produces out-of-sample prediction errors. We view this as a strengthening of the independence conditions that motivated the out-of-bag construction of $\hat{F}_E(e \mid x)$.

While we do not prove that our practical estimator $\hat{F}_E(e \mid x)$ is uniformly consistent, we nonetheless consider it noteworthy for two reasons. From an applied perspective, it is more data-efficient than our stringent estimator and thus more reasonable for applications with only modest sample sizes. From a theoretical perspective, the contrast between $\hat{F}_E(e \mid x)$ and the estimator introduced in this section highlights aspects of the dependence structure of random forests that we believe merit further investigation. Future research into such topics may close the gap between the two versions of our estimator and contribute more generally to a deeper understanding of tree-based algorithms.

6.1 Stringent Estimator of the Conditional Prediction Error Distribution

The algorithm for computing our more stringent estimator of $F_E(e \mid x)$ is outlined below. Note that, in what follows, we redefine some earlier notation rather than introduce new
symbols in order to reduce notational complexity; an effort has been made to be explicit whenever notation is redefined.

1. Partition the training data evenly into three subsets, arbitrarily labeled $I$, $J$, and $K$. Let $n$ denote the sample size of each subset, as opposed to the sample size of the full training set.

2. Grow one random forest with $B$ trees on $I$; label it the “first random forest.”

3. Grow one random forest with $B$ trees using $J$ and the covariates in $K$; in other words, do not consider the responses of the units in $K$ when splitting tree nodes. This is similar to the honest double-sample regression tree algorithm of Wager and Athey (2018), but here the training data were split into subsets before resampling or subsampling. Label this random forest the “second random forest.”

4. Compute the errors of the first random forest’s predictions of the $n$ units in $K$:
   \[
   E_i := Y_i - \hat{\varphi}(X_i),
   \]
   where $\hat{\varphi}$ denotes the first random forest estimator.

5. For a target $x$, compute the weight of each of the $n$ units in $K$ given by the second random forest:
   \[
   v_i(x) := \frac{1}{B} \sum_{b=1}^{B} \frac{\# \{ Z_i \in D_{n,b}^* : X_i \in R_{\ell(x, \theta_b)} \}}{\sum_{j=1}^{n} \# \{ Z_j \in D_{n,b}^* : X_j \in R_{\ell(x, \theta_b)} \}},
   \]
   where $D_{n,b}^*$ denotes the bootstrap set of units from $K$ whose covariates were used in the construction of the $b^{th}$ tree of the second random forest, and $R_{\ell(x, \theta_b)}$ denotes the rectangular subspace corresponding to the terminal node of the $b^{th}$ tree of the second random forest in which $x$ falls.

6. Letting $\hat{F}_E(e \mid x)$ now denote our stringent estimator rather than the estimator given by (1) in Section 4, define
   \[
   \hat{F}_E(e \mid x) := \sum_{i=1}^{n} v_i(x) \mathbb{1}(E_i \leq e), \quad e \in \mathbb{R}. \tag{2}
   \]

6.2 Consistency

We prove consistency of $\hat{F}_E(e \mid x)$ as defined in (2) under the following set of assumptions, many of which are from Meinshausen (2006). First, we make an assumption about the covariate distribution.

**Assumption 1.** $X$ has the uniform distribution over $[0, 1]^p$.

Assumption 1 is largely for notational convenience. More generally, one could assume that the density of $X$ is positive and bounded.

We also make a set of assumptions on the way the observations in $K$ are used in the construction of the second random forest. For a tree in the second random forest grown with parameter vector $\theta$, let $k_\theta(\ell) := \# \{ Z_i \in D_n^* : X_i \in R_{\ell(x, \theta)} \}$ denote the number of units from the bootstrap sample $D_n^*$ of $K$ in the terminal node containing $x$. 

15
Assumption 2.

(a) The proportion of observations from $D_n^*$ in any given node, relative to all observations from $D_n^*$, is decreasing in $n$—that is, $\max_{\ell, \theta} k_\theta(\ell) = o(n)$. The minimum number of observations from $D_n^*$ in a node is increasing in $n$—that is, $1/\min_{\ell, \theta} k_\theta(\ell) = o(1)$.

(b) The probability that variable $m \in \{1, \ldots, p\}$ is chosen for a given split point is bounded from below for every node by a positive constant.

(c) When a node is split, the proportion of observations belonging to $D_n^*$ in the original node that fall into each of the resulting sub-nodes is bounded from below by a positive constant.

The conditions given by Assumption 2 are adapted from assumptions used to prove consistency of quantile regression forests (Meinshausen, 2006). Tree construction algorithms that satisfy these properties or variants of them have been referred to in recent random forest literature as “regular,” “balanced,” or “random-split” (Wager and Athey, 2018; Athey et al., 2019; Friedberg et al., 2019).

Next, we assume that the distribution of prediction errors is sufficiently smooth.

Assumption 3. $F_E(e \mid X = x)$ is Lipschitz continuous with parameter $L$. That is, for all $x, x' \in [0, 1]^p$,

$$\sup_e |F_E(e \mid X = x) - F_E(e \mid X = x')| \leq L\|x - x'\|_1.$$ 

As Wager and Athey (2018) note, all existing results on pointwise consistency of random forests have required an analogous smoothness condition in the distribution of interest, including Biau (2012), Meinshausen (2006), and Wager and Athey (2018).

Additionally, we assume that the distribution of prediction errors is strictly monotone so that consistency of quantile estimates follows from consistency of distribution estimates.

Assumption 4. $F_E(e \mid X = x)$ is strictly monotone in $e$ for all $x \in [0, 1]^p$.

We also assume that the random forest is stable in the following sense.

Assumption 5. There exists $B_0 \in \mathbb{N}$ such that the predicted response $\hat{\varphi}(X)$ by a random forest with $B \geq B_0$ trees satisfies $\hat{\varphi}(X) \xrightarrow{p} \varphi(X)$ as $n \to \infty$, where $\varphi(X)$ is a random variable that depends only on $X$ and satisfies $-\infty < \varphi(X) < \infty$ a.s.

It may help one’s intuition to imagine that $\varphi(X) = \mathbb{E}[Y \mid X]$, in which case Assumption 5 simply states that the random forest is consistent; however, $\varphi(X)$ need not be the conditional mean response. Note also that Assumption 5 does not require stability as defined by Bühlmann and Yu (2002), since here the convergence does not have to be pointwise. Stability—and, in particular, consistency—of random forests is an ongoing area of research. Scornet et al. (2015) prove consistency of the original random forest algorithm of Breiman (2001) when the underlying data follow an additive regression model. Wager and Walther (2016) prove consistency of adaptively grown random forests, including forests built using CART-like algorithms, in high-dimensional settings.

Finally, we make an assumption about the behavior of the weights given by the second random forest relative to the predictions of the first random forest. For any $\delta > 0$, define the event $\mathcal{M}_i(\delta) := \{ |\hat{\varphi}(X_i) - \varphi(X_i)| < \delta \}$. 

Assumption 6. For all \( x \in [0,1]^p \), there exists \( \delta_0 > 0 \) such that, for any \( \delta \in (0, \delta_0) \),
\[
\mathbb{E}[v_i(x) | \mathcal{M}_i(\delta)] = O(n^{-1}) \quad \text{and} \quad \mathbb{E}[v_i(x) | \neg \mathcal{M}_i(\delta)] = O(n^{-1}).
\]

Assumption 6 further characterizes the stability of the random forest and the underlying population distribution. It states that the expected weight of an observation in \( K \) is of order \( 1/n \) whether stability has been realized for the observation or not. Notice that Assumption 6 is satisfied if \( \mathbb{E}[v_i(x) | \mathcal{M}_i(\delta)] > \mathbb{E}[v_i(x) | \neg \mathcal{M}_i(\delta)] \) for all \( n \) and Assumption 5 holds since the weights must be nonnegative and \( \mathbb{E}[v_i(x)] = 1/n \). Note also that the bounding constant can vary by \( \delta \in (0, \delta_0) \).

Under these assumptions, we prove that \( \hat{F}_E(e | x) \) is a uniformly consistent estimator for the true conditional prediction error distribution \( F_E(e | x) \).

**Theorem 1.** Under Assumptions 1-6,
\[
\sup_{e \in \mathbb{R}} \left| \hat{F}_E(e | x) - F_E(e | x) \right| \overset{p}{\to} 0, \quad n \to \infty
\]
pointwise for every \( x \in [0,1]^p \) and \( B \geq B_0 \).

The proof is deferred to the appendix, but we outline the broad steps here to highlight the role of our assumptions and algorithm for \( \hat{F}_E(e | x) \).

**Proof Outline**

1. Following Meinshausen (2006), we break the absolute difference into one term reflecting heterogeneity in the conditional prediction error distribution across the covariate space and one term resembling a variance-type quantity; the remainder of the proof shows that both terms converge to zero in probability.

2. We show that the expected value of both terms is zero. In this step, we rely on the fact that \( v_i(x) \) and \( \bar{v}_i \) are independent conditionally on \( X_i \) because we use separate random forests for weighting and prediction, with both the weights and the predictions computed independently of the responses of the units in \( K \).

3. We show that the variance of both terms converges to zero. Here, the main obstacle is the covariances between the weighted indicator terms in (2); although our algorithm for \( \hat{F}_E(e | x) \) removes some dependencies by construction, each term in (2) still depends on the same two random forests. To show that the covariances nonetheless vanish, we condition on the stability of the random forest predictions to substitute \( \hat{\varphi}(X_i) \), which depends on the first random forest, for \( \varphi(X_i) \), which depends only on \( X_i \). This removes enough of the remaining dependence structure to reduce the covariances. We repeatedly invoke Assumptions 2, 5, and 6, as well as the continuity of the conditional response distribution, to eliminate leftover terms that arise throughout this step. We also use Assumption 3 and Lemma 2 of Meinshausen (2006) to address the heterogeneity of the conditional prediction error distribution.

7. **Conclusion**

We propose a unified framework for random forest prediction error estimation based on a novel estimator for the conditional prediction error distribution. Under this framework,
useful parameters and measures of uncertainty can be estimated by simply plugging in the estimated conditional prediction error distribution. By contrast, these quantities previously each had to be estimated by different, and in some cases not obviously compatible, algorithms. We demonstrate the unified nature of our approach by deriving, to our knowledge, the first estimator for the conditional mean squared prediction error of random forests, as well as estimators for conditional bias and conditional prediction intervals that are competitive with, and in some cases outperform, existing methods.

We believe that one advantage of our framework is its generality. While this paper discusses our work in the context of CART, our estimators can be readily adapted to other bagged, tree-based estimators with different splitting criteria and subsampling rules. More broadly, we believe that our general approach of weighting out-of-sample prediction errors by their similarity to the test point of interest with respect to the estimator is applicable to a wide range of estimators with suitably defined metrics for similarity, even those not based on decision trees. While beyond the scope of this paper, future work into such extensions may prove fruitful.

Acknowledgments

The authors gratefully acknowledge the Pomona College Summer Undergraduate Research Program and the Pomona College Kenneth Cooke Summer Research Fund for their support of this research. This material is based upon work supported by the National Science Foundation under Grant No. 1745640.

References

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

Gérard Biau. Analysis of a Random Forests Model. *Journal of Machine Learning Research*, 13:1063–1095, 2012.

Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of Random Forests and Other Averaging Classifiers. *Journal of Machine Learning Research*, 9:2015–2033, 2008.

Leo Breiman. Out-of-bag Estimation. Technical report, University of California, Berkeley, Dept. of Statistics, 1996.

Leo Breiman. Using Adaptive Bagging to Debias Regressions. Technical report, University of California, Berkeley, Dept. of Statistics, 1999.

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

Leo Breiman. Manual on Setting up, Using, and Understanding Random Forests. Technical report, University of California, Berkeley, Dept. of Statistics, 2002.

Peter Bühlmann and Bin Yu. Analyzing Bagging. *Annals of Statistics*, 30(4):927–961, 2002.
Random Forest Prediction Error Estimation

Antonio Criminisi, Jamie Shotton, Duncan Robertson, and Ender Konukoglu. Regression Forests for Efficient Anatomy Detection and Localization in CT Studies. *International MICCAI Workshop on Medical Computer Vision*, pages 106–117, 2010.

Dheeru Dua and Casey Graff. UCI Machine Learning Repository, 2019. URL http://archive.ics.uci.edu/ml.

Yun Fang, Peirong Xu, Jialiang Yang, and Yufang Qin. A Quantile Regression Forest Based Method to Predict Drug Response and Assess Prediction Reliability. *PLoS ONE*, 13(10): e0205155, 2018.

Rina Friedberg, Julie Tibshirani, Susan Athey, and Stefan Wager. Local Linear Forests. *arXiv preprint arXiv:1807.11408v3*, 2019.

Jerome H. Friedman. Multivariate Adaptive Regression Splines. *Annals of Statistics*, 19(1):1–67, 1991.

Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5):11891232, 2001.

Indrayudh Ghosal and Giles Hooker. Boosting Random Forests to Reduce Bias; One-step Boosted Forest and its Variance Estimate. *arXiv preprint arXiv:1803.08000v2*, 2018.

R. Grimm, T. Behrens, M. Märker, and H. Elsenbeer. Soil Organic Carbon Concentrations and Stocks on Barro Colorado Island—Digital Soil Mapping Using Random Forests Analysis. *Geoderma*, 146:102–113, 2008.

Giles Hooker and Lucas Mentch. Bootstrap Bias Corrections for Ensemble Methods. *Statistics and Computing*, 28(1):77–86, 2018.

Ulf Johansson, Henrik Boström, Tuve Löfström, and Henrik Linusson. Regression Conformal Prediction With Random Forests. *Machine Learning*, 97:155–176, 2014.

Jing Lei and Larry Wasserman. Distribution-free Prediction Bands for Non-parametric Regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:71–96, 2014.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113:1094–1111, 2018.

Friedrich Leisch and Evgenia Dimitriadou. mlbench: Machine Learning Benchmark Problems. 2010. R package version 2.1-1.

Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002.

Ranjan Maitra and Volodymyr Melnykov. Simulating Data to Study Performance of Finite Mixture Modeling and Clustering Algorithms. *Journal of Computational and Graphical Statistics*, 19(2):354–376, 2010.
Nicolai Meinshausen. Quantile Regression Forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

Volodymyr Melnykov, Wei-Chen Chen, and Ranjan Maitra. MixSim: An R Package for Simulating Data to Study Performance of Clustering Algorithms. *Journal of Statistical Software*, 51(12):1–25, 2012.

Lucas Mentch and Giles Hooker. Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. *Journal of Machine Learning Research*, 17:1–41, 2016.

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of Random Forests. *Annals of Statistics*, 43(4):1716–1741, 2015.

Joseph Sexton and Petter Laake. Standard Errors for Bagged and Random Forest Estimators. *Computational Statistics & Data Analysis*, 53(3):801–811, 2009.

Benjamin M. Van Doren and Kyle G. Horton. A Continental System for Forecasting Bird Migration. *Science*, 361:1115–1118, 2018.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics With S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

Stefan Wager and Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, 113:1228–1242, 2018.

Stefan Wager and Guenther Walther. Adaptive Concentration of Regression Trees, With Application to Random Forests. *arXiv preprint arXiv:1503.06388v3*, 2016.

Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife. *Journal of Machine Learning Research*, 15:1625–1651, May 2014.

Chih-Lin Wei, Gilbert T. Rowe, Elva Escobar-Briones, Antje Boetius, Thomas Soltwedel, M. Julian Caley, Yousria Soliman, Falk Huettmann, Fangyuan Qu, Zishan Yu, C. Roland Pitcher, Richard L. Haedrich, Mary K. Wicksten, Michael A. Rex, Jeffrey G. Baguley, Jyotsna Sharma, Roberto Danovaro, Ian R. MacDonald, Clifton C. Nunnally, Jody W. Deming, Paul Montagna, Mélanie Lévesque, Jan Marcin Weslawski, Maria Wlodarska-Kowalczyk, Baban S. Ingole, Brian J. Bett, David S. M. Billett, Andrew Yool, Bodil A. Bluhm, Katrin Iken, and Bhavani E. Narayanaswamy. Global Patterns and Predictions of Seafloor Biomass Using Random Forests,. *PLoS ONE*, 5(12):e15323, 2010.

Guoyi Zhang and Yan Lu. Bias-corrected Random Forests in Regression. *Journal of Applied Statistics*, 39(1):151–160, 2012.
Appendix

Proof of Proposition 1. Fix \( x \in \mathcal{X} \). Since \( \theta_1, \ldots, \theta_B \) are i.i.d. and independent of \( D_n \), the weak law of large numbers implies that, conditionally on \( D_n \),

\[
\frac{1}{B} \sum_{b=1}^{B} \sum_{j=1}^{n} w_j(x, \theta_b) Y_j \overset{p}{\to} \mu := \mathbb{E} \left[ \sum_{j=1}^{n} w_j(x, \theta) Y_j \mid D_n \right], \quad B \to \infty,
\]

where the expectation is taken over \( \theta \). Then, by Slutsky’s Theorem, \( E \mid x \overset{d}{\to} \mu - Y \). An analogous argument implies that \( E^* \mid x \overset{d}{\to} \mu - Y \) as well. Recall that (pointwise) convergence in distribution implies uniform convergence for continuous cumulative distribution functions. Since \( F_Y(\cdot \mid x) \) is continuous by assumption, \( F_{\mu - Y}(\cdot \mid x) \) is continuous as well. Therefore,

\[
\lim_{B \to \infty} \sup_{e \in \mathbb{R}} |F_{E^*}(e \mid x) - F_E(e \mid x)| \leq \lim_{B \to \infty} \sup_{e \in \mathbb{R}} |(F_{E^*}(e \mid x) - F_{\mu - Y}(e \mid x)) + (F_E(e \mid x) - F_{\mu - Y}(e \mid x))| \leq \lim_{B \to \infty} \sup_{e \in \mathbb{R}} |F_{E^*}(e \mid x) - F_{\mu - Y}(e \mid x)| + \lim_{B \to \infty} \sup_{e \in \mathbb{R}} |F_E(e \mid x) - F_{\mu - Y}(e \mid x)| = 0,
\]

which completes the proof.

Before proving Theorem 1, we first establish additional notation and a useful lemma. First, let \( \Omega_v \) denote the set of observations—\( \mathcal{J} \cup \mathcal{K} \)—and parameters that fully define the second random forest (Step 3 of Section 6.1). Second, let \( M_n \) denote the maximum possible value of \( v_i(x) \), which is decreasing in \( n \) by Assumption 2. Finally, for any \( \delta > 0 \), let \( \gamma_n := \text{Pr}(\neg M_i(\delta)) \) denote the probability that stability is not realized for the first random forest’s prediction of the \( i \)th training unit in \( \mathcal{K} \), which is decreasing in \( n \) by Assumption 5. Having defined the necessary notation for the remainder of the proofs, we now proceed to the lemma.

Lemma 1. Under Assumptions 1-6, we have the following asymptotic results for any \( \delta \in (0, \delta_0) \):

\[
\gamma_n \sum_{i=1}^{n} \mathbb{E} [v_i(x) \mid \mathcal{M}_i(\delta)] \to 0, \quad n \to \infty;
\]

\[
\gamma_n \sum_{i=1}^{n} \mathbb{E} [v_i(x) \mid \neg \mathcal{M}_i(\delta)] \to 0, \quad n \to \infty; \text{ and}
\]

\[
\sum_{i=1}^{n} \mathbb{E} [v_i(x)^2 \mid \mathcal{M}_i(\delta)] \to 0, \quad n \to \infty.
\]
**Proof of Lemma 1.** Fix $\delta \in (0, \delta_0)$ and let $\epsilon > 0$. Assumption 6 implies that there exists $K > 0$ and $N_1 \in \mathbb{N}$ so that, for $n \geq N_1$, $\mathbb{E} [v_i(x) \mid \mathcal{M}_i(\delta)] \leq K/n$ and $\mathbb{E} [v_i(x) \mid -\mathcal{M}_i(\delta)] \leq K/n$. Fix that value of $K$. Since the random forest is stable by Assumption 5, there exists $N_2 \in \mathbb{N}$ so that, for $n \geq N_2$, $\gamma_n < \epsilon/K$. Moreover, by Assumption 2, the minimum number of observations in each node is growing, so the maximum possible weight $M_n$ given to any one unit is decreasing in $n$. Thus, there exists $N_3 \in \mathbb{N}$ so that, for $n \geq N_3$, $M_n < \epsilon/K$. Therefore, for $n \geq \max\{N_1, N_2, N_3\}$,

$$
\gamma_n \sum_{i=1}^{n} \mathbb{E} [v_i(x) \mid \mathcal{M}_i(\delta)] < \frac{\epsilon}{K} \sum_{i=1}^{n} \frac{K}{n} = \epsilon,
$$

$$\gamma_n \sum_{i=1}^{n} \mathbb{E} [v_i(x) \mid -\mathcal{M}_i(\delta)] < \frac{\epsilon}{K} \sum_{i=1}^{n} \frac{K}{n} = \epsilon,$n

and

$$\sum_{i=1}^{n} \mathbb{E} [v_i(x)^2 \mid \mathcal{M}_i(\delta)] \leq M_n \sum_{i=1}^{n} \mathbb{E} [v_i(x) \mid \mathcal{M}_i(\delta)] < \frac{\epsilon}{K} \sum_{i=1}^{n} \frac{K}{n} = \epsilon.$$

This completes the proof. \hfill \Box

We now prove Theorem 1.

**Proof of Theorem 1.** Fix $x \in [0, 1]^p$ and $B \geq B_0$. Let the random variables $U_i, i = 1, \ldots, n$, be defined as the quantiles of $E_i$ given $X_i$:

$$U_i := F_E(E_i \mid X_i).$$

Notice that, since $E_i$ follows the distribution of $E \mid X_i$, $U_i \sim \text{Unif}[0, 1]$. Additionally, Assumption 4 implies that the event $\{E_i \leq e\}$ is equivalent to the event $\{U_i \leq F_E(e \mid X_i)\}$. Using this equivalence, we have

$$\hat{F}_E(e \mid x) = \sum_{i=1}^{n} v_i(x)\mathbb{1}(E_i \leq e)$$

$$= \sum_{i=1}^{n} v_i(x)\mathbb{1}(U_i \leq F_E(e \mid X_i))$$

$$= \sum_{i=1}^{n} v_i(x)\mathbb{1}(U_i \leq F_E(e \mid x)) + \sum_{i=1}^{n} v_i(x)(\mathbb{1}(U_i \leq F_E(e \mid X_i)) - \mathbb{1}(U_i \leq F_E(e \mid x))),$$

so $|\hat{F}_E(e \mid x) - F_E(e \mid x)|$ is bounded above by

$$|\hat{F}_E(e \mid x) - F_E(e \mid x)| \leq \sum_{i=1}^{n} v_i(x)\mathbb{1}(U_i \leq F_E(e \mid x)) - F_E(e \mid x)$$

$$+ \sum_{i=1}^{n} v_i(x)(\mathbb{1}(U_i \leq F_E(e \mid X_i)) - \mathbb{1}(U_i \leq F_E(e \mid x))).$$

As mentioned in Meinshausen (2006), the first term on the right side of (3) can be thought of as a variance-type term, while the second term can be thought of as reflecting the shift in the underlying error distribution across the covariate space. In the next two subsections, we show that each term converges to zero in probability.
Bounding the Variance Term

Taking the supremum over \( e \) in the first term on the right side of (3) yields

\[
\sup_{e \in \mathbb{R}} \left| \sum_{i=1}^{n} v_i(x) 1(U_i \leq F_E(e \mid x)) - F_E(e \mid x) \right| = \sup_{z \in [0,1]} \left| \sum_{i=1}^{n} v_i(x) 1(U_i \leq z) - z \right|.
\]

It suffices to prove that, for all \( z \in [0,1] \),

\[
\left| \sum_{i=1}^{n} v_i(x) 1(U_i \leq z) - z \right| = o_p(1). \tag{4}
\]

Since the forest used for the weights \( v_i(x) \) is built on a separate subset of the training data from the forest used for the predictions \( \hat{\varphi}(X_i) \), and the prediction of the \( i \)th observation in \( \mathcal{K} \) does not depend on the other \( n - 1 \) observations in \( \mathcal{K} \), conditioning on \( X_i \) yields the necessary independence to evaluate the expectation of the weighted average inside (4):

\[
\mathbb{E} \sum_{i=1}^{n} v_i(x) 1(U_i \leq z) = \sum_{i=1}^{n} \mathbb{E} \left[ \mathbb{E} [v_i(x) 1(U_i \leq z) \mid X_i] \mid X_i \right] = z \mathbb{E} \sum_{i=1}^{n} v_i(x) = z. \tag{5}
\]

The remainder of this subsection is devoted to showing that the variance of the weighted average in (4) converges to zero. Since the variance of a summation is equal to the summation of the covariances, we have

\[
\text{Var} \left( \sum_{i=1}^{n} v_i(x) 1(U_i \leq z) \right) = \sum_{i=1}^{n} \text{Var}(v_i(x) 1(U_i \leq z)) + \sum_{i \neq j} \text{Cov}(v_i(x) 1(U_i \leq z), v_j(x) 1(U_j \leq z)). \tag{6}
\]

We first show that the sum of variances on the right side of (6) converges to zero.
SUM OF VARIANCES

By the law of total variance,
\[
\sum_{i=1}^{n} \text{Var}(v_i(x) 1(U_i \leq z)) \\
= \sum_{i=1}^{n} \text{Var}(E[v_i(x) 1(U_i \leq z) \mid \Omega_v \{Y_i\}]) + E[\text{Var}(v_i(x) 1(U_i \leq z) \mid \Omega_v \{Y_i\})] \\
= \sum_{i=1}^{n} \text{Var}(v_i(x) \text{Pr}(U_i \leq z \mid X_i)) + E[v_i(x)^2 \text{Var}(1(U_i \leq z) \mid X_i)] \\
= \sum_{i=1}^{n} z^2 \text{Var}(v_i(x)) + z(1-z)E[v_i(x)^2] \\
\leq \sum_{i=1}^{n} \text{Var}(v_i(x)) + E[v_i(x)^2].
\] (7)

Notice that
\[
\sum_{i=1}^{n} \text{Var}(v_i(x)) \leq \sum_{i=1}^{n} E[v_i(x)^2] \leq M_n \sum_{i=1}^{n} v_i(x) = M_n.
\] (8)

Since the minimum number of observations in each node is growing by Assumption 2, the maximum possible weight given to any observation is decreasing in \(n\)—that is, \(M_n \rightarrow 0\).

Thus, plugging the bound given by (8) into (7) yields the desired result:
\[
\lim_{n \rightarrow \infty} \sum_{i=1}^{n} \text{Var}(v_i(x) 1(U_i \leq z)) \leq \lim_{n \rightarrow \infty} 2M_n = 0.
\]

SUM OF COVARIANCES

Next, we show that the covariance term in (6) converges to zero as well. To better define our objective, observe that, by definition of covariance,
\[
\sum_{i \neq j} \text{Cov}(v_i(x) 1(U_i \leq z), v_j(x) 1(U_j \leq z)) \\
= \sum_{i \neq j} E[v_i(x)v_j(x) 1(U_i \leq z) 1(U_j \leq z)] - E[v_i(x) 1(U_i \leq z)]E[v_j(x) 1(U_j \leq z)] \\
\rightarrow \sum_{i \neq j} E[v_i(x)v_j(x) 1(U_i \leq z) 1(U_j \leq z)] - z^2,
\]
so we want to show that
\[
\lim_{n \rightarrow \infty} \left| \sum_{i \neq j} E[v_i(x)v_j(x) 1(U_i \leq z) 1(U_j \leq z)] - z^2 \right| = 0.
\]

Let \(\epsilon > 0\). By uniform continuity of the conditional response distribution, there exists \(\delta_1 > 0\) so that
\[
|y_1 - y_2| < \delta_1 \implies |F_Y(y_1 \mid x) - F_Y(y_2 \mid x)| < \epsilon/3.
\] (9)
Fix $\delta < \min \{\delta_0, \delta_1\}$. Then, by Lemma 1, there exists $N \in \mathbb{N}$ so that, for $n \geq N$, the leftover terms in the following steps of the proof sum to at most $\epsilon/3$; for concision, we note these terms where they appear and then cite Lemma 1 to drop them.

**Conditioning on Stability**

We use the law of total expectation to condition on stability of the random forest prediction of each observation, noting that leftover terms converge to zero by Lemma 1:

$$
\sum_{i \neq j} \mathbb{E}[v_i(x)v_j(x)1(U_i \leq z)1(U_j \leq z)] - z^2
= \sum_{i \neq j} (1 - \gamma_n)\mathbb{E}[v_i(x)v_j(x)1(U_i \leq z)1(U_j \leq z) \mid M_i(\delta)] + \gamma_n\mathbb{E}[v_i(x)v_j(x)1(U_i \leq z)1(U_j \leq z) \mid \neg M_i(\delta)] - z^2
\leq \sum_{i \neq j} \mathbb{E}[v_i(x)v_j(x)1(U_i \leq z)1(U_j \leq z) \mid M_i(\delta)] - z^2
+ \gamma_n \sum_{i \neq j} \left( \mathbb{E}[v_i(x)v_j(x)1(U_i \leq z)1(U_j \leq z) \mid \neg M_i(\delta)] + \mathbb{E}[v_i(x)v_j(x)1(U_i \leq z)1(U_j \leq z) \mid M_i(\delta)] \right)
\leq \sum_{i \neq j} \mathbb{E}[v_i(x)v_j(x)1(U_i \leq z)1(U_j \leq z) \mid M_i(\delta)] - z^2
+ \gamma_n \sum_{i=1}^n \mathbb{E}[v_i(x) \mid \neg M_i(\delta)] + \gamma_n \sum_{i=1}^n \mathbb{E}[v_i(x) \mid M_i(\delta)]
\rightarrow \sum_{i \neq j} \mathbb{E}[v_i(x)v_j(x)1(U_i \leq z)1(U_j \leq z) \mid M_i(\delta)] - z^2.
$$

(10)

**Independence of the Error Terms Given Stability**

Next, we use the assumed stability of the $i^{th}$ observation to achieve independence of the $i^{th}$ and $j^{th}$ error terms, then evaluate the $j^{th}$ error term. Without loss of generality of whether the $\delta$ is added or subtracted, we can bound (10) by applying the stability condition, then use the tower property conditioning on the $j^{th}$ covariate:

$$
\sum_{i \neq j} \mathbb{E}[v_i(x)v_j(x)1(U_i \leq z)1(U_j \leq z) \mid M_i(\delta)] - z^2
\leq \sum_{i \neq j} \mathbb{E}[v_i(x)v_j(x)1(\varphi(X_i) - \delta - Y_i \leq F_E^{-1}(z \mid X_i))1(U_j \leq z) \mid M_i(\delta)] - z^2
= \sum_{i \neq j} \mathbb{E}[\mathbb{E}[v_i(x)v_j(x)1(\varphi(X_i) - \delta - Y_i \leq F_E^{-1}(z \mid X_i)) \mid X_j, M_i(\delta)]
\cdot \mathbb{P}(U_j \leq z \mid X_j, M_i(\delta)) \mid M_i(\delta)] - z^2.
$$

(11)
We evaluate the conditional probability in (11) by exploiting the fact that stability of the \(i\)th observation is independent of \(X_j\), so \(\Pr(\mathcal{M}_i(\delta) \mid X_j) = \Pr(\mathcal{M}_i(\delta)) = 1 - \gamma_n\). In particular, the law of total probability, the triangle inequality, and Assumption 5 imply that

\[
|\Pr(U_j \leq z \mid X_j, \mathcal{M}_i(\delta)) - z| = \left| \frac{\Pr(U_j \leq z \mid X_j) - \Pr(U_j \leq z \mid X_j, \neg \mathcal{M}_i(\delta)) \gamma_n}{1 - \gamma_n} - z \right|
\]

\[
= \left| \frac{z}{1 - \gamma_n} - \frac{\gamma_n}{1 - \gamma_n} \Pr(U_j \leq z \mid X_j, \neg \mathcal{M}_i(\delta)) - z \right|
\]

\[
\leq \left| \frac{z}{1 - \gamma_n} - z \right| + \frac{\gamma_n}{1 - \gamma_n} \Pr(U_j \leq z \mid X_j, \neg \mathcal{M}_i(\delta))
\]

\[
\leq \frac{\gamma_n}{1 - \gamma_n} + \frac{\gamma_n}{1 - \gamma_n}
\]

Moreover, Lemma 1 implies that

\[
\frac{2\gamma_n}{1 - \gamma_n} \sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x) \mathbb{1}(\varphi(X_i) - \delta - Y_i \leq F_E^{-1}(z \mid X_i)) \mid \mathcal{M}_i(\delta)]
\]

\[
\leq \frac{2\gamma_n}{1 - \gamma_n} \sum_{i=1}^{n} \mathbb{E}[v_i(x) \mid \mathcal{M}_i(\delta)]
\]

\[
\rightarrow 0.
\]

We therefore substitute into (11) our upper bound on \(\Pr(U_j \leq z \mid X_j, \mathcal{M}_i(\delta))\) given by (12) via the triangle inequality, then apply (13) to eliminate the leftover \(2\gamma_n/(1 - \gamma_n)\) term, ultimately bounding (11) above by

\[
z \sum_{i \neq j} \mathbb{E}[v_i(x)v_j(x) \mathbb{1}(\varphi(X_i) - \delta - Y_i \leq F_E^{-1}(z \mid X_i)) \mid \mathcal{M}_i(\delta)] - z, \tag{14}
\]

discounting leftover terms. Since \(z \leq 1\), we drop the \(z\) outside of the absolute value in (14) moving forward. Next, we eliminate \(v_j(x)\) terms by applying the triangle inequality and Lemma 1 as follows:

\[
\sum_{i \neq j} \mathbb{E}[v_i(x)v_j(x) \mathbb{1}(\varphi(X_i) - \delta - Y_i \leq F_E^{-1}(z \mid X_i)) \mid \mathcal{M}_i(\delta)] - z
\]

\[
= \sum_{i=1}^{n} \mathbb{E} [v_i(x)(1 - v_i(x)) \mathbb{1}(\varphi(X_i) - \delta - Y_i \leq F_E^{-1}(z \mid X_i)) \mid \mathcal{M}_i(\delta)] - z
\]

\[
\leq \sum_{i=1}^{n} \mathbb{E} [v_i(x) \mathbb{1}(\varphi(X_i) - \delta - Y_i \leq F_E^{-1}(z \mid X_i)) \mid \mathcal{M}_i(\delta)] - z + \sum_{i=1}^{n} \mathbb{E} [v_i(x)^2 \mid \mathcal{M}_i(\delta)]
\]

\[
\rightarrow \sum_{i=1}^{n} \mathbb{E} [v_i(x) \mathbb{1}(\varphi(X_i) - \delta - Y_i \leq F_E^{-1}(z \mid X_i)) \mid \mathcal{M}_i(\delta)] - z. \tag{15}
\]
PROXIMITY OF $z$ CONDITIONAL ON STABILITY

In this subsection, we express the $z$ term inside the absolute value of (15) as a conditional expectation similar to the one inside the absolute value of (15). In particular, we replace $z$ with the expectation given by (5), decompose the expectation using the law of total expectation into expectations conditional on stability and instability, and eliminate leftover terms using Lemma 1 to obtain

$$
\sum_{i=1}^{n} \mathbb{E}[v_i(x) \mathbb{1}(\varphi(X_i) - \delta - Y_i \leq F_E^{-1}(z | X_i)) | \mathcal{M}_i(\delta)] - z
$$

$$
= \sum_{i=1}^{n} \mathbb{E}[v_i(x) \mathbb{1}(\varphi(X_i) - \delta - Y_i \leq F_E^{-1}(z | X_i)) | \mathcal{M}_i(\delta)] - \sum_{i=1}^{n} \mathbb{E}[v_i(x) \mathbb{1}(U_i \leq z)]
$$

$$
\leq \sum_{i=1}^{n} \mathbb{E}[v_i(x) \mathbb{1}(\varphi(X_i) - \delta - Y_i \leq F_E^{-1}(z | X_i)) | \mathcal{M}_i(\delta)] - \sum_{i=1}^{n} \mathbb{E}[v_i(x) \mathbb{1}(U_i \leq z) | \mathcal{M}_i(\delta)]
$$

$$
+ \gamma_n \sum_{i=1}^{n} \mathbb{E}[v_i(x) \mathbb{1}(U_i \leq z) | -\mathcal{M}_i(\delta)] + \gamma_n \sum_{i=1}^{n} \mathbb{E}[v_i(x) \mathbb{1}(U_i \leq z) | \mathcal{M}_i(\delta)]
$$

$$
\rightarrow \sum_{i=1}^{n} \mathbb{E}[v_i(x) \mathbb{1}(\varphi(X_i) - \delta - Y_i \leq F_E^{-1}(z | X_i)) | \mathcal{M}_i(\delta)] - \sum_{i=1}^{n} \mathbb{E}[v_i(x) \mathbb{1}(U_i \leq z) | \mathcal{M}_i(\delta)].
$$

(16)

By linearity of expectation and the assumed stability, (16) is bounded above by

$$
\sum_{i=1}^{n} \mathbb{E}[v_i(x) (\mathbb{1}(\varphi(X_i) - \delta - Y_i \leq F_E^{-1}(z | X_i)) - \mathbb{1}(\varphi(X_i) + \delta - Y_i \leq F_E^{-1}(z | X_i))) | \mathcal{M}_i(\delta)].
$$

(17)

CONTINUITY OF THE CONDITIONAL CDF

Next, we show that the two indicators in (17) are close to each other in expectation by continuity of the CDF of $Y$ conditional on $X$. First, notice that using the tower property to condition on $X_i$ achieves independence of the indicator functions from $v_i(x)$ and obviates the conditioning on $\mathcal{M}_i(\delta)$. Therefore, (17) is equivalent to

$$
\sum_{i=1}^{n} \mathbb{E}[\mathbb{E}[v_i(x) | X_i] (\Pr(\varphi(X_i) - \delta - Y_i \leq F_E^{-1}(z | X_i) | X_i)
$$

$$
- \Pr(\varphi(X_i) + \delta - Y_i \leq F_E^{-1}(z | X_i) | X_i)) | \mathcal{M}_i(\delta)].
$$

(18)

By uniform continuity of the conditional CDF of $Y$ as applied in (9), the difference between the conditional probabilities in (18) is bounded above by $\epsilon/3$, so (18) is bounded above by

$$
\frac{\epsilon}{3} \sum_{i=1}^{n} \mathbb{E}[v_i(x) | \mathcal{M}_i(\delta)].
$$

(19)
Finally, for \( n \) large enough that \( \gamma_n < 1/2, \) \( \mathbb{E}[v_i(x) \mid \mathcal{M}_i(\delta)] < 2/n \) by the law of total expectation since the weights must be nonnegative and \( \mathbb{E}[v_i(x)] = 1/n. \) Thus, (19) can be bounded above by

\[
\frac{\epsilon}{3} \sum_{i=1}^{n} \frac{2}{n} = \frac{2\epsilon}{3}
\]

Recalling that the leftover terms we have dropped throughout these steps sum to \( \epsilon/3, \) we conclude that

\[
\left| \sum_{i \neq j} \mathbb{E}[v_i(x)v_j(x) \mathbb{1}(U_i \leq z) \mathbb{1}(U_j \leq z)] - z^2 \right| < \epsilon,
\]

as desired. Thus, we have shown (4).

**Bounding the Shift Term**

Next, we show convergence of the shift term in (3); again, it suffices to show convergence for all \( e \in \mathbb{R}: \)

\[
\left\| \sum_{i=1}^{n} v_i(x)(\mathbb{1}(U_i \leq F_E(e \mid X_i)) - \mathbb{1}(U_i \leq F_E(e \mid x))) \right\| = o_p(1).
\]

As an intermediate result, we first show that

\[
\sum_{i=1}^{n} v_i(x)(\mathbb{1}(U_i \leq F_E(e \mid X_i)) - \mathbb{1}(U_i \leq F_E(e \mid x))) \xrightarrow{p} \sum_{i=1}^{n} v_i(x)(F_E(e \mid X_i) - F_E(e \mid x)).
\]  

(20)

By the triangle inequality, the union bound, and our convergence result (4), we can reduce the task of showing (20) to simply showing that

\[
\sum_{i=1}^{n} v_i(x)\mathbb{1}(U_i \leq F_E(e \mid X_i)) \xrightarrow{p} \sum_{i=1}^{n} v_i(x)F_E(e \mid X_i),
\]

or, equivalently,

\[
\sum_{i=1}^{n} v_i(x)\mathbb{1}(U_i \leq F_E(e \mid X_i)) - \sum_{i=1}^{n} v_i(x)F_E(e \mid X_i) \xrightarrow{p} 0.
\]  

(21)

We do so by showing that the left side of (21) has expectation zero and decreasing variance. Since \( v_i(x) \) and \( \mathbb{1}(U_i \leq F_E(e \mid X_i)) \) are independent conditional on \( X_i, \) and \( \Pr(U_i \leq F_E(e \mid X_i) \mid X_i) = F_E(e \mid X_i), \) a direct application of the tower property conditioning on \( X_i \) yields the identity

\[
\mathbb{E}[v_i(x)\mathbb{1}(U_i \leq F_E(e \mid X_i))] = \mathbb{E}[v_i(x)F_E(e \mid X_i)].
\]  

(22)

Applying (22) and linearity of expectation yields

\[
\mathbb{E}\left[ \sum_{i=1}^{n} v_i(x)\mathbb{1}(U_i \leq F_E(e \mid X_i)) - \sum_{i=1}^{n} v_i(x)F_E(e \mid X_i) \right] = 0.
\]
It thus remains to be shown that
\[
\lim_{n \to \infty} \text{Var} \left( \sum_{i=1}^{n} v_i(x) \mathbb{1}(U_i \leq F_E(e \mid X_i)) - \sum_{i=1}^{n} v_i(x)F_E(e \mid X_i) \right) = 0.
\]

We can decompose the variance of the summation into the sum of covariances:
\[
\text{Var} \left( \sum_{i=1}^{n} v_i(x) \mathbb{1}(U_i \leq F_E(e \mid X_i)) - \sum_{i=1}^{n} v_i(x)F_E(e \mid X_i) \right)
= \sum_{i=1}^{n} \text{Var} (v_i(x)( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)))
+ \sum_{i \neq j} \text{Cov} (v_i(x)( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)), v_j(x)( \mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j))).
\]
(23)

We show that each term in (23) converges to zero in turn.

**SUM OF VARIANCES**

We can decompose the sum of variances in (23) using the law of total variance:
\[
\sum_{i=1}^{n} \text{Var} (v_i(x)( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)))
= \sum_{i=1}^{n} \text{Var} (\mathbb{E}[v_i(x)( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)) \mid \Omega_v \setminus \{Y_i\}])
+ \mathbb{E}[\text{Var} (v_i(x)( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)) \mid \Omega_v \setminus \{Y_i\}]).
\]
(24)

By noting that \(v_i(x)\) is a constant given \(\Omega_v \setminus \{Y_i\}\) and applying an argument similar to the one yielding (22), we can reduce the variance-of-expectation term in (24) to
\[
\sum_{i=1}^{n} \text{Var} (\mathbb{E}[v_i(x)( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)) \mid \Omega_v \setminus \{Y_i\}])
= \sum_{i=1}^{n} \text{Var} (v_i(x)\mathbb{E}[( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)) \mid \Omega_v \setminus \{Y_i\}])
= \sum_{i=1}^{n} \text{Var} (v_i(x)\mathbb{E}[( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)) \mid X_i])
= \sum_{i=1}^{n} \text{Var} (v_i(x)(\Pr(U_i \leq F_E(e \mid X_i) \mid X_i) - \mathbb{E}[F_E(e \mid X_i) \mid X_i]))
= \sum_{i=1}^{n} \text{Var} (v_i(x)(F_E(e \mid X_i) - F_E(e \mid X_i)))
= 0.
\]
Moreover, we can reduce the expectation-of-variance term in (24) using Assumption 2 to note that the maximum possible weight $M_n$ given to an observation converges to zero as $n$ increases:

\[
\sum_{i=1}^{n} \mathbb{E} \left[ \text{Var} \left( v_i(x) \left( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i) \right) | \Omega_v \setminus \{Y_i\} \right) \right] \\
= \sum_{i=1}^{n} \mathbb{E} \left[ v_i(x)^2 \text{Var} \left( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i) \mid X_i \right) \right] \\
\leq \sum_{i=1}^{n} \mathbb{E} \left[ v_i(x)^2 \right] \\
\leq M_n \sum_{i=1}^{n} \mathbb{E} \left[ v_i(x) \right] \\
= M_n \rightarrow 0.
\]

Thus, we have shown that the sum of variances in (23) converges to zero—i.e., that

\[
\sum_{i=1}^{n} \text{Var} \left( v_i(x) \left( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i) \right) \right) \rightarrow 0
\]

as $n \rightarrow \infty$.

**Sum of Covariances**

Next, we show that the sum of covariances in (23) converges to zero. First, we rewrite the covariance in terms of expectations:

\[
\sum_{i \neq j} \text{Cov} \left( v_i(x) \left( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i) \right), v_j(x) \left( \mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j) \right) \right) \\
= \sum_{i \neq j} \mathbb{E} \left[ v_i(x)v_j(x) \left( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i) \right) \left( \mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j) \right) \right] \\
- \mathbb{E} \left[ v_i(x) \left( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i) \right) \right] \mathbb{E} \left[ v_j(x) \left( \mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j) \right) \right] \\
= \sum_{i \neq j} \mathbb{E} \left[ v_i(x)v_j(x) \left( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i) \right) \left( \mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j) \right) \right],
\]

where the last equality follows by (22). We therefore seek to show that

\[
\lim_{n \rightarrow \infty} \left| \sum_{i \neq j} \mathbb{E} \left[ v_i(x)v_j(x) \left( \mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i) \right) \left( \mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j) \right) \right] \right| = 0.
\]

Let $\epsilon > 0$. As before, the uniform continuity of the conditional response distribution implies that there exists $\delta_1 > 0$ so that

\[
|y_1 - y_2| < \delta_1 \quad \Rightarrow \quad |F_Y(y_1 \mid x) - F_Y(y_2 \mid x)| < \epsilon/3.
\]

(25)
Fix \( \delta < \min \{ \delta_0, \delta_1 \} \). Then, by Lemma 1, there exists \( N \in \mathbb{N} \) so that, for \( n \geq N \), the leftover terms in the following steps of the proof sum to at most \( \epsilon / 3 \); for concision, we note these terms where they appear and then cite Lemma 1 to drop them.

**CONDITIONING ON STABILITY**

We condition on stability using the law of total expectation to obtain:

\[
\sum_{i \neq j} \mathbb{E} \left[ v_i(x)v_j(x) (\mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)) (\mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \right] \\
\leq \sum_{i \neq j} \mathbb{E} \left[ v_i(x)v_j(x) (\mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)) (\mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \right] | \mathcal{M}_i(\delta) \\
+ \gamma_n \sum_{i \neq j} \mathbb{E} \left[ v_i(x)v_j(x) (\mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)) (\mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \right] | \mathcal{M}_i(\delta) \\
+ \gamma_n \sum_{i \neq j} \mathbb{E} \left[ v_i(x)v_j(x) (\mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)) (\mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \right] \rightarrow \mathcal{M}_i(\delta) .
\]

(26)

We can reduce the second term of (26) using Lemma 1 and Jensen’s inequality:

\[
\gamma_n \sum_{i \neq j} \mathbb{E} \left[ v_i(x)v_j(x) (\mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)) (\mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \right] | \mathcal{M}_i(\delta) \\
= \gamma_n \sum_{i=1}^{n} \mathbb{E} \left[ v_i(x) (\mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)) \sum_{j \neq i} v_j(x) (\mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \right] | \mathcal{M}_i(\delta) \\
\leq \gamma_n \sum_{i=1}^{n} \mathbb{E} \left[ v_i(x) (\mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)) \sum_{j \neq i} v_j(x) (\mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \right] | \mathcal{M}_i(\delta) \\
\leq \gamma_n \sum_{i=1}^{n} \mathbb{E} \left[ v_i(x) (\mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)) \right] | \mathcal{M}_i(\delta) \\
\leq \gamma_n \sum_{i=1}^{n} \mathbb{E} \left[ v_i(x) \right] | \mathcal{M}_i(\delta) \\
\rightarrow 0.
\]

We can reduce the third term of (26) by the same argument, so (26) is bounded above by

\[
\sum_{i \neq j} \mathbb{E} \left[ v_i(x)v_j(x) (\mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)) (\mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \right] | \mathcal{M}_i(\delta) ,
\]

31
discounting leftover terms. Distributing terms inside the expectation yields

\[ \sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x)1(U_i \leq F_E(e \mid X_i))(1(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \mid \mathcal{M}_i(\delta)] \]

\[ - \sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x)F_E(e \mid X_i)(1(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \mid \mathcal{M}_i(\delta)] \] .

(27)

**Applying Stability**

As in our proof of (4), we now apply the assumed stability of \( \hat{\varphi}(X_i) \). In this case, we apply it to the first summation in (27); in particular, we show that

\[ \sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x)1(U_i \leq F_E(e \mid X_i))(1(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \mid \mathcal{M}_i(\delta)] \]

\[ - \sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x)1(\varphi(X_i) - Y_i + \delta \leq e)(1(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \mid \mathcal{M}_i(\delta)] \] \leq \frac{2e}{3} .

(28)

Our proof of this claim is as follows. We can rearrange terms inside the absolute value of (28) to obtain

\[ \sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x)1(U_i \leq F_E(e \mid X_i))(1(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \mid \mathcal{M}_i(\delta)] \]

\[ - \sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x)1(\varphi(X_i) - Y_i \leq e)(1(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \mid \mathcal{M}_i(\delta)] \]

\[ = \sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x)(1(\varphi(X_i) - Y_i \leq e) - 1(\varphi(X_i) - Y_i + \delta \leq e))(1(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \mid \mathcal{M}_i(\delta)] .

(29)

Next, because, conditional on \( \mathcal{M}_i(\delta) \), \( 1(\varphi(X_i) - Y_i \leq e) - 1(\varphi(X_i) - Y_i + \delta \leq e) \geq 0 \), we can bound (29) above by taking all of the \( 1(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j) \) terms to be
the fact that
By uniform continuity of the conditional distribution of $Y$
above by
this fact and applying the tower property to condition on $X_i$,
terms as follows:
Next, we note that, conditional on $\mathcal{M}_i(\delta)$, $\mathbb{1}(\varphi(X_i) - Y_i - \delta \leq e) \geq \mathbb{1}(\hat{\varphi}(X_i) - Y_i \leq e)$. Using
this fact and applying the tower property to condition on $X_i$, we can bound (30) above by

$$\sum_{i=1}^{n} \mathbb{E} \left[ \mathbb{E} [v_i(x) \mid X_i] \left( \Pr(\varphi(X_i) - Y_i - \delta \leq e \mid X_i) - \Pr(\hat{\varphi}(X_i) - Y_i + \delta \leq e \mid X_i) \right) \right] \mathcal{M}_i(\delta).$$

(31)

By uniform continuity of the conditional distribution of $Y$ given $X$ as applied in (25) and
the fact that $\mathbb{E} [v_i(x) \mid \mathcal{M}_i(\delta)] < 2/n$ for $n$ large enough that $\gamma_n < 1/2$, (31) can be bounded
above by

$$\frac{\epsilon}{3} \sum_{i=1}^{n} \mathbb{E} [v_i(x) \mid \mathcal{M}_i(\delta)] < \frac{\epsilon}{3} \sum_{i=1}^{n} \frac{2}{n} = \frac{2\epsilon}{3}.$$

Thus, we have shown (28). Applying this result, we can bound (27) above and rearrange
terms as follows:

$$\sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x)\mathbb{1}(\varphi(X_i) - Y_i + \delta \leq e)(\mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \mid \mathcal{M}_i(\delta)]$$

$$- \sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x)F_E(e \mid X_i)(\mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \mid \mathcal{M}_i(\delta)] \geq \frac{2\epsilon}{3}$$

$$= \sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x)(\mathbb{1}(\varphi(X_i) - Y_i + \delta \leq e) - F_E(e \mid X_i))(\mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \mid \mathcal{M}_i(\delta)] \geq \frac{2\epsilon}{3}.$$

(32)
FROM CONDITIONAL EXPECTATIONS TO UNCONDITIONAL EXPECTATIONS

Next, we use the law of total expectation, Jensen’s inequality, and Lemma 1 to show that the conditional expectation in (32) is close to the corresponding unconditional expectation:

\[
\begin{align*}
\sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x) & (1(\varphi(X_i) - Y_i + \delta \leq e) - F_E(e | X_i))(1(U_j \leq F_E(e | X_j)) - F_E(e | X_j))] \\
& - \sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x) (1(\varphi(X_i) - Y_i + \delta \leq e) - F_E(e | X_i))(1(U_j \leq F_E(e | X_j)) - F_E(e | X_j))] \\
& = \left( \sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x) (1(\varphi(X_i) - Y_i + \delta \leq e) - F_E(e | X_i))(1(U_j \leq F_E(e | X_j)) - F_E(e | X_j))] \\
& \quad - \gamma_n \mathbb{E} [v_i(x)v_j(x) (1(\varphi(X_i) - Y_i + \delta \leq e) - F_E(e | X_i))(1(U_j \leq F_E(e | X_j)) - F_E(e | X_j))] \right) \\
& \quad \left( \frac{1}{1 - \gamma_n} \right) \\
& \leq \frac{\gamma_n}{1 - \gamma_n} \sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x) | 1(\varphi(X_i) - Y_i + \delta \leq e) - F_E(e | X_i)| 1(U_j \leq F_E(e | X_j)) - F_E(e | X_j)] \\
& + \frac{\gamma_n}{1 - \gamma_n} \sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x) | 1(\varphi(X_i) - Y_i + \delta \leq e) - F_E(e | X_i)| 1(U_j \leq F_E(e | X_j)) - F_E(e | X_j)] \\
& \quad \cdot 1(U_j \leq F_E(e | X_j)) - F_E(e | X_j) | -M_i(\delta)] \\
& \leq \frac{\gamma_n}{1 - \gamma_n} + \frac{\gamma_n}{1 - \gamma_n} \sum_{i=1}^{n} \mathbb{E} [v_i(x) | -M_i(\delta)] \\
& \rightarrow 0.
\end{align*}
\]

Applying this with the triangle inequality, we thus have that (32) is bounded above by

\[
\left| \sum_{i \neq j} \mathbb{E} [v_i(x)v_j(x)(1(\varphi(X_i) - Y_i + \delta \leq e) - F_E(e | X_i))(1(U_j \leq F_E(e | X_j)) - F_E(e | X_j))] \right| + \frac{2\epsilon}{3}.
\]

(33)

CONCLUSION

Finally, we apply the tower property, noting that 1(U_j \leq F_E(e | X_j)) - F_E(e | X_j) is independent of v_i(x)v_j(x)(1(\varphi(X_i) - Y_i + \delta \leq e) - F_E(e | X_i)) conditional on X_j and that
\[
\mathbb{E} \left[ \mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j) \mid X_j \right] = 0, \text{ to find that}
\]
\[
\sum_{i \neq j} \mathbb{E} \left[ v_i(x) v_j(x) (\mathbb{1}(\varphi(X_i) - Y_i + \delta \leq e) - F_E(e \mid X_i)) (\mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \right] = 0,
\]
so (33) reduces to \(2\epsilon / 3\). Recalling that the leftover terms we have dropped throughout these steps sum to \(\epsilon / 3\), we conclude that
\[
\sum_{i \neq j} \mathbb{E} \left[ v_i(x) v_j(x) (\mathbb{1}(U_i \leq F_E(e \mid X_i)) - F_E(e \mid X_i)) (\mathbb{1}(U_j \leq F_E(e \mid X_j)) - F_E(e \mid X_j)) \right] < \epsilon,
\]
so the sum of covariances in (23) converges to zero, as desired. Thus, we have shown (20).

By Lipschitz continuity of the conditional prediction error distribution (Assumption 3), it therefore remains to be shown that
\[
\sum_{i=1}^{n} v_i(x) \|X_i - x\|_1 = o_p(1). \tag{34}
\]
This follows from Lemma 2 of Meinshausen (2006). In particular, recall that
\[
v_i(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{\# \{ Z_i \in \mathcal{D}_{n,b}^* \} \mathbb{1}(X_i \in R_{\ell(x, \theta_b)}^*)}{\sum_{j=1}^{n} \# \{ Z_j \in \mathcal{D}_{n,b}^* \} \mathbb{1}(X_j \in R_{\ell(x, \theta_b)}^*)},
\]
so showing (34) is equivalent to showing that
\[
\frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} \frac{\# \{ Z_i \in \mathcal{D}_{n,b}^* \} \mathbb{1}(X_i \in R_{\ell(x, \theta_b)}^*)}{\sum_{j=1}^{n} \# \{ Z_j \in \mathcal{D}_{n,b}^* \} \mathbb{1}(X_j \in R_{\ell(x, \theta_b)}^*)} \|X_i - x\|_1 \overset{p}{\to} 0.
\]
Therefore, it suffices to show that, for a single tree,
\[
\sum_{i=1}^{n} \frac{\# \{ Z_i \in \mathcal{D}_{n,b}^* \} \mathbb{1}(X_i \in R_{\ell(x, \theta)}^*)}{\sum_{j=1}^{n} \# \{ Z_j \in \mathcal{D}_{n,b}^* \} \mathbb{1}(X_j \in R_{\ell(x, \theta)}^*)} \|X_i - x\|_1 \overset{p}{\to} 0.
\]
Following the argument in the proof of Theorem 1 of Meinshausen (2006), we can decompose the rectangular subspace \(R_{\ell(x, \theta)}\) \(\subseteq [0, 1]^p\) of leaf \(\ell(x, \theta)\) of the tree into the intervals \(I(x, m, \theta) \subseteq [0, 1]\) for \(m = 1, \ldots, p\):
\[
R_{\ell(x, \theta)} = \otimes_{m=1}^{p} I(x, m, \theta).
\]
Note that \(X_i \notin I(x, m, \theta)\) implies that \(\mathbb{1}(X_i \in R_{\ell(x, \theta)}^*) = 0\). Thus, it suffices to show that \(\max_m |I(x, m, \theta)| = o_p(1)\), which Lemma 2 of Meinshausen (2006) accomplishes. \(\Box\)