Asymptotic properties of high-dimensional random forests

Author: Chien-Ming Chi, Patrick Vossler, Yingying Fan, and Jinchi Lv

Presenter: Beining Wu

July 5th, 2022

Outline

- Random forests: a quick review
  - Notations
  - Model settings
  - Generic approach
  - CART-split criterion
  - Bias-variance decomposition
- Preliminaries to proof
  - Technical notations
  - Critical conditions: SID
- Bounding the squared bias
  - Relaxed tree growing rules
  - Modifying sample tree
- Bounding the estimation variance
- Main results and extensions
  - Convergence rates and interpretation
  - Sharper rate for binary feature model
  - Relevant features

Outline

- Random forests: a quick review
  - Notations
  - Model settings
  - Generic approach
  - CART-split criterion
  - Bias-variance decomposition
- Preliminaries to proof
  - Technical notations
  - Critical conditions: SID
- Bounding the squared bias
  - Relaxed tree growing rules
  - Modifying sample tree
- Bounding the estimation variance
- Main results and extensions
  - Convergence rates and interpretation
  - Sharper rate for binary feature model
  - Relevant features

Notations

- $(\Omega, \mathcal{F}, \mathbb{P})$: Underlying probability space
- $\#S$: Cardinality of a finite set S
- $|t| = \sup t - \inf t$: Length of an interval
- $A_{1:k}$ for $A_1, \ldots, A_k$
- $a_n = o(b_n)$ if $\lim_n a_n/b_n = 0$
- $a_n = O(b_n)$ if $\limsup_n |a_n|/|b_n| < \infty$
- $[n]$ for $\{1, \ldots, n\}$. Often, $[n]$ represents the samples and $[p]$ represents the features.

Model settings

- Observation: $(\mathbf{x}_i, y_i) \in [0,1]^p \times \mathbb{R}$. i.i.d.
- Underlying model:

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

  where m is the unknown regression function and ε_i is the independent noise.
- Goal: learn the regression function and predict $m(\boldsymbol{X})$ for $\boldsymbol{X}$ i.i.d. to $\mathbf{x}_i$.

Question: How is the quality of our prediction function $\widehat{m}(\boldsymbol{X})$ in terms of the L^2 prediction error?

Generic approach of random forest algorithm

Intuition: If the regression function is smooth, then it's reasonable to use the sample label mean of the near points.

$$\Downarrow$$

Partition the feature space into small disjoint regions $A_1, \ldots, A_m$. Then using the sample in the same region to give an estimate

$$\widehat{m}(\mathbf{c}) = \sum_{j=1}^{m} \mathbf{1}_{\mathbf{c} \in A_j} \frac{\sum_{i:\mathbf{x}_i \in A_j} y_i}{\#\{i : \mathbf{x}_i \in A_j\}}$$

One-at-a-time partition might miss the information? Recursively refine the partition! **Decision tree.**

Generic approach of random forest algorithm

Random forest algorithm: Each time recursively partition the feature space into tree-structured disjoint cells. Give an individual tree estimate. Resample the columns and observations to form a group of individual trees, i.e., a forest. Average out to give a random forests estimate.

Concepts

- Cell: $\mathbf{t} = t_1 \times t_2 \times \dots t_p \subset [0,1]^p$, t_i is interval.
- Split: (j, c). j is the feature to split. c is the location.
- Available features for a cell: $\Theta \subset [p]$. This set specifies the available features for splitting.

Generic approach of random forest algorithm

Before building a tree, assume that we have access to $\Theta_{1:k}$, where $\Theta_k := \{\Theta_{k,1}, \Theta_{k,2}, \ldots, \Theta_{k,2^{k-1}}\}$. Here two subscript means the available set for one cell, single subscript means the collection of the available sets for one level, and $\Theta_{1:k}$ means the collection of all the available sets (for every cells in a level-k tree).

Generic approach of random forest algorithm

Procedure

- At beginning: root cell $\mathbf{t}_0 = [0,1]^p$.
- Assume that we have built $k-1$ levels. For the l cell in the $k-1$-th level, decide the split $(j,c), j \in \Theta_{k,l}$ based on **certain criterion**.
- After split: $\mathbf{t} \to \mathbf{t}(j,c) = t_1 \times \cdots \times t_{j-1} \times (t_j \cap [0,c)) \times \cdots t_p$ (left daughter) and $\mathbf{t} \setminus \mathbf{t}(j,c)$ (right daughter).
- Repeat to build a tree of level k.

[Use a graph to illustrate.]

Generic approach of random forest algorithm

Splitting criterion: Could be data-dependent or deterministic. By data-dependent we mean (j, c) should be chosen according to the data. By deterministic we mean the observed data won't affect the choice of (j, c).

We also refer to the splitting criterion as the **tree growing rule**, and use

- T: deterministic rule
- $\widehat{T}$: data-dependent rule

to represent the abstract growing rules.

A **branch** is the a k-tuple $(\mathbf{t}_1, \dots, \mathbf{t}_k)$.
A **tree** is the collection of all the branches, namely

$$\{(\mathbf{t}_1, \dots, \mathbf{t}_k) : \mathbf{t}_j \text{ is the daughter cell of } \mathbf{t}_{j-1}, \forall j \leq k\}.$$

Generic approach of random forest algorithm

A level-k tree has 2^k end cells, each corresponds to one and only one branches.

Given a deterministic tree growing rule T and the all the available sets $\Theta_{1:k}$, we can immediately construct the tree, denoted as $T(\Theta_{1:k})$.

If the tree growing rule is data-dependent, then given the observed data $\mathcal{X}_n = \{(\mathbf{x}_i, y_i)\}_{i \in [n]}$, we can grow the sample tree $\widehat{T}(\Theta_{1:k})$

Generic approach of random forest algorithm

Tree estimate: Let $\mathcal{X}_n = \{(\mathbf{x}_i, y_i)\}_{i\in[n]}$. For fixed $\mathbf{c} \in [0,1]^p$ and $\Theta_{1:k}$,

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

which is the sample conditional average of the responses on the cell that is same to $\mathbf{c}$.

Column subsampling: Randomize $\Theta_{1:k}$. $\boldsymbol{\Theta}_{1:k}$ is the random element on the collection of all the possible $\Theta_{1:k}$. Random forest estimate of $m(\mathbf{c})$ is

$$\mathbb{E}[\widehat{m}_{T(\boldsymbol{\Theta}_{1:k})}(\mathbf{c}, \mathcal{X}_n)|\mathcal{X}_n] = \sum_{\Theta_{1:k}} \mathbb{P}\big(\bigcap_s \{\boldsymbol{\Theta}_s = \Theta_s\}\big)\widehat{m}_{T(\Theta_{1:k})}(\mathbf{c}, \mathcal{X}_n). \qquad (1)$$

Generic approach of random forest algorithm

In practice we use independent column sampling, which means each $\boldsymbol{\Theta}_{k,l} \subset [p]$ is drawn without replacement from a black box with p features. Let $\gamma_0 \in (0, 1]$ be the sample rate, we need to do $2^k - 1$ times of independent sampling, each has $\binom{p}{\lceil \gamma_0 p \rceil}$ possibilities., therefore

$$\mathbb{P}\big(\bigcap_s \{\boldsymbol{\Theta}_s = \Theta_s\}\big) = \binom{p}{\lceil \gamma_0 p \rceil}^{1-2^k}.$$

Generic approach of random forest algorithm

Row subsampling: Resample the observations. $A = \{a_1, \ldots, a_B\}$ with $a_i \subset [n]$ independently drawn from $[n]$ without replacement and $\#a_i = \lceil bn \rceil, b \in (0, 1]$. Then we replace the full-sample conditional average with the a-subsample version.

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

The random forests estimate given A is then defined as

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

CART-split criterion

The (sample) **classification and regression tree, CART** criterion is a data-dependent splitting criterion. Given a cell $\mathbf{t}$, available set $\Theta \subset [p]$ and subsample indices $a \subset [n]$,

$$(\widehat{j}, \widehat{c}) = \operatorname*{argmin}_{j \in \Theta, c \in \{\mathbf{x}_i^j, i \in a\}} \left[\sum_{i \in P_L \cap a} (\overline{y}_L - y_i)^2 + \sum_{i \in P_R \cap a} (\overline{y}_R - y_i)^2 \right].$$

Where $P_L = \{\mathbf{x}_i : \mathbf{x}_i^j < c\}$ and $P_R = \{\mathbf{x}_i : \mathbf{x}_i^j \geq c\}$, and

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

If there is only one or zero points in $\mathbf{t}$, then we choose the split randomly.

CART-split criterion

For data-dependent growing rule $\widehat{T}$, the tree we construct depends on the data and the subsample indices. So we use $\widehat{T}_a(\Theta_{1:k})$ to represent the sample tree.

$$\widehat{m}_{\widehat{T}_a}(\Theta_{1:k}\mathbf{c}, \mathcal{X}_n) = \sum_{(\mathbf{t}_1,\ldots,\mathbf{t}_k)\in\widehat{T}_a(\Theta_{1:k})} \mathbf{1}_{\mathbf{c}\in\mathbf{t}_k}\left(\frac{\sum_{i:\mathbf{x}_i\in\mathbf{t}_k} y_i}{\#\{i : \mathbf{x}_i \in \mathbf{t}_k\}}\right). \tag{2}$$

Analogously, we can define $\widehat{m}_{\widehat{T}_a,a}$ by replacing the part in the bracket.

Notice: when use $\widehat{T}_a$, we emphasize that the tree construction depends on the subsamples we use. When use $\hat{m}_a$, we emphasize that the conditional average is calculated with the subsamples.

CART-split criterion

Theoretical CART Theoretical CART criterion is the deterministic counterpart of the sample CART criterion. This criterion exploits the oracle knowledge of $m(\boldsymbol{X})$, as

$$\begin{aligned}(j^*, c^*) = \operatorname*{argmin}_{j \in \Theta, c \in \mathbf{t}_j} \, & \mathbb{P}(\boldsymbol{X} \in \mathbf{t}' | \boldsymbol{X} \in \mathbf{t}) \operatorname{var}(m(X) | \boldsymbol{X} \in \mathbf{t}') \\ & + \mathbb{P}(\boldsymbol{X} \in \mathbf{t}'' | \boldsymbol{X} \in \mathbf{t}) \operatorname{var}(m(X) | \boldsymbol{X} \in \mathbf{t}'').\end{aligned}$$

It's the population version of the sample CART. We will mention this criterion later.

Bias-variance decomposition

The main topic of the paper is

How does sample-CART random forest behaves in terms of prediction?

We start from the bias-variance decomposition. The L^2 prediction error of random forests estimation is

$$\mathbb{E}\left[\left(m(\boldsymbol{X}) - B^{-1}\sum_{a\in A}\mathbb{E}[\widehat{m}_{\widehat{T}_a(\boldsymbol{\Theta}_{1:k}),a}(\boldsymbol{X},\mathcal{X}_n)|\mathcal{X}_n]\right)^2\right].$$

We may omit the row subsampling for simplicity.

Bias-variance decomposition

For a tree growing rule T and $\Theta_{1:k}$, we now define

$$m^*_{T(\Theta_{1:k})}(\mathbf{c}) = \sum_{(\mathbf{t}_1,\ldots,\mathbf{t}_k)\in T(\Theta_{1:k})} \mathbf{1}_{\mathbf{c}\in\mathbf{t}_k}\mathbb{E}[m(\boldsymbol{X})|\boldsymbol{X}\in\mathbf{t}_k].$$

which is the population version of $\widehat{m}_{T(\Theta_{1:k})}(\mathbf{c})$. We can bound the prediction error as

$$\mathbb{E}\left[\Big(m(\boldsymbol{X}) - \mathbb{E}[\widehat{m}_{\widehat{T}}(\boldsymbol{\Theta}_{1:k},\boldsymbol{X},\mathcal{X}_n)|\boldsymbol{X},\mathcal{X}_n]\Big)^2\right]$$
$$\leq 2\underbrace{\mathbb{E}\left[\Big(m(\boldsymbol{X}) - m^*_{\widehat{T}}(\boldsymbol{\Theta}_{1:k},\boldsymbol{X})\Big)^2\right]}_{\text{Squared bias}} + 2\underbrace{\mathbb{E}\left[\Big(m^*_{\widehat{T}}(\boldsymbol{\Theta}_{1:k},\boldsymbol{X}) - \widehat{m}_{\widehat{T}}(\boldsymbol{\Theta}_{1:k},\boldsymbol{X},\mathcal{X}_n)\Big)^2\right]}_{\text{Estimation variance}}.$$

We now turn to bound two terms respectively.

Outline

Random forests: a quick review

- Notations
- Model settings
- Generic approach
- CART-split criterion
- Bias-variance decomposition

Preliminaries to proof

- Technical notations
- Critical conditions: SID

Bounding the squared bias

- Relaxed tree growing rules
- Modifying sample tree

Bounding the estimation variance

Main results and extensions

- Convergence rates and interpretation
- Sharper rate for binary feature model
- Relevant features

Technical notations

For a given cell $\mathbf{t}$ and its daughter cells $\mathbf{t}', \mathbf{t}''$. We define the following quantity.

$$\begin{aligned}
(\mathbf{I})_{\mathbf{t},\mathbf{t}'} =& \mathbb{P}(\boldsymbol{X} \in \mathbf{t}' | \boldsymbol{X} \in \mathbf{t}) \, \mathrm{var}(m(\boldsymbol{X}) | \boldsymbol{X} \in \mathbf{t}') \\
&+ \mathbb{P}(\boldsymbol{X} \in \mathbf{t}' | \boldsymbol{X} \in \mathbf{t}) \, \mathrm{var}(m(\boldsymbol{X}) | \boldsymbol{X} \in \mathbf{t}''), \\
(\mathbf{II})_{\mathbf{t},\mathbf{t}'} =& \mathbb{P}(\boldsymbol{X} \in \mathbf{t}' | \boldsymbol{X} \in \mathbf{t}) \big(\mathbb{E}[m(\boldsymbol{X}) | \boldsymbol{X} \in \mathbf{t}'] - \mathbb{E}[m(\boldsymbol{X}) | \boldsymbol{X} \in \mathbf{t}] \big)^2 \\
&+ \mathbb{P}(\boldsymbol{X} \in \mathbf{t}'' | \boldsymbol{X} \in \mathbf{t}) \big(\mathbb{E}[m(\boldsymbol{X}) | \boldsymbol{X} \in \mathbf{t}''] - \mathbb{E}[m(\boldsymbol{X}) | \boldsymbol{X} \in \mathbf{t}] \big)^2.
\end{aligned}$$

By the conditional variance decomposition formula, we have

$$\underbrace{\mathrm{var}(m(\boldsymbol{X}) | \boldsymbol{X} \in \mathbf{t})}_{\text{Conditional total bias}} = \underbrace{(\mathbf{I})_{\mathbf{t},\mathbf{t}'}}_{\text{Conditional remaining bias}} + \underbrace{(\mathbf{II})_{\mathbf{t},\mathbf{t}'}}_{\text{Conditional bias decrease}} .$$

Technical notations

Interpretations: The conditional variance actually characterizes the prediction error using m^*. Define the local version of m^*:

$$f_1(\boldsymbol{X}) = \mathbf{1}_{\boldsymbol{X}\in\mathbf{t}}\mathbb{E}[m(\boldsymbol{X})|\boldsymbol{X}\in\mathbf{t}],$$

or finer version

$$f_2(\boldsymbol{X}) = \mathbf{1}_{\boldsymbol{X}\in\mathbf{t}'}\mathbb{E}[m(\boldsymbol{X})|\boldsymbol{X}\in\mathbf{t}'] + \mathbf{1}_{\boldsymbol{X}\in\mathbf{t}''}\mathbb{E}[m(\boldsymbol{X})|\boldsymbol{X}\in\mathbf{t}''].$$

Then we have

$$\begin{aligned}\mathbb{E}\big[(m(\boldsymbol{X}) - f_1(\boldsymbol{X}))^2|\boldsymbol{X}\in\mathbf{t}\big] &= \mathrm{var}(m(\boldsymbol{X})|\boldsymbol{X}\in\mathbf{t}),\\ \mathbb{E}\big[(m(\boldsymbol{X}) - f_2(\boldsymbol{X}))^2\big|\boldsymbol{X}\in\mathbf{t}] &= (\mathbf{I})_{\mathbf{t},\mathbf{t}'}.\end{aligned}$$

Technical notations

Intuition

- Good splitting should have large conditional bias decrease. That's how CART derive, recall that theoretical CART choose

$$(j^*, c^*) = \underset{j \in \Theta, c \in [0,1]}{\operatorname{argmax}} \, (\mathbf{II})_{\mathbf{t}, \mathbf{t}(j,c)}.$$

- To guarantee the global performance, we should expect that for every cell we can find some split that sufficiently decrease the conditional total bias.

Main condition: sufficient impurity decrease

Based on the observation before, we have

Condition 1 (Sufficient impurity decrease, SID).

The function m is said to follow the SID condition if there exists a constant $\alpha \geq 1$ such that for every cell $\mathbf{t}$,

$$\operatorname{var}(m(\boldsymbol{X})|\boldsymbol{X} \in \mathbf{t}) \leq \alpha_1 \sup_{j \in [p], c \in t_j} (\mathbf{II})_{\mathbf{t}, \mathbf{t}(j,c)}.$$

Main condition: sufficient impurity decrease

Interpretations

- We can always find appropriate splitting method to decrease the approximation bias by a proportion of α_1^{-1}.
- Roughly speaking, theoretical CART criterion should be able to decrease the prediction error of m^* at geometric rate.

Applicability of the SID conditions

The original paper provides a sequence of examples such that SID condition holds.

Specifically, they showed that the SID condition allows for the dependent features, sparse quadratic regression models and some additive models for some cases.

Connection to sparsity For s^*-sparse quadratic regression model, the authors proved that it satisfy the SID condition with $\alpha_1 = cs^*$ for some c. This means, if a model contains more active features, then a split on single feature would reduce less conditional bias.

Regularity conditions

We list rest of the conditions here.

Condition 2.

The features are continuously distributed and have density $f(\mathbf{x})$ that is bounded away from 0 and ∞.

Condition 3.

There exists some positive constants q and K_0 such that $p = O(n^{K_0})$ and $\mathbb{E}[|\varepsilon|^q] < \infty$. Moreover, ε is symmetrically distributed.

Condition 4.

The regression function $m(\mathbf{x})$ is bounded, in the sense that for some $M_0 > 0$, $|m(\mathbf{x})| < M_0, \forall \mathbf{x} \in [0,1]^p$.

Outline

- Random forests: a quick review
  - Notations
  - Model settings
  - Generic approach
  - CART-split criterion
  - Bias-variance decomposition
- Preliminaries to proof
  - Technical notations
  - Critical conditions: SID
- Bounding the squared bias
  - Relaxed tree growing rules
  - Modifying sample tree
- Bounding the estimation variance
- Main results and extensions
  - Convergence rates and interpretation
  - Sharper rate for binary feature model
  - Relevant features

Bounding the squared bias

We consider the first term in the bias-variance decomposition, namely

$$\mathbb{E}\left[\left(m(\boldsymbol{X}) - m^*_{\widehat{T}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X})\right)^2\right].$$

Difficulty: $\widehat{T}$ depends on $\mathcal{X}_n$, has random boundaries.

Our strategy

- First relax to a class of deterministic criterion, and derive uniform upper bound.
- Slightly modify the sample tree, and prove that the modified tree realization belongs to the class above with high probability.

Relaxed tree growing rules

The following condition characterizes a class of splitting criterion and associated growing rules

Condition 5.

There exists some $\varepsilon \geq 0, \alpha_2 \geq 1$, such that, for any sets of available features $\Theta_{1:k}$, the corresponding tree $T(\Theta_{1:k})$ satisfy: for every branches $(\mathbf{t}_1, \ldots, \mathbf{t}_k)$ and every $1 \leq l \leq k$, we have

1. If $(\mathbf{II})_{\mathbf{t}_{l-1}, \mathbf{t}_l} < \varepsilon$, then $\sup_{j \in \Theta_l, c} (\mathbf{II})_{\mathbf{t}_{l-1}, \mathbf{t}_{l-1}(j,c)} \leq \alpha_2 \varepsilon$;
2. If $(\mathbf{II})_{\mathbf{t}_{l-1}, \mathbf{t}_l} \geq \varepsilon$, then $\sup_{j \in \Theta_l, c} (\mathbf{II})_{\mathbf{t}_{l-1}, \mathbf{t}_{l-1}(j,c)} \leq \alpha_2 (\mathbf{II})_{\mathbf{t}_{l-1}, \mathbf{t}_l}$.

Here $\Theta_l \subset [p]$ is the available feature set corresponds to $\mathbf{t}_{l-1}$, with slightly abused notation.

Relaxed tree growing rules

Observation: if $\varepsilon = 0$ and $\alpha_2 = 1$, this exactly characterized the theoretical CART criterion. So it can be seen as the relaxation of the theoretical CART criterion.

Theorem 3.1.

Assume that Condition 1 holds with $\alpha_1 \geq 1$, $\operatorname{var}(m(\boldsymbol{X})) < \infty$, *and the tree growing rule* T *satisfies Condition 5 with some integer* $k > 0$, $\varepsilon \geq 0$, *and* $\alpha_2 \geq 1$. *Then for each* $0 < \gamma_0 \leq 1$, *we have*

$$\mathbb{E}\Big(m(\boldsymbol{X}) - m_T^*\left(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X}\right)\Big)^2 \leq \alpha_1\alpha_2\varepsilon + \left(1 - \gamma_0(\alpha_1\alpha_2)^{-1}\right)^k \operatorname{var}(m(\boldsymbol{X})).$$

Modifying sample tree

The sample CART tree does not necessarily follow the condition 5, but we may expect that it's very close to it because it's the sample version of the theoretical CART tree.

Modification Given sample tree $\widehat{T}(\Theta_{1:k})$ and $\zeta > 0$,

- For each branch such that $\mathbb{P}(\boldsymbol{X} \in \mathbf{t}_{k-1}) < \zeta$ define $l_0 = \min\{l - 1 : \mathbb{P}(\boldsymbol{X} \in \mathbf{t}_{l-1}) < \zeta\}$.
- For those branches, trim off the subtree rooted from $\mathbf{t}_{l_0}$.
- From each cutting point, grow the tree back with the theoretical CART criterion.

Modifying sample tree

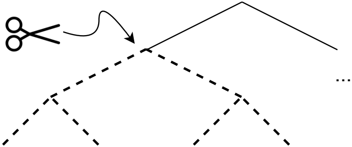

Figure: Trimming off the sample cells with small probability

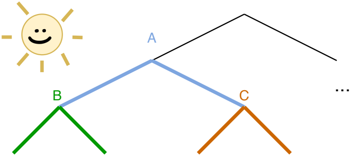

Figure: Grow back the tree with the theoretical CART

Modifying sample tree

We call this modified sample tree as **semi-sample tree**, and denote it with $\widehat{T}_\zeta$. Next theorem asserts that this modified tree satisfy the Condition 5 with high probability.

Theorem 3.2.

Assume that Conditions 2–4 hold and let $\alpha_2 > 1$, $0 < \eta < \frac{1}{8}$, $c > 0$, and δ with $2\eta < \delta < \frac{1}{4}$ be given. Then there exists an $\mathcal{X}_n$-measurable event $\boldsymbol{U}_n$ with $\mathbb{P}(\boldsymbol{U}_n^c) = o(n^{-1})$ such that conditional on $\mathcal{X}_n$, on event $\boldsymbol{U}_n$ and for all large n, $\widehat{T}_\zeta$ with $\zeta = n^{-\delta}$ satisfies Condition 5 with $k = \lfloor c \log_2 n \rfloor$, $\varepsilon = n^{-\eta}$, and α_2.

Bounding the squared bias

With Theorem 3.1 and Theorem 3.2, we're now able to establish the upper bound for the mean squared bias.

Lemma 1.

Assume that Conditions 1–4 hold and let $0 < \gamma_0 \leq 1$, $\alpha_2 > 1$, $0 < \eta < \frac{1}{8}$, δ *with* $2\eta < \delta < \frac{1}{4}$, *and* $c > 0$ *be given. Then, for all large* n *and each* $1 \leq k \leq c \log_2 n$,

$$\begin{aligned}\mathbb{E}\left[\left(m(\boldsymbol{X}) - m^*_{\widehat{T}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X})\right)^2\right] \leq & 8M_0^2 n^{-\delta} 2^k + 2\alpha_1\alpha_2 n^{-\eta} \\ & + 2M_0^2(1 - \gamma_0(\alpha_1\alpha_2)^{-1})^k + 2n^{-1}.\end{aligned}$$

Essentially, the first term of RHS comes from sample tree modification, and the last term comes from $\mathbf{U}_n^c$.

Outline

Random forests: a quick review

- Notations
- Model settings
- Generic approach
- CART-split criterion
- Bias-variance decomposition

Preliminaries to proof

- Technical notations
- Critical conditions: SID

Bounding the squared bias

- Relaxed tree growing rules
- Modifying sample tree

Bounding the estimation variance

Main results and extensions

- Convergence rates and interpretation
- Sharper rate for binary feature model
- Relevant features

Bounding the estimation variance

We now turn to bound the second term, namely

$$\mathbb{E}\left[\left(m^*_{\widehat{T}}(\mathbf{\Theta}_{1:k},\boldsymbol{X}) - \widehat{m}_{\widehat{T}}(\mathbf{\Theta}_{1:k},\boldsymbol{X},\mathcal{X}_n)\right)^2\right].$$

The deviation of m^* and $\widehat{m}$ is nothing but the difference of (conditional) sample average and population mean.

Main obstacle: randomness of $\widehat{T}$. We cannot derive a uniform concentration bound for uncountably many cells.

Gridding the feature space

This technical obstacle may provide some motivation. We can solve uniform concentration problem for some random tree with finite possibilities. **So, can we slightly modify the sample tree to get a easy-to-tackle class?**

The answer is true. The uncountability of the sample tree comes from the continuity of feature space. So we can consider discretizing the feature space. A natural approach is grdding.

Grdding the feature space

Let ρ_1 be a positive constant and $b_i = i/\lceil n^{1+\rho_1} \rceil$ with $0 \leq i \leq \lceil n^{1+\rho_1} \rceil$.

For the j-th coordinate, we put $\lceil n^{1+\rho_1} \rceil$ hyperplanes, each perpendicular to the j-th axis and intersects the axis at b_i. This is equivalent to cutting a unit hypercube into disjoint hypercubes with edge length equals to $1/\lceil n^{1+\rho_1} \rceil$.

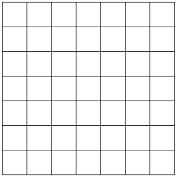

Figure: Gridding in the unit square

Tree modification

We want to modify the original sample-tree into the tree in the grid above, basically with the smallest non-singular cells equal to the hypercubes in the grid.

Assume that we have constructed the sample tree $T(\Theta_{1:k})$. For each cell $\mathbf{t}$, we move the boundaries of the cell to the nearest hyperplanes to form a new cell. For example, if $\mathbf{t} = t_1 \times t_2 \times \cdots \times t_p$, then we modify $t_j = (l_j, u_j)$ into $t_j^{\#} = (l_j', u_j')$. Here $l_j' = b_{i_0}$ with

$$i_0 = \operatorname*{argmin}_i |b_i - l_j|.$$

And $u_j^{\#}$ is defined analogously. Then we form the new cell as $\mathbf{t}^{\#} = t_1^{\#} \times \cdots \times t_p^{\#}$.

Tree modification

Key observation: The modification keeps the tree structure. If $\mathbf{t}', \mathbf{t}''$ are the daughter cells of $\mathbf{t}$, then $\mathbf{t}'^{\#}$ and $\mathbf{t}''^{\#}$ are the daughter cells of $\mathbf{t}^{\#}$.

If the original tree is grown using rule T, then we can find a corresponding rule $T^{\#}$, which grows tree in the grid.

Bounding estimation variance

Here we give the upper bound for the estimation variance using $T^{\#}$, which is a simple application of AM-QM inequality.

$$\underbrace{L^2 \text{ distance of } m^*_{\widehat{T}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X}) \text{ and } m^*_{\widehat{T}^{\#}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X})}_{\text{Controlled with Lemma ??}}$$

$$+ \underbrace{L^2 \text{ distance of } m^*_{\widehat{T}^{\#}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X}) \text{ and } \widehat{m}_{\widehat{T}^{\#}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X}, \mathcal{X}_n)}_{\text{Controlled with Theorem 4.1}}$$

$$+ \underbrace{L^2 \text{ distance of } \widehat{m}_{\widehat{T}^{\#}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X}, \mathcal{X}_n) \text{ and } \widehat{m}_{\widehat{T}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X}, \mathcal{X}_n)}_{\text{Controlled with Lemma 2}}.$$

Bounding the estimation variance

Next result gives the upper bound for the second term, which is the application of the uniform concentration over finite grid-based trees

Theorem 4.1.

Assume that Conditions 2–4 hold and let $0 < \eta < \frac{1}{4}$ and $0 < c < \frac{1}{4}$ be given. Then for all large n and each $1 \leq k \leq c \log_2 n$, we have

$$\mathbb{E}\left\{\sup_{T} \mathbb{E}\left[\left(m^*_{T\#}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X}) - \widehat{m}_{T\#}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X}, \mathcal{X}_n)\right)^2 \,\middle|\, \boldsymbol{\Theta}_{1:k}, \mathcal{X}_n\right]\right\} \leq n^{-\eta},$$

where the supremum is over all possible tree growing rules. Note that due to the use of the grid, the supremum can be simplified to a max over a finitely many tree growing rules.

Bounding the estimation variance

And the following lemma characterizes the deviance of $\widehat{T}$ and $\widehat{T}^{\#}$.

Lemma 2.

Assume that Conditions 2–4 hold and let $1/2 < \Delta < 1$ and $c > 0$ be given. Then there exists some constant $C > 0$ such that for all large n and each $1 \leq k \leq c \log_2 n$,

$$\mathbb{E}\left(m^*_{\widehat{T}^{\#}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X}) - m^*_{\widehat{T}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X})\right)^2 \leq C2^k n^{\Delta - 1} \tag{3}$$

and

$$\mathbb{E}\left(\widehat{m}_{\widehat{T}^{\#}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X}, \mathcal{X}_n) - \widehat{m}_{\widehat{T}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X}, \mathcal{X}_n)\right)^2 \leq C2^k n^{\Delta - 1}. \tag{4}$$

Bounding the estimation variance

Combine the arguments above, we see that

Lemma 3.

Assume that Conditions 2–4 hold and let $0 < \eta < 1/4$, $0 < c < 1/4$, *and* $\nu > 0$ *be given. Then there exists some constant* $C > 0$ *such that for all large* n *and each* $1 \leq k \leq c \log_2 n$,

$$\mathbb{E}\left(m^*_{\widehat{T}}(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k, \boldsymbol{X}) - \widehat{m}_{\widehat{T}}(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k, \boldsymbol{X}, \mathcal{X}_n)\right)^2 \leq n^{-\eta} + C 2^k n^{-\frac{1}{2}+\nu}. \quad (5)$$

Outline

Random forests: a quick review

- Notations
- Model settings
- Generic approach
- CART-split criterion
- Bias-variance decomposition

Preliminaries to proof

- Technical notations
- Critical conditions: SID

Bounding the squared bias

- Relaxed tree growing rules
- Modifying sample tree

Bounding the estimation variance

Main results and extensions

- Convergence rates and interpretation
- Sharper rate for binary feature model
- Relevant features

Convergence rate I

We're now able to establish formal convergence rate using the lemmas given above.

Theorem 5.1.

Assume that Conditions 1–4 hold and let $0 < b \leq 1$, $0 < \gamma_0 \leq 1$, $\alpha_2 > 1$, $0 < \eta < 1/8$, $0 < c < 1/4$, *and* $\delta > 0$ *be given with* $2\eta < \delta < \frac{1}{4}$*. Let* $A = \{a_1, \ldots, a_B\}$ *with* $\# a_i = \lceil bn \rceil$ *for* $i = 1, \cdots, B$ *and* $a \in A$ *be given. Then, there exists some* $C > 0$ *such that for all large* n *and each* $1 \leq k \leq c \log_2 \lceil bn \rceil$,

$$\begin{aligned}
&\mathbb{E}\Big(m(\boldsymbol{X}) - \mathbb{E}\Big(\widehat{m}_{\widehat{T}_a, a}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X}, \mathcal{X}_n) \;\Big|\; \boldsymbol{X}, \mathcal{X}_n\Big)\Big)^2 \\
&\quad \leq C\Big(\alpha_1(\lceil bn \rceil)^{-\eta} + (1 - \gamma_0(\alpha_1\alpha_2)^{-1})^k + (\lceil bn \rceil)^{-\delta + c}\Big).
\end{aligned}$$

Convergence rate

In addition, when we also aggregate over row subsamples (i.e., over $a \in A$), we have

$$\mathbb{E}\Big(m(\boldsymbol{X}) - \frac{1}{B}\sum_{a\in A}\mathbb{E}\Big(\widehat{m}_{\widehat{T}_a,a}(\boldsymbol{\Theta}_{1:k},\boldsymbol{X},\mathcal{X}_n) \;\Big|\; \boldsymbol{X},\mathcal{X}_n\Big)\Big)^2$$
$$\leq C\Big(\alpha_1(\lceil bn\rceil)^{-\eta} + (1-\gamma_0(\alpha_1\alpha_2)^{-1})^k + (\lceil bn\rceil)^{-\delta+c}\Big).$$

This is a simple corollary of last result, with Jensen's inequality.

Notice: The bound here is too conservative and is independent of the subsample parameter B.

Interpretations of the rate

How each factor influence the rate?

Corollary 5.1.

Under all the conditions of Theorem 5.1, for all large n and each $1 \le k \le c\log_2 n$, it holds for the two terms on the RHS of the decomposition that

$$\text{Squared bias} := \mathbb{E}\Big(m(\boldsymbol{X}) - m^*_{\widehat{T}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X})\Big)^2$$

$$\le O\Big(n^{-\eta} + \underbrace{(1-\gamma_0(\alpha_1\alpha_2)^{-1})^k}_{\text{Main term of bias}}\Big) + \underbrace{O(n^{-\delta+c})}_{\text{Uninteresting error}}$$

$$\text{Estimation variance} := \mathbb{E}\Big(m^*_{\widehat{T}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X}) - \widehat{m}_{\widehat{T}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X}, \mathcal{X}_n)\Big)^2$$

$$\le O(n^{-\eta}) + \underbrace{O(n^{-\delta+c})}_{\text{Uninteresting error}} \quad .$$

Interpretations of the rate

Remark.

- If we choose γ_0 near 1, then main term of the bias will decrease faster with k. This is because larger γ_0 tends to do more splitting on the relevant features.
- The uninteresting error comes from the technical approximation error, namely the modification of sample-tree.
- In the derivation of the upper bound of these two terms, we implicitly use the condition that $p = O(n^{K_0})$.

Convergence rate (cont'd)

If we set $\eta = \frac{1}{8} - \epsilon$, $\delta = \frac{1}{4} - \epsilon$, $c = \frac{1}{8}$, and $k = \lfloor \frac{1}{8} \log_2(n) \rfloor$ in Theorem 5.1, we obtain more informative convergence rate as shown in Corollary 5.2 below.

Corollary 5.2.

Assume that Conditions 1–4 hold and let $0 < \epsilon < \frac{1}{8}$, $\alpha_1 \geq 1$, $\alpha_2 > 1$, *and* $0 < \gamma_0 \leq 1$ *be given. For all large* n *and tree height* $k = \lfloor \frac{1}{8} \log_2 n \rfloor$,

$$\sup_{m(\boldsymbol{X}) \in \mathsf{SID}(\alpha_1)} \left[\mathbb{E}\Big(m(\boldsymbol{X}) - \mathbb{E}\Big(\widehat{m}_{\widehat{T}}(\boldsymbol{\Theta}_{1:k}, \boldsymbol{X}, \mathcal{X}_n) \;\Big|\; \boldsymbol{X}, \mathcal{X}_n \Big) \Big)^2 \right]$$

$$\leq O\Big(n^{-\frac{1}{8}+\epsilon} + \underbrace{(1 - \gamma_0(\alpha_1\alpha_2)^{-1})^{\lfloor \frac{\log_2(n)}{8} \rfloor}}_{\text{Main term of bias}} \Big) \leq O\Big(n^{-\frac{1}{8}+\epsilon} + n^{-\frac{\log_2(e)}{8} \times \frac{\gamma_0}{\alpha_1\alpha_2}} \Big),$$

Relevant features

In order for (sample or theoretical) CART criterion to sufficiently reduce the conditional bias, there should be some crucial features on which splitting can work.

Original SID Condition 1 put no restrictions on the features. Here we consider a more restricted version,

Condition 6.

For a feature subset $S_0 \subset [p]$ and $\alpha_2 \geq 1$, the SID2 condition is true if, for any cell $\mathbf{t} \subset [0,1]^p$.

$$\mathrm{var}(m(\boldsymbol{X})|\boldsymbol{X} \in \mathbf{t}) \leq \sup_{j \in S_0, c \in t_j} (\mathbf{II})_{\mathbf{t}, \mathbf{t}(j,c)}.$$

Relevant features

Interpretation

- The SID2 condition is stronger than SID, in the sense that if SID2 condition is true for S_0 and α_2, then the SID2 condition is true for any $S_1 \supset S_0$.
- If SID2 condition is true for some subset $S_0 \subset [p]$, then a simple application of Theorem 5.1 indicates that even if we don't split on the rest features, the consistency is still effective.

Relevant features

Definition 5.1 (Relevant features).

A feature j is said to be relevant feature if, for some constant $\iota > 0$ we have

$$\mathbb{E}[\mathrm{var}(m(\boldsymbol{X})|X_s, s \neq j)] > \iota.$$

If we exclude a relevant feature j and use the rest of the features to approximate the original regression function, then there is a positive gap in terms of the L^2 error.

Relevant features

How relevant feature influence the performance of random forests algorithm is revealed in the theorem below

Theorem 5.2.

Assume that the Condition 3 4 is satisfied, and the relevant feature j is not included in the training process of random forests, i.e., the available sets $\mathbf{\Theta}_{k,l}$ never includes j, then we have

$$\mathbb{E}\Big(m(\boldsymbol{X}) - \frac{1}{B}\sum_{a\in A}\mathbb{E}\Big(\widehat{m}_{\widehat{T}_a,a}(\mathbf{\Theta}_{1:k}, \boldsymbol{X}, \mathcal{X}_n) \;\Big|\; \boldsymbol{X}, \mathcal{X}_n\Big)\Big)^2 \geq \iota.$$

Conclusion: Consistency requires splitting on the relevant features.

Thanks for attending!