07 Random Forests

Random forests are extremely popular in practice and have shown impressive accuracy across a variety of problems. They were first proposed by Breiman [2001] and are essentially averaged regression trees combined via bagging (with some extra randomness). We therefore start from regression trees and bagging.

1 Regression Trees

Regression trees [Breiman et al., 2017] are an example of a partitioning estimator, where the feature space $\mathcal{X}$ is split up into a disjoint partition

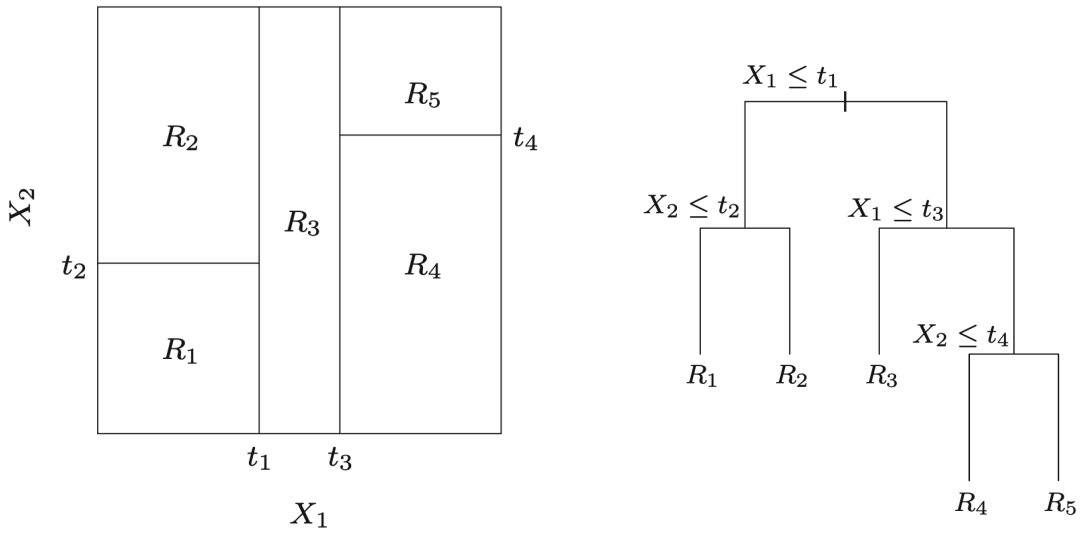
$$P_n = \{A_1, .A_2, ..., \}$$

and then for $A_n(x)$ the cell of P_n containing x, we estimate the regression function at a point x with

$$\widehat{\mu}(x) = \frac{\sum_i \mathbb{1}\{X_i \in A_n(x)\} Y_i}{\sum_i \mathbb{1}\{X_i \in A_n(x)\}},$$

i.e., a simple average of outcomes in that same cell. In their Theorem 4.2, Györfi et al. [2002] show that a partitioning estimator is universally consistent as long as the cells in the partition shrink to zero (so the bias goes to zero), while the number of cells in a bounded set is small, so that the cells don't shrink too quickly (so the variance also goes to zero). Importantly, this is shown in Theorem 4.2 for *non-data-dependent* partitions P_n, e.g., if A_j are intervals of size h in the $d = 1$ case. For data-dependent partitions see Chapter 13 of Györfi et al. [2002].

For regression trees, the partition is chosen a particular way, namely via recursive binary splitting. At each step of the recursive binary splitting process, one chooses a feature $j \in \{1, ..., d\}$ and a split point $t \in \mathbb{R}$, resulting in partitions of the form shown below (left plot), which can be represented visually with trees (right plot).



This makes trees very appealing from the perspective of interpretability. The bottom nodes of the tree are called leaves.

In order to decide on particular splits (j, t) at each step, it is common to use a squared error criterion (note optimizing across all possible trees is typically computationally infeasible, so one resorts to a stepwise greedy algorithm). Specifically, let A be a generic cell (which at step one would be the whole feature space), and let $|A|$ denote the number of data points in A. Then the split criterion is

$$\widehat{R}(j,t) = \frac{1}{|A|}\sum_i \mathbb{1}(X_i \in A)(Y_i - \overline{Y}_A)^2 \\ - \frac{1}{|A|}\sum_i \mathbb{1}(X_i \in A)\Big\{Y_i - \overline{Y}_{A_L}\mathbb{1}(X_j < t) - \overline{Y}_{A_R}\mathbb{1}(X_j \geq t)\Big\}^2$$

where for a set J

$$\overline{Y}_J = \frac{1}{|J|}\sum_i \mathbb{1}(X_i \in J)Y_i$$

is just the average outcome in that set, and

$$A_L = \{x \in A : x_j < t\}$$
$$A_R = \{x \in A : x_j \geq t\}$$

are the partitions resulting from a split at (j, t). The criterion $\widehat{R}(j, t)$ is therefore just comparing the error of an estimator that is constant across A to one that splits at (j, t) and estimates with the mean to the left/right (note it will always be non-negative since it is an in-sample error criterion). The best split is

$$\arg\max \widehat{R}(j,t)$$

i.e., the one that gives the best error across all possible features j and cutpoints t. Then the process is repeated with $A = A_L^*, A_R^*$, i.e., the partitions resulting from the best split, and so on until some stopping criterion is met.

Typically the above splitting procedure continues until all leaves contain no fewer than some number of datapoints (call it s = node size), which is the main tuning parameter at this point. You can expect that if the leaves have fewer observations (i.e., s is small), the tree will be less biased but highly variable, while if the leaves have more observations (i.e., s large) the tree will have more bias but less variance. Sometimes trees are grown until some node size threshold is reached, and then they are pruned back. We will not discuss this, since it is less common in random forest implementations of trees.

There is a crucial difference between the tree estimator described above, and the estimators we have discussed previously based on local/global modeling. Namely, the above tree estimator essentially double dips, using the outcome data twice: once to pick the partition, and once to estimate the regression function within cells of the partition. Thus the tree estimator should really be written as

$$\widehat{\mu}(x) = \frac{\sum_i \mathbb{1}\{X_i \in A_n(x; D^n)\}Y_i}{\sum_i \mathbb{1}\{X_i \in A_n(x; D^n)\}},$$

so the cell $A_n(x) = A_n(x; D^n)$ actually depends on the whole dataset through maximizing the split criterion above. This aspect, sometimes called adaptivity or outcome-adaptivity, makes the analysis of trees challenging.

2 Bagging

A deep tree with many leaves may have small bias but high variance and some instability (e.g., changing a few observations can lead to very different trees). Bagging (short for bootstrap aggregation) is a generic method proposed by Breiman [1996] to stabilize any base estimator. Note stability may be desired for its own sake (e.g., for purposes of reproducibility), or it may be a prerequisite for some other goal, e.g., generalization [Shalev-Shwartz et al., 2010], or avoiding sample splitting [Chen et al., 2022].

The idea behind bagging is quite simple and intuitive. We draw m samples from a dataset D^n, recompute the estimator, repeat B times, and then average the B estimators to arrive at a final bagged estimator. Classical bagging amounts to drawing a sample of size m with replacement, but there are other variants, e.g., sampling without replacement, or Poissonized bagging where m is drawn from a Poisson distribution, etc. To formalize things, in what follows we use the presentation of Soloff et al. [2024].

Let S_n denote the set of all sequences of indices in $\{1, ..., n\}$. Then a particular sequence $r \in S_n$ is called a bag, and given a bag and the original dataset $D^n = (Z_1, ..., Z_n)$ we can define a new dataset $D_r = \{Z_i : i \in r\}$. The general form of bagging starts with a distribution Q over S_n, and draws bags randomly $r_1, ..., r_B \sim Q$. For classical bagging using sampling with replacement, for example, Q would be a uniform distribution over all sequences of length m. Therefore a bagged regression estimator can be written as

$$\widehat{\mu}(x) = \frac{1}{B} \sum_{b=1}^{B} \widehat{\mu}(x; D_{r_b})$$

where $\widehat{\mu}(x; D_r)$ is the estimator fit to the dataset D_r. It is also useful for theoretical analysis to consider a limiting version of bagging where $B \to \infty$, called derandomized bagging, given by

$$\widehat{\mu}(x) = \sum_{r} \widehat{\mu}(x; D_r) Q(r).$$

Soloff et al. [2024] show the following stability result.

Theorem 1. *Assume $\widehat{\mu}(x; D_r) \in [0, 1]$. Then classical bagging is (ϵ, δ)-stable, i.e., the derandomized bagging estimator satisfies*

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\Big\{ |\widehat{\mu}(x) - \widehat{\mu}_{-i}(x)| > \epsilon \Big\} \leq \delta$$

for all datasets D^n and points x, as long as $\delta\epsilon^2 \geq \frac{1}{4(n-1)} \frac{p}{1-p}$ for $p = \mathbb{P}(i \in r) = 1 - (1 - 1/n)^m$.

Soloff et al. [2024] show a similar result holds for more general forms of bagging, for finite B, unbounded outcomes, etc. Note that (ϵ, δ)-stability ensures that dropping one observation from a dataset cannot change the estimator by more than ϵ, for at least $1 - \delta$ fraction of the observations. There are many other ways to define algorithmic stability; see for example [Shalev-Shwartz et al., 2010] for more details.

Random forest implementations in practice typically use classical bagging (sampling with replacement) with $m = n$. Here the condition on δ and ϵ simplifies to $\delta\epsilon^2 \geq \frac{1}{4(n-1)}$, so one could take for example $\delta = 1/3$ and $\epsilon = 1/\sqrt{n}$ as long as $n \geq 4$, i.e., no more than a $1/3$ of points could lead to leave-one-out estimates that differ by more than $1/\sqrt{n}$. Alternatively one could take $\delta = 1/n$ and $\epsilon = 1/\sqrt{3}$, in which case at most one point could lead to leave-one-out differences of more than $1/\sqrt{3}$, or $\delta = 1/(3\sqrt{n})$ and $\epsilon = 1/n^{1/4}$, etc.

Interestingly, Soloff et al. [2024] also show the (ϵ, δ) condition is tight up to constants, by constructing an algorithm which is provably not stable when $\delta\epsilon^2 < \frac{C}{n}\frac{p}{1-p}$.

3 Random Forests

Now having discussed trees and bagging, we are ready to define and describe random forests. Here we mostly follow "A random forest guided tour" by Biau and Scornet [2016].

Random forests are simply bagged trees (typically with $m = n$ samples drawn with replacement), and with an additional piece of randomness put in: within the cell of each tree, splits are considered not among all features, but instead among random subsets of features. Typically at each cell $q = d/3$ features are chosen at random for possible splits. Thus there are three main tuning parameters for random forests:

1. m = the number of randomly sampled data points for each tree,
2. q = the number of randomly selected features at each split,
3. s = node size discussed earlier, i.e., the minimum leaf size.

The default choices in `randomForest` in R are $m = n$, $q = d/3$, and $s = 5$. The random forest algorithm is displayed below.

Algorithm 1: Breiman's forests algorithm

(1) Grow M trees as follows:
  (a) Prior to the j-th tree construction, select uniformly with replacement, a_n data points among $\mathcal{D}_n$. Only these a_n observations are used in the tree construction.
  (b) Consider the cell $[0,1]^d$.
  (c) Select uniformly without replacement `mtry` coordinates among $\{1, \ldots, d\}$.
  (d) Select the split maximizing the CART-split criterion (see below for details) along the pre-selected `mtry` directions.
  (e) Cut the cell at the selected split.
  (f) Repeat $(c) - (e)$ for the two resulting cells until each cell of the tree contains less than `nodesize` observations.
  (g) For a query point $\mathbf{x}$, the j-th tree outputs the average of the Y_i falling into the same cell as $\mathbf{x}$.
(2) For a query point $\mathbf{x}$, Breiman's forest outputs the average of the predictions given by the M trees.

The bth tree in the forest can be written as

$$\widehat{\mu}_b(x) = \frac{\sum_{i \in D_b^n} \mathbb{1}\{X_i \in A_n(x;, \Theta_b, D_b^n)\} Y_i}{\sum_{i \in D_b^n} \mathbb{1}\{X_i \in A_n(x; \Theta_b, D_b^n)\}},$$

where D_b^n is the bth random sample of D^n and Θ_b is a random variable indicating which features to split on (e.g., indices of splitting candidate directions in each cell). Therefore the overall random forest estimator is given by

$$\widehat{\mu}(x) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\mu}_b(x).$$

As with bagging, it is common to consider a limiting version of the estimator as $B \to \infty$ and averaged over randomness in Θ,

$$\widehat{\mu}(x) = \sum_{r,\theta} \frac{\sum_{i \in D_r^n} \mathbb{1}\{X_i \in A_n(x;, \theta, D_r^n)\} Y_i}{\sum_{i \in D_r^n} \mathbb{1}\{X_i \in A_n(x; \theta, D_r^n)\}} Q(r) P(\theta)$$

4 Theoretical results

Random forests are notoriously difficult to analyze due to the greedy and data-dependent splitting they use in each tree, as well as the bagging. Thus there are some important gaps in using theoretical analysis to explain the impressive practical performance of random forests. Most theoretical work on random forests considers stylized versions of random forests, which do not exactly align with Breiman's proposal, and which is the version implemented most commonly in practice.

One exception is Scornet et al. [2015] who gave the following consistency result for Breiman's version of random forests under an additive model assumption. They use the number of leaves in each tree ℓ as a tuning parameter, instead of node size s. (Note assumption 2 can easily be replaced with bounded density assumptions, and 3 with bounded outcomes/variance assumptions).

Theorem 2. *Assume:*

1. $\mu(x) = \sum_{j=1}^{d} \mu_j(X_j)$, *with each* μ_j *continuous,*
2. X *is uniformly distributed over* $[0,1]^d$,
3. $Y \mid X = x \sim N(\mu(x), \sigma^2)$.

Then the limiting random forest estimator $\widehat{\mu}$ *(using subsampling without replacement) satisfies*

$$\lim_{n \to \infty} \mathbb{E}\|\widehat{\mu} - \mu\|^2 = 0$$

if $m \to \infty$, $\ell \to \infty$, *and* $\ell(\log m)^9/m \to 0$.

The main idea in the proof is the following. First, Scornet et al. [2015] analyze an oracle random forest estimator that uses a population MSE version of the splitting criterion, rather than an empirical estimate. Thus the partitions do not depend on the data anymore, and this estimator can essentially be studied as a partitioning estimator as in Györfi et al. [2002]. Then Scornet et al. [2015] show this oracle forest is sufficiently close to the data-dependent version.

A common simplification in analyzing random forests is to do away with the data-dependent partitioning; these are called *purely random forests*. One example is the *centered forest*, where there is no resampling (trees are repeatedly fit to the whole dataset) and splits in each tree are done entirely randomly, by selecting a feature at random and cutting at the middle of that feature. A similar variant is the *uniform random forest* where cuts are done at random instead of in the middle.

Here we give more details on centered forests, and ultimately show a convergence rate result from Biau [2012] Crucially, for centered forests, the partitions in each tree are random and do not depend on outcome information. Therefore each tree estimator is given by

$$\widehat{\mu}_j(x) = \frac{\sum_i \mathbb{1}\{X_i \in A_n(x, \Theta_j)\} Y_i}{\sum_i \mathbb{1}\{X_i \in A_n(x, \Theta_j)\}},$$

and the random forest estimator is given by

$$\widehat{\mu}(x) = \lim_{B \to \infty} \frac{1}{B} \sum_{j=1}^{B} \widehat{\mu}_j(x) = \sum_{i=1}^{n} \mathbb{E}_\Theta\{w_i(x; X^n, \Theta)\} Y_i$$

where the weight is, for $N_n(x; \Theta) = \sum_i \mathbb{1}\{X_i \in A_n(x, \Theta)\}$ the number of units in the leaf containing x,

$$w_i(x; X^n, \Theta) = \frac{\mathbb{1}\{X_i \in A_n(x, \Theta)\}}{N_n(x; \Theta)} \mathbb{1}\{N_n(x; \Theta) > 0\}.$$

Recall Θ is a random variable independent of (X^n, Y^n) that indicates which variables to split on. Biau [2012] also assumes that feature j is selected for splitting with probability $p_j = (1/s)(1 + \xi_j)$ for $j \in S$ and $p_j = \xi_j$ otherwise.

The main tuning parameter is ℓ, which determines the number of leaves; specifically there are $\log_2 \ell$ levels of splitting, leading to $2^{\log_2 \ell} = \ell$ leaves. When the features are approximately uniform, this leads to approximately n/ℓ observations per leaf.

For these centered forests Biau [2012] showed universal consistency as long as $p_j \log \ell \to \infty$ (so the number of leaves increases and bias shrinks) and $\ell/n \to 0$ (so the number of observations per leaf grows and variance shrinks), and gave the following rate of convergence result.

Theorem 3. *Assume:*

1. *X is uniformly distributed over $[0, 1]^d$,*
2. *$\mu(x)$ is Lipschitz, and s-sparse, i.e., $\mathbb{E}(Y \mid X = x) = \mathbb{E}(Y \mid V = v)$ for $V \subseteq X$, $dim(V) = s \leq d$,*
3. *$var(Y \mid X = x) \leq \sigma^2$.*

Let S denote the indices of the variables V. Also assume the feature selection probabilities equal $p_j = (1/S)(1 + \xi_j)$ for $j \in S$ and $p_j = \xi_j$ for $j \notin S$, where $\xi_j \to 0$ as $n \to \infty$. Then the centered random forest estimator satisfies

$$\|\widehat{\mu} - \mu\| = O_{\mathbb{P}}\left(n^{\frac{-1}{2+s(8/3)\log 2}}\right)$$

if $\xi_j \log n \to 0$ for $j \in S$ and the number of leaves satisfies $\ell \sim n^{1/(1+0.75/s \log 2)}$.

4.1 Derivation of bias in Biau [2012] rate result

It is perhaps not surprising that the variance can be bounded by a multiple of ℓ/n, given that there are $\sim n/\ell$ observations in each of ℓ leaves. The bias is a bit more surprising and interesting. For it the main term is

$$
\begin{aligned}
\mathbb{E}\Big[\sum_{i=1}^n w_i(X;X^n,\Theta)\Big\{\mu(X_i)-\mu(X)\Big\}^2\Big] &= \mathbb{E}\left[\sum_{i=1}^n w_i(X;X^n,\Theta)\Big\{\mu(V_i)-\mu(V)\Big\}^2\right] \\
&\lesssim \mathbb{E}\left[\sum_{i=1}^n w_i(X;X^n,\Theta)\|V_i-V\|^2\right] \\
&\le n\mathbb{E}\left[w_1(X;X^n,\Theta)\|V_1-V\|^2\right] \\
&= n\mathbb{E}\left[\frac{\mathbb{1}\{X_1\in A_n(X;\Theta)\}}{N_n(X;\Theta)}\mathbb{1}\{N_n(X;\Theta)>0\}\|V_1-V\|^2\right] \\
&= n\mathbb{E}\left[\|V_1-V\|^2\frac{\mathbb{1}\{X_1\in A_n(X;\Theta)\}}{1+\sum_{i>1}\mathbb{1}\{X_i\in A_n(X;\Theta)\}}\right] \\
&= n\mathbb{E}\left(\|V_1-V\|^2\mathbb{1}\{X_1\in A_n(X;\Theta)\}\mathbb{E}\left[\frac{1}{1+\sum_{i>1}\mathbb{1}\{X_i\in A_n(X;\Theta)\}}\;\Big|\;X_1,X,\Theta\right]\right) \\
&\le \ell\;\mathbb{E}\Big[\|V_1-V\|^2\mathbb{1}\{X_1\in A_n(X;\Theta)\}\Big] \\
&= \ell\sum_{j\in S}\mathbb{E}\Big[(X_{j1}-X_j)^2\mathbb{1}\{X_1\in A_n(X;\Theta)\}\Big] \\
&= \ell\sum_{j\in S}\mathbb{E}\Big[(X_{j1}-X_j)^2\mathbb{1}\{X_{j1}\in A_{nj}(X;\Theta)\}\prod_{t\neq j}\mathbb{1}\{X_{t1}\in A_{nt}(X;\Theta)\}\Big] \\
&= \ell\sum_{j\in S}\mathbb{E}\left(\mathbb{E}\Big[(X_{j1}-X_j)^2\mathbb{1}\{X_{j1}\in A_{nj}(X;\Theta)\}\mid X\Big]\prod_{t\neq j}\mathbb{1}\{X_{t1}\in A_{nt}(X;\Theta)\}\right) \\
&\le \ell\sum_{j\in S}\mathbb{E}\left[\lambda\{A_{nj}(X;\Theta)\}^3\prod_{t\neq j}\mathbb{1}\{X_{t1}\in A_{nt}(X;\Theta)\}\right] \\
&= \sum_{j\in S}\mathbb{E}\Big[\lambda\{A_{nj}(X;\Theta)\}^2\Big] \\
&= \sum_{j\in S}\mathbb{E}\Big[2^{-2K_{nj}(X,\Theta)}\Big] = \sum_{j\in S}\mathbb{E}\Big\{2^{-2\mathrm{Bin}(\log_2\ell,p_j)}\Big\} \\
&= \sum_{j\in S}(1-p_j+p_j/4)^{\log_2\ell} \\
&\le \sum_{j\in S}\exp(-(3/4)p_j\log_2\ell) = \sum_{j\in S}\exp\Big\{-(3/4)(1/s)(1+\xi_j)\log\ell/\log 2\Big\} \\
&= \sum_{j\in S}\ell^{-3(1+\xi_j)/4s\log 2} \le s\ell^{-3(1+\xi_{\max})/4s\log 2}
\end{aligned}
$$

In the above, Biau [2012] uses the following key facts: (i) $\sum_{i>1}\mathbb{1}\{X_i\in A_n(x;\Theta)\}$ is conditionally Binomial$(n-1,\lambda\{A_n(x;\Theta)\})$, (ii) the Lebesgue measure of the leaf is $\lambda\{A_n(x;\Theta)\}=1/\ell$, and (iii) that $K_{nj}(x,\Theta)=$ the number of times the leaf $A_n(x;\Theta)$ is split on the jth feature is conditionally Binomial$(\log_2\ell,p_j)$.

4.2 Asymptotic normality

A number of authors have proved asymptotic normality results for variants of random forests estimators. For example, the following comes from Mentch and Hooker [2016], who use results for incomplete infinite-order U-statistics (in the below, their k_n is our subsample size m, their m_n is our number of subsamples B):

Theorem 2 *Let $U_{\omega;n,k_n,m_n}$ be a random kernel U-statistic of the form defined in equation (5) such that $U^*_{\omega;n,k_n,m_n}$ satisfies Condition 1 and suppose that $\mathbb{E}h^2_{k_n}(Z_1,...,Z_{k_n}) < \infty$ for all n, $\lim \frac{k_n}{\sqrt{n}} = 0$, and $\lim \frac{n}{m_n} = \alpha$. Then, letting β index the subsamples, so long as $\lim \zeta_{1,k_n} \neq 0$ and*

$$\lim_{n\to\infty} \mathbb{E}\left(h^{(\omega)}_{k_n}(Z_{\beta_1},...,Z_{\beta_{k_n}}) - \mathbb{E}_\omega h^{(\omega)}_{k_n}(Z_{\beta_1},...,Z_{\beta_{k_n}})\right)^2 \neq \infty,$$

$U_{\omega;n,k_n,m_n}$ is asymptotically normal and the limiting distributions are the same as those provided in Theorem 1.

This result is powerful since it applies to generic outcome-adaptive random forest estimators, though it does require quite strict subsampling with $m \ll \sqrt{n}$. Further, notice that it only says the estimator (properly scaled) converges in distribution to a Gaussian, and does not characterize the mean of the estimator or how it relates to the true regression function, e.g., under smoothness assumptions, nor how the variance scales with sample size, etc.

Wager and Athey [2018] show a similar result, though for stylized "honest" trees where the splitting criterion $\widehat{R}(j,t)$ is computed on a separate independent sample, and where the subsampling size is chosen in a quite particular way, depending on a regularity parameter α controlling what fraction of observations is left on each side of every split:

Theorem 1. *Suppose that we have n independent and identically distributed training examples $Z_i = (X_i, Y_i) \in [0,1]^d \times \mathbb{R}$. Suppose moreover that the features are independently and uniformly distributed[6] $X_i \sim U([0,1]^d)$, that $\mu(x) = \mathbb{E}\left[Y \,\middle|\, X = x\right]$ and $\mu_2(x) = \mathbb{E}\left[Y^2 \,\middle|\, X = x\right]$ are Lipschitz-continuous, and finally that $\mathrm{Var}\left[Y \,\middle|\, X = x\right] > 0$ and $\mathbb{E}[|Y - \mathbb{E}[Y \,|\, X = x]|^{2+\delta} \,|\, X = x] \leq M$ for some constants $\delta, M > 0$, uniformly over all $x \in [0,1]^d$. Given this data-generating process, let T be an honest, α-regular with $\alpha \leq 0.2$, and symmetric random-split tree in the sense of Definitions 2, 3, 4, and 5, and let $\hat{\mu}_n(x)$ be the estimate for $\mu(x)$ given by a random forest with base learner T and a subsample size s_n. Finally, suppose that the subsample size s_n scales as*

$$s_n \asymp n^\beta \quad \text{for some} \quad \beta_{\min} := 1 - \left(1 + \frac{d}{\pi}\frac{\log\left(\alpha^{-1}\right)}{\log\left((1-\alpha)^{-1}\right)}\right)^{-1} < \beta < 1. \tag{14}$$

Then, random forest predictions are asymptotically Gaussian:

$$\frac{\hat{\mu}_n(x) - \mu(x)}{\sigma_n(x)} \Rightarrow \mathcal{N}(0,1) \text{ for a sequence } \sigma_n(x) \to 0. \tag{15}$$

References

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

G. Biau and E. Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

L. Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Routledge, 2017.

Q. Chen, V. Syrgkanis, and M. Austern. Debiased machine learning without sample-splitting for stable estimators. *Advances in Neural Information Processing Systems*, 35:3096–3109, 2022.

L. Györfi, M. Kohler, A. Krzykaz, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

L. Mentch and G. Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(26):1–41, 2016.

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

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

J. A. Soloff, R. F. Barber, and R. Willett. Bagging provides assumption-free stability. *Journal of Machine Learning Research*, 25(131):1–35, 2024.

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