

## Topic 20: Random Forest

by Sai Zhang

**Key points:** .

**Disclaimer:** The note is built on Prof. [Jinchi Lv](#)'s lectures of the course at USC, DSO 607, High-Dimensional Statistics and Big Data Problems.

## 20.1 Motivation

Denote by  $m(\mathbf{X})$  the measurable nonparametric regression function with  $p$ -dimensional random vector  $\mathbf{X}$  taking values in  $[0, 1]^p$ . The Random Forest algorithm aims to learn the regression function in a non-parametric way based on the observations  $\mathbf{x}_i \in [0, 1]^p$ ,  $y_i \in \mathbb{R}$ ,  $i = 1, \dots, n$ , from the model

$$y_i = m(\mathbf{x}_i) + \epsilon_i$$

where  $\mathbf{X}$ ,  $\mathbf{x}_i$ ,  $\epsilon_i$ ,  $i = 1, \dots, n$  are independent, and  $\{\mathbf{x}_i\}$  and  $\{\epsilon_i\}$  are two sequences of identically distributed random variables.  $\mathbf{x}_i$  is distributed identically as  $\mathbf{X}$ .

**Why Random Forest (RF)?** RF has gained significant popularity due to its

- **High accuracy:** RF consistently rank among the top performer, often surpassing more complex models
- **Robustness:** RF are less subject to overfitting due to the ensemble nature leveraging multiple decision trees
- **Interpretability:** RF provide rankings of feature importance

As illustrated in Figure 20.1, in a level-2 tree, each node (cell) defines the point where the current cell split and new cells are produced. The sets of features eligible for splitting cells at level  $k - 1$  are denoted as  $\Theta_k := \{\Theta_{k,1}, \dots, \Theta_{k,2^{k-1}}\}$ , where  $\Theta_{k,s} \subset \{1, \dots, p\}$ .



Figure 20.1: Level-2 Tree Example

Given any  $T$  (and the associated splitting criterion) and  $\Theta_{1:k}$ , the tree estimate denoted as  $\hat{m}_{T(\Theta_{1:k})}$  for a test

point  $\mathbf{c} \in [0, 1]^p$  is defined as

$$\hat{m}_{T(\Theta_{1:k})}(\mathbf{c}, \mathcal{X}_n) := \sum_{(\mathbf{t}_1, \dots, \mathbf{t}_k) \in T(\Theta_{1:k})} \mathbf{1}_{\mathbf{c} \in \mathbf{t}_k} \left( \frac{\sum_{i \in \{i: \mathbf{x}_i \in \mathbf{t}_k\}} y_i}{\#\{i: \mathbf{x}_i \in \mathbf{t}_k\}} \right)$$

where  $\mathcal{X}_n := \{\mathbf{x}_i, y_i\}_{i=1}^n$ , the fraction is defined as 0 when no sample is in the cell  $\mathbf{t}_k$ , and  $\mathbf{1}_{\mathbf{c} \in \mathbf{t}_k}$  is an indicator function = 1 if  $\mathbf{c} \in \mathbf{t}_k$  and = 0 otherwise.

## 20.2 Chi et al. (2022): High Dimensional RFs

Following [Chi et al. \(2022\)](#), for a RF model where

- a sequence of distinct  $\Theta_{1:k}$  results in a distinct tree
- every set of available features  $\Theta_{l,s}$ ,  $l = 1, \dots, k$ ;  $s = 1, \dots, 2^{l-1}$

**Column subsampling** Define a **column subsampling** procedure:  $\Theta_{l,s}, \forall l, s$  has  $\lceil \gamma_0 p \rceil$  distinct integers among  $1, \dots, p$ , with  $\lceil \cdot \rceil$  the ceiling function for some  $0 < \gamma_0 \leq 1$ .  $\gamma_0$  is the predetermined constant parameter of column subsampling. Introduce the boldface random mappings  $\Theta_{1:k}$ , which are independent and uniformly distributed over all possible  $\Theta_{1:k}$  for all integer  $k$ . Then random forests estimate for  $\mathbf{c}$  with observations  $\mathcal{X}_n$  is given by

$$\mathbb{E}(\hat{m}_{T(\Theta_{1:k})}(\mathbf{c}, \mathcal{X}_n) \mid \mathcal{X}_n) = \sum_{\Theta_{1:k}} \mathbb{P} \left( \bigcap_{s=1}^k \{\Theta_s = \Theta_s\} \right) \hat{m}_{T(\Theta_{1:k})}(\mathbf{c}, \mathcal{X}_n)$$

The expectation is taken over sets of available features.

**Observation resampling** Let  $A = \{a_1, \dots, a_B\}$  be a set of subsamples with each  $a_i$  consisting of  $\lceil bn \rceil$  observations (indices) drawn without replacement from  $\{1, \dots, n\}$  for some positive integer  $B$  and  $0 < b \leq 1$ ; in addition, each  $a_i$  is independent of model training. The default values of  $B$  and  $b$  are 500 and 0.632<sup>1</sup>. Then the tree estimate using subsample  $a$  is define as

$$\hat{m}_{T(\Theta_{1:k}),a}(\mathbf{c}, \mathcal{X}_n) := \sum_{(\mathbf{t}_1, \dots, \mathbf{t}_k) \in T(\Theta_{1:k})} \mathbf{1}_{\mathbf{c} \in \mathbf{t}_k} \left( \frac{\sum_{i \in a \cap \{i: \mathbf{x}_i \in \mathbf{t}_k\}} y_i}{\#(a \cap \{i: \mathbf{x}_i \in \mathbf{t}_k\})} \right)$$

the random forests estimate given  $A$  is then

$$B^{-1} \sum_{a \in A} \mathbb{E}[\hat{m}_{T,a}(\Theta_{1:k}, \mathbf{c}, \mathcal{X}_n) \mid \mathcal{X}_n] := B^{-1} \sum_{a \in A} \mathbb{E}[\hat{m}_{T(\Theta_{1:k}),a}(\mathbf{c}, \mathcal{X}_n) \mid \mathcal{X}_n]$$

**CART-split criterion** Given a cell  $\mathbf{t}$ , a subset of observation indices  $a$  and a set of available features  $\Theta \subset \{1, \dots, p\}$ , the CART-split is defined as

$$(\hat{j}, \hat{c}) = \arg \min_{j \in \Theta, c \in \{x_{ij}: \mathbf{x}_i \in \mathbf{t}, i \in a\}} \left[ \sum_{i \in a \cap P_L} (\bar{y}_L - y_i)^2 + \sum_{i \in a \cap P_R} (\bar{y}_R - y_i)^2 \right] \quad (20.1)$$

<sup>1</sup>Or,  $b = 1$  but observations are drawn with replacement.

where

$$P_L := \{i : \mathbf{x}_i \in \mathbf{t}, x_{ij} < c\} \quad P_R := \{i : \mathbf{x}_i \in \mathbf{t}, x_{ij} \geq c\}$$

$$\bar{y}_L := \sum_{i \in a \cap P_L} \frac{y_i}{\#(a \cap P_L)} \quad \bar{y}_R := \sum_{i \in a \cap P_R} \frac{y_i}{\#(a \cap P_R)}$$

The CART-split criterion conditional on the sample is a deterministic splitting criterion; conditioning on another sample leads to another deterministic splitting criterion. Define  $\hat{T}_a$  as the sample tree growing rule that is associated with a splitting criterion following Eq. (20.1), the tree estimates using  $\hat{T}_a$  can be similarly defined as

$$\hat{m}_{\hat{T}_a(\Theta_{1:k})}(\mathbf{c}, \mathcal{X}_n) := \sum_{(\mathbf{t}_1, \dots, \mathbf{t}_k) \in \hat{T}_a(\Theta_{1:k})} \mathbf{1}_{\mathbf{c} \in \mathbf{t}_k} \left( \frac{\sum_{i \in \{i: \mathbf{x}_i \in \mathbf{t}_k\}} y_i}{\#\{i : \mathbf{x}_i \in \mathbf{t}_k\}} \right)$$

the definition is the same for  $\hat{m}_{\hat{T}_a, a}$ . Then the random forests estimate for a test point  $\mathbf{c} \in [0, 1]^p$  is given by

$$B^{-1} \sum_{a \in A} \mathbb{E} \left( \hat{m}_{\hat{T}_a, a}(\Theta_{1:k}, \mathbf{c}, \mathcal{X}_n) \mid \mathcal{X}_n \right)$$

where the average and conditional expectation correspond to the sample and column subsamplings respectively, and they are interchangeable.

**Bias-variance decomposition** For a tree growing rule  $T$  and  $\Theta_{1:k}$ , the population version is defined as

$$m_{T(\Theta_{1:k})}^*(\mathbf{c}) := \sum_{(\mathbf{t}_1, \dots, \mathbf{t}_k) \in T(\Theta_{1:k})} \mathbf{1}_{\mathbf{c} \in \mathbf{t}_k} \mathbb{E}(m(\mathbf{X}) \mid \mathbf{X} \in \mathbf{t}_k) \quad (20.2)$$

for each test point  $\mathbf{c} \in [0, 1]^p$ . And the  $\mathbb{L}^2$  prediction loss for random forests is defined as

$$\mathbb{E} \left[ m(\mathbf{X}) - B^{-1} \sum_{a \in A} \mathbb{E} \left( \hat{m}_{\hat{T}_a, a}(\Theta_{1:k}, \mathbf{X}, \mathcal{X}_n) \mid \mathbf{X}, \mathcal{X}_n \right) \right]^2 \quad (20.3)$$

if we use the full sample  $a = \{1, \dots, n\}$ , and denote  $\hat{T}_a$  and  $\hat{m}_{\hat{T}_a, a}$  as  $\hat{T}$  and  $\hat{m}_{\hat{T}}$ , the sample subsampling and average  $B^{-1} \sum_{a \in A} (\cdot)$  in the random forests estimate are no longer needed, then Eq.(20.3) can be simplified as

$$\mathbb{E} \left[ m(\mathbf{X}) - \mathbb{E}(\hat{m}_{\hat{T}}(\Theta_{1:k}, \mathbf{X}, \mathcal{X}_n) \mid \mathbf{X}, \mathcal{X}_n) \right]^2$$

By Jensen's inequality and Cauchy-Schwarz inequality,

$$\begin{aligned} & \frac{1}{2} \mathbb{E} \left[ m(\mathbf{X}) - \mathbb{E}(\hat{m}_{\hat{T}}(\Theta_{1:k}, \mathbf{X}, \mathcal{X}_n) \mid \mathbf{X}, \mathcal{X}_n) \right]^2 \\ & \leq \underbrace{\mathbb{E} \left[ m(\mathbf{X}) - m_{\hat{T}}^*(\Theta_{1:k}, \mathbf{X}) \right]^2}_{\text{approximation error (squared bias)}} + \underbrace{\mathbb{E} \left[ m_{\hat{T}}^*(\Theta_{1:k}, \mathbf{X}) - \hat{m}_{\hat{T}}(\Theta_{1:k}, \mathbf{X}, \mathcal{X}_n) \right]^2}_{\text{estimation variance}} \end{aligned}$$

## Consistency of RF Models

For a cell  $\mathbf{t}$  and its two daughter cells  $\mathbf{t}'$  and  $\mathbf{t}''$ , define

$$\begin{aligned} \text{(I)}_{\mathbf{t}, \mathbf{t}'} &:= \mathbb{P}(\mathbf{X} \in \mathbf{t}' \mid \mathbf{X} \in \mathbf{t}) \text{Var}(m(\mathbf{X}) \mid \mathbf{X} \in \mathbf{t}') + \mathbb{P}(\mathbf{X} \in \mathbf{t}'' \mid \mathbf{X} \in \mathbf{t}) \text{Var}(m(\mathbf{X}) \mid \mathbf{X} \in \mathbf{t}'') \\ \text{(II)}_{\mathbf{t}, \mathbf{t}'} &:= \mathbb{P}(\mathbf{X} \in \mathbf{t}' \mid \mathbf{X} \in \mathbf{t}) [\mathbb{E}(m(\mathbf{X}) \mid \mathbf{X} \in \mathbf{t}') - \mathbb{E}(m(\mathbf{X}) \mid \mathbf{X} \in \mathbf{t})]^2 \\ &+ \mathbb{P}(\mathbf{X} \in \mathbf{t}'' \mid \mathbf{X} \in \mathbf{t}) [\mathbb{E}(m(\mathbf{X}) \mid \mathbf{X} \in \mathbf{t}'') - \mathbb{E}(m(\mathbf{X}) \mid \mathbf{X} \in \mathbf{t})]^2 \end{aligned}$$

and  $(\mathbb{I})_{\mathbf{t}, \mathbf{t}'}$  and  $(\mathbb{III})_{\mathbf{t}, \mathbf{t}'}$  are defined similarly.

in this context, we assume the following regularity conditions:

- **Absolutely continuous distribution**:  $f$ , the density function of  $\mathbf{X}$ , is bounded away from 0 and  $\infty$
- **Covariates and model errors**: assume  $p = O(n^{K_0})$  for  $K_0 > 0$ , and there is a symmetric distribution around 0 for  $\epsilon_1$ , s.t.  $\mathbb{E} |\epsilon_1|^q < \infty$  for sufficiently large  $q > 0$
- **Bounded regression functions**:  $\sup_{\mathbf{c} \in [0,1]^p} |m(\mathbf{c})| \leq M_0$ , for some  $M_0 > 0$
- **Sufficient impurity decrease**:  $\exists \alpha_1 \geq 1$  s.t.  $\forall \mathbf{t} = t_1 \times \dots \times t_p$ ,

$$\text{Var} [m(\mathbf{X}) \mid \mathbf{X} \in \mathbf{t}] \leq \alpha_1 \sup_{j \in \{1, \dots, p\}, c \in t_j} (\mathbb{III})_{\mathbf{t}, \mathbf{t}(j, c)}$$

where

- $(\mathbb{III})_{\mathbf{t}, \mathbf{t}'}$ : conditional bias decrease (or conditional impurity decrease)
- $\text{Var} [m(\mathbf{X}) \mid \mathbf{X} \in \mathbf{t}]$ : conditional **total** bias,  $\text{Var} [m(\mathbf{X}) \mid \mathbf{X} \in \mathbf{t}] = (\mathbb{I})_{\mathbf{t}, \mathbf{t}'} + (\mathbb{III})_{\mathbf{t}, \mathbf{t}'}$
- **Intuition**: having large conditional bias decrease on each cell is a desired property for achieving a good control of the squared bias of random forests estimate

**Sufficient impurity decrease (SID)** Define the functional class

$$\text{SID}(\alpha) := \{m(\mathbf{X}) : m(\mathbf{X}) \text{ satisfies SID with } \alpha_1 \leq \alpha\}$$

the size of  $\text{SID}(\alpha)$  is **non-decreasing** in  $\alpha \geq 1$ : if  $m(\mathbf{X}) \in \text{SID}(\alpha - c)$  for some  $\alpha - c \geq 1$  and  $c > 0$ , then  $m(\mathbf{X}) \in \text{SID}(\alpha)$ <sup>2</sup>.

Under the regularity conditions mentioned above, we have the following theorem

#### Theorem 20.2.1: Consistency

Let  $0 < b \leq 1, 0 < \gamma_0 \leq 1, \alpha_2 > 1, 0 < \eta < 1/8, 0 < c < 1/4$  and  $\delta > 0, 2\eta < \delta < \frac{1}{4}$ . Let  $A = \{a_1, \dots, a_B\}$  with  $\#a_i = \lceil bn \rceil$  for  $i = 1, \dots, B$  and  $a \in A$ . Then  $\exists C > 0$  s.t. for all large  $n$  and each  $1 \leq k \leq c \log_2 \lceil bn \rceil$ ,

$$\begin{aligned} & \mathbb{E} \left[ m(\mathbf{X}) - \mathbb{E} \left( \hat{m}_{\hat{T}_{a,a}}(\Theta_{1:k}, \mathbf{X}, \mathcal{X}_n) \mid \mathbf{X}, \mathcal{X}_n \right) \right]^2 \\ & \leq C \left[ \alpha_1 (\lceil bn \rceil)^{-\eta} + \left( 1 - \gamma_0 (\alpha_1 \alpha_2)^{-1} \right)^k + (\lceil bn \rceil)^{-\delta+c} \right] \end{aligned}$$

In addition, when aggregate over row subsamples (over  $a \in A$ ), get

$$\begin{aligned} & \mathbb{E} \left[ m(\mathbf{X}) - \frac{1}{B} \mathbb{E} \left( \hat{m}_{\hat{T}_{a,a}}(\Theta_{1:k}, \mathbf{X}, \mathcal{X}_n) \mid \mathbf{X}, \mathcal{X}_n \right) \right]^2 \\ & \leq C \left[ \alpha_1 (\lceil bn \rceil)^{-\eta} + \left( 1 - \gamma_0 (\alpha_1 \alpha_2)^{-1} \right)^k + (\lceil bn \rceil)^{-\delta+c} \right] \end{aligned}$$

Here, the feature dimensionality  $p$  and tree height  $k$  decide the number of all possible cells when growing trees.

<sup>2</sup>Many popular regression functions belong to this functional class.

**Bias-variance decomposition** Under Theorem (20.2.1), both bias and variance depend implicitly on  $p$  through  $n$  in the upper bounds.

**Proposition 20.2.2: Bias-Variance Decomposition**

Under Thm (20.2.1), for all large  $n$  and  $1 \leq k \leq c \log_2 n$ , it holds that

$$\begin{aligned} \text{Squared bias} &:= \underbrace{\mathbb{E} \left[ m(\mathbf{X}) - m_{\hat{T}}^*(\Theta_{1:k}, \mathbf{X}) \right]^2}_{\text{approximation error}} \\ &\leq O \left( n^{-\eta} + \underbrace{\left( 1 - \gamma_0 (\alpha_1 \alpha_2)^{-1} \right)^k}_{\text{Main term of bias}} \right) + \underbrace{O \left( n^{-\delta+c} \right)}_{\text{Uninteresting error}} \end{aligned}$$

where  $n^{-\eta}$  upper-bounds the error caused by the sample CART-splits. Under theoretical CART-splits,  $n^{-\eta}$  vanishes and  $\alpha_2 = 1$ . and

$$\begin{aligned} \text{Estimatio Variance} &:= \mathbb{E} \left[ m_{\hat{T}}^*(\Theta_{1:k}, \mathbf{X}) - \hat{m}_{\hat{T}}(\Theta_{1:k}, \mathbf{X}, \mathcal{X}_n) \right]^2 \\ &\leq O(n^{-\eta}) + \underbrace{O(n^{-\delta+c})}_{\text{Uninteresting error}} \end{aligned}$$

where the upper bound is conservative since we establish a uniform upper bound for the variances of **individual** trees.

**Relevant features** Under the regularity conditions, for some cells, only the splits along the relevant feature directions can reduce a sufficient amount of bias. To be precise, introduce a variance of SID with some  $S_0 \subset \{1, \dots, p\}$  below

- **Sufficient impurity decrease 2**:  $\exists \alpha_1 \geq 1$  s.t. for each cell  $\mathbf{t} = t_1 \times \dots \times t_p$

$$\text{Var}(m(\mathbf{X}) \mid \mathbf{X} \in \mathbf{t}) \leq \alpha_1 \sup_{j \in S_0, c \in t_j} (\mathbb{I})_{\mathbf{t}, \mathbf{t}(j,c)}$$

When the regularity conditions on the underlying regression function and SID are assumed, SID2 holds only if  $S_0$  includes all relevant features

**Definition 20.2.3: Relevant Features**

A feature  $j$  is said to be relevant for regression function  $m(\mathbf{X})$  if and only if there exists some constant  $\iota > 0$  s.t.

$$\mathbb{E} \left[ \text{Var}(m(\mathbf{X}) \mid X_s, s \in \{1, \dots, p\} \setminus \{j\}) \right] > \iota$$

Then, the magnitude of the  $\mathbb{L}_2$  loss when a relevant feature is left out during the model training

**Theorem 20.2.4:  $\mathbb{L}_2$  Loss of Missing A Relevant Feature**

Under the condition of **covariates and model errors** and **bounded regression functions**, some relevant feature  $j$  is not involved in the random forests model training procedure, then

$$\mathbb{E} \left[ m(\mathbf{X}) - \frac{1}{B} \sum_{a \in A} \mathbb{E} \left( \hat{m}_{\hat{T}_a, a}(\boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_k, \mathbf{X}, \mathcal{X}_n) \mid \mathbf{X}, \mathcal{X}_n \right) \right]^2 \geq \iota$$

## References

Chien-Ming Chi, Patrick Vossler, Yingying Fan, and Jinchi Lv. Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*, 50(6):3415–3438, 2022.