A guided tour in random forests

E. Scornet

Internal meeting - ANR GATSBII - November 2026





Summary

1. Random Forests construction
2. Several random forest models
3. Number of trees
4. Centered forests
5. Mondrian Forests
6. Median forests
7. Breiman forests
8. Conclusion

Statistical framework

Consider (X_i, Y_i) i.i.d. copies of the pair (X, Y)

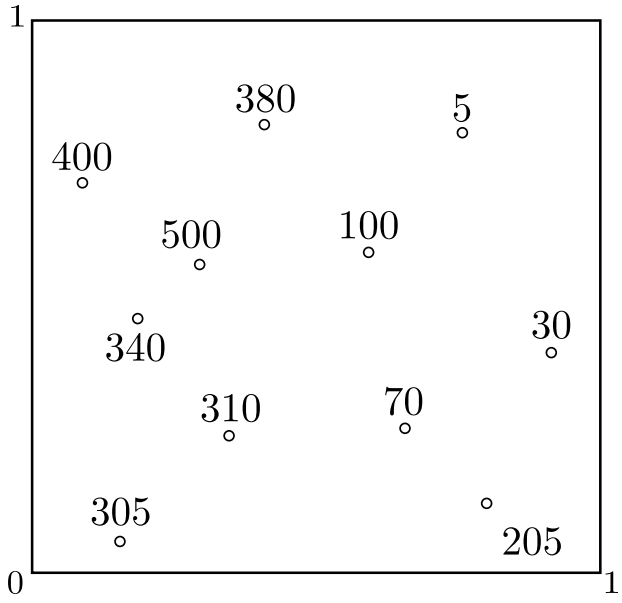
- **Input variable** $X \in \mathcal{X} = [0,1]^d$
- **Output variable** $Y \in \mathbb{R}$.
- **Regression model**: $Y = m(\mathbf{X}) + \varepsilon$.

Goal: estimation of m using **random forests**.



- Non-parametric method
- Based on bagging and random feature selections
- Aggregate the predictions of M trees

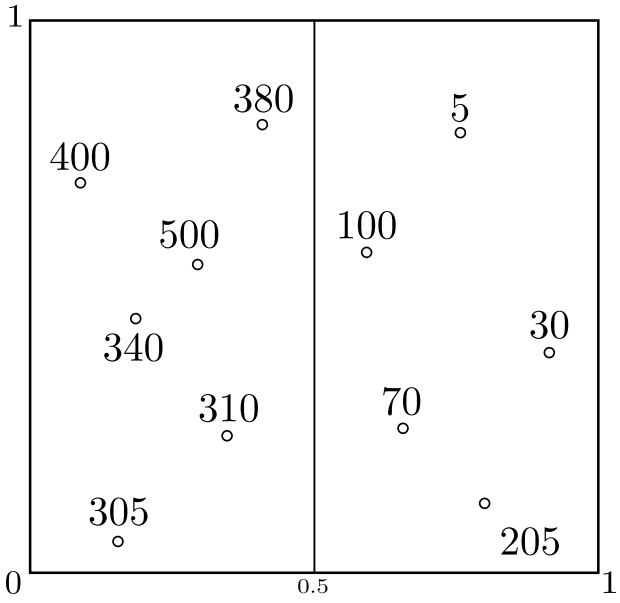
Construction of Decision trees - regression



$k = 0$

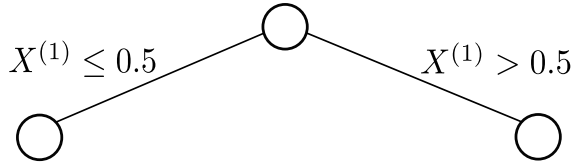


Construction of Decision trees - regression

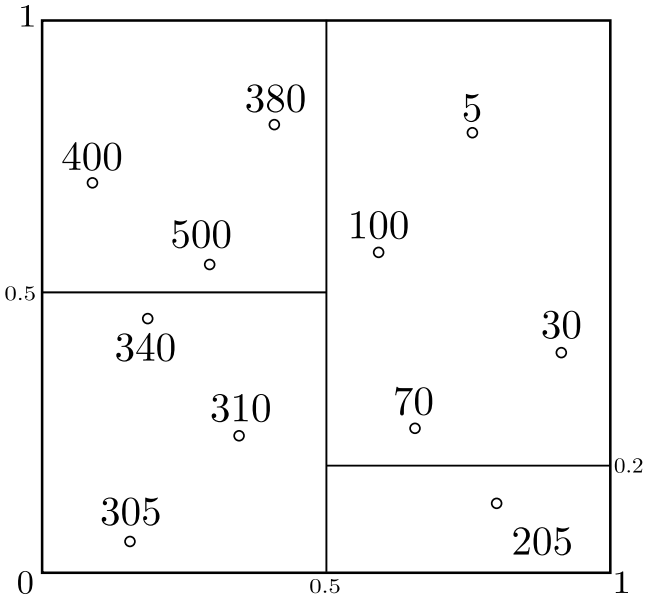


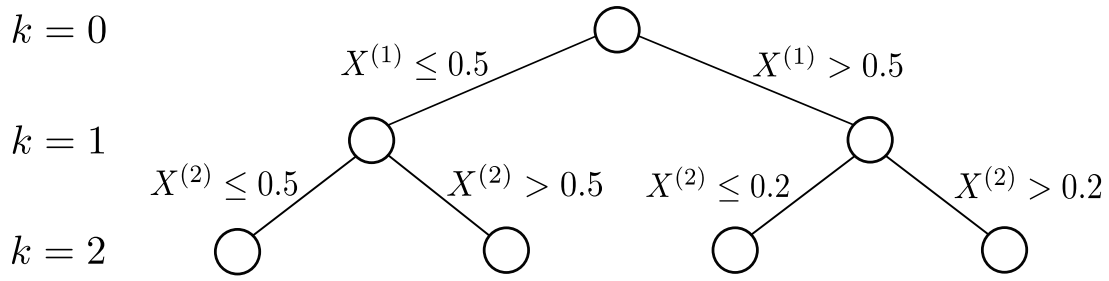
$k = 0$

$k = 1$

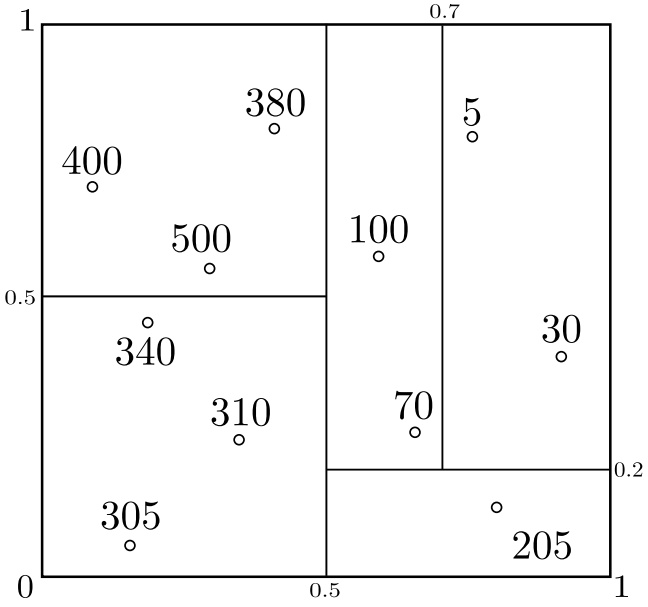


Construction of Decision trees - regression





Construction of Decision trees - regression

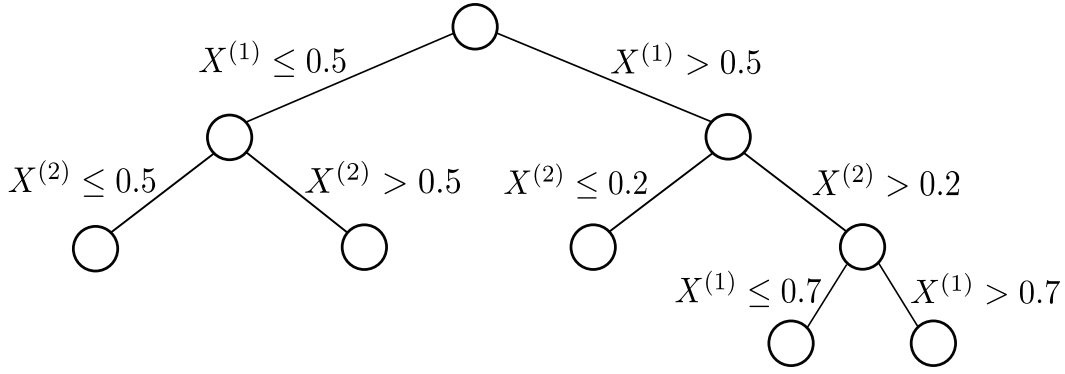


$k = 0$

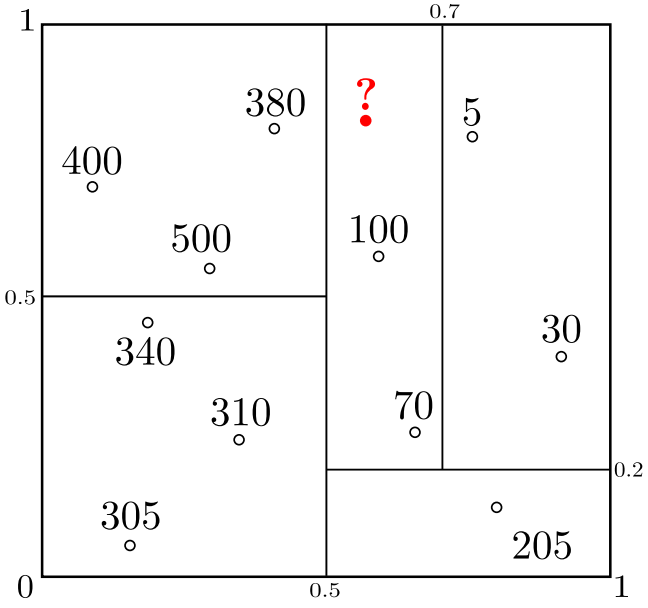
$k = 1$

$k = 2$

$k = 3$



Construction of Decision trees - regression

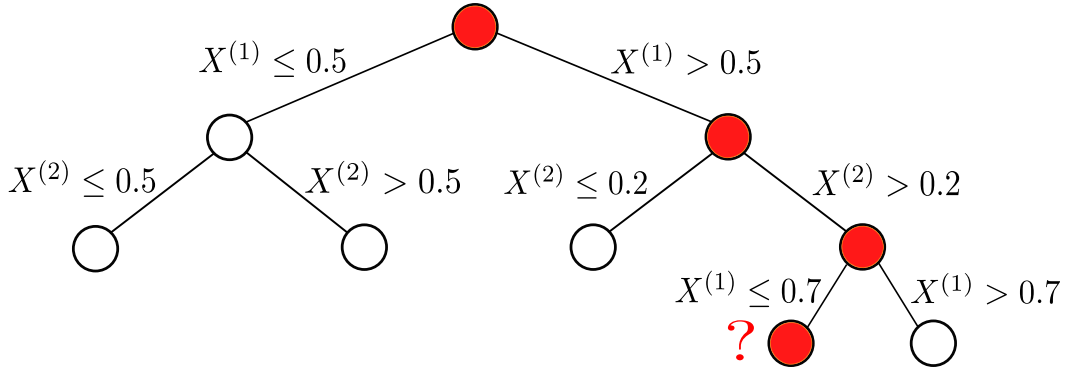


$k = 0$

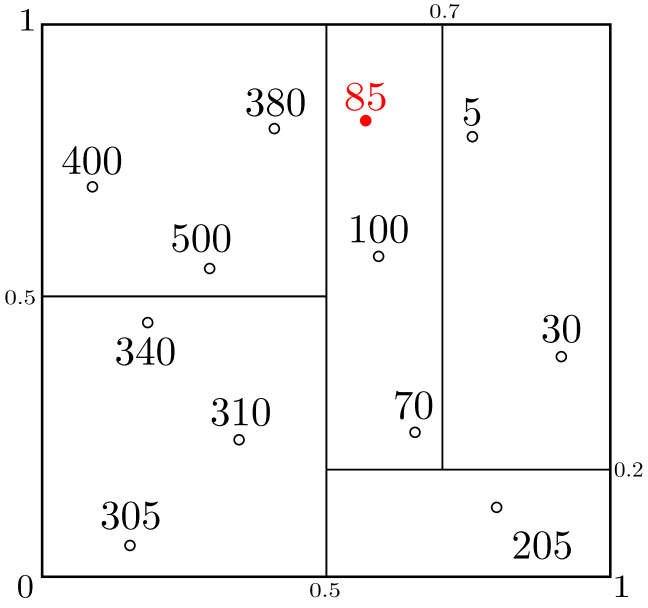
$k = 1$

$k = 2$

$k = 3$



Construction of Decision trees - regression

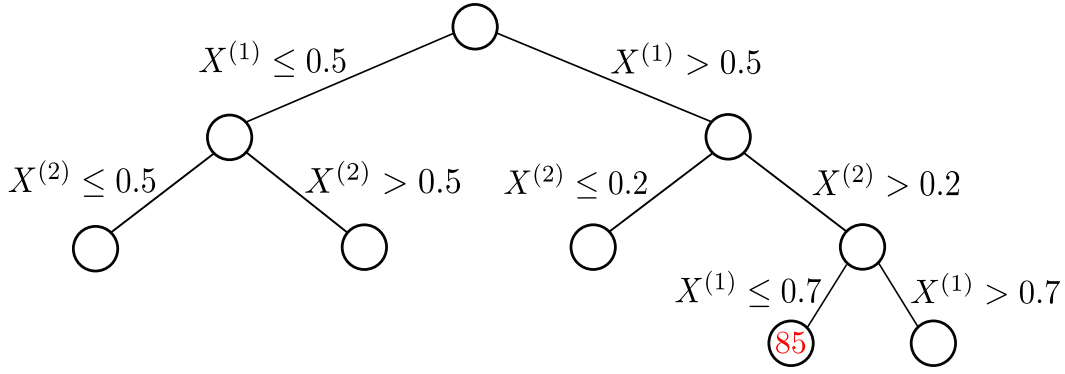


$k = 0$

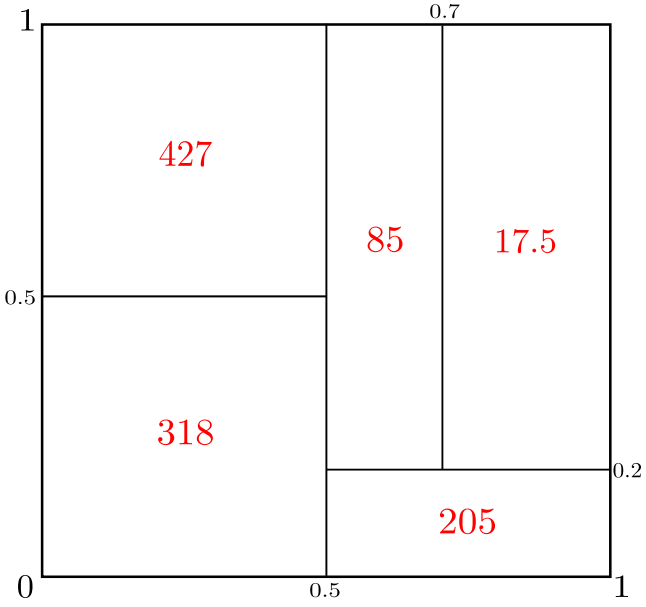
$k = 1$

$k = 2$

$k = 3$



Construction of Decision trees - regression

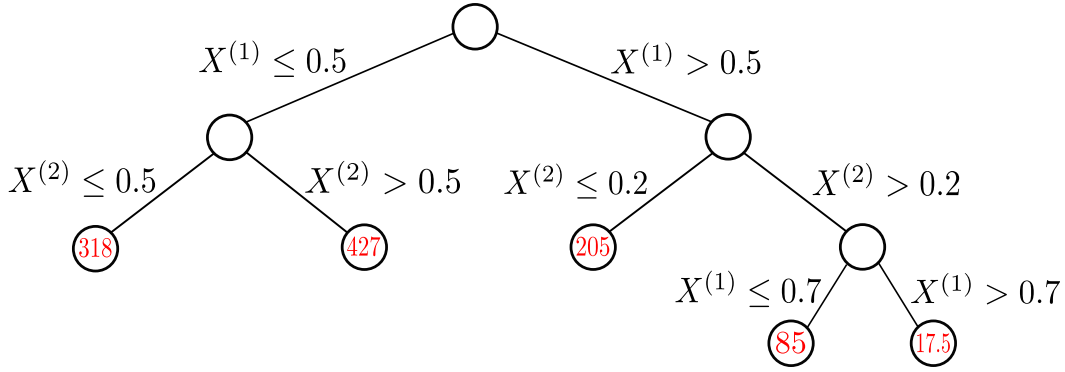


$k = 0$

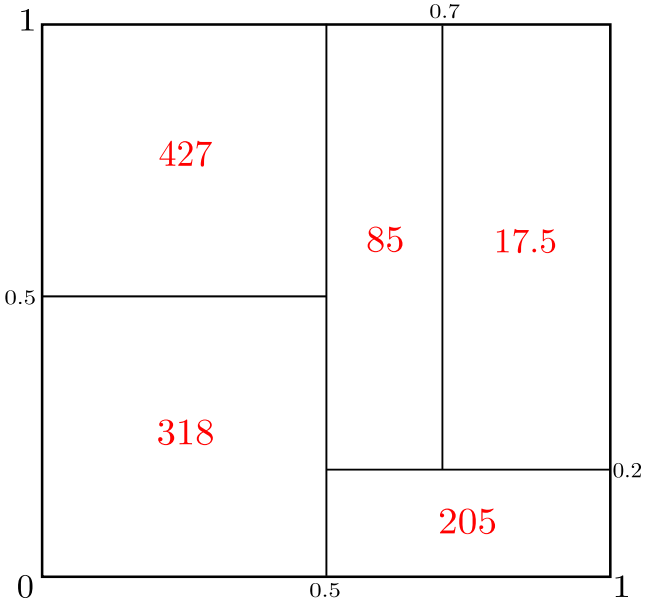
$k = 1$

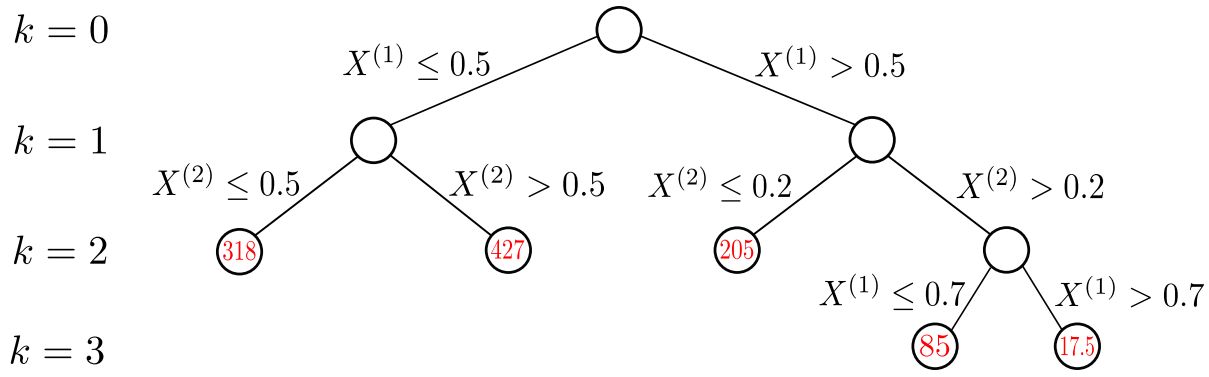
$k = 2$

$k = 3$



Construction of Decision trees - regression





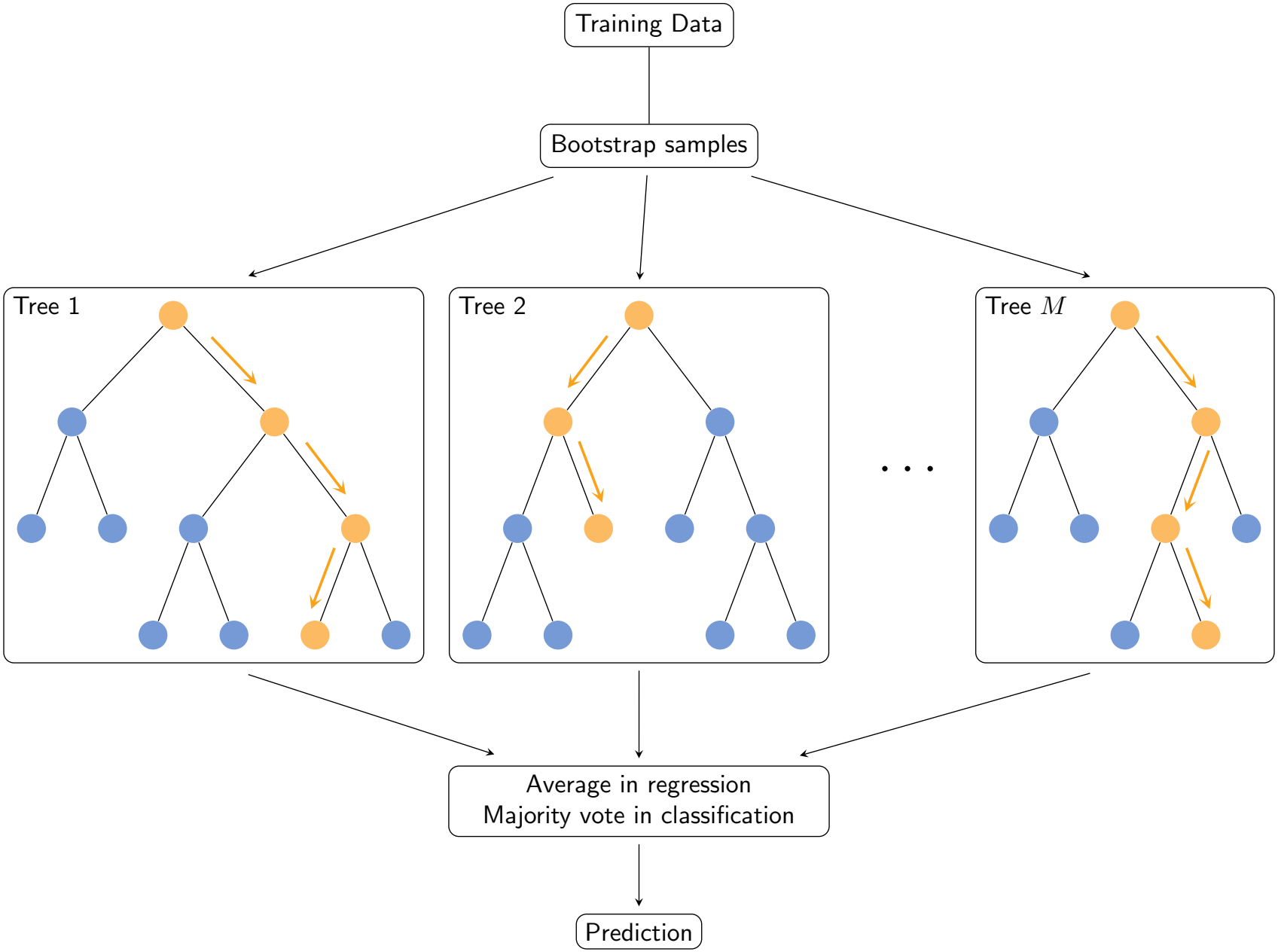
Decrease in impurity for a split (j, s):

$$\Delta Imp(j, s; A) = Imp(A) - p_L Imp(A_L) - p_R Imp(A_R),$$

where p_L (resp. p_R) is the fraction of observations in A that fall into A_L (resp. A_R). The best split $(j^\star, s^\star)$ is then chosen as

$$(j^\star, s^\star) \in \text{argmax}_{j,s} \Delta Imp(j, s; A).$$

A classical random forest



Theory of RF (Breiman, 2001): literature review

- **Simplified RF versions**, whose construction is **independent of the dataset**.

  (Biau et al., 2008; Biau, 2012; Genuer, 2012; Arlot and Genuer, 2014; Scornet, 2016b; Mourtada et al., 2020; Klusowski, 2021)

Theory of RF (Breiman, 2001): literature review

- **Simplified RF versions**, whose construction is **independent of the dataset**.

  (Biau et al., 2008; Biau, 2012; Genuer, 2012; Arlot and Genuer, 2014; Scornet, 2016b; Mourtada et al., 2020; Klusowski, 2021)

- Analysis of more data-dependent forests:
  - **Asymptotic normality** of random forests (Mentch and Hooker, 2016; Wager and Athey, 2018),
  - **Variable importance** (Louppe et al., 2013; Li et al., 2019; Scornet, 2023),
  - **(Rate of) consistency** (Scornet et al., 2015; Wager and Walther, 2015; Klusowski and Tian, 2024).

Theory of RF (Breiman, 2001): literature review

- **Simplified RF versions**, whose construction is **independent of the dataset**.

  (Biau et al., 2008; Biau, 2012; Genuer, 2012; Arlot and Genuer, 2014; Scornet, 2016b; Mourtada et al., 2020; Klusowski, 2021)

- Analysis of more data-dependent forests:
  - **Asymptotic normality** of random forests (Mentch and Hooker, 2016; Wager and Athey, 2018),
  - **Variable importance** (Louppe et al., 2013; Li et al., 2019; Scornet, 2023),
  - **(Rate of) consistency** (Scornet et al., 2015; Wager and Walther, 2015; Klusowski and Tian, 2024).

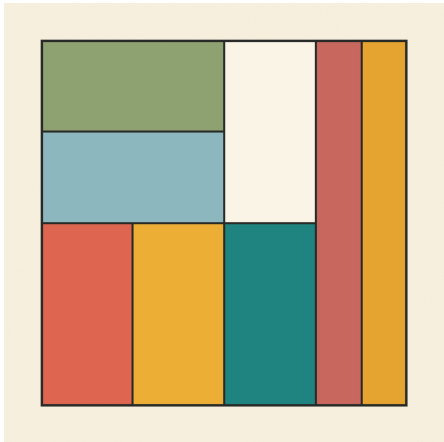
- Literature review on random forests:
  - **Methodological review** (Criminisi et al., 2011; Boulesteix et al., 2012),
  - **Theoretical review** (Biau and Scornet, 2016; Scornet and Hooker, 2026)

Summary

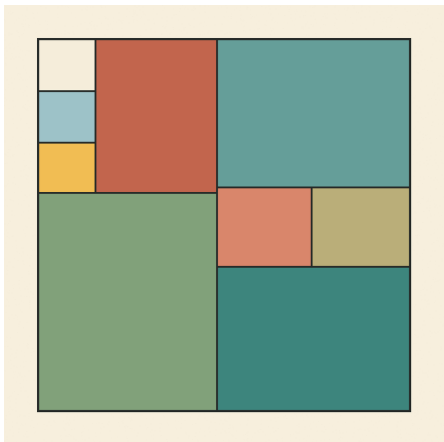
1. Random Forests construction
2. Several random forest models
3. Number of trees
4. Centered forests
5. Mondrian Forests
6. Median forests
7. Breiman forests
8. Conclusion

Different types of forests

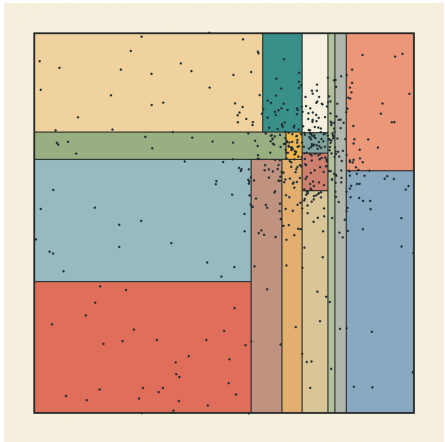
Purely RF	Semi-adaptive RF	Adaptive RF
Independent of X_i and Y_i	Independent of Y_i	Dependent on X_i and Y_i



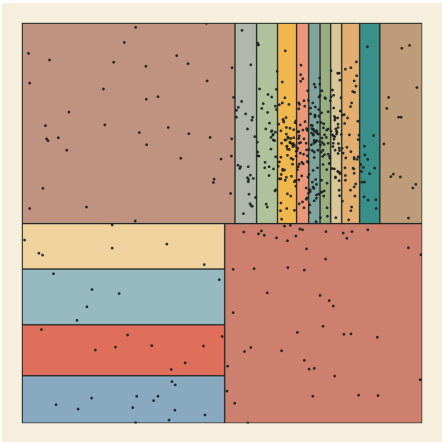
Centered RF



Mondrian RF



Median RF



Breiman RF

Summary of random forests properties

	Centred	Mondrian	Median	Breiman
Data-dependent				
Fully-developed				
Consistency				
Rate (Lipschitz)				
Benefits of				
Number of trees				
Subsampling				
Feature random.				
Fully-developed				

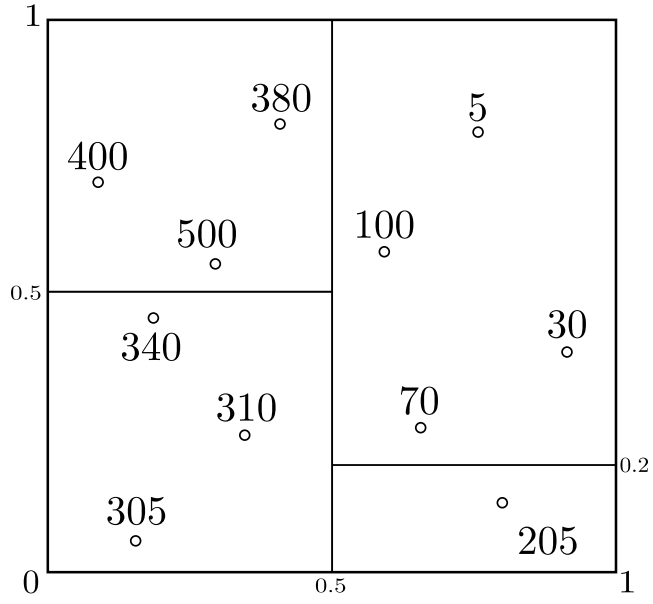
Summary of random forests properties

	Centred	Mondrian	Median	Breiman
Data-dependent	✗	✗	$\simeq$	✓
Fully-developed	✗	✗	✓	✓
Consistency				
Rate (Lipschitz)				
Benefits of				
Number of trees				
Subsampling				
Feature random.				
Fully-developed				

Summary

1. Random Forests construction
2. Several random forest models
3. Number of trees
4. Centered forests
5. Mondrian Forests
6. Median forests
7. Breiman forests
8. Conclusion

A tree

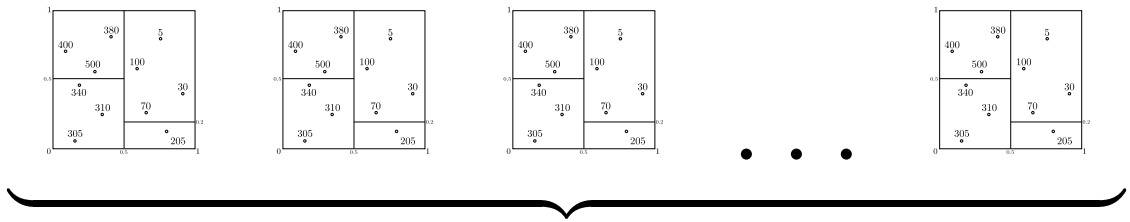


- Tree estimate:

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

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

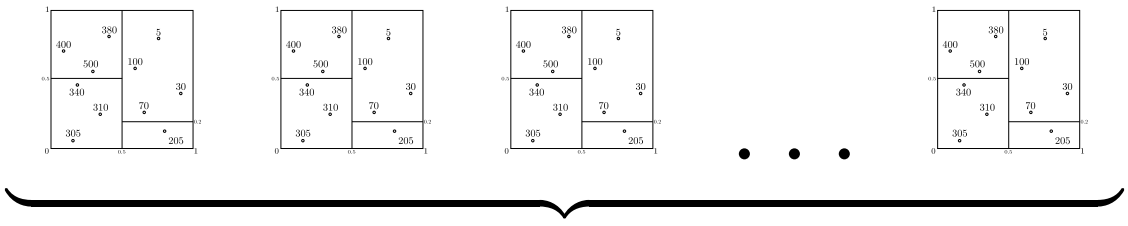
A finite forest



- M-Finite forest 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

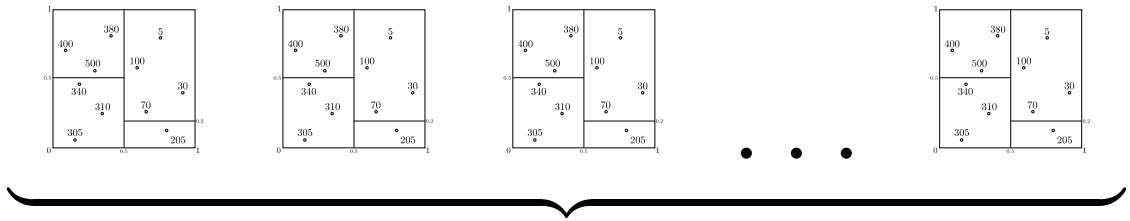


- M-Finite forest estimate :

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

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

Toward infinite forest



- M-Finite forest 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, 2016a)

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,

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

A single tree versus a forest

Theorem

We have

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

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.

Summary

1. Random Forests construction
2. Several random forest models
3. Number of trees
4. Centered forests
5. Mondrian Forests
6. Median forests
7. Breiman forests
8. Conclusion

A single tree



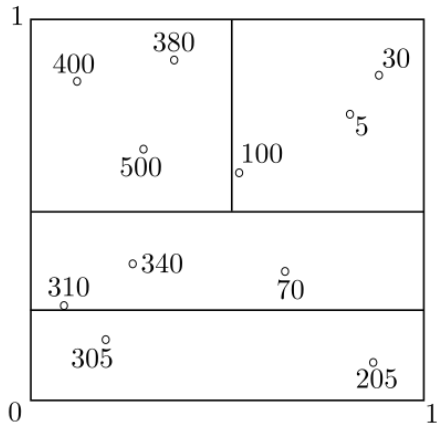
For a tree whose construction is independent of data, if

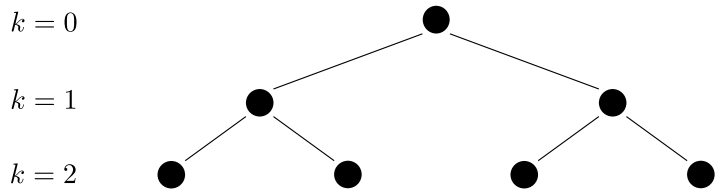
1. $\mathrm{diam}(A_n(\mathbf{X})) \to 0$, in probability;
2. $N_n(A_n(\mathbf{X})) \to \infty$, in probability;

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 $n \to \infty$.

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

Centered forests versus

Intermediate quantity $\bar{m}^{cc}_{\infty,n}(\mathbf{X}) = \sum_{i=1}^{n} W_{ni}(\mathbf{X}) m(\mathbf{X}_i)$.

Estimation error (Biau, 2012)

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

Approximation error (Biau, 2012)

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

- Smaller variance due to averaging over more observations
- Randomness of tree partition harms the approximation error of the forest

Centered forests versus cubic partitions

Intermediate quantity $\bar{m}^{cc}_{\infty,n}(\mathbf{X}) = \sum_{i=1}^{n} W_{ni}(\mathbf{X}) m(\mathbf{X}_i)$.

Estimation error (Biau, 2012)

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

Approximation error (Biau, 2012)

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

- Smaller variance due to averaging over more observations
- Randomness of tree partition harms the approximation error of the forest

Centered forests versus cubic partitions

Intermediate quantity $\bar{m}^{cc}_{\infty,n}(\mathbf{X}) = \sum_{i=1}^{n} W_{ni}(\mathbf{X}) m(\mathbf{X}_i)$.

Estimation error (Biau, 2012)

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

Approximation error (Biau, 2012)

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

- Smaller variance due to averaging over more observations
- Randomness of tree partition harms the approximation error of the forest

Consistency of centred random forest

Intermediate quantity

$$\bar{m}^{cc}_{\infty,n}(\mathbf{X}) = \sum_{i=1}^{n} W_{ni}(\mathbf{X}) m(\mathbf{X}_i)$$

Estimation error (Biau, 2012)

Under proper assumptions on the regression model,

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

Approximation error (Biau, 2012)

Under proper assumptions on the regression model,

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

Consistency of centred random forest

Intermediate quantity

$$\bar{m}^{cc}_{\infty,n}(\mathbf{X}) = \sum_{i=1}^{n} W_{ni}(\mathbf{X}) m(\mathbf{X}_i)$$

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,

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

Approximation error (Biau, 2012)

Under proper assumptions on the regression model,

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

Consistency of centred random forest

Intermediate quantity

$$\bar{m}^{cc}_{\infty,n}(\mathbf{X}) = \sum_{i=1}^{n} W_{ni}(\mathbf{X}) m(\mathbf{X}_i)$$

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,

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

Approximation error (Biau, 2012)

Under proper assumptions on the regression model,

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

Consistency of centred random forest

Intermediate quantity

$$\bar{m}^{cc}_{\infty,n}(\mathbf{X}) = \sum_{i=1}^{n} W_{ni}(\mathbf{X}) m(\mathbf{X}_i)$$

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,

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

Approximation error (Biau, 2012)

Under proper assumptions on the regression model,

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

Consistency of centred random forest

Intermediate quantity

$$\bar{m}^{cc}_{\infty,n}(\mathbf{X}) = \sum_{i=1}^{n} W_{ni}(\mathbf{X}) m(\mathbf{X}_i)$$

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,

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

Approximation error (Biau, 2012)

Under proper assumptions on the regression model,

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

Consistency of centred random forest

Intermediate quantity

$$\bar{m}^{cc}_{\infty,n}(\mathbf{X}) = \sum_{i=1}^{n} W_{ni}(\mathbf{X}) m(\mathbf{X}_i)$$

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,

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

Approximation error (Biau, 2012)

Under proper assumptions on the regression model,

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

Risk of centered forests

Assume that $\mathbf{X}$ is uniformly distributed on $[0,1]^d$, $\|m\|_\infty < \infty$ and $\max_j \|\partial_j m\|_\infty < \infty$. Then,

$$\mathbb{E}\left[m^{cc}_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\right]^2 \lesssim \sigma^2 \frac{2^{k_n}}{n k_n^{1/2}} + L^2 . 2^{-\frac{0.75 k_n}{d \log 2}} + \|m\|^2_\infty e^{-n/2^{k_n}}$$
$$\lesssim n^{-r},$$

with

- $r = \frac{1}{(4/3) d \log 2 + 1}$ (Breiman, 2004; Biau, 2012)

- Non minimax
- Optimality of bias and variance for Lispchitz functions

Risk of centered forests

Assume that $\mathbf{X}$ is uniformly distributed on $[0,1]^d$, $\|m\|_\infty < \infty$ and $\max_j \|\partial_j m\|_\infty < \infty$. Then,

$$\mathbb{E}\left[m^{cc}_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\right]^2 \lesssim \sigma^2 \frac{2^{k_n}}{n k_n^{1/2}} + L^2 . 2^{-\frac{k_n}{d \log 2}} + \|m\|_\infty^2 e^{-n/2^{k_n}}$$
$$\lesssim n^{-r},$$

with

- $r = \frac{1}{(4/3) d \log 2 + 1}$ (Breiman, 2004; Biau, 2012)
- $r = \frac{1}{d \log 2 + 1}(1+\delta)$ with $\delta \to 0$ as $d \to \infty$ (Klusowski, 2021)

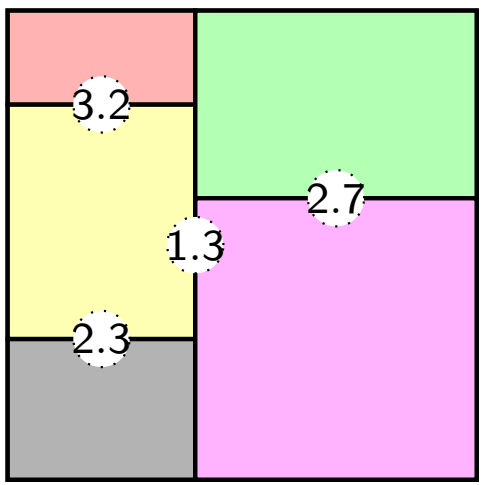
- Non minimax
- Optimality of bias and variance for Lispchitz functions

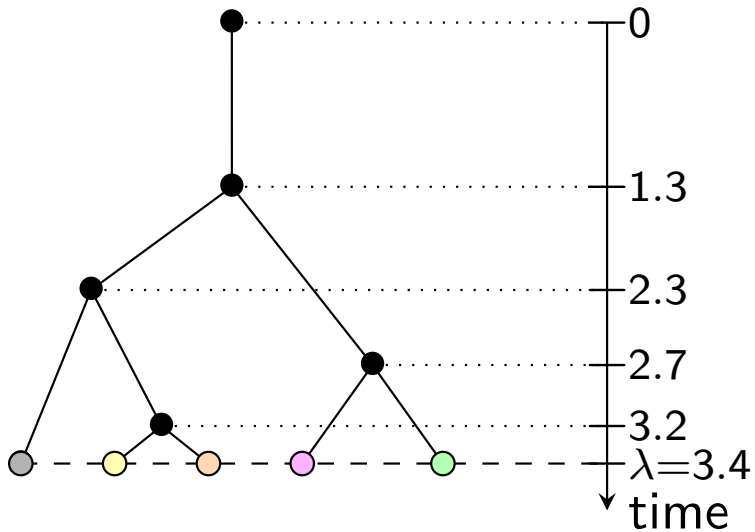
Summary

1. Random Forests construction
2. Several random forest models
3. Number of trees
4. Centered forests
5. Mondrian Forests
6. Median forests
7. Breiman forests
8. Conclusion

The Mondrian process (Roy and Teh, 2008)

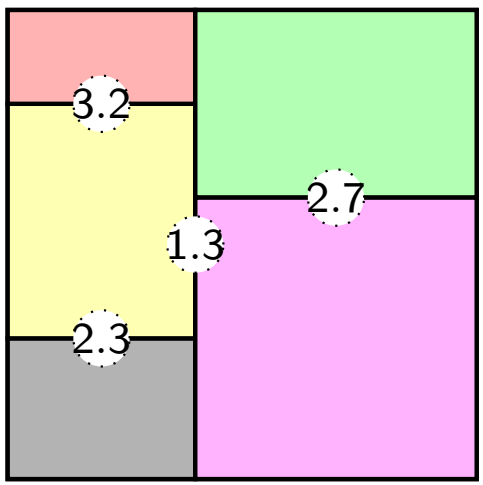
- $\mathrm{MP}(\lambda, C)$: distribution on recursive, axis-aligned partitions of $C = \prod_{j=1}^{d}[a_j, b_j] \subset \mathbb{R}^d$ (= trees).
- $\lambda > 0$ "lifetime" = complexity parameter.

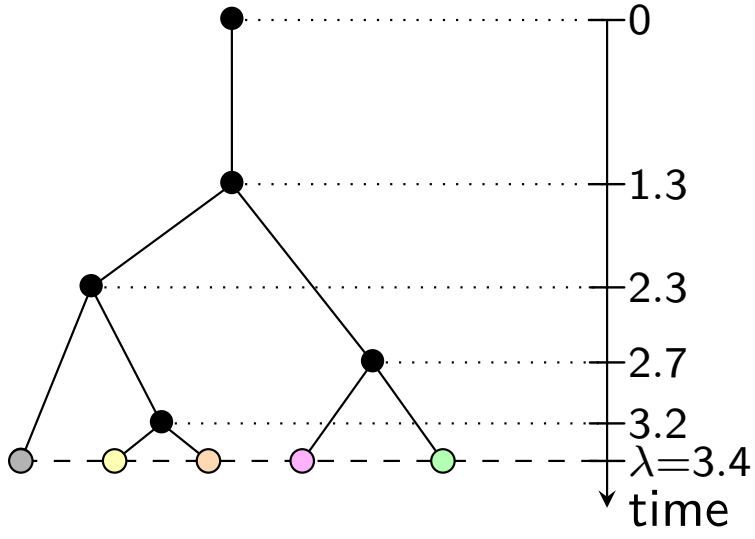




The distribution $\mathrm{MP}(\lambda, C)$

- **Start** with cell C (root), formed at time $\tau_C = 0$.
- **Sample** time till split $E \sim (|C|)$ with $|C| := \sum_{j=1}^{d}(b_j - a_j)$
- **If** $\tau_C + E \leqslant \lambda$,
  - split C in $C_L = \{x \in C : x_J \leqslant S_J\}$ 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 $S_J | J \sim ([a_J, b_J])$
  - 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 Lakshminarayanan et al. (2014) 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.
- Purely random forests: construction independent of the dataset.
- No depth parameter but a lifetime parameter λ: tree depth increases with λ.

Theoretical results

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

Theorem Mourtada et al. (2020)

Assume $\mathbb{V}(Y|X) \leqslant \sigma^2 < \infty$ and that the regression function m is L-Lipschitz. Then:

$$\mathbb{E}[(m_{\lambda,M,n}(\mathbf{X}) - m(\mathbf{X}))^2] \leqslant \frac{4dL^2}{\lambda^2} + \frac{(1+\lambda)^d}{n}\left(2\sigma^2 + 9\|m\|_\infty^2\right).$$

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

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

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 $M \geqslant 1$ (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 C^2 regularity

Theorem Mourtada et al. (2020)

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

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

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

$$\mathbb{E}[(Y - m_{\lambda,M,n}(\mathbf{X})^2 \mathbb{1}_{[\varepsilon,1-\varepsilon]^d}] = O(n^{-4/(d+4)})$$

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 $(1+\lambda)^d$.

Summary

1. Random Forests construction
2. Several random forest models
3. Number of trees
4. Centered forests
5. Mondrian Forests
6. Median forests
7. Breiman forests
8. Conclusion

Construction of Median forests

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 contains exactly $\texttt{nodesize} = 1$ observation.

Consistency of median forests

Assumption **(H1)**

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

Theorem Scornet (2016a)

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

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

- 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 (subsampling).

Proof of Theorem (1)

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

Lemma 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

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

So,

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

which tends to zero by assumption.

Construction of Median forests

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 and Scornet (2018)

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

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

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 and Scornet (2018)

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

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

- 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 and Scornet (2018)

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

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

$$\leqslant 2\sigma^2 \frac{2^k}{a_n} + dL^2 C_1 \left(1 - \frac{3}{4d}\right)^k.$$

- 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 and Scornet (2018)

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

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

$$\leqslant 2\sigma^2 \frac{2^k}{a_n} + dL^2 C_1 \left(1 - \frac{3}{4d}\right)^k.$$

- 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 and Scornet (2018)

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

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

- 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 and Scornet (2018)

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

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

- 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 and Scornet (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 $a_n \geqslant 2^{k_n^\star}$,

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

Rate of consistency

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

Theorem Duroux and Scornet (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 $a_n \geqslant 2^{k_n^\star}$,

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

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 and Scornet (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 $a_n \geqslant 2^{k_n^\star}$,

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

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 $a_n/n \to 0$.

No need for tuning them both at the same time.

Interpolation and non-parametric methods

- Non-parametric learning

  No fixed number of parameters a priori

Interpolation and non-parametric methods

- Non-parametric learning

  No fixed number of parameters a priori

- Nearest neighbour predictor
  - ✓ Simplest interpolator

Interpolation and non-parametric methods

- Non-parametric learning

  No fixed number of parameters a priori

- Nearest neighbour predictor
  - ✓ Simplest interpolator
  - ✗ Inconsistent (apart from the noiseless setting) i.e. [Biau et al. 2015]

$$\text{Risk}(f^{1NN}) \underset{n \to +\infty}{\not\longrightarrow} \text{Risk}(f^\star)$$

Interpolation and non-parametric methods

- Non-parametric learning

  No fixed number of parameters a priori

- Nearest neighbour predictor
  - ✓ Simplest interpolator
  - ✗ Inconsistent (apart from the noiseless setting) i.e. [Biau et al. 2015]

$$\text{Risk}(f^{1NN}) \underset{n \to +\infty}{\not\longrightarrow} \text{Risk}(f^\star)$$

- Local-means estimator: $f(x) = \dfrac{\sum_{i=1}^n Y_i K\left(\frac{\|x - X_i\|}{h}\right)}{\sum_{i=1}^n K\left(\frac{\|x - X_i\|}{h}\right)}$ with

$$K(x) = \frac{1}{\|x\|^p}$$

Interpolation and non-parametric methods

- Non-parametric learning

  No fixed number of parameters a priori

- Nearest neighbour predictor
  - ✓ Simplest interpolator
  - ✗ Inconsistent (apart from the noiseless setting) i.e. [Biau et al. 2015]

$$\text{Risk}(f^{1NN}) \xrightarrow[n\to+\infty]{\not\longrightarrow} \text{Risk}(f^\star)$$

- Local-means estimator: $f(x) = \dfrac{\sum_{i=1}^n Y_i K\left(\frac{\|x-X_i\|}{h}\right)}{\sum_{i=1}^n K\left(\frac{\|x-X_i\|}{h}\right)}$ with

$K(x) = \frac{1}{\|x\|^p}$

  - ✓ Interpolator
  - ✓ Consistent

[Devroye et al. 1998]
[Belkin et al. (2019)]

Consistency of singular kernels

Belkin et al. (2019) consider Nadaraya-Watson predictors of the form

$$f_{a,h,n}(x) = \frac{\sum_{i=1}^{n} Y_i K_a\left(\frac{\|x-X_i\|}{h}\right)}{\sum_{i=1}^{n} K_a\left(\frac{\|x-X_i\|}{h}\right)},$$

with singular kernels $K_a(x) = \|x\|^{-a} \mathbb{1}_{\|x\| \leqslant 1}$.

Consistency of singular kernels

Belkin et al. (2019) consider Nadaraya-Watson predictors of the form

$$f_{a,h,n}(x) = \frac{\sum_{i=1}^{n} Y_i K_a\left(\frac{\|x-X_i\|}{h}\right)}{\sum_{i=1}^{n} K_a\left(\frac{\|x-X_i\|}{h}\right)},$$

with singular kernels $K_a(x) = \|x\|^{-a} \mathbb{1}_{\|x\| \leqslant 1}$.

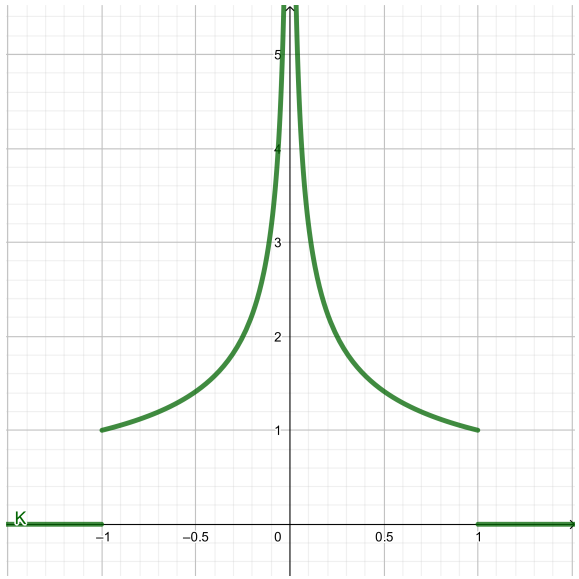


Figure: Singular kernel above for $a = 0.5$

Consistency of singular kernels

Belkin et al. (2019) consider Nadaraya-Watson predictors of the form

$$f_{a,h,n}(x) = \frac{\sum_{i=1}^n Y_i K_a\left(\frac{\|x-X_i\|}{h}\right)}{\sum_{i=1}^n K_a\left(\frac{\|x-X_i\|}{h}\right)},$$

with singular kernels $K_a(x) = \|x\|^{-a}\mathbb{1}_{\|x\|\leqslant 1}$.

Regression model: $Y = f^\star(X) + \varepsilon$ with

- $\mathbb{E}[\varepsilon^2|X] \leqslant \sigma^2$ a.s.
- $X \sim \mathcal{U}([0,1]^d)$
- and $f^\star$ Lipschitz.

Theorem (Belkin et al. (2019) - A specific case)

Let $0 < a < d/2$. Letting $h_n = n^{-1/(2+d)}$, we have

$$Risk(f_{a,h_n,n}) \leqslant Cn^{-2/(d+2)}.$$

Interpolation

- To analyze the strict interpolation case, we have to consider semi-adaptive RF

Semi-adaptive median RF

1. Median tree
   - The original n observations are used to build each 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 contains exactly $\texttt{nodesize} = 1$ observation.
2. Median RF: aggregation of median trees

Semi-adaptive RF: median RF

Theorem (Arnould et al., 2023)

Suppose that m has bounded partial derivatives and that n is a power of two. Then, the interpolating Median RF $^{\mathrm{MedRF}}$ verifies:

$$\mathbb{E}\big[m_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\big]^2 \leqslant C_1 d \left(\sum_{\ell=1}^{d} ||\partial_\ell m||_\infty^2\right) \left(1 - \frac{3}{4d}\right)^{\log_2 n} + \sigma^2 C_{2,d} (\log_2 n)^{-(d-1)/2},$$

where C_1 and $C_{2,d}$ are explicit constants.

Semi-adaptive RF: median RF

Theorem (Arnould et al., 2023)

Suppose that m has bounded partial derivatives and that n is a power of two. Then, the interpolating Median RF $^{\mathrm{MedRF}}$ verifies:

$$\mathbb{E}\big[m_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\big]^2 \leqslant C_1 d \left(\sum_{\ell=1}^{d} ||\partial_\ell m||_\infty^2\right) \left(1 - \frac{3}{4d}\right)^{\log_2 n} + \sigma^2 C_{2,d} (\log_2 n)^{-(d-1)/2},$$

where C_1 and $C_{2,d}$ are explicit constants.

- Interpolating (median) RF are consistent in a noisy setting (first result).
- Each tree is not consistent but the forest is (due to the randomization of splitting directions).
- First result to highlight the asymptotic benefit of split randomization (making the forest consistent).

Unreachable minimax rate for purely random forests

Let $Y = m(\mathbf{X}) + \varepsilon$ with $\mathbb{E}(\varepsilon) = 0$, $\text{var}(\varepsilon) = \sigma^2$ and $\mathbf{X} \sim \mathcal{U}([0,1]^d)$.

Theorem Lin and Jeon (2006)

Consider an estimator $\hat{m}_{\ell,n}$ resulting from a nonadaptive random forest with terminal node size at most ℓ. Then, there exists $\Lambda_d > 0$ such that for any n,

$$\mathbb{E}[(m(\mathbf{X}) - \hat{m}_{\ell,n}(\mathbf{X}))^2] \geqslant \Lambda_d \frac{(\log n)^{-(d-1)}}{\ell}.$$

Summary

1. Random Forests construction
2. Several random forest models
3. Number of trees
4. Centered forests
5. Mondrian Forests
6. Median forests
7. Breiman forests
8. Conclusion

Construction of Breiman forests

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.

Oracle inequality for CART

Let $\mathcal{G}^1 = \{g(\mathbf{x}) = g_1(x_1) + \ldots, +g_d(x_d)\}$, where $g_j : \mathbb{R} \to \mathbb{R}$ are measurable, with

$$\|g\|_{TV} = \sum_{j=1}^{d} TV(g_j), \quad \text{and} \quad TV(g_j) = \int |g_j'(u)| du.$$

Theorem - Klusowski and Tian (2024)

Assume that the noise is sub-Gaussian, then $\forall K \geqslant 1$,

$$0.5\mathbb{E}\Big[(\widehat{m}_K^{CART}(\mathbf{X}) - m(\mathbf{X}))^2\Big] \leqslant \inf_{g(\cdot)\in\mathcal{G}^1} \Big\{\mathbb{E}[(g(\mathbf{X}) - m(\mathbf{X}))^2] + \frac{\|g\|_{\mathrm{TV}}^2}{K+3} + C_1 \frac{2^K \log^2(N)\log(Nd)}{N}\Big\}$$

where $C_1 > 0$ depends on $\|m\|_\infty$ and σ^2 (sub-Gaussian variance).

Oracle inequality for CART

Theorem - Klusowski and Tian (2024)

Assume that the noise is sub-Gaussian, then $\forall K \geqslant 1$,

$$0.5\mathbb{E}\left[(\widehat{m}_K^{CART}(\mathbf{X}) - m(\mathbf{X}))^2\right] \leqslant \inf_{g(\cdot)\in\mathcal{G}^1} \left\{\mathbb{E}[(g(\mathbf{X}) - m(\mathbf{X}))^2] + \frac{\|g\|_{\mathrm{TV}}^2}{K+3} + C_1 \frac{2^K \log^2(N)\log(Nd)}{N}\right\}$$

where $C_1 > 0$ depends on $\|m\|_\infty$ and σ^2 (sub-Gaussian variance).

- Consistency of generic decision trees for additive functions.
- Upper bound on the rate of consistency for CART $(\log n)^{-1}$.

Oracle inequality for CART

Theorem - Klusowski and Tian (2024)

Assume that the noise is sub-Gaussian, then $\forall K \geqslant 1$,

$$0.5\mathbb{E}\left[(\widehat{m}_K^{RF}(\mathbf{X}) - m(\mathbf{X}))^2\right] \leqslant \inf_{g(\cdot)\in\mathcal{G}^1} \left\{\mathbb{E}[(g(\mathbf{X}) - m(\mathbf{X}))^2] + \frac{d}{mtry}\frac{\|g\|_{\mathrm{TV}}^2}{K+3} + C_1\frac{2^K \log^2(N)\log(Nd)}{N}\right\}$$

where $C_1 > 0$ depends on $\|m\|_\infty$ and σ^2 (sub-Gaussian variance).

- Consistency of generic decision trees for additive functions.
- Upper bound on the rate of consistency for CART $(\log n)^{-1}$.
- Theorem 7.1: split randomization (mtry) cannot damage too much the predictive performances of RF.

Oracle inequality for CART

Theorem - Klusowski and Tian (2024)

Assume that the noise is sub-Gaussian, then $\forall K \geqslant 1$,

$$0.5\mathbb{E}\left[(-m(\mathbf{X}))^2\right] \leqslant \inf_{g(\cdot)\in\mathcal{G}^1} \left\{\mathbb{E}[(g(\mathbf{X}) - m(\mathbf{X}))^2] + \frac{\|g\|_{\mathrm{TV}}^2}{K+3} + C_1 \frac{2^K \log^2(N)\log(Nd)}{N}\right\}$$

where $C_1 > 0$ depends on $\|m\|_\infty$ and σ^2 (sub-Gaussian variance).

- Consistency of generic decision trees for additive functions.
- Upper bound on the rate of consistency for CART $(\log n)^{-1}$.
- Corollary 7.2: consistency of RF in sparse high-dimensional settings, where the dimension d is allowed to grow subexponentially with n.

Lower bound for additive models

- Input $\mathbf{X} \in [0,1]^d$
- Additive model $m(\mathbf{x}) = \sum_{j=1}^{d} m_j(x_j)$
- Noise $\sigma^2 = \mathbb{E}[\varepsilon^2 | \mathbf{X} = \mathbf{x}]$

Consider an honest tree $\hat{m}(\cdot, \mathfrak{p})$ where $\mathfrak{p}$ is the admissible partition:

$$\mathcal{A} = \{\text{data-independent partitions } \mathfrak{p} \ s.t. \forall C \in \mathfrak{p}, \mathbb{P}(\mathbf{X} \in C) \geqslant 1/n\}$$

Theorem 5.1 Tan et al. (2022)

Assume that for all $j = 1, \ldots, d$, $m_j \in C^1([0,1])$ and

$$\inf_{t \in [0,1]} |m_j'(t)| \geqslant \beta_0 > 0 \qquad \text{for all } j \in S.$$

Assume that X has a continuous density q such that $q \geqslant q_{min} > 0$. Then

$$\mathbb{E}\left[(\hat{m}(\mathbf{X}, \mathfrak{p}) - m(\mathbf{X}))^2\right] \geqslant d \left(\frac{\beta_0^2 q_{\min}}{12}\right)^{\frac{d}{d+2}} \left(\frac{\sigma^2}{4n}\right)^{\frac{2}{d+2}}.$$

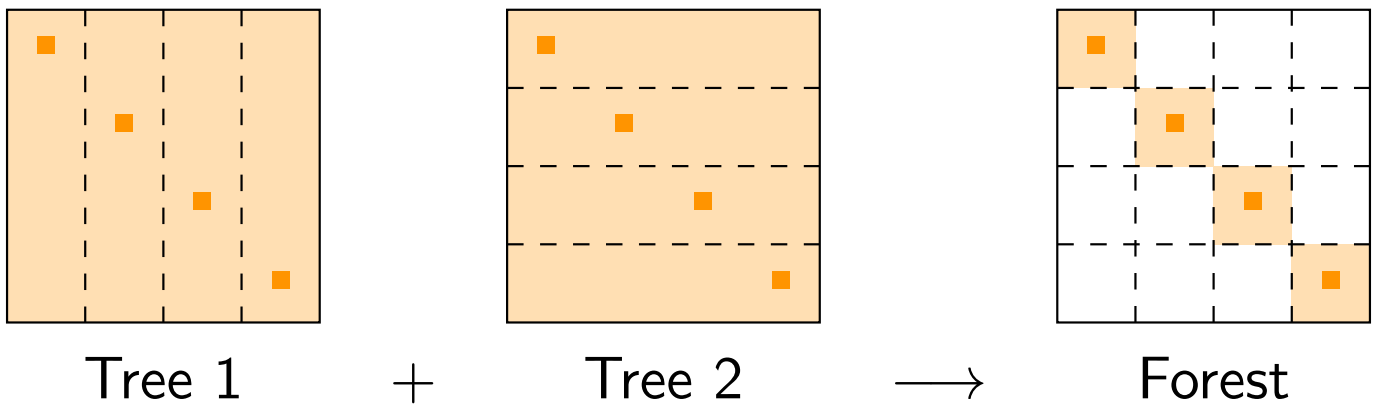
Vitesse minimax pour un modèle additif avec fonctions C^1 : $n^{-2/3}$.

Breiman RF and interpolation

- ✗ Theoretical analysis of interpolating Breiman RF consistency: out of reach for now
- ▶ Study of the interpolation zone instead!
- ▶ Partition of the RF $\equiv$ intersection of the partitions of the trees of the RF

Breiman RF and interpolation

- ✗ Theoretical analysis of interpolating Breiman RF consistency: out of reach for now
- ▶ Study of the interpolation zone instead!
- ▶ Partition of the RF $\equiv$ intersection of the partitions of the trees of the RF



Interpolation zone

Area of the space where the prediction relies on only one point of the dataset

Breiman RF: volume of the interpolation zone

Proposition Arnould et al. (2023)

Consider an infinite Breiman forest constructed without bootstrap, with `mtry` fixed to 1. Then, the volume of its interpolation zone Z_n verifies

$$\mathbb{E}\left[\mathrm{vol}(Z_n)\right] \leqslant \frac{1}{n^{d-1}}(1-2^{-n})^d$$

- The risk can be decomposed as

$$\begin{aligned}&\mathrm{Risk}(f_n(X)) - \mathrm{Risk}(f^\star)\\ =&\underbrace{\mathrm{Risk}((f_n(X)-f^\star(X))\mathbb{1}_{X\in Z_n})}_{\geqslant \sigma^2\mathrm{vol}(Z_n)} + \mathrm{Risk}((f_n(X)-f^\star(X))\mathbb{1}_{X\notin Z_n})\end{aligned}$$

- Necessary condition for consistency: $\mathrm{vol}(Z_n) \to 0$ as $n \to \infty$
- For most points of the space, more than one point are involved in the prediction of the RF.

Summary

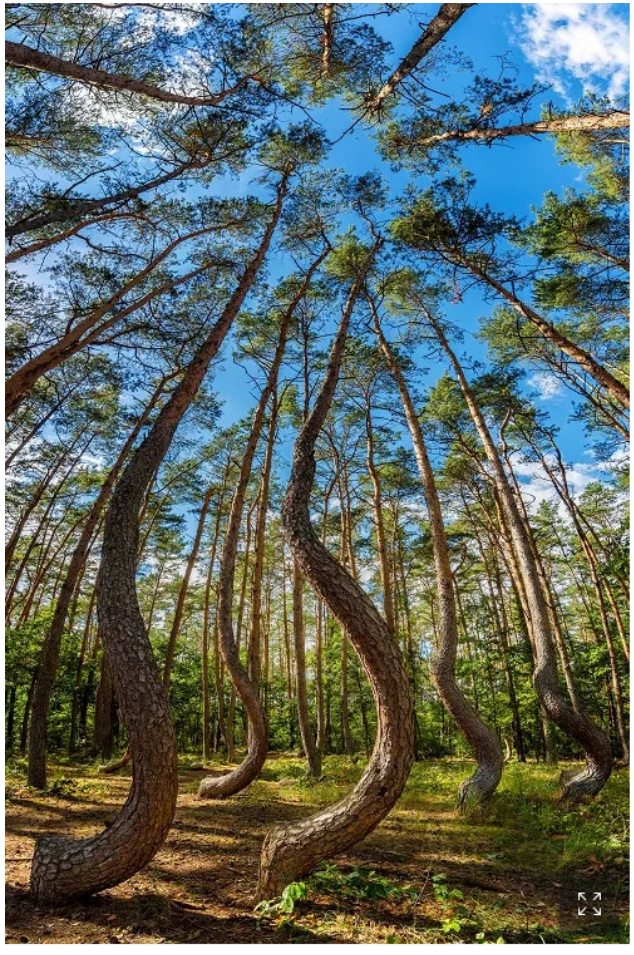
1. Random Forests construction
2. Several random forest models
3. Number of trees
4. Centered forests
5. Mondrian Forests
6. Median forests
7. Breiman forests
8. Conclusion

Summary of random forests properties

	Centred	Mondrian	Median	Breiman
Data-dependent	✗	✗	$\simeq$	✓
Fully-developed	✗	✗	✓	✓
Consistency	✓	✓	✓	✓[1]
Rate (Lipschitz)	$n^{-1/(1+d\log 2)}$	$n^{-2/(d+2)}$	$(\log n)^{-(d-1)/2}$	$(\log n)^{-1}$ (meas. funct.)
Benefits of				
Number of trees	Var. reduct.	C^2-minimax	Consistency	✗
Subsampling	NA	NA	✓	✗
Feature random.	✓	✓	✓	✗
Fully-developed	NA	NA	✗	✗

[1] Additive models

Thank you for your attention!



More information: Biau and Scornet (2016), Scornet and Hooker (2026)

References I

S. Arlot and R. Genuer. Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*, 2014.

G. Biau. Analysis of a random forests model. *JMLR*, 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. *JMLR*, 9(9), 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(1):5–32, 2001.

L. Breiman. *Consistency for a simple model of random forests*. Technical Report 670, UC Berkeley, 2004.

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.

References II

R. Duroux and E. Scornet. Impact of subsampling and tree depth on random forests. *ESAIM: Probability and Statistics*, 22:96–128, 2018.

R. Genuer. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24(3):543–562, 2012.

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

S. Gaïffas J. Mourtada and E. Scornet. Minimax optimal rates for mondrian trees and forests. *Annals of Statistics*, 2020.

J. Klusowski. Sharp analysis of a simple model for random forests. In *International Conference on Artificial Intelligence and Statistics*, pages 757–765. PMLR, 2021.

J. Klusowski and P. Tian. Large scale prediction with decision trees. *Journal of the American Statistical Association*, 119(545):525–537, 2024.

E. Scornet L. Arnould, C. Boyer. Is interpolation benign for random forest regression? In *International Conference on Artificial Intelligence and Statistics*, pages 5493–5548. PMLR, 2023.

References III

B. Lakshminarayanan, D. Roy, and Y.W. Teh. Mondrian forests: Efficient online random forests. *Advances in neural information processing systems*, 27, 2014.

X. Li, Y. Wang, S. Basu, K. Kumbier, and B. Yu. A debiased mdi feature importance measure for random forests. *NeurIPS*, 2019.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

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.

A. Rakhlin M. Belkin and A. Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd international conference on artificial intelligence and statistics*, pages 1611–1619. PMLR, 2019.

L. Mentch and G. Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *JMLR*, 2016.

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

References IV

E. Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500, 2016b.

E. Scornet. Trees, forests, and impurity-based variable importance in regression. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*. Institut Henri Poincaré, 2023.

E. Scornet and G. Hooker. Theory of random forests. *Annual Review of Statistics and Its Application*, 13(1):99–121, 2026.

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

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 2018.

S. Wager and G. Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.

A. Agarwal Y.S. Tan and B. Yu. A cautionary tale on fitting decision trees to data from additive models: generalization lower bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 9663–9685. PMLR, 2022.