Halte à la déforestation : luttons contre l'élagage des forêts aléatoires

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

Mini-cours IHP Janvier 2018

Background on 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.

Background on 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).

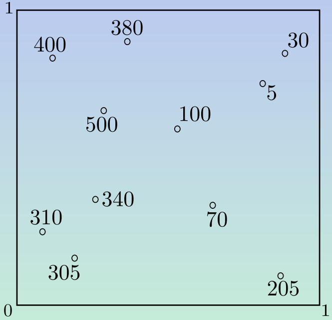
We assume that

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

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

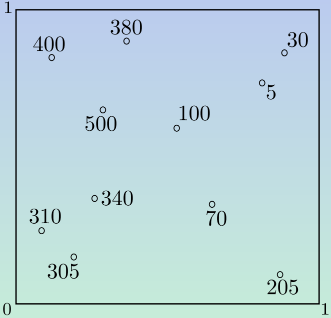
How to build a tree?

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



How to build a tree?

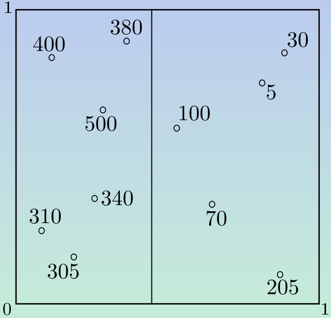
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

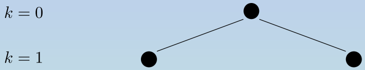


$k = 0$

How to build a tree?

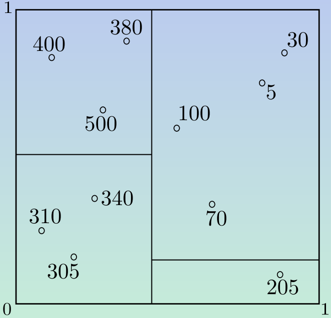
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

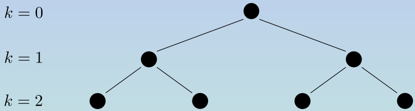




How to build a tree?

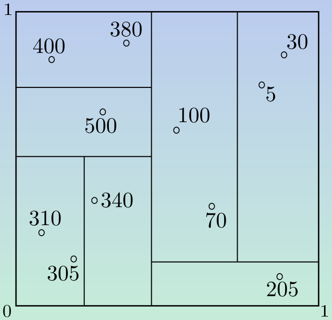
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

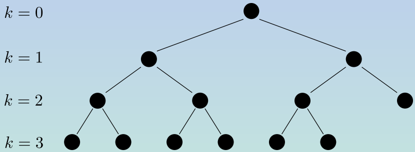




How to build a tree?

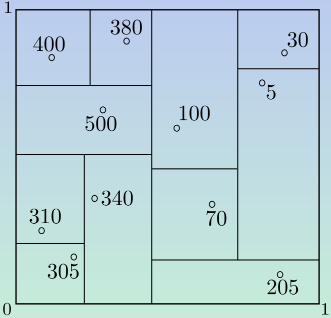
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

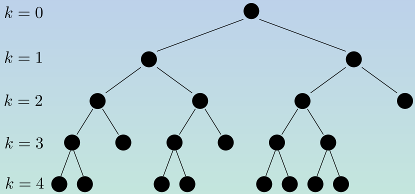




How to build a tree?

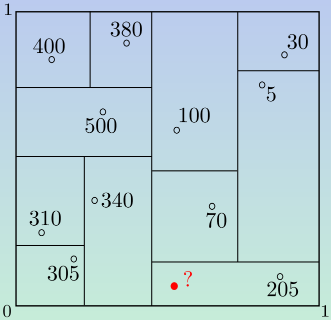
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

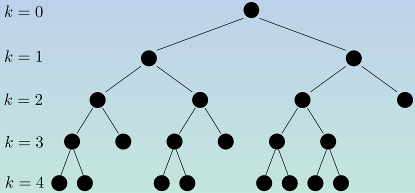




How to build a tree?

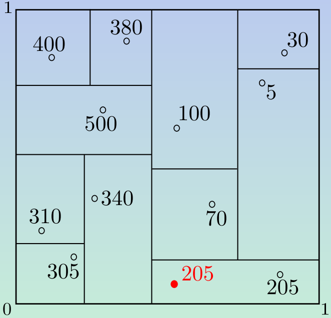
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

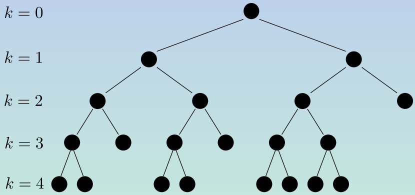




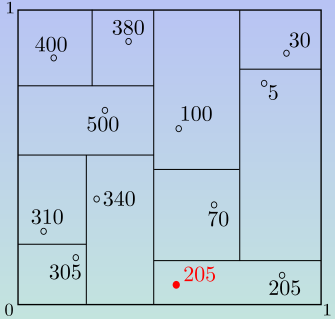
How to build a tree?

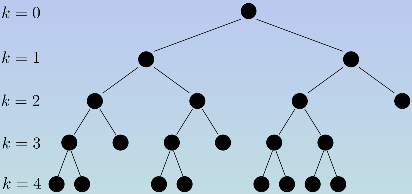
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.





How to build a tree?





Breiman Random forests are defined by

1. A splitting rule : minimize the square loss.
2. A stopping rule : leave exactly one point in each cell.

How to perform splits of Breiman's forests?

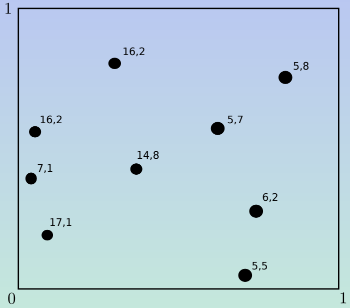
For a cut direction $j \in \{1, \dots, d\}$ and a split position $z \in [0,1]$, the criterion takes the form

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

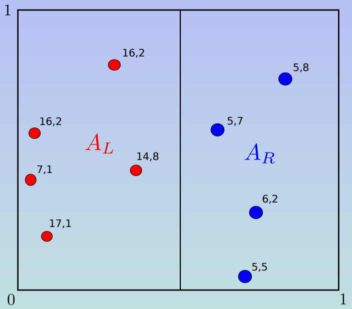
where

- $A_L = \{\mathbf{x} \in A : \mathbf{x}^{(j)} < z\}$ 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

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

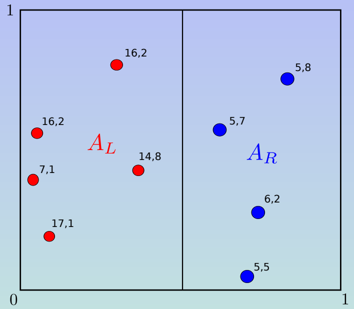


An example: $j = 1$ and $z = 0.5$.



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

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

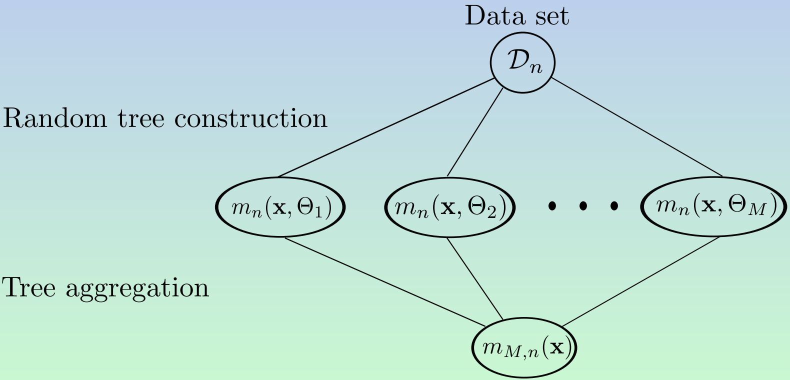


$$L_n(1, 0.5) = \frac{1}{N_n(A)} \sum_{i=1}^{n} \Big(Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(1)} < 0.5} - \underbrace{\bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(1)} \geq 0.5}}_{\text{Average on } A_R} \Big)^2,$$

Construction of random forests

Randomness in tree construction

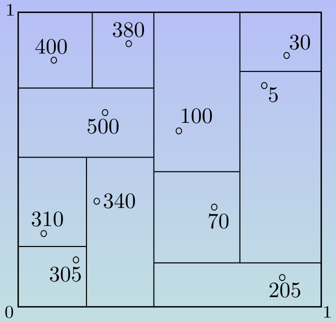
- Resample the data set via bootstrap;
- At each node, preselect a subset of m_{try} variables eligible for splitting.



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 [Wager, 2014, Mentch and Hooker, 2015].
  - 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].

A tree

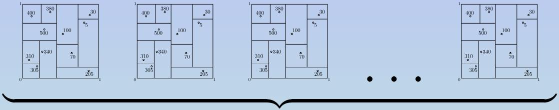


- Tree estimate:

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

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

A finite forest

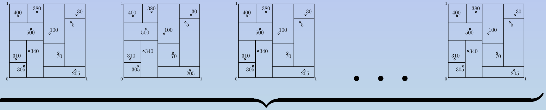


M-Finite forest

- M-Finite forest estimate :

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

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 $\Theta_1, \ldots, \Theta_M$.

A forest is not worse than a single tree.

Theorem

We have

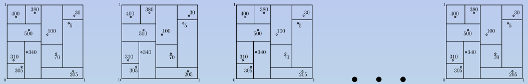
$$\mathbb{E}[m(\mathbf{X}) - m_{M,n}(\mathbf{X}, \Theta_1, \ldots, \Theta_M)]^2 \leq \mathbb{E}[m(\mathbf{X}) - m_n(\mathbf{X}, \Theta)]^2,$$

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

Proof.

Jensen's inequality. $\square$

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,

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

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
		

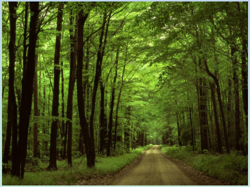
Different types of forests

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
		

A single tree

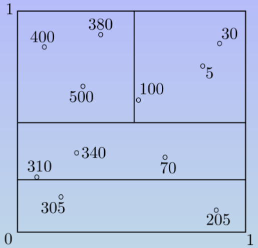
For a tree whose construction is independent of data, if

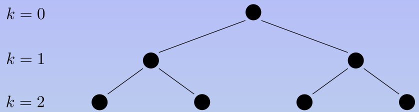
1. $\mathrm{diam}(A_n(\mathbf{X})) \to 0$, in probability;
2. $N_n(A_n(\mathbf{X})) \to \infty$, in probability;

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]

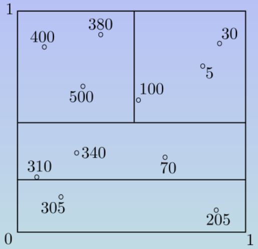
Consider a totally non adaptive forest of level k. Assume that

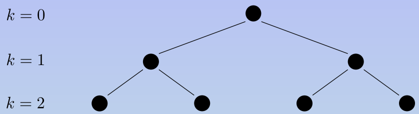
$$\mathrm{diam}(A_n(\mathbf{X}, \Theta)) \to 0, \quad \text{in probability.}$$

Then, providing $k \to \infty$ and $2^k/n \to 0$, the infinite random forest is consistent, that is

$$R(m_{\infty,n}) \to 0 \quad \text{as } n \to \infty.$$

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.

$\rightarrow$ Forest consistency results from the consistency of each tree.

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 $\mathbb{E}|g(\mathbf{X})| < \infty$,
$$\mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(\mathbf{X})|g(\mathbf{X}_i)|\Big] \leq C\mathbb{E}|g(\mathbf{X})|, \quad \text{for all } n \geq 1.$$
2. For all $a > 0$, $\sum_{i=1}^{n} W_{ni}(\mathbf{X}) \mathbb{1}_{\|\mathbf{X}_i - \mathbf{X}\| > a} \to 0$, in probability.
3. $\max_{1 \leq i \leq n} W_{ni}(\mathbf{X}) \to 0$, 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 $a > 0$,

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

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

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

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

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

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 $n\to\infty$.

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,

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

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,

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

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,

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

Approximation error [Biau, 2012]

Under proper assumptions on the regression model,

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

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,

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

Approximation error [Biau, 2012]

Under proper assumptions on the regression model,

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

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,

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

Approximation error [Biau, 2012]

Under proper assumptions on the regression model,

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

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.
- At each cell, select randomly `mtry` coordinates among $\{1, \ldots, d\}$.
- Split at the location that minimizes the square loss.
- 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.
- At each cell, select randomly `mtry` coordinates among $\{1, \ldots, d\}$.
- Split at the location that minimizes the square loss.
- 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.
- At 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 contains exactly $\texttt{nodesize} = 1$ observation.

Theorem [Scornet, 2016]

Assume that **(H1)** is satisfied. 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) 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,

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

To check (3), observe that in the subsampling step, there are exactly $\binom{a_n-1}{n-1}$ 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

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

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.

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.
- At each cell, select randomly `mtry` coordinates among $\{1, \ldots, d\}$.
- Split at the location that minimizes the square loss.
- 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.
- At each cell, select randomly `mtry` coordinates among $\{1, \ldots, d\}$.
- Split at the location that minimizes the square loss.
- Stop when each cell contains less than `nodesize` observations.

Modified Breiman tree

- Select a_n observations without replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- At each cell, select randomly `mtry` coordinates among $\{1, \ldots, d\}$.
- Split at the location that minimizes the square loss.
- Stop when the number of cells is exactly t_n.

Assumption (H1)

Additive regression model:

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

where

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

Consistency

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):

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

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

(iii) For all $L > 0$,

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

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 $t_n(\log a_n)^9/a_n \to 0$.

Theorem [Scornet et al., 2015]

Assume that **(H1)** and **(H2.1)** are satisfied and let $t_n = a_n$. Then, provided $a_n \to \infty$ and $a_n \log n / n \to 0$, 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 result for fully developed forest;
- Importance of subsampling;
- One major drawback: **(H2)** seems impossible to verify.

Sparsity and random forests

- Assume that

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

  for some $S < d$.
- Denote by $j_{1,n}(\mathbf{X}), \dots, j_{k,n}(\mathbf{X})$ the first k cut directions used to construct the cell containing $\mathbf{X}$.

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 $1 \leq q \leq k$,

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

Conclusion

- Centred forests: their consistency results from the consistency of each tree.

  $\rightarrow$ No benefits from using a forest instead of a single tree.

- Median forests: the aggregation process can turn inconsistent trees into a consistent forest.

  $\rightarrow$ Benefits from using a random forest compared to a single tree.

- Breiman forests: consistent as well as CART procedure. The splitting criterion asymptotically selects relevant features.

  $\rightarrow$ Good performance in high-dimensional settings.

Definition of a modified Mondrian tree

Consider the following random tree of parameter $\lambda > 0$:

- For the root node, we let $\tau = 0$ and $A = [0,1]^d$.
- For each cell $A = \prod_{j=1}^{d}[a^j, b^j]$, the selected splitting dimension $j \in \{1, \ldots, d\}$ and location s are chosen as follows:

$$j^\star = \operatorname*{argmin}_{1 \leq j \leq d} \frac{T^j}{b^j - a^j}, \qquad s^\star = U([a^{j^\star}, b^{j^\star}]),$$

  where T^j for $j = 1, \ldots, d$ are independent random variable distributed as $\mathrm{Exp}(1)$. The cell A is then split at time

$$\tau_A = \tau + T^{j^\star}/(b^{j^\star} - a^{j^\star}),$$

  where τ is the splitting time of the direct ancestor of A.
- All splits performed at time larger than λ are removed from the tree.
- Finally, the observations are used to compute the average in each cell.

Theorem: Minimax rates for Lipschitz functions

Assume that the regression function

$$\begin{aligned} m : &[0,1]^d \to \mathbb{R} \\ &\mathbf{x} \mapsto \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] \end{aligned}$$

is Lipschitz on $[0,1]^d$. Let m_n be the Mondrian Forest regressor, with a lifetime sequence that satisfies $\lambda_n \asymp n^{1/(d+2)}$. Then, the following upper bound holds

$$\mathbb{E}\big[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\big]^2 \leq Cn^{-2/(d+2)}.$$

which corresponds to the minimax rate over the set of Lipschitz functions.

Proposition: Cell diameter

Let $\mathbf{x} \in [0,1]^d$, and let $D_\lambda(\mathbf{x})$ be the diameter of the cell $A_\lambda(\mathbf{x})$ containing $\mathbf{x}$ in a partition $\mathrm{MP}(\lambda, [0,1]^d)$. If $\lambda \to \infty$, then $D_\lambda(x) \to 0$ in probability. More precisely, for every $\delta, \lambda > 0$, we have

$$\mathbb{P}(D_\lambda(\mathbf{x}) \geq \delta) \leq d\left(1 + \frac{\lambda\delta}{\sqrt{d}}\right) \exp\left(-\frac{\lambda\delta}{\sqrt{d}}\right)$$

and

$$\mathbb{E}\left[D_\lambda(\mathbf{x})^2\right] \leq \frac{4d}{\lambda^2}.$$

Proposition: Number of cells

If K_λ denotes the number of cells in a tree partition $\mathrm{MP}(\lambda, [0,1]^d)$, we have $\mathbb{E}[K_\lambda] = (1+\lambda)^d$.

Theorem: Minimax rates for $\mathscr{C}^2$ functions

Assume that $\mathbf{X}$ is uniformly distributed on $[0,1]^d$ and that the regression function

$$\begin{aligned} m :& [0,1]^d \to \mathbb{R} \\ & \mathbf{x} \mapsto \mathbb{E}[Y|\mathbf{X}=\mathbf{x}] \end{aligned}$$

is $\mathscr{C}^2$ on $[0,1]^d$. Let m_n be the infinite Mondrian Forest estimate, with a lifetime sequence that satisfies $\lambda_n \asymp n^{1/(d+4)}$. Then, for every $\epsilon > 0$,

$$\mathbb{E}\left[(m_{\infty,n}(\mathbf{X}) - m(\mathbf{X}))^2 | \mathbf{X} \in (\epsilon, 1-\epsilon)^d\right] \leq C n^{-4/(d+4)},$$

which corresponds to the minimax rate over the set of $\mathscr{C}^2$ functions.

Theoretical difficulties for studying random forests

The infinite random forests estimate takes the form

$$m_{\infty,n}(\mathbf{x}) = \mathbb{E}_{\Theta}\left[m_n(\mathbf{x}, \Theta)\right].$$

Theoretical difficulties for studying random forests

The infinite random forests estimate takes the form

$$m_{\infty,n}(\mathbf{x}) = \mathbb{E}_{\Theta}\left[\sum_{i=1}^{n} Y_i \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x},\Theta)}}{N_n(\mathbf{x},\Theta)}\right].$$

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

Theoretical difficulties for studying random forests

The infinite random forests estimate takes the form

$$m_{\infty,n}(\mathbf{x}) = \sum_{i=1}^{n} Y_i \mathbb{E}_{\Theta}\left[\frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x},\Theta)}}{N_n(\mathbf{x},\Theta)}\right],$$

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

Theoretical difficulties for studying random forests

The infinite random forests estimate takes the form

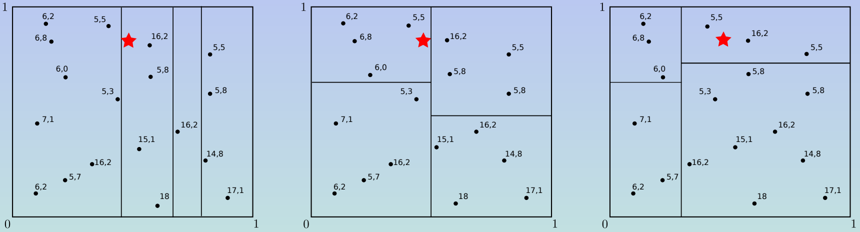
$$m_{\infty,n}(\mathbf{x}) = \sum_{i=1}^{n} Y_i \mathbb{E}_{\Theta}\left[\frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x},\Theta)}}{N_n(\mathbf{x},\Theta)}\right],$$

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

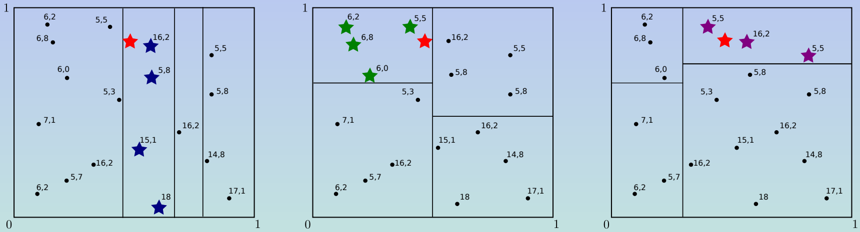
Two different difficulties:

- The tree dependency on the random variable Θ is unknown.
- The number of points in each cell is unknown.

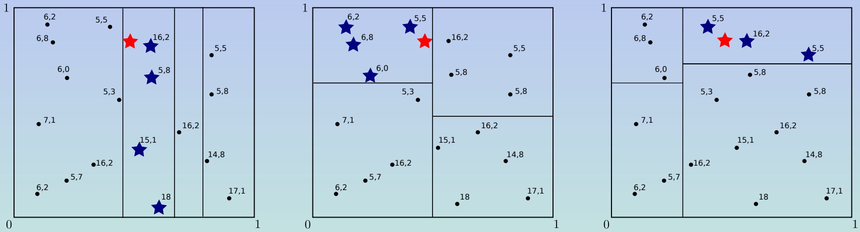
Kernel based on Random Forests (KeRF)



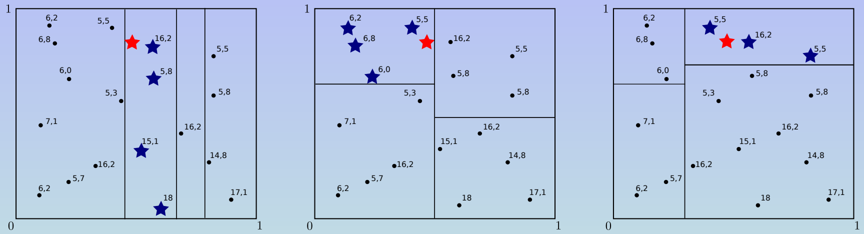
Kernel based on Random Forests (KeRF)



Kernel based on Random Forests (KeRF)



Kernel based on Random Forests (KeRF)



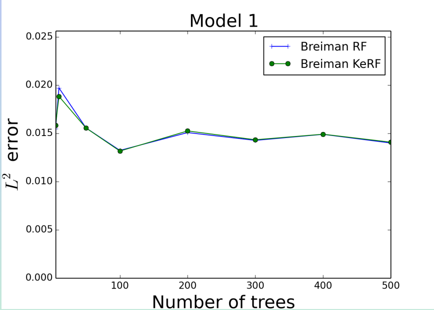
Infinite KeRF estimate:

$$\widetilde{m}_{\infty,n}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K_k(\mathbf{x}, \mathbf{X}_i)}{\sum_{j=1}^{n} K_k(\mathbf{x}, \mathbf{X}_j)},$$

where $K_k(\mathbf{x}, \mathbf{X}_i) = \mathbb{P}_\Theta\left[\mathbf{X}_i \in A_n(\mathbf{x}, \Theta)\right]$.

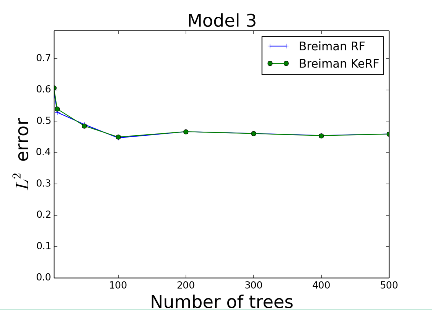
Breiman KeRF vs Breiman random forests

$n = 800, d = 50$



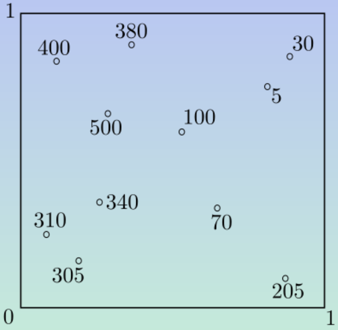
$$Y = X_1^2 + \exp(-X_2^2)$$

$n = 600, d = 100$



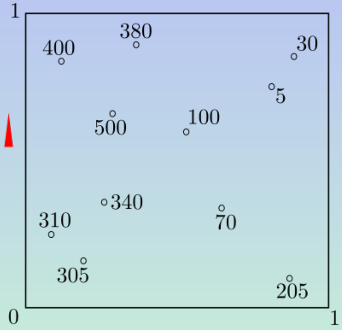
$$Y = -\sin(2X_1) + X_2^2 + X_3 - \exp(-X_4) + \mathcal{N}(0, 0.5)$$

A simple model: the centred forest



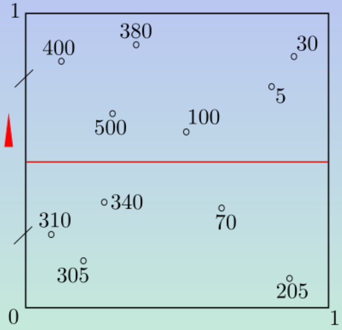
$k = 0$

A simple model: the centred forest



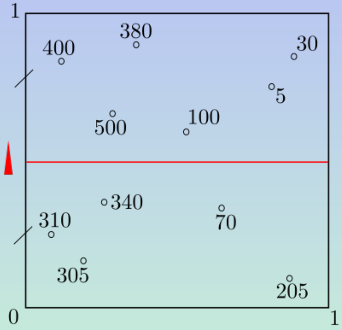
$k = 0$

A simple model: the centred forest



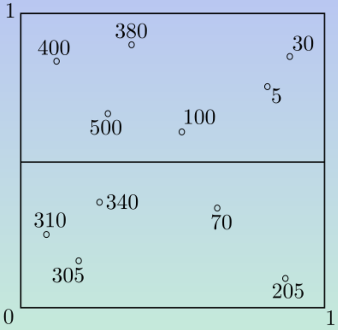
$k = 0$

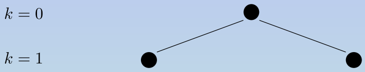
A simple model: the centred forest



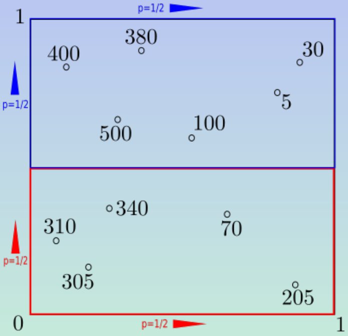
$k = 0$

A simple model: the centred forest



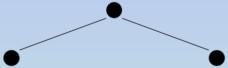


A simple model: the centred forest

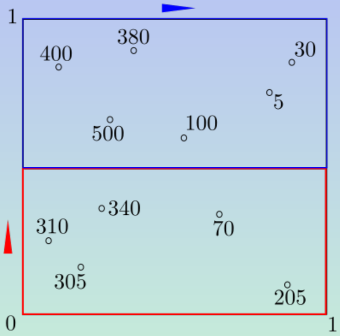


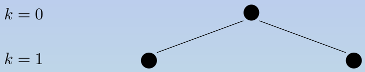
$k = 0$

$k = 1$

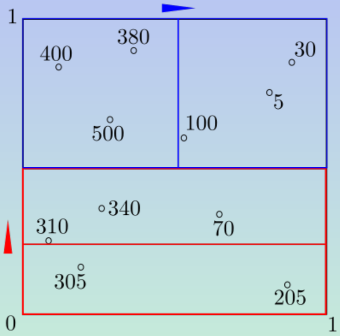


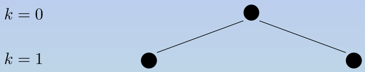
A simple model: the centred forest



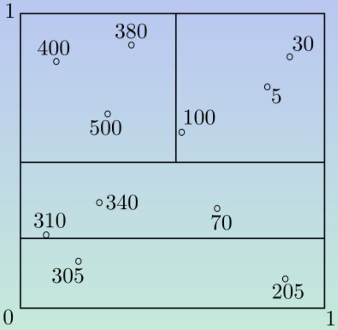


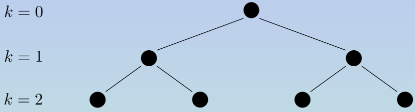
A simple model: the centred forest





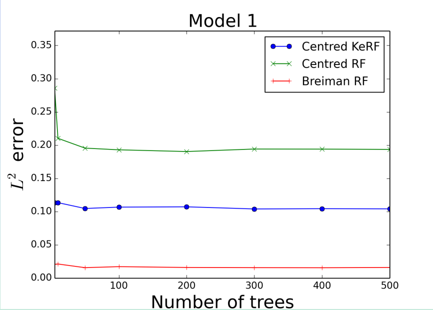
A simple model: the centred forest





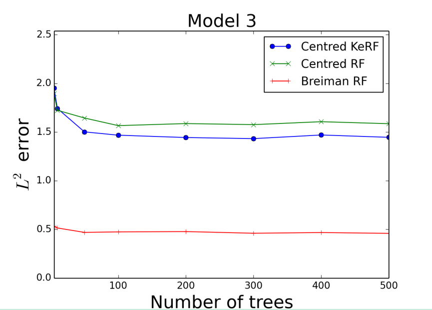
Centred KeRF vs centred random forests

$n = 800, d = 50$



$$Y = X_1^2 + \exp(-X_2^2)$$

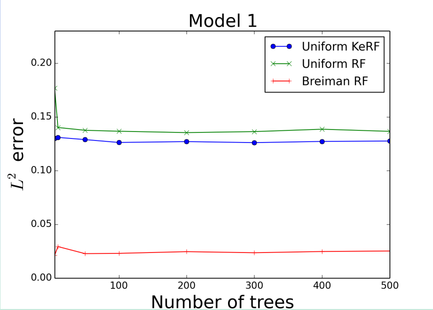
$n = 600, d = 100$



$$Y = -\sin(2X_1) + X_2^2 + X_3 - \exp(-X_4) + \mathcal{N}(0, 0.5)$$

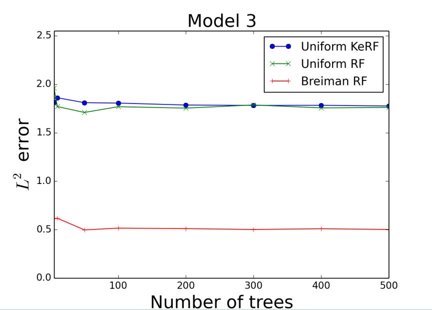
Uniform KeRF vs uniform random forests

$n = 800, d = 50$



$$Y = X_1^2 + \exp(-X_2^2)$$

$n = 600, d = 100$



$$Y = -\sin(2X_1) + X_2^2 + X_3 - \exp(-X_4) + \mathcal{N}(0, 0.5)$$

Analyzing KeRF estimates

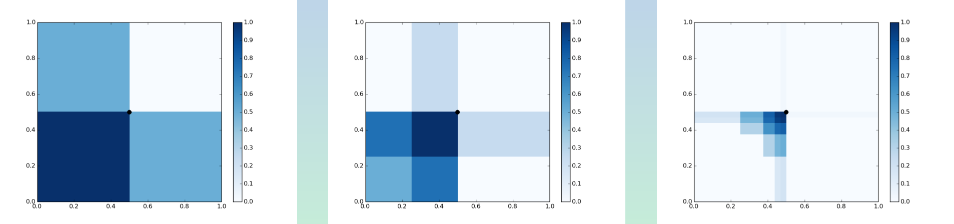
Infinite KeRF estimate: $\widetilde{m}_{\infty,n}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K_k(\mathbf{x},\mathbf{X}_i)}{\sum_{j=1}^{n} K_k(\mathbf{x},\mathbf{X}_j)}$

- Local averaging estimate and thus easier to analyze.
- One common assumption on kernel estimate is that $K_k(\mathbf{x},\mathbf{z}) = K(\frac{\mathbf{x}-\mathbf{z}}{k})$ which is not verified here.
- Generally, $K_k(\mathbf{x},\mathbf{X}_i)$ cannot be made explicit (due to the complexity of partitioning). But it can be computed for centred/uniform random forests.

Centred forests

For all $\mathbf{x}, \mathbf{z} \in [0,1]^d$,

$$K_k^{cc}(\mathbf{x}, \mathbf{z}) = \sum_{\substack{k_1, \ldots, k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1! \ldots k_d!} \left(\frac{1}{d}\right)^k \prod_{m=1}^{d} \mathbb{1}_{\lceil 2^{k_m} x_m \rceil = \lceil 2^{k_m} z_m \rceil}.$$

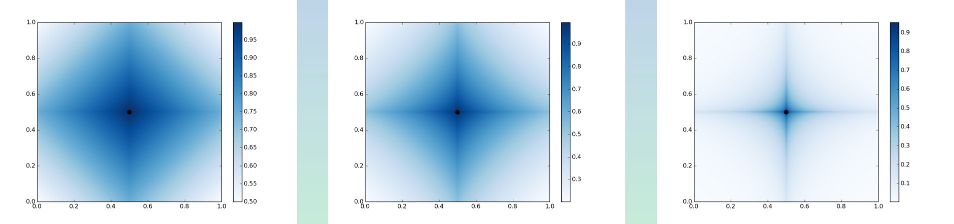


Representations of $\mathbf{z} \mapsto K_k^{cc}((0.5, 0.5), \mathbf{z})$ for $k = 1, 2, 5$

Uniform forests

For all $\mathbf{z} \in [0,1]^d$,

$$K_k^{uf}(0,\mathbf{z}) = \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1!\ldots k_d!} \left(\frac{1}{d}\right)^k \prod_{m=1}^d z_m \sum_{j=k_m}^{\infty} \frac{(-\log z_m)^j}{j!}.$$



Representations of $\mathbf{z} \mapsto K_k^{uf}\big(0, (z_1 - 0.5, z_2 - 0.5)\big)$ for $k = 1, 2, 5$

Summary of KeRF

- Interpretable form (kernel estimate):

$$\widetilde{m}_{\infty,n}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K_k(\mathbf{x}, \mathbf{X}_i)}{\sum_{j=1}^{n} K_k(\mathbf{x}, \mathbf{X}_j)}.$$

- The kernel function $K_k(\mathbf{x}, \mathbf{X}_i) = \mathbb{P}_\Theta\left[\mathbf{X}_i \in A_n(\mathbf{x}, \Theta)\right]$ is related to the shape of partitions
- KeRF are close to random forests in terms of prediction accuracy.
- **But** explicit expression for Breiman KeRF is difficult to obtain.

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.

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. Asymptotic theory for random forests. arXiv:1405.0352, 2014.

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.

Merci pour votre attention !