A walk in random forests

Erwan Scornet (École Polytechnique),
joint work with
Gérard Biau (University Paris 6),
Stéphane Gaïffas (École Polytechnique),
Jaouad Mourtada (École Polytechnique),
Jean-Philippe Vert (Institut Curie)

I Taller Interinstitucional de Ciencia de Datos e Inteligencia Artificial
October 2019

Random forests

Random forests are a class of algorithms used to solve regression and classification problems

- They are often used in applied fields since they handle high-dimensional settings.
- They have good predictive power and can outperform state-of-the-art methods.

Random forests

Random forests are a class of algorithms used to solve regression and classification problems

- They are often used in applied fields since they handle high-dimensional settings.
- They have good predictive power and can outperform state-of-the-art methods.

But mathematical properties of random forests remain a bit magical.

General framework of the presentation

Regression setting

We are given a training set $\mathcal{D}_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ where the pairs $(X_i, Y_i) \in [0,1]^d \times \mathbb{R}$ are *i.i.d.* distributed as (X, Y).

We assume that

$$Y = m(\mathbf{X}) + \varepsilon.$$

We want to build an estimate of the regression function m using random forest algorithm.

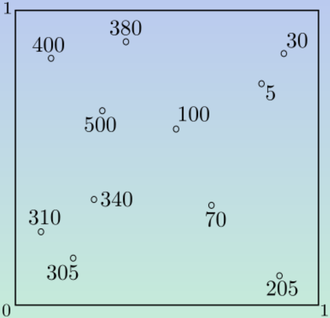
1. Day 1: Discovering random forests
   - Construction of random forests
   - Infinite forest
   - Centred Forests
   - Median forests

2. Day 2: Rate of consistency
   - Rate of consistency for centred forests
   - Rate of consistency for Median Forests
   - Minimax rates for Mondrian Forests

3. Day 3: Breiman's forests
   - Consistency of Breiman forests

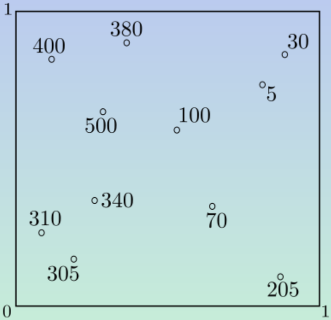
How to build a tree?

- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.



How to build a tree?

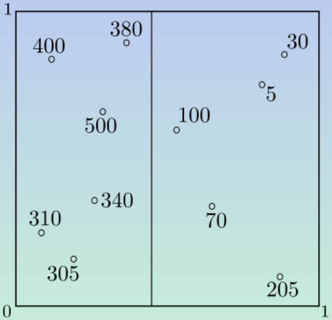
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

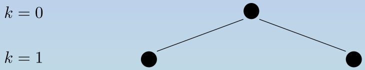


$k = 0$

How to build a tree?

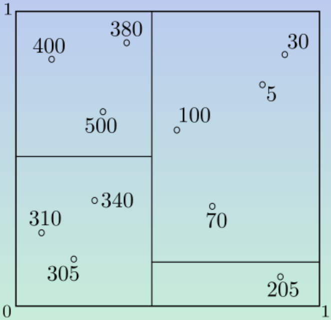
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

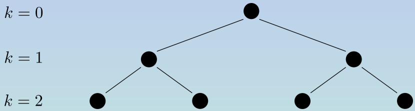




How to build a tree?

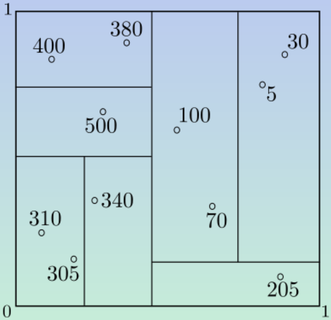
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

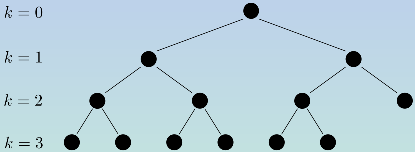




How to build a tree?

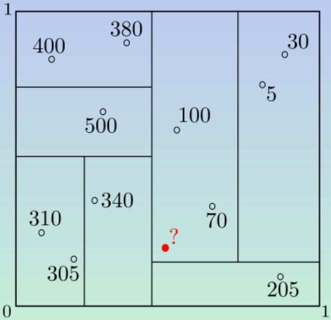
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

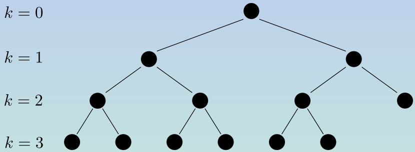




How to build a tree?

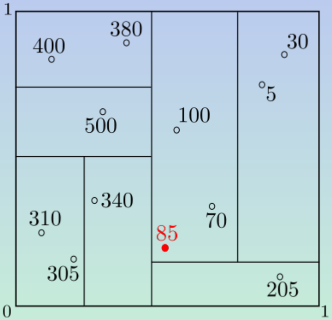
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

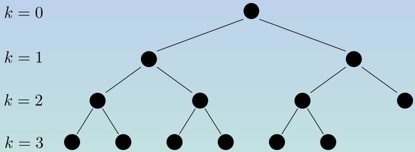




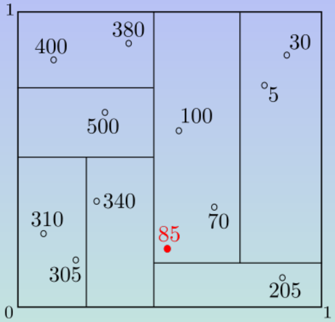
How to build a tree?

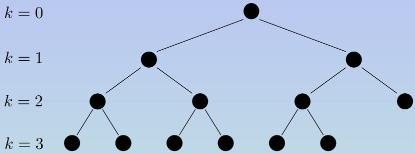
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.





How to build a tree?





Breiman Random forests are defined by

1. A splitting rule : minimize the variance within the resulting cells.
2. A stopping rule : stop when each cell contains less than `nodesize` $= 2$ observations.

How to perform splits?

For a split direction $j \in \{1, \ldots, d\}$ and a split position $z \in [0, 1]$, the criterion takes the form

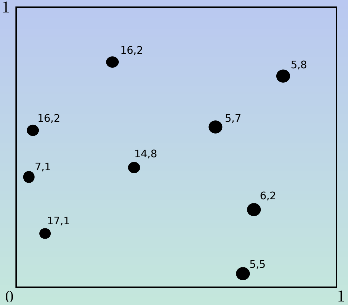
$$L_n(j, z) = \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(j)} < z} - \bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(j)} \geq z} \right)^2,$$

where

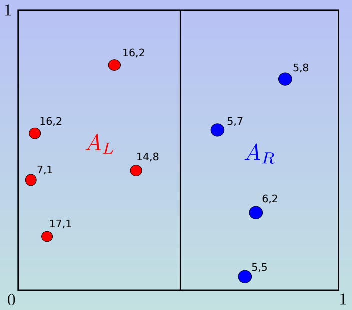
- $A_L = \{\mathbf{x} \in A : \mathbf{x}^{(j)} < z\}$ and $A_R = \{\mathbf{x} \in A : \mathbf{x}^{(j)} \geq z\}$
- $\bar{Y}_A$ is the average of the Y_i's belonging to A.
- $N_n(A)$ is the number of points in A

How to perform splits of Breiman's forests?

An example: $j = 1$ and $z = 0.5$.



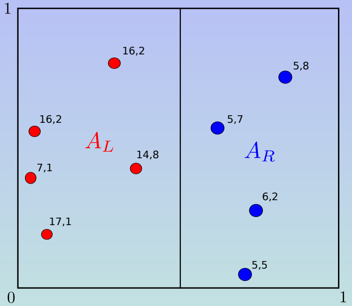
An example: $j = 1$ and $z = 0.5$.



$$L_n(1, 0.5) = \frac{1}{N_n(A)} \sum_{i=1}^{n} \Big(Y_i - \underbrace{\bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(1)} < 0.5}}_{\text{Average on } A_L} - \bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(1)} \geq 0.5} \Big)^2,$$

How to perform splits of Breiman's forests?

An example: $j = 1$ and $z = 0.5$.

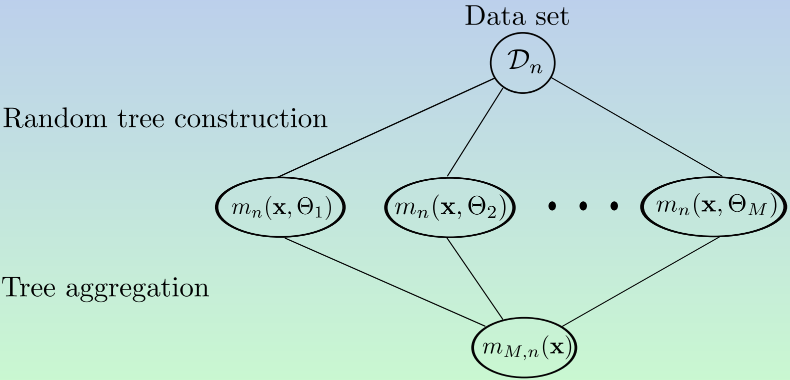


$$L_n(1, 0.5) = \frac{1}{N_n(A)} \sum_{i=1}^{n} \Big(Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(1)} < 0.5} - \underbrace{\bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(1)} \geq 0.5}}_{\text{Average on } A_R} \Big)^2,$$

Construction of random forests

Randomness in tree construction

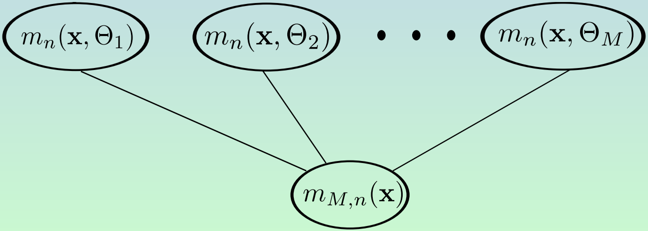
- Resampling the data set via bootstrap;
- For each cell:
  - Preselecting a subset of m_{try} variables, eligible for splitting.



Construction of Breiman forests

Breiman tree

- Select a_n observations with replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- For each cell,
  - Select randomly `mtry` coordinates among $\{1, \ldots, d\}$;
  - Choose the best split along previous direction, the one minimizing the CART criterion.
- Stop when each cell contains less than `nodesize` observations.



Literature

- Random forests were created by Breiman [2001].

- Many theoretical results focus on simplified version on random forests, whose construction is independent of the dataset.
  [Biau et al., 2008, Biau, 2012, Genuer, 2012, Zhu et al., 2012, Arlot and Genuer, 2014]

- Analysis of more data-dependent forests:
  - Asymptotic normality of random forests [Mentch and Hooker, 2015, Wager and Athey, 2017],
  - Variable importance [Louppe et al., 2013],
  - Rate of consistency [Wager and Walther, 2015].

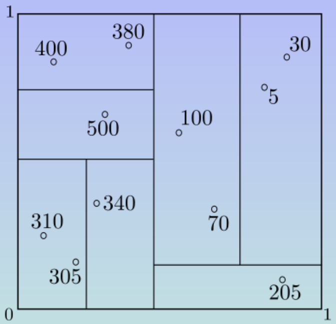
- Literature review on random forests:
  - Methodological review [Criminisi et al., 2011, Boulesteix et al., 2012],
  - Theoretical review [Biau and Scornet, 2016]

1. Day 1: Discovering random forests
   - Construction of random forests
   - Infinite forest
   - Centred Forests
   - Median forests

2. Day 2: Rate of consistency
   - Rate of consistency for centred forests
   - Rate of consistency for Median Forests
   - Minimax rates for Mondrian Forests

3. Day 3: Breiman's forests
   - Consistency of Breiman forests

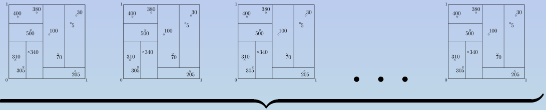
A tree



- Tree estimate:

$$m_n(\mathbf{x}, \Theta) = \sum_{i=1}^{n} \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta)}}{N_n(\mathbf{x}, \Theta)} Y_i$$

where $N_n(\mathbf{x}, \Theta)$ is the number of points in the cell $A_n(\mathbf{x}, \Theta)$.

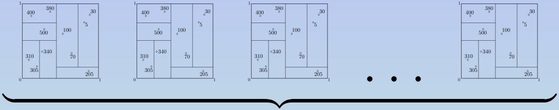


M-Finite forest

- M-Finite forest estimate :

$$m_{M,n}(\mathbf{x}, \Theta_1, \dots, \Theta_M) = \frac{1}{M} \sum_{m=1}^{M} m_n(\mathbf{x}, \Theta_m).$$

A finite forest



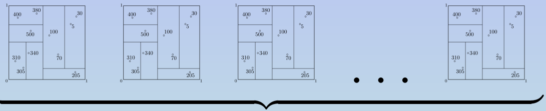
M-Finite forest

- M-Finite forest estimate :

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) = \frac{1}{M} \sum_{m=1}^{M} m_n(\mathbf{x}, \Theta_m).$$

Conditionally on $\mathcal{D}_n$, the estimate $m_{M,n}$ depends on $\Theta_1, \ldots, \Theta_M$.

Toward infinite forest



M-Finite forest

- M-Finite forest estimate :

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) = \frac{1}{M} \sum_{m=1}^{M} m_n(\mathbf{x}, \Theta_m) \underset{M \to \infty}{\to} \underbrace{\mathbb{E}_{\Theta}\left[m_n(\mathbf{x}, \Theta)\right]}_{m_{\infty,n}(\mathbf{x})}$$

Finite forest versus infinite forest

Infinite forest is better than finite forest.

(H1) One has

$$Y = m(\mathbf{X}) + \varepsilon,$$

where ε is a centered Gaussian noise with finite variance σ^2, independent of $\mathbf{X}$.

Theorem [Scornet, 2016]

Assume that **(H2)** is satisfied. Then, for all $M, n \in \mathbb{N}^\star$,

$$R(m_{M,n}) = R(m_{\infty,n}) + \frac{1}{M}\,\mathbb{E}_{\mathbf{X},\mathcal{D}_n}\Big[\mathbb{V}_\Theta\left[m_n(\mathbf{X},\Theta)\right]\Big].$$

In particular,

$$0 \leq R(m_{M,n}) - R(m_{\infty,n}) \leq \frac{8}{M} \times \big(\|m\|_\infty^2 + \sigma^2(1 + 4\log n)\big).$$

A single tree versus a forest

Theorem

We have

$$\mathbb{E}[m_{M,n}(\mathbf{X}, \Theta_1, \ldots, \Theta_M) - m(\mathbf{X})]^2 \leq \mathbb{E}[m_n(\mathbf{X}, \Theta) - m(\mathbf{X})]^2,$$

that is the risk of a forest is lower than the risk of each individual tree that composed the forest.

Proof.

Jensen's inequality. □

A forest is not worse than a single tree.

Different types of forests

Centred forest		

Different types of forests

Centred forest		
Independent of X_i and Y_i		

Different types of forests

Centred forest		
Independent of X_i and Y_i		
		

Different types of forests

Centred forest		Breiman's forests
Independent of X_i and Y_i		
		

Different types of forests

Centred forest		Breiman's forests
Independent of X_i and Y_i		Dependent on X_i and Y_i
		

Different types of forests

Centred forest		Breiman's forests
Independent of X_i and Y_i		Dependent on X_i and Y_i
		

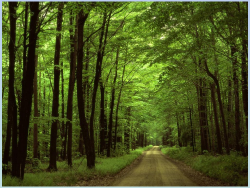
Different types of forests

Centred forest	Median forests	Breiman's forests
Independent of X_i and Y_i		Dependent on X_i and Y_i
		

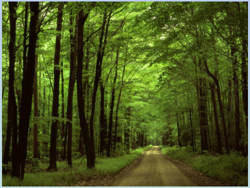
Different types of forests

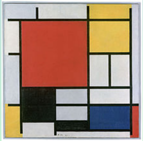
Centred forest	Median forests	Breiman's forests
Independent of X_i and Y_i	Independent of Y_i	Dependent on X_i and Y_i
		

Different types of forests

Centred forest	Median forests	Breiman's forests
Independent of X_i and Y_i	Independent of Y_i	Dependent on X_i and Y_i
		

Different types of forests

Centred forest	Median forests	Breiman's forests
Independent of X_i and Y_i	Independent of Y_i	Dependent on X_i and Y_i
		



1. Day 1: Discovering random forests
   - Construction of random forests
   - Infinite forest
   - Centred Forests
   - Median forests

2. Day 2: Rate of consistency
   - Rate of consistency for centred forests
   - Rate of consistency for Median Forests
   - Minimax rates for Mondrian Forests

3. Day 3: Breiman's forests
   - Consistency of Breiman forests

A single tree

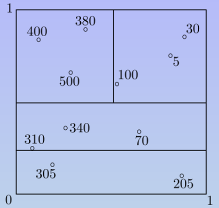
For a tree whose construction is independent of data, if

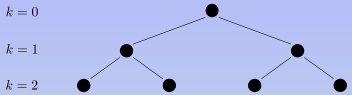
1. $\mathrm{diam}(A_n(\mathbf{X})) \to 0$, in probability;
2. $N_n(A_n(\mathbf{X})) \to \infty$, in probability;

then the tree is consistent, that is

$$\lim_{n\to\infty} \mathbb{E}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 = 0.$$

Consistency of purely random forests





Theorem [Biau et al., 2008]

Consider a totally non adaptive forest of level k. Assume that

$$\mathrm{diam}(A_n(\mathbf{X}, \Theta)) \to 0, \quad \text{in probability.}$$

Then, providing $k \to \infty$ and $n/2^k \to \infty$, the infinite random forest is consistent, that is $R(m_{\infty,n}) \to 0$ as $n \to \infty$.

- $\rightarrow$ Forest consistency results from the consistency of each tree.
- $\rightarrow$ Trees are not fully developed.

Stone Theorem

Consider an estimate of the form

$$m_n(\mathbf{x}) = \sum_{i=1}^{n} W_{ni}(\mathbf{x}) Y_i.$$

Theorem [Stone, 1977]

Assume that the weights W_{ni} are nonnegative and sum to one. Then the estimate m_n is consistent if and only if:

1. There is constant C such that, for every measurable function $g : [0,1]^d \to \mathbb{R}$ with $\mathbb{E}|g(\mathbf{X})| < \infty$,

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(\mathbf{X})|g(\mathbf{X}_i)|\right] \leq C\mathbb{E}|g(\mathbf{X})|, \quad \text{for all } n \geq 1.$$

2. For all $a > 0$, $\sum_{i=1}^{n} W_{ni}(\mathbf{X}) \mathbb{1}_{\|\mathbf{X}_i - \mathbf{X}\| > a} \to 0, \quad \text{in probability.}$
3. $\max_{1 \leq i \leq n} W_{ni}(\mathbf{X}) \to 0, \quad \text{in probability}$

For a tree estimate

$$m_n(\mathbf{x}) = \sum_{i=1}^{n} Y_i \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x},\Theta)}}{N_n(\mathbf{x},\Theta)}$$

that is

$$W_{ni}(\mathbf{x}) = \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x},\Theta)}}{N_n(\mathbf{x},\Theta)}.$$

1 is ok.

2 To check condition (2), note that, for all $a > 0$,

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{X})\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty > a}\right] = \mathbb{E}\left[\sum_{i=1}^{n} \frac{\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}}{N_n(\mathbf{X},\Theta)}\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty > a}\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{n} \frac{\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}}{N_n(\mathbf{X},\Theta)}\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty > a} \times \mathbb{1}_{\mathrm{diam}(A_n(\mathbf{X},\Theta))\geq a/2}\right],$$

because $\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty > a}\mathbb{1}_{\mathrm{diam}(A_n(\mathbf{X},\Theta))< a/2} = 0$. Thus,

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{X})\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty > a}\right] \leq \mathbb{E}\left[\mathbb{1}_{\mathrm{diam}(A_n(\mathbf{X},\Theta))\geq a/2} \times \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty > a}\right]$$

$$\leq \mathbb{P}\left[\mathrm{diam}(A_n(\mathbf{X},\Theta)) \geq a/2\right],$$

which tends to zero, as $n \to \infty$, by assumption.

Proof of (3)

The tree partition has 2^k cells, denoted by $A_1, \ldots, A_{2^k}$. For $1 \leq i \leq 2^k$, let N_i be the number of points among $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$ falling into A_i. Finally, set $\mathcal{S} = \{\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n\}$. Since these points are independent and identically distributed, fixing the set $\mathcal{S}$ (but not the order of the points) and Θ, the probability that $\mathbf{X}$ falls in the i-th cell is $N_i/(n+1)$. Thus, for every fixed $t > 0$,

$$
\begin{aligned}
\mathbb{P}\Big[N_n(\mathbf{X}, \Theta) < t\Big] &= \mathbb{E}\Big[\mathbb{P}\Big[N_n(\mathbf{X}, \Theta) < t \Big| \mathcal{S}, \Theta\Big]\Big] \\
&= \mathbb{E}\left[\sum_{i : N_i < t+1} \frac{N_i}{n+1}\right] \\
&\leq \frac{2^k}{n+1} t.
\end{aligned}
$$

Thus, by assumption, $N_n(\mathbf{X}, \Theta) \to \infty$ in probability, as $n \to \infty$.

Proof of (3)

At last, to prove (3), note that,

$$\begin{aligned}
\mathbb{E}\left[\max_{1\leq i\leq n} W_{ni}^{\infty}(\mathbf{X})\right] &\leq \mathbb{E}\left[\max_{1\leq i\leq n} \frac{\mathbb{1}_{\mathbf{X}_i\in A_n(\mathbf{X},\Theta)}}{N_n(\mathbf{X},\Theta)}\right]\\
&\leq \mathbb{E}\left[\frac{\mathbb{1}_{N_n(\mathbf{X},\Theta)>0}}{N_n(\mathbf{X},\Theta)}\right]\\
&\to 0 \quad \text{as } n\to\infty,
\end{aligned}$$

since $N_n(\mathbf{X},\Theta)\to\infty$ in probability, as $n\to\infty$.

1. Day 1: Discovering random forests
   - Construction of random forests
   - Infinite forest
   - Centred Forests
   - Median forests

2. Day 2: Rate of consistency
   - Rate of consistency for centred forests
   - Rate of consistency for Median Forests
   - Minimax rates for Mondrian Forests

3. Day 3: Breiman's forests
   - Consistency of Breiman forests

Construction of Median forests

Breiman tree

- Select a_n observations with replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- For each cell,
  - Select randomly `mtry` coordinates among $\{1, \dots, d\}$;
  - Choose the best split along previous direction, the one minimizing the CART criterion.
- Stop when each cell contains less than `nodesize` observations.

Construction of Median forests

Breiman tree

- Select a_n observations with replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- For each cell,
  - Select randomly `mtry` coordinates among $\{1, \ldots, d\}$;
  - Choose the best split along previous direction, the one minimizing the CART criterion.
- Stop when each cell contains less than `nodesize` observations.

Median tree

- Select a_n observations without replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- For each cell,
  - Select randomly `mtry` $= 1$ coordinate among $\{1, \ldots, d\}$;
  - Split at the location of the empirical median of X_i.
- Stop when each cell contains exactly `nodesize` $= 1$ observation.

Consistency of median forests

Assumption (H1)

The model writes $Y = m(\mathbf{X}) + \varepsilon$, where ε is a centred noise such that $\mathbb{V}[\varepsilon|\mathbf{X} = \mathbf{x}] \leq \sigma^2$, $\mathbf{X}$ has a density on $[0, 1]^d$ and m is continuous.

Theorem [S.(2016)]

Grant Assumption **(H1)**. Then, provided $a_n \to \infty$ and $a_n/n \to 0$, median forests are consistent, i.e.,

$$\lim_{n\to\infty} \mathbb{E}\left[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\right]^2 = 0.$$

Remarks

- Good trade-off between simplicity of centred forests and complexity of Breiman's forests.
- First consistency results for fully grown trees.
- Each tree is not consistent but the forest is, because of subsampling.

Proof of Theorem (1)

Condition (i) in Stone's Theorem is satisfied since the regression function is uniformly continuous and $\mathrm{Var}[Y|\mathbf{X} = \mathbf{x}] \leq \sigma^2$ [see remark after Stone theorem in Györfi et al., 2002].

Lemme 1

Assume that $\mathbf{X}$ has a density over $[0, 1]^d$, with respect to the Lebesgue measure. Thus, the median tree satisfies, for all γ,

$$\mathbb{P}\left[\mathrm{diam}(A_n(\mathbf{X}, \Theta)) > \gamma\right] \underset{n\to\infty}{\to} 0.$$

Proof of Theorem (2)

To check (3), observe that in the subsampling step, there are exactly $\binom{n-1}{a_n-1}$ choices to pick a fixed observation $\mathbf{X}_i$. Since $\mathbf{x}$ and $\mathbf{X}_i$ belong to the same cell only if $\mathbf{X}_i$ is selected in the subsampling step, we see that

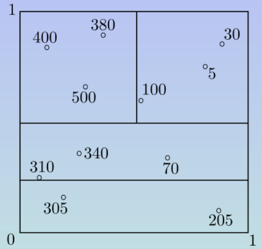
$$\mathbb{P}_\Theta\left[\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i\right] \leq \frac{\binom{n-1}{a_n-1}}{\binom{n}{a_n}} = \frac{a_n}{n}.$$

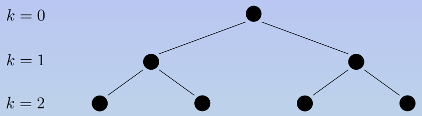
So,

$$\mathbb{E}\left[\max_{1\leq i\leq n} W_{ni}(\mathbf{X})\right] \leq \mathbb{E}\left[\max_{1\leq i\leq n} \mathbb{P}_\Theta\left[\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i\right]\right] \leq \frac{a_n}{n},$$

which tends to zero by assumption.

Centered forests





Theorem (Biau [2012])

Under proper regularity hypothesis, provided $k \to \infty$ and $n/2^k \to \infty$, the centred random forest is consistent.

Consistency of centred random forest

Estimation error [Biau, 2012]

Under proper assumptions on the regression model,

$$\mathbb{E}\left[m^{cc}_{\infty,n}(\mathbf{X}) - \bar{m}^{cc}_{\infty,n}(\mathbf{X})\right]^2 \leq C\sigma^2 \frac{2^{k_n}}{n k_n^{1/2}}$$

Approximation error [Biau, 2012]

Under proper assumptions on the regression model,

$$\mathbb{E}\left[\bar{m}^{cc}_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\right]^2 \leq 2dL^2.2^{-\frac{0.75k_n}{d\log 2}} + \|m\|^2_\infty e^{-n/2^{k_n}}$$

Consistency of centred random forest

If the forest is fully grown, that is, if $k_n = \lfloor \log_2 n \rfloor$

Estimation error [Biau, 2012]

Under proper assumptions on the regression model,

$$\mathbb{E}\left[m_{\infty,n}^{cc}(\mathbf{X}) - \bar{m}_{\infty,n}^{cc}(\mathbf{X})\right]^2 \leq C\sigma^2 \frac{2^{k_n}}{n k_n^{1/2}}$$

Approximation error [Biau, 2012]

Under proper assumptions on the regression model,

$$\mathbb{E}\left[\bar{m}_{\infty,n}^{cc}(\mathbf{X}) - m(\mathbf{X})\right]^2 \leq 2dL^2.2^{-\frac{0.75k_n}{d\log 2}} + \|m\|_\infty^2 e^{-n/2^{k_n}}$$

Consistency of centred random forest

If the forest is fully grown, that is, if $k_n = \lfloor \log_2 n \rfloor$

Estimation error [Biau, 2012]

Under proper assumptions on the regression model,

$$\mathbb{E}\left[m^{cc}_{\infty,n}(\mathbf{X}) - \bar{m}^{cc}_{\infty,n}(\mathbf{X})\right]^2 \leq C\sigma^2 \frac{2^{k_n}}{n k_n^{1/2}}$$

Approximation error [Biau, 2012]

Under proper assumptions on the regression model,

$$\mathbb{E}\left[\bar{m}^{cc}_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\right]^2 \leq 2dL^2.2^{-\frac{0.75k_n}{d\log 2}} + \|m\|^2_\infty e^{-n/2^{k_n}}$$

Consistency of centred random forest

If the forest is fully grown, that is, if $k_n = \lfloor \log_2 n \rfloor$

Estimation error [Biau, 2012]

Under proper assumptions on the regression model,

$$\mathbb{E}\left[m_{\infty,n}^{cc}(\mathbf{X}) - \bar{m}_{\infty,n}^{cc}(\mathbf{X})\right]^2 \leq C\sigma^2 (\log_2 n)^{-1/2}$$

Approximation error [Biau, 2012]

Under proper assumptions on the regression model,

$$\mathbb{E}\left[\bar{m}_{\infty,n}^{cc}(\mathbf{X}) - m(\mathbf{X})\right]^2 \leq 2dL^2.2^{-\frac{0.75k_n}{d\log 2}} + \|m\|_\infty^2 e^{-n/2^{k_n}}$$

Consistency of centred random forest

If the forest is fully grown, that is, if $k_n = \lfloor \log_2 n \rfloor$

Estimation error [Biau, 2012]

Under proper assumptions on the regression model,

$$\mathbb{E}\left[m_{\infty,n}^{cc}(\mathbf{X}) - \bar{m}_{\infty,n}^{cc}(\mathbf{X})\right]^2 \leq C\sigma^2(\log_2 n)^{-1/2}$$

Approximation error [Biau, 2012]

Under proper assumptions on the regression model,

$$\mathbb{E}\left[\bar{m}_{\infty,n}^{cc}(\mathbf{X}) - m(\mathbf{X})\right]^2 \leq 2dL^2 2^{-\frac{0.75k_n}{d\log 2}} + \|m\|_\infty^2 e^{-n/2^{k_n}}$$

Consistency of centred random forest

If the forest is fully grown, that is, if $k_n = \lfloor \log_2 n \rfloor$

Estimation error [Biau, 2012]

Under proper assumptions on the regression model,

$$\mathbb{E}\left[m_{\infty,n}^{cc}(\mathbf{X}) - \bar{m}_{\infty,n}^{cc}(\mathbf{X})\right]^2 \leq C\sigma^2(\log_2 n)^{-1/2}$$

Approximation error [Biau, 2012]

Under proper assumptions on the regression model,

$$\mathbb{E}\left[\bar{m}_{\infty,n}^{cc}(\mathbf{X}) - m(\mathbf{X})\right]^2 \leq 2dL^2 n^{-\frac{0.75}{d \log 2}} + \|m\|_\infty^2 \times 1$$

Construction of Breiman/Median forests

Breiman tree

- Select a_n observations with replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- For each cell,
  - Select randomly `mtry` coordinates among $\{1, \ldots, d\}$;
  - Choose the best split along previous direction, the one minimizing the CART criterion.
- Stop when each cell contains less than `nodesize` observations.

Construction of Breiman/Median forests

Breiman tree

- Select a_n observations with replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- For each cell,
  - Select randomly `mtry` coordinates among $\{1, \ldots, d\}$;
  - Choose the best split along previous direction, the one minimizing the CART criterion.
- Stop when each cell contains less than `nodesize` observations.

Median tree

- Select a_n observations without replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- For each cell,
  - Select randomly $\texttt{mtry} = 1$ coordinate among $\{1, \ldots, d\}$;
  - Split at the location of the empirical median of X_i.
- Stop when each cell has been cut k times (i.e., $\texttt{nodesize} \simeq \lfloor a_n/2^k \rfloor$).

Rate of consistency

Assumption (H1)

The regression model writes $Y = m(\mathbf{X}) + \varepsilon$, where $\mathbf{X}$ is uniformly distributed on $[0,1]^d$, m is L-Lipschitz, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Theorem [Duroux, S.(2018)]

Grant **(H1)**. For all n, for all $\mathbf{x} \in [0,1]^d$, if $a_n \geq 2^k$, then

$$\mathbb{E}\big[m_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\big]^2 \leq 2\sigma^2 \frac{2^k}{n} + dL^2 C_1 \left(1 - \frac{3}{4d}\right)^k.$$

Rate of consistency

Assumption (H1)

The regression model writes $Y = m(\mathbf{X}) + \varepsilon$, where $\mathbf{X}$ is uniformly distributed on $[0,1]^d$, m is L-Lipschitz, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Theorem [Duroux, S.(2018)]

Grant **(H1)**. For all n, for all $\mathbf{x} \in [0,1]^d$, if $a_n \geq 2^k$, then

$$\mathbb{E}\big[m_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\big]^2 \leq 2\sigma^2 \frac{2^k}{n} + dL^2 C_1 \left(1 - \frac{3}{4d}\right)^k.$$

- Right-hand side: Estimation error + approximation error.

Rate of consistency

Assumption (H1)

The regression model writes $Y = m(\mathbf{X}) + \varepsilon$, where $\mathbf{X}$ is uniformly distributed on $[0,1]^d$, m is L-Lipschitz, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Theorem [Duroux, S.(2018)]

Grant **(H1)**. For all n, for all $\mathbf{x} \in [0,1]^d$, if $a_n \geq 2^k$, then

$$\begin{aligned}\mathbb{E}\left[m_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\right]^2 &\leq 2\sigma^2 \frac{2^k}{n} + dL^2 C_1 \left(1 - \frac{3}{4d}\right)^k. \\ &\leq 2\sigma^2 \frac{2^k}{a_n} + dL^2 C_1 \left(1 - \frac{3}{4d}\right)^k.\end{aligned}$$

- Right-hand side: Estimation error + approximation error.

Rate of consistency

Assumption (H1)

The regression model writes $Y = m(\mathbf{X}) + \varepsilon$, where $\mathbf{X}$ is uniformly distributed on $[0,1]^d$, m is L-Lipschitz, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Theorem [Duroux, S.(2018)]

Grant **(H1)**. For all n, for all $\mathbf{x} \in [0,1]^d$, if $a_n \geq 2^k$, then

$$\begin{aligned}\mathbb{E}\left[m_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\right]^2 &\leq 2\sigma^2 \frac{2^k}{n} + dL^2 C_1 \left(1 - \frac{3}{4d}\right)^k . \\ &\leq 2\sigma^2 \frac{2^k}{a_n} + dL^2 C_1 \left(1 - \frac{3}{4d}\right)^k .\end{aligned}$$

- Right-hand side: Estimation error + approximation error.
- Consistency result for each tree if $2^k / a_n \to 0$ and $k \to \infty$.

Rate of consistency

Assumption (H1)

The regression model writes $Y = m(\mathbf{X}) + \varepsilon$, where $\mathbf{X}$ is uniformly distributed on $[0,1]^d$, m is L-Lipschitz, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Theorem [Duroux, S.(2018)]

Grant **(H1)**. For all n, for all $\mathbf{x} \in [0,1]^d$, if $a_n \geq 2^k$, then

$$\mathbb{E}\big[m_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\big]^2 \leq 2\sigma^2 \frac{2^k}{n} + dL^2 C_1 \left(1 - \frac{3}{4d}\right)^k.$$

- Right-hand side: Estimation error + approximation error.
- Consistency result for each tree if $2^k / a_n \to 0$ and $k \to \infty$.

Rate of consistency

Assumption (H1)

The regression model writes $Y = m(\mathbf{X}) + \varepsilon$, where $\mathbf{X}$ is uniformly distributed on $[0,1]^d$, m is L-Lipschitz, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Theorem [Duroux, S.(2018)]

Grant **(H1)**. For all n, for all $\mathbf{x} \in [0,1]^d$, if $a_n \geq 2^k$, then

$$\mathbb{E}\left[m_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\right]^2 \leq 2\sigma^2 \frac{2^k}{n} + dL^2 C_1 \left(1 - \frac{3}{4d}\right)^k.$$

- Right-hand side: Estimation error + approximation error.
- Consistency result for each tree if $2^k/a_n \to 0$ and $k \to \infty$.
- The upper bound is independent of a_n.

Rate of consistency

Let $\beta = 1 - 3/(4d)$ and $k_n^\star = (\ln(n) + C_2)/(\ln 2 - \ln \beta)$.

Theorem [Duroux, S.(2018)]

Assume that **(H1)** is satisfied. Consider a median forest of level $k = k_n^\star$. For all n, for all $\mathbf{x} \in [0,1]^d$, if $a_n \geq 2^{k_n^\star}$,

$$\mathbb{E}\big[m_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\big]^2 \leq C n^{\frac{\ln \beta}{\ln 2 - \ln \beta}}.$$

Rate of consistency

Let $\beta = 1 - 3/(4d)$ and $k_n^\star = (\ln(n) + C_2)/(\ln 2 - \ln \beta)$.

Theorem [Duroux, S.(2018)]

Assume that **(H1)** is satisfied. Consider a median forest of level $k = k_n^\star$. For all n, for all $\mathbf{x} \in [0,1]^d$, if $a_n \geq 2^{k_n^\star}$,

$$\mathbb{E}\left[m_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\right]^2 \leq C n^{\frac{\ln \beta}{\ln 2 - \ln \beta}}.$$

The previous theorem holds in particular for two regimes:

- $a_n = n$: the median forest is not fully grown $(2^{k_n^\star}/n \to 0)$ and the whole data set is used to build each tree.
- $a_n = 2^{k_n^\star}$: the median forest is fully grown and the subsampling rate $a_n/n \to 0$.

Rate of consistency

Let $\beta = 1 - 3/(4d)$ and $k_n^\star = (\ln(n) + C_2)/(\ln 2 - \ln\beta)$.

Theorem [Duroux, S.(2018)]

Assume that **(H1)** is satisfied. Consider a median forest of level $k = k_n^\star$. For all n, for all $\mathbf{x} \in [0,1]^d$, if $a_n \geq 2^{k_n^\star}$,

$$\mathbb{E}\big[m_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\big]^2 \leq Cn^{\frac{\ln\beta}{\ln 2 - \ln\beta}}.$$

The previous theorem holds in particular for two regimes:

- $a_n = n$: the median forest is not fully grown $(2^{k_n^\star}/n \to 0)$ and the whole data set is used to build each tree.
- $a_n = 2^{k_n^\star}$: the median forest is fully grown and the subsampling rate $a_n/n \to 0$.

No need for tuning them both at the same time.

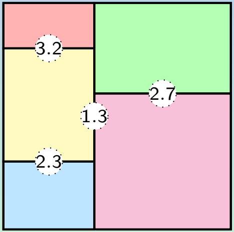
1. Day 1: Discovering random forests
   - Construction of random forests
   - Infinite forest
   - Centred Forests
   - Median forests

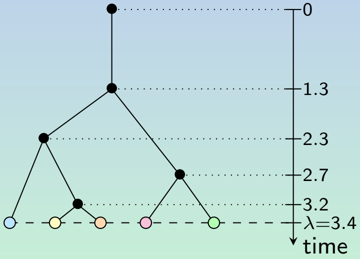
2. Day 2: Rate of consistency
   - Rate of consistency for centred forests
   - Rate of consistency for Median Forests
   - Minimax rates for Mondrian Forests

3. Day 3: Breiman's forests
   - Consistency of Breiman forests

The Mondrian process (Roy and Teh, 2008)

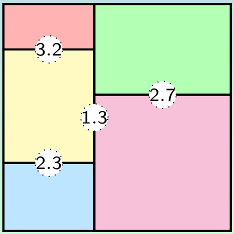
- $\mathrm{MP}(\lambda, C)$: distribution on recursive, axis-aligned partitions of $C = \prod_{j=1}^{d}[a_j, b_j] \subset \mathbb{R}^d$ (= trees).
- $\lambda > 0$ "lifetime" = complexity parameter.

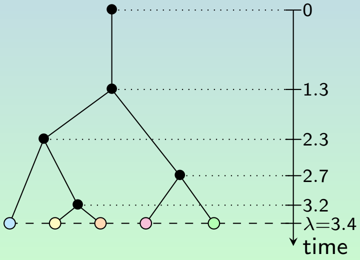




The distribution $\mathrm{MP}(\lambda, C)$

- **Start** with cell C (root), formed at time $\tau_C = 0$.
- **Sample** time till split $E \sim \mathrm{Exp}(|C|)$ with $|C| := \sum_{j=1}^{d}(b_j - a_j)$
- **If** $\tau_C + E \leq \lambda$,
  - split C in $C_L = \{x \in C : x_J \leq S_J\}$ and $C_R = C \setminus C_L$:
    - split coordinate $J \in \{1, \ldots, d\}$ with $\mathbb{P}(J = j) = \frac{b_j - a_j}{|A|}$,
    - split threshold $S_J | J \sim \mathcal{U}([a_J, b_J])$
  - Apply the procedure to $(C_L, \tau_C + E), (C_R, \tau_C + E)$.
- **Else** don't split C (which becomes a leaf of the tree).





Mondrian forests

- Introduced in [1] for computational reasons: predictions updated efficiently with new sample point (online algorithm).
- Approximately: sample independent partitions $\Pi_\lambda^{(1)}, \ldots, \Pi_\lambda^{(M)} \sim \mathrm{MP}(\lambda, [0,1]^d)$, fit them and average their predictions.
- No theoretical analysis of the algorithm.
- Choice of the parameter λ ?

[1] Lakshminarayanan, Roy, Teh. Mondrian forests: Efficient online random forests. In *NIPS*, 2014.

Denote $m_{\lambda,M,n}$ the (randomized) Mondrian forest estimator with M trees and parameter λ Let

$$(\mathbf{H})\ \mathrm{Var}(Y|X) \leq \sigma^2 < \infty \ a.s.$$

Theorem (Mourtada, Gaïffas, S.)

Assume (**H**) *and that the regression function m is L-Lipschitz. Then:*

$$\mathcal{R}(m_{\lambda,M,n}) \leq \frac{4dL^2}{\lambda^2} + \frac{(1+\lambda)^d}{n}\left(2\sigma^2 + 9\|m\|_\infty^2\right). \quad (1)$$

In particular, the choice $\lambda := \lambda_n \asymp n^{1/(d+2)}$ gives

$$\mathcal{R}(m_{\lambda,M,n}) = O(n^{-2/(d+2)}), \quad (2)$$

which is the minimax optimal rate for the estimation of a Lipschitz function in dimension d.

"Forest effect": influence of the number of trees

- The above result is true for every $M \geq 1$ (number of trees): in particular, a single tree is already optimal for the estimation of a Lipschitz function in dimension d.
- In practice, forests with $M \gg 1$ perform better than trees.
- How to account for this ? Do we gain something by randomizing partitions ?
- When is M "large enough" ?

Improved rates under $\mathscr{C}^2$ regularity

Theorem (Mourtada, Gaïffas, S.)

Assume **(H)**, *m of class $\mathscr{C}^2$, and that* $\mathbf{X}$ *has a positive, Lipschitz density on $[0,1]^d$. Then, for every $\varepsilon > 0$:*

$$\mathcal{R}_{[\varepsilon,1-\varepsilon]^d}(m_{\lambda,M,n}) = O\Big(\frac{1}{M\lambda^2} + \frac{1}{\lambda^4} + \frac{e^{-\lambda\varepsilon}}{\lambda^3} + \frac{(1+\lambda)^d}{n}\Big)$$

For $\lambda := \lambda_n \asymp n^{1/(d+4)}$ and $M := M_n \gtrsim n^{2/(d+4)}$, this implies

$$\mathcal{R}_{[\varepsilon,1-\varepsilon]^d}(m_{\lambda,M,n}) = O(n^{-4/(d+4)})$$

which is the optimal rate for twice differentiable m in dimension d. Without conditioning, we get $O(n^{-3/(d+3)})$ (boundary effect). By contrast, Mondrian trees do not exhibit improved rates.

Remark: Similar result obtained by Arlot and Genuer (2014) in dimension 1 for another variant of Random forests.

Proof ideas

- Bias-variance decomposition: standard decomposition in approximation error + estimation error.
- Exact geometric properties (local and global) of Mondrian partitions are directly available, without reasoning conditionally on the graph structure / on earlier splits.
- Restriction property: enables to obtain the exact distribution of the cell $C_\lambda(x)$ of $x \in [0,1]^d$ in the partition $\Pi_\lambda \sim \mathrm{MP}(\lambda, [0,1]^d)$ (4 lines proof).
- By modifying the distribution of the Mondrian and using the one-dimensional case, one can show that the expected number of leaves in Π_λ is $(1+\lambda)^d$.

Online implementation and adaptivity to smoothness

- If $m : x \mapsto \mathbb{E}[Y \mid X = x]$ is α-Hölder ($\alpha \in (0, 1]$), optimal rate $\mathcal{R}(\widehat{f}_{\lambda,n}) = O(n^{-2\alpha/(d+2\alpha)})$ for $\lambda \asymp n^{-1/(d+2\alpha)}$.
- In practice, α is unknown. How to choose λ ?
- Exponentially weighted aggregation over the class of all finite labeled subtrees of the "infinite Mondrian" Π_∞. BUT: infinite tree (sampled from the start ??) + number of subtrees exponential in the number of nodes.
- Extension properties of Mondrian + efficient algorithm for branching process prior ("Context Tree Weighting": one weight per node) $\Longrightarrow$ online and efficient exact algorithm ($O(\log n)$ update, $O(n \log n)$ training time, $O(\log n)$ prediction).
- Resulting $\widehat{m}_n$ is adaptive to α: $\mathcal{R}(\widehat{m}_n) = \widetilde{O}(n^{-2\alpha/(d+2\alpha)})$.

Construction of Breiman forests

Breiman tree

- Select a_n observations with replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- For each cell,
  - Select randomly `mtry` coordinates among $\{1, \ldots, d\}$;
  - Choose the best split along previous direction, the one minimizing the CART criterion.
- Stop when each cell contains less than `nodesize` observations.

Construction of Breiman forests

Breiman tree

- Select a_n observations with replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- For each cell,
  - Select randomly `mtry` coordinates among $\{1, \ldots, d\}$;
  - Choose the best split along previous direction, the one minimizing the CART criterion.
- Stop when each cell contains less than `nodesize` observations.

Our Breiman tree

- Select a_n observations without replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- For each cell,
  - Select randomly `mtry` coordinates among $\{1, \ldots, d\}$;
  - Choose the best split along previous direction, the one minimizing the CART criterion.
- Stop when the number of cells is exactly t_n.

Additive regression model

Assumption (H1)

The regression model is

$$Y = \sum_{j=1}^{d} m_j(\mathbf{X}^{(j)}) + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with ε independent of $\mathbf{X}$; $\mathbf{X}$ is uniformly distributed on $[0, 1]^d$; each model component m_j is continuous.

Consistency

Assumption (H1)

The regression model is $Y = \sum_{j=1}^{d} m_j(\mathbf{X}^{(j)}) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with ε independent of $\mathbf{X}$; $\mathbf{X}$ is uniformly distributed on $[0,1]^d$; each model component m_j is continuous.

Theorem [Scornet et al., 2015]

Assume that **(H1)** is satisfied. Then, provided $a_n \to \infty$ and $t_n(\log a_n)^9/a_n \to 0$, random forests are consistent, i.e.,

$$\lim_{n\to\infty} \mathbb{E}\left[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\right]^2 = 0.$$

Remarks

- First consistency result for Breiman's original forest.
- Consistency of CART.

Sketch of proof

$$\Delta(m, A) = \sup_{\mathbf{x},\mathbf{x}' \in A} |m(\mathbf{x}) - m(\mathbf{x}')|.$$

Furthermore, we denote by $A_n(\mathbf{X}, \Theta)$ the cell of a tree built with random parameter Θ that contains the point $\mathbf{X}$.

Proposition

Assume that **(H1)** holds. Then, for all $\rho, \xi > 0$, there exists $N \in \mathbb{N}^\star$ such that, for all $n > N$,

$$\mathbb{P}\left[\Delta(m, A_n(\mathbf{X}, \Theta)) \leq \xi\right] \geq 1 - \rho.$$

Theoretical splitting criterion for a split (j, z):

$$\begin{aligned} L^\star(j,z) = \mathbb{V}[Y|\mathbf{X} \in A] &- \mathbb{P}[\mathbf{X}^{(j)} < z \,|\, \mathbf{X} \in A]\ \mathbb{V}[Y|\mathbf{X}^{(j)} < z, \mathbf{X} \in A] \\ &- \mathbb{P}[\mathbf{X}^{(j)} \geq z \,|\, \mathbf{X} \in A]\ \mathbb{V}[Y|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A]. \end{aligned}$$

- Assume that **(H1)** is satisfied. Then, for all $\mathbf{x} \in [0,1]^p$,

$$\Delta(m, A_k^\star(\mathbf{x}, \Theta)) \to 0, \quad \text{almost surely, as } k \to \infty.$$

- Assume that **(H1)** is satisfied. Fix $\mathbf{x} \in [0,1]^p$, $k \in \mathbb{N}^\star$, and let $\xi > 0$. Then $L_{n,k}(\mathbf{x}, \cdot)$ is stochastically equicontinuous on $\bar{\mathcal{A}}_k^\xi(\mathbf{x})$, that is, for all $\alpha, \rho > 0$, there exists $\delta > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\left[\sup_{\substack{\|\mathbf{d}_k - \mathbf{d}'_k\|_\infty \le \delta \\ \mathbf{d}_k, \mathbf{d}'_k \in \bar{\mathcal{A}}_k^\xi(\mathbf{x})}} |L_{n,k}(\mathbf{x}, \mathbf{d}_k) - L_{n,k}(\mathbf{x}, \mathbf{d}'_k)| > \alpha \right] \le \rho.$$

- Assume that **(H1)** is satisfied. Fix $\xi, \rho > 0$ and $k \in \mathbb{N}^\star$. Then there exists $N \in \mathbb{N}^\star$ such that, for all $n \ge N$,

$$\mathbb{P}\left[d_\infty(\hat{\mathbf{d}}_{k,n}(\mathbf{X}, \Theta), \mathcal{A}_k^\star(\mathbf{X}, \Theta)) \le \xi \right] \ge 1 - \rho.$$

We let $\mathcal{F}_n(\Theta)$ be the set of all functions $f : [0,1]^d \to \mathbb{R}$ piecewise constant on each cell of the partition $\mathcal{P}_n(\Theta)$

Theorem [Györfi et al., 2002]

Let m_n and $\mathcal{F}_n(\Theta)$ be as above. Assume that

(i) $\lim_{n\to\infty} \beta_n = \infty$,

(ii) $\lim_{n\to\infty} \mathbb{E}\left[\inf_{\substack{f\in\mathcal{F}_n(\Theta)\\ \|f\|_\infty \le \beta_n}} \mathbb{E}_{\mathbf{X}} \left[f(\mathbf{X}) - m(\mathbf{X})\right]^2 \right] = 0$,

(iii) For all $L > 0$,

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{\substack{f\in\mathcal{F}_n(\Theta)\\ \|f\|_\infty \le \beta_n}} \left| \frac{1}{a_n} \sum_{i\in\mathcal{I}_{n,\Theta}} \left[f(\mathbf{X}_i) - Y_{i,L}\right]^2 - \mathbb{E}\left[f(\mathbf{X}) - Y_L\right]^2 \right| \right] = 0.$$

Then

$$\lim_{n\to\infty} \mathbb{E}\left[T_{\beta_n} m_n(\mathbf{X},\Theta) - m(\mathbf{X})\right]^2 = 0.$$

Sketch of proof

According to the Proposition

Proposition

Assume that **(H1)** holds. Then, for all $\rho, \xi > 0$, there exists $N \in \mathbb{N}^\star$ such that, for all $n > N$,

$$\mathbb{P}\left[\Delta(m, A_n(\mathbf{X}, \Theta)) \leq \xi\right] \geq 1 - \rho.$$

the statement (ii) holds.
The second one is true because the complexity of the partition is controlled by the condition $t_n(\log a_n)^9/a_n \to 0$.

Sparsity and Breiman's forests

Assumption

Assume that,

$$Y = \sum_{\ell=1}^{S} m_\ell(\mathbf{X}^{(\ell)}) + \varepsilon,$$

for some $S < d$ and that, for each cell, the best split is selected among all d variables

Proposition [Scornet et al., 2015]

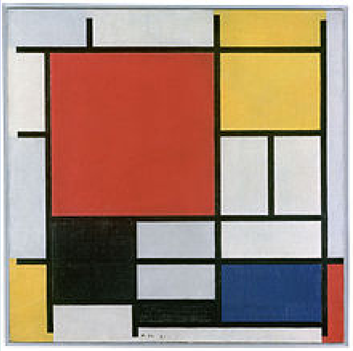
Let $k \in \mathbb{N}^\star$ and $\xi > 0$. Under appropriate assumptions, with probability $1 - \xi$, for all n large enough, we have, for all $1 \leq q \leq k$,

$$j_{q,n}(\mathbf{X}) \in \{1, \ldots, S\},$$

where $j_{1,n}(\mathbf{X}), \ldots, j_{k,n}(\mathbf{X})$ are the first k splitting directions used to construct the cell containing $\mathbf{X}$

Conclusion

- Centred forests: Trees and forests are consistent.
- Median forests
  - A single tree is not consistent but the forest is.
    $\rightarrow$ Benefits from using a random forest compared to a single tree.
  - Asymptotic to optimize subsampling size and tree depth.
    $\rightarrow$ No need to tune them both at the same time.
- Mondrian forests
  - Achieve minimax rate for both $\mathscr{C}^1$ and $\mathscr{C}^2$ functions
    $\rightarrow$ For $\mathscr{C}^2$ functions, a tree is not optimal but the forest is.
- Breiman forests
  - Trees and forests are consistent and relevant feature selection
    $\rightarrow$ Good performance in high-dimensional settings.

Thank you!

S. Arlot and R. Genuer. Analysis of purely random forests bias. 2014.

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

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

G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9:2015–2033, 2008.

A.-L. Boulesteix, S. Janitza, J. Kruppa, and I.R. König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2:493–507, 2012.

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

A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7:81–227, 2011.

R. Genuer. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24:543–562, 2012.

References II

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

G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts. Understanding variable importances in forests of randomized trees. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 431–439, 2013.

L. Mentch and G. Hooker. Ensemble trees and CLTs: Statistical inference for supervised learning. *Journal of Machine Learning Research, in press*, 2015.

E. Scornet. On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146:72–83, 2016.

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

C. Stone. Consistent nonparametric regression. *The annals of Statistics*, 5(4): 595–645, 1977.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

S. Wager and G. Walther. Adaptive concentration of regression trees, with application to random forests. 2015.

R. Zhu, D. Zeng, and M.R. Kosorok. Reinforcement learning trees. 2012.