The art of data science via Mondrian forests

Erwan Scornet (Ecole Polytechnique)

Joint work with

Stéphane Gaïffas (University Paris 7),

Jaouad Mourtada (Ecole Polytechnique),

IHES - January 2019

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).

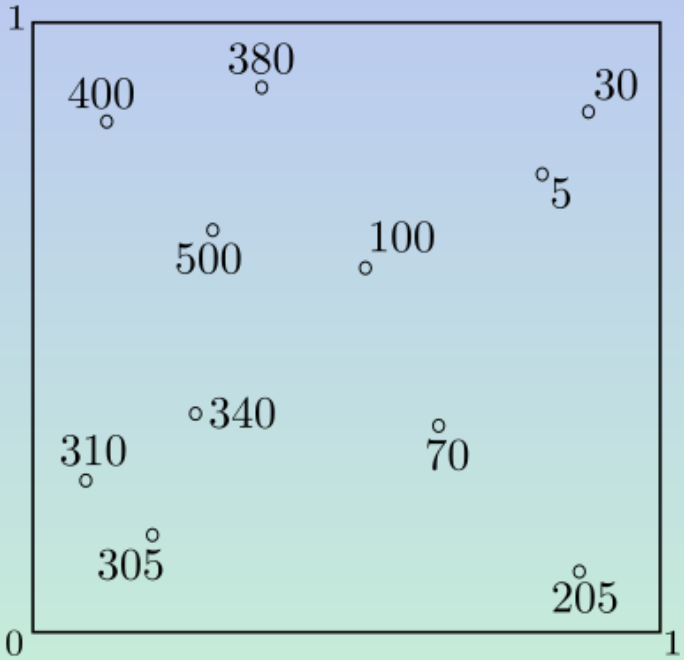
We assume that

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

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

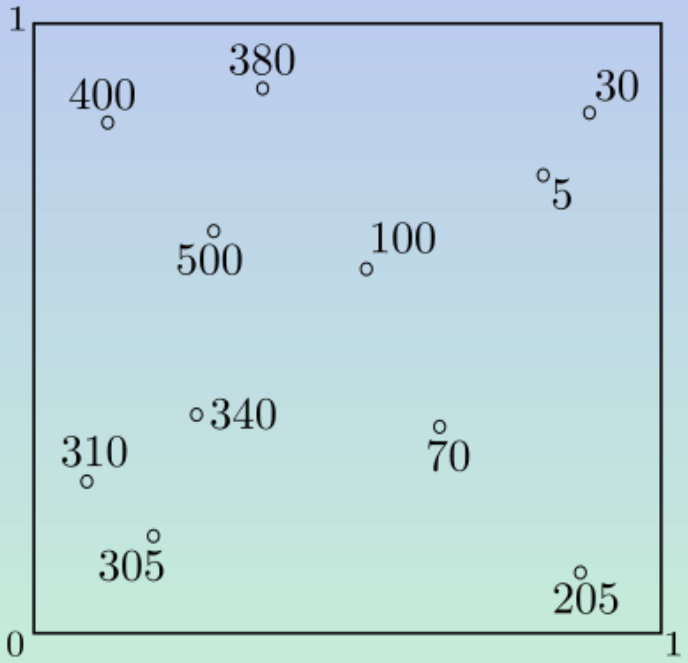
How to build a tree?

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



How to build a tree?

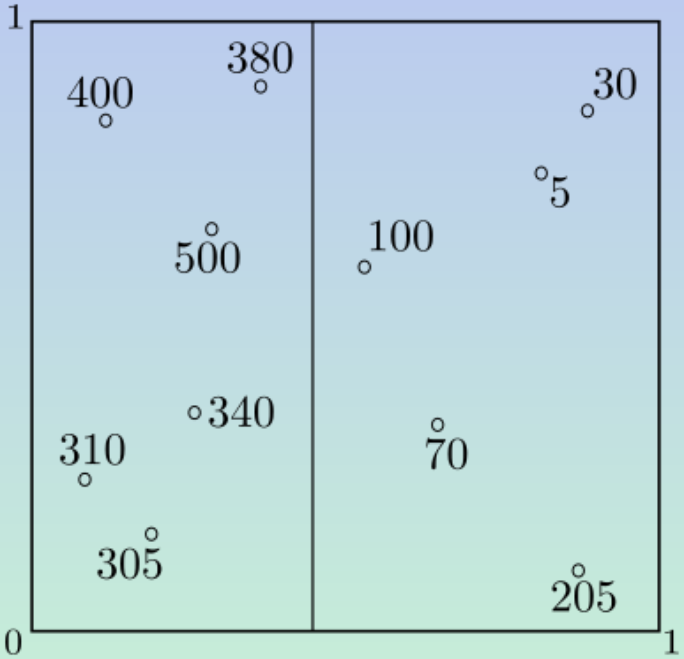
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

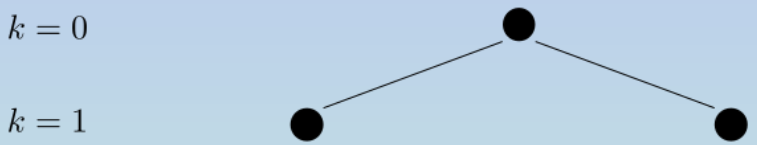


$k = 0$

How to build a tree?

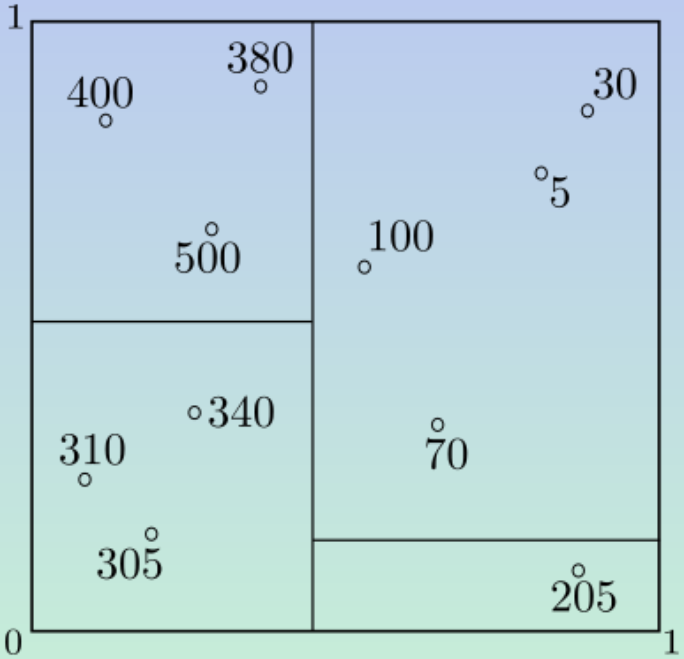
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

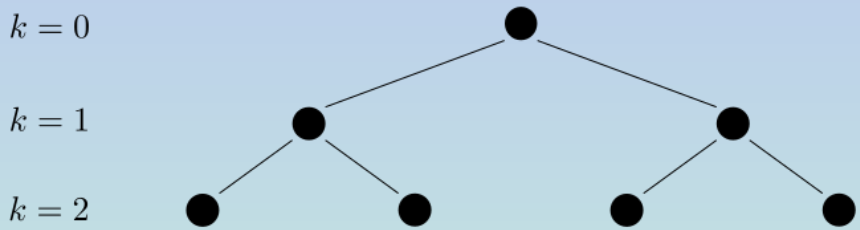




How to build a tree?

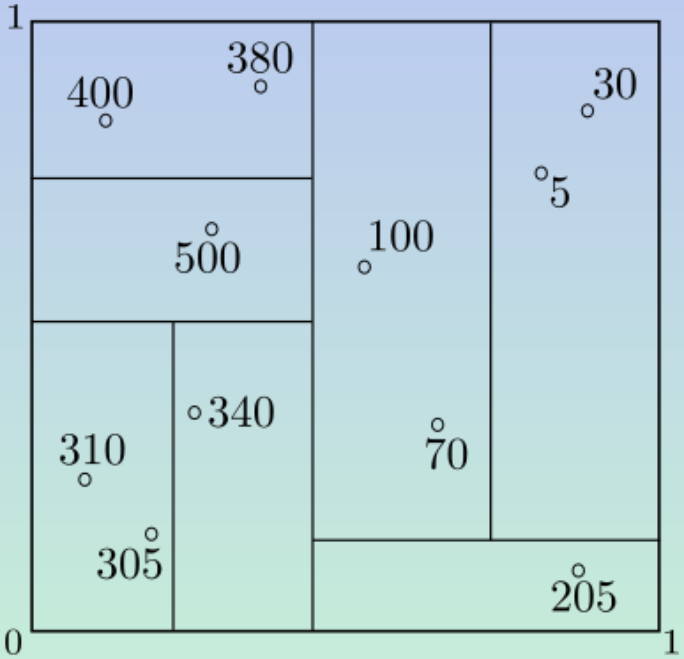
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

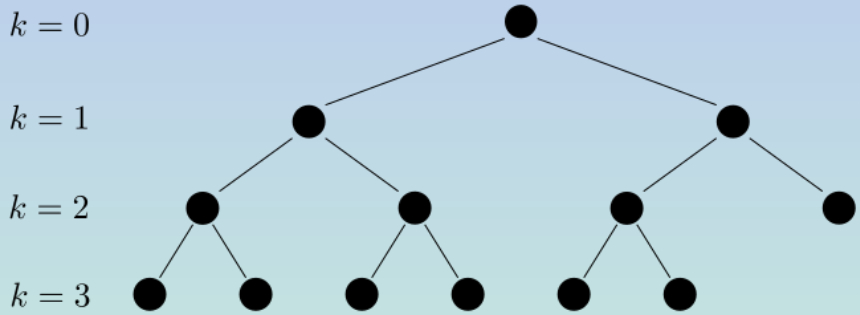




How to build a tree?

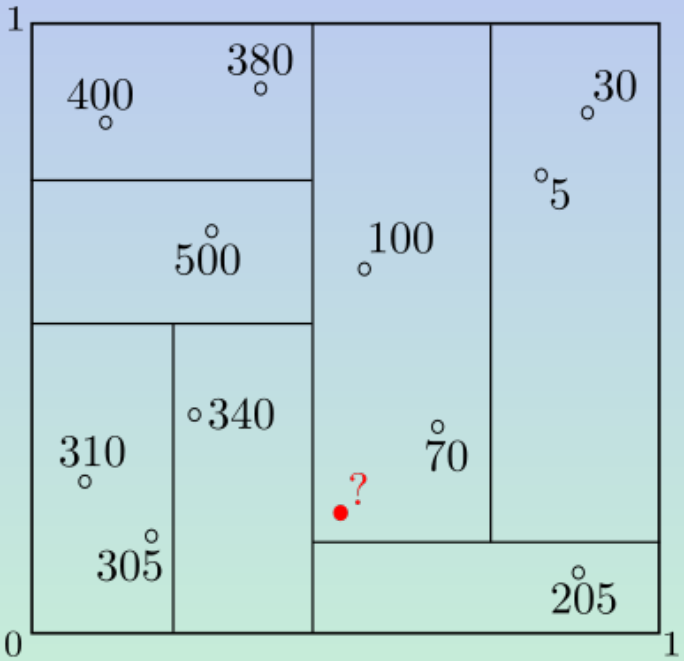
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

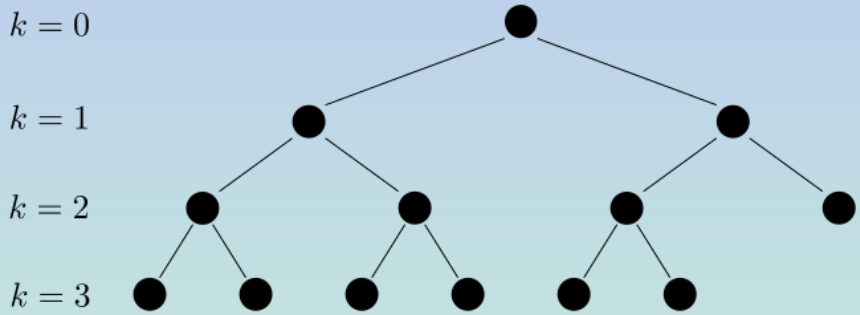




How to build a tree?

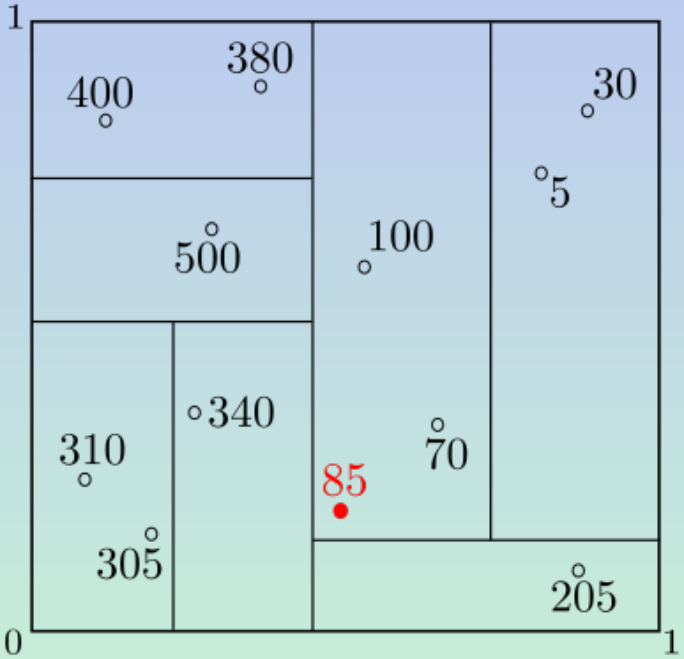
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

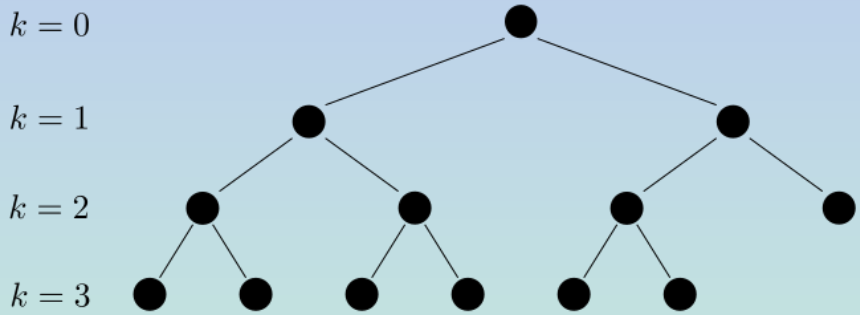




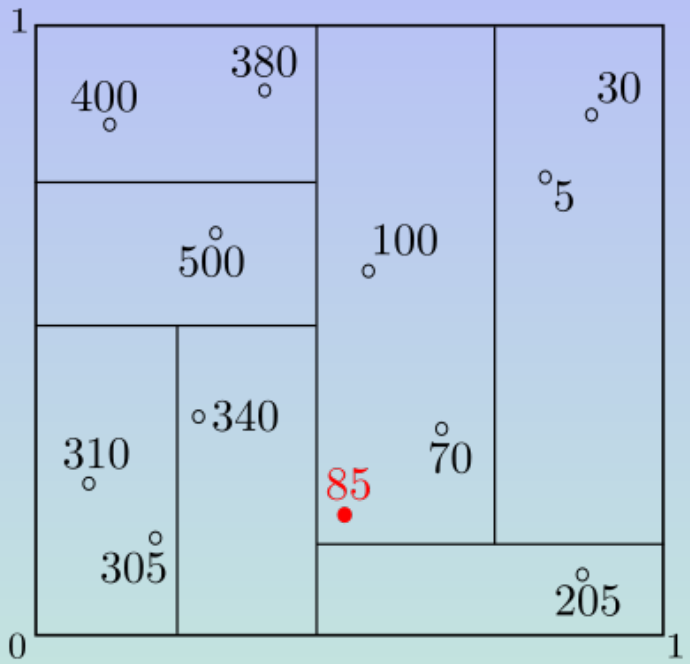
How to build a tree?

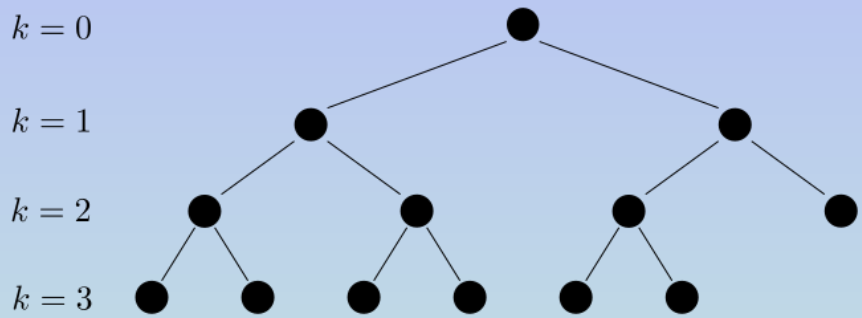
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.





How to build a tree?





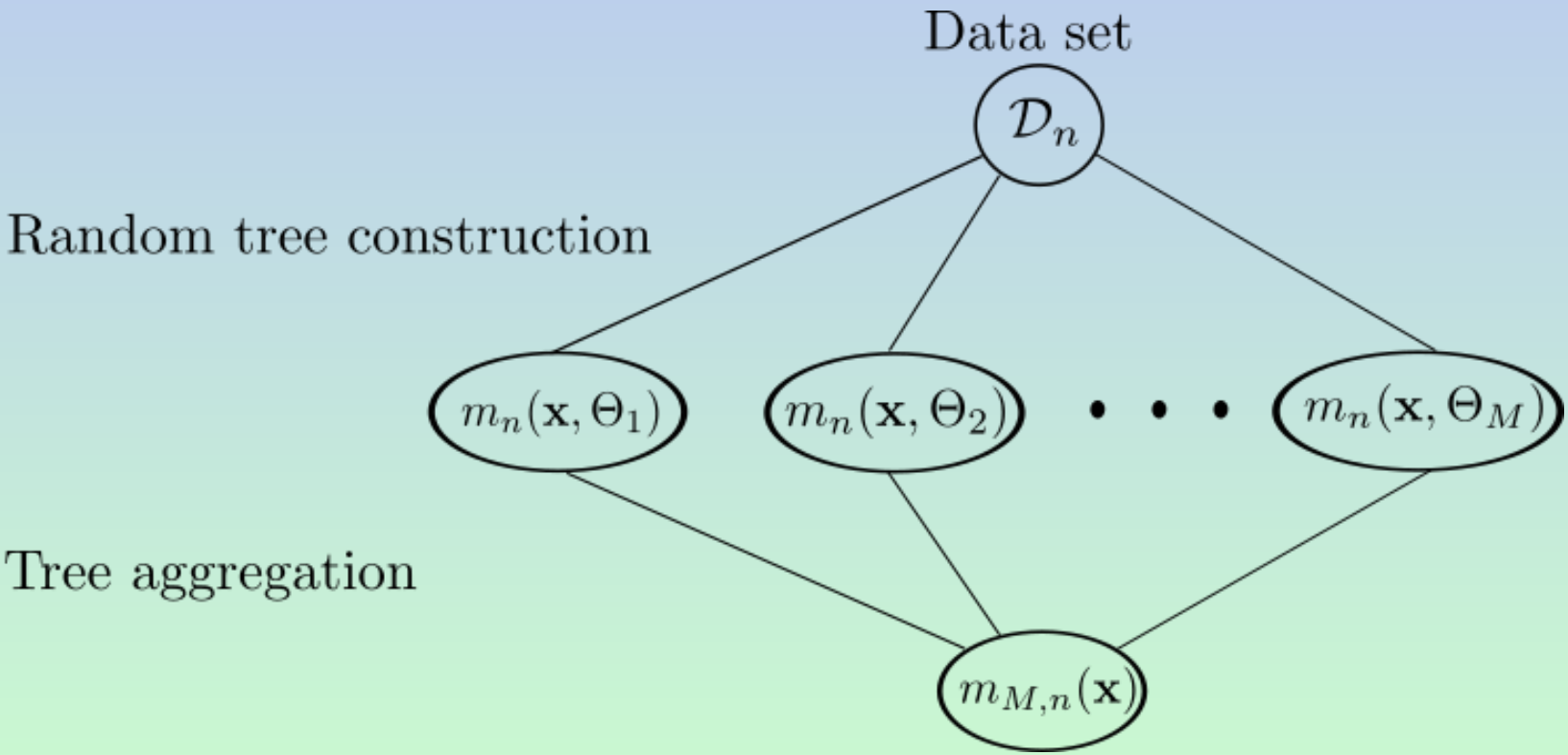
Breiman Random forests are defined by

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

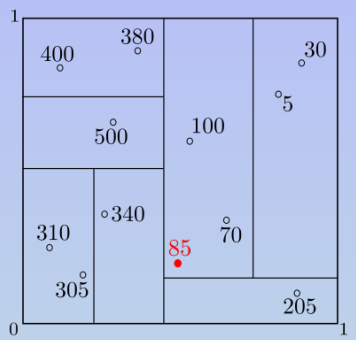
Construction of random forests

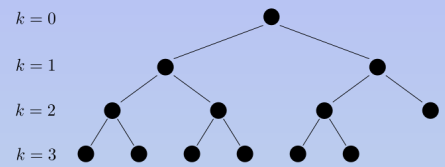
Randomness in tree construction

- Resample the data set via bootstrap;
- At each node, preselect a subset of `mtry` variables eligible for splitting.



Construction of Breiman forests





Breiman tree

- Select a_n observations with replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- 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.

Literature

- Random forests were created by Breiman [2001].
- Many theoretical results focus on simplified version on random forests, whose construction is independent of the dataset.
  [Biau et al., 2008, Biau, 2012, Genuer, 2012, Zhu et al., 2012, Arlot and Genuer, 2014].
- Analysis of more data-dependent forests:
  - Asymptotic normality of random forests [Mentch and Hooker, 2015, Wager and Athey, 2017].
  - Variable importance [Louppe et al., 2013, Kazemitabar et al., 2017].
  - 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].

Different types of forests

Centred forest		

Different types of forests

Centred forest		
Independent of X_i and Y_i		

Different types of forests

Centred forest		
Independent of X_i and Y_i		
		

Different types of forests

Centred forest		Breiman's forests
Independent of X_i and Y_i		
		

Different types of forests

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

Different types of forests

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

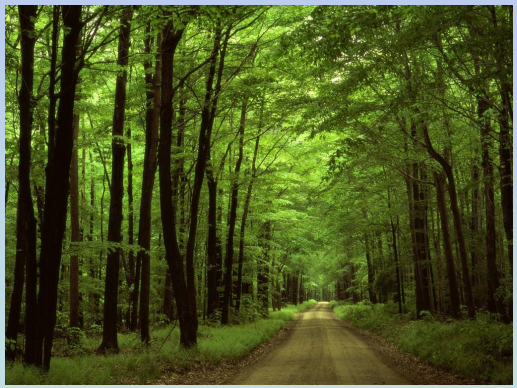
Different types of forests

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

Different types of forests

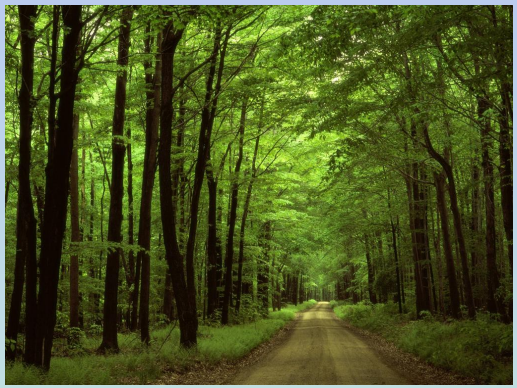
Centred forest	Median forests	Breiman's forests
Independent of X_i and Y_i	Independent of Y_i	Dependent on X_i and Y_i
		

Different types of forests

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

Different types of forests

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



Tree consistency

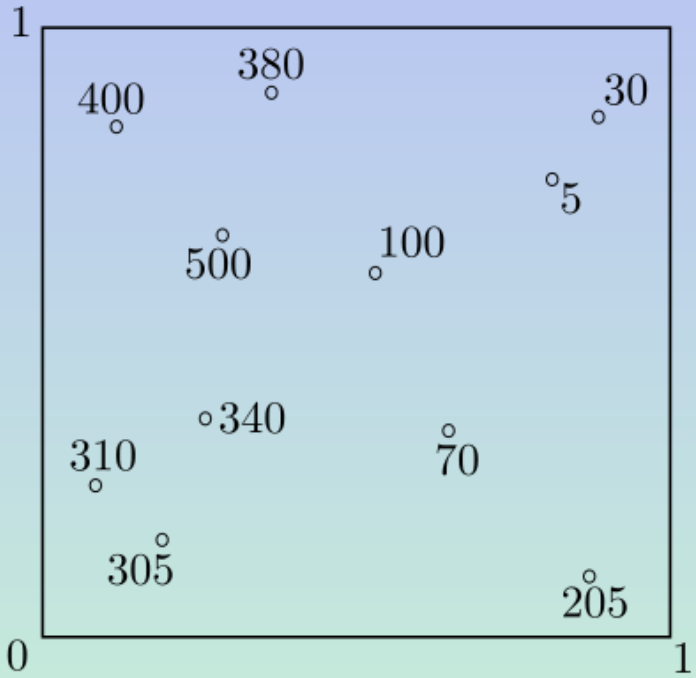
For a tree whose construction is independent of data, if

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

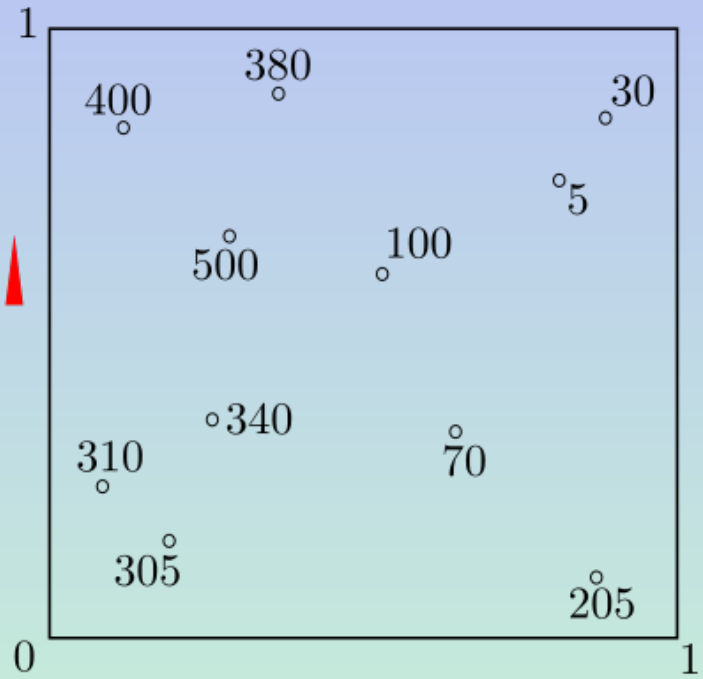
$$\lim_{n \to \infty} \mathbb{E}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 = 0.$$

Centered forests



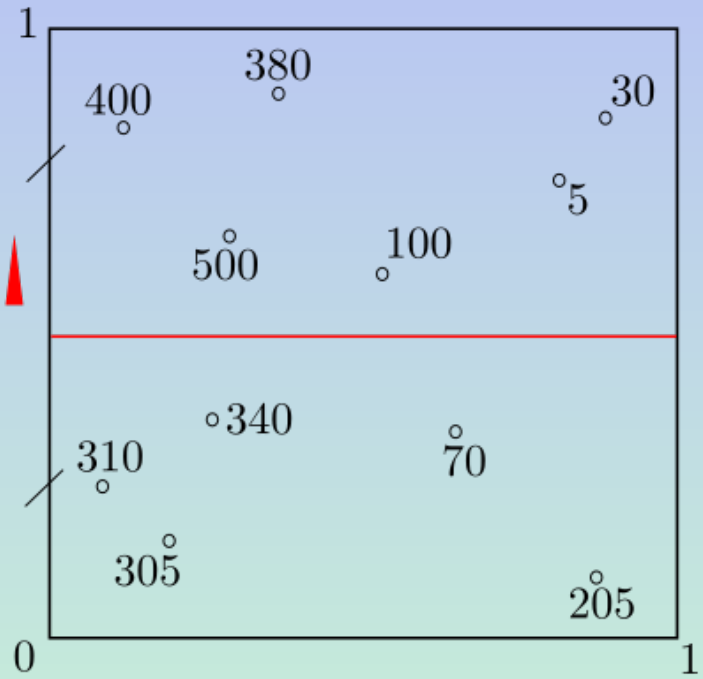
$k = 0$

Centered forests



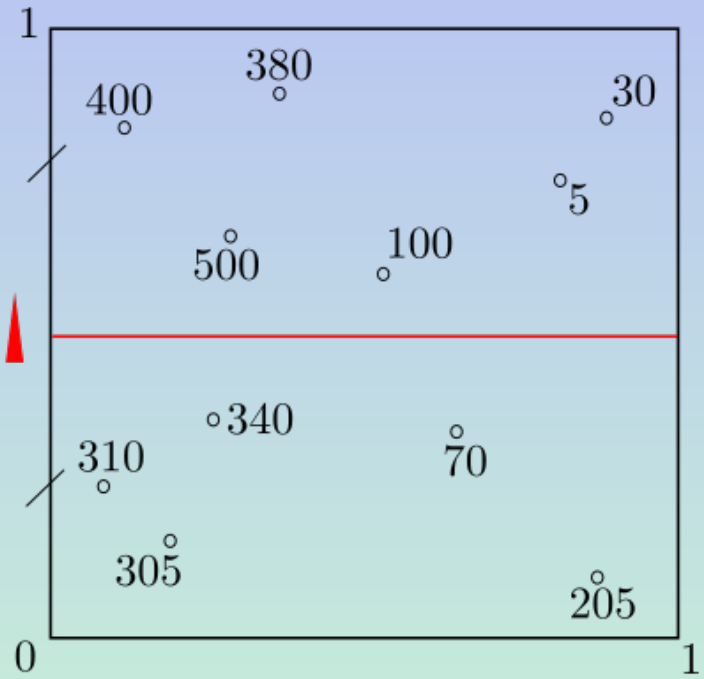
$k = 0$

Centered forests



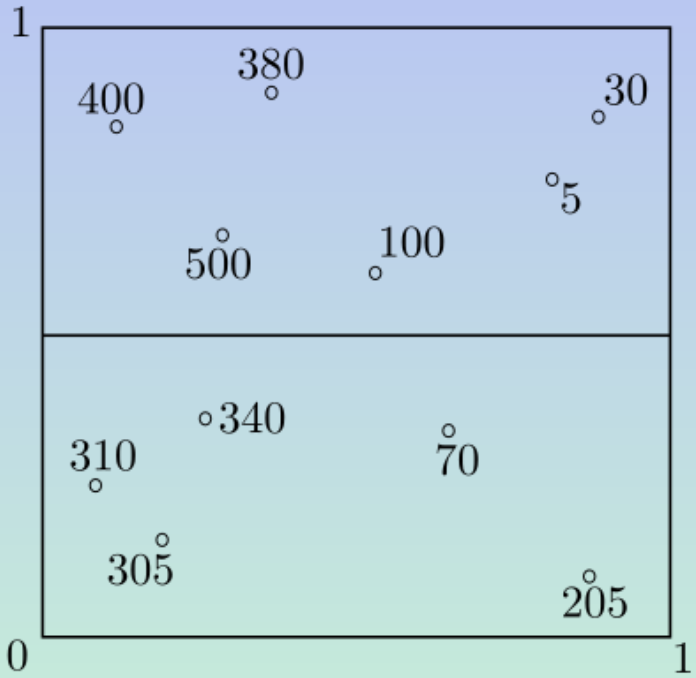
$k = 0$

Centered forests



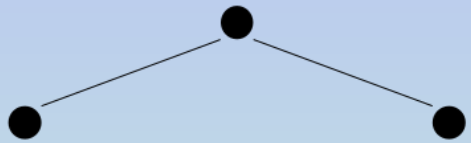
$k = 0$

Centered forests

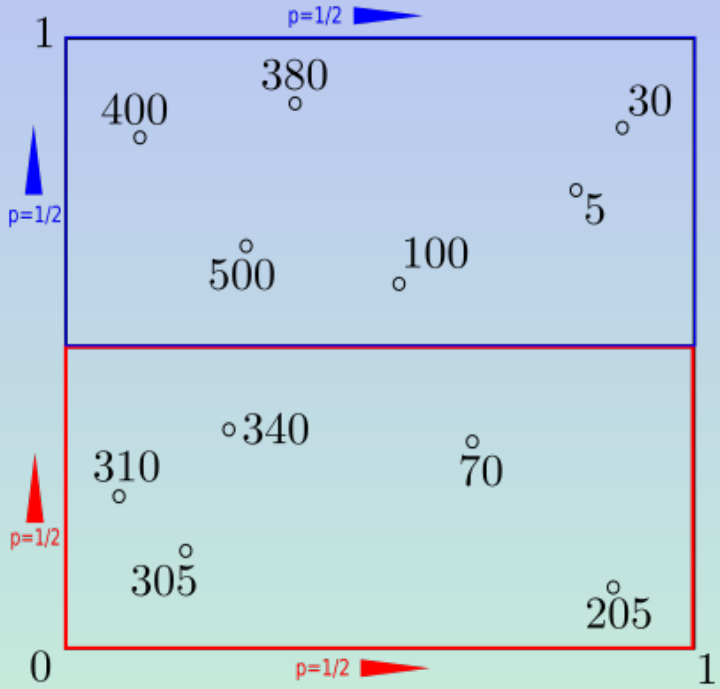


$k = 0$

$k = 1$

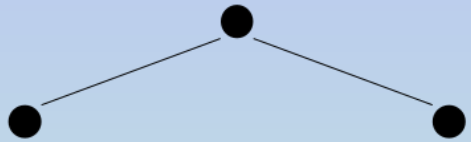


Centered forests

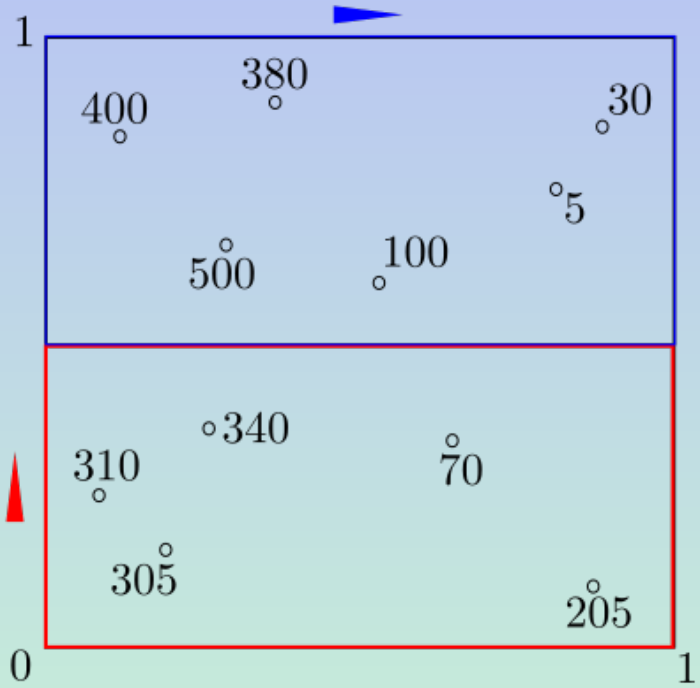


$k = 0$

$k = 1$

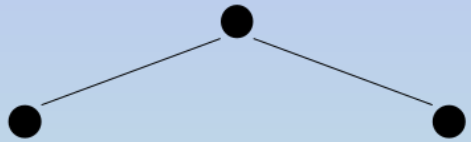


Centered forests

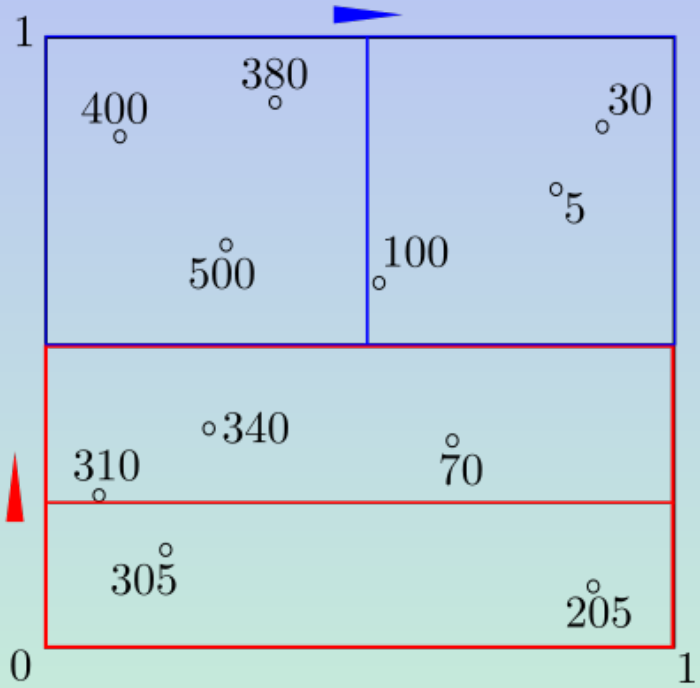


$k = 0$

$k = 1$

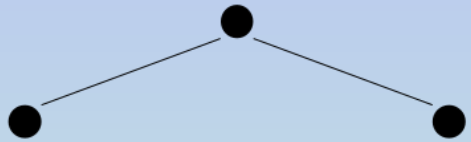


Centered forests

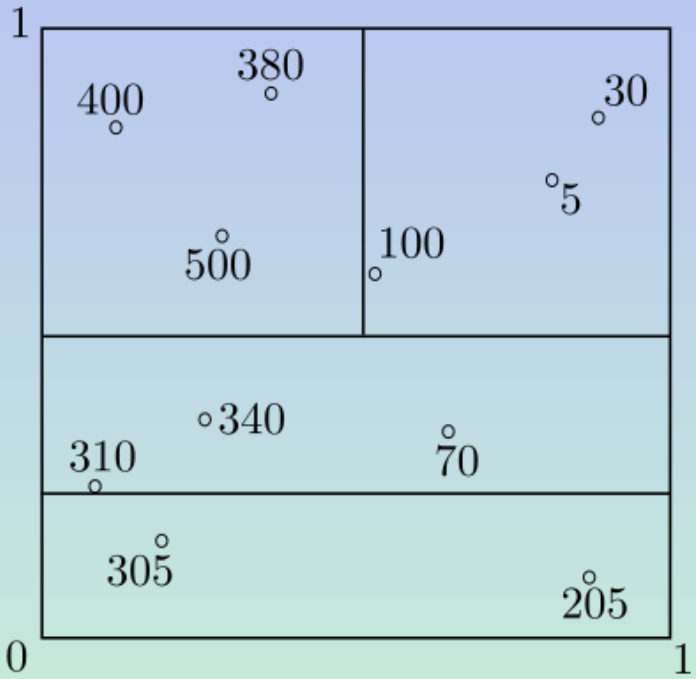


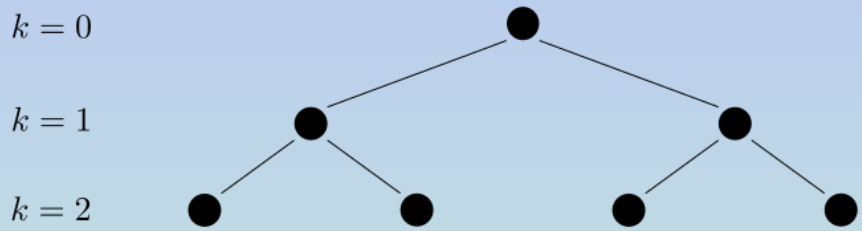
$k = 0$

$k = 1$

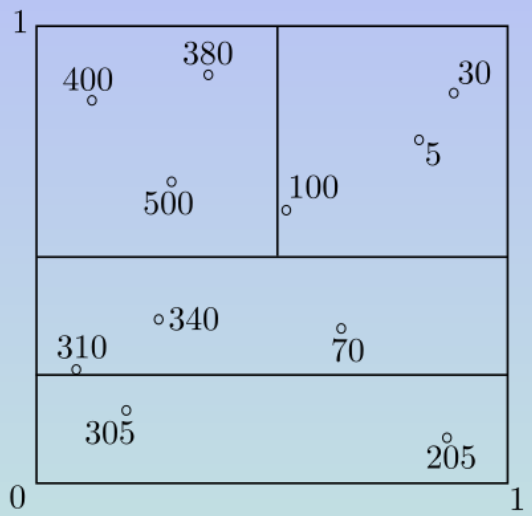


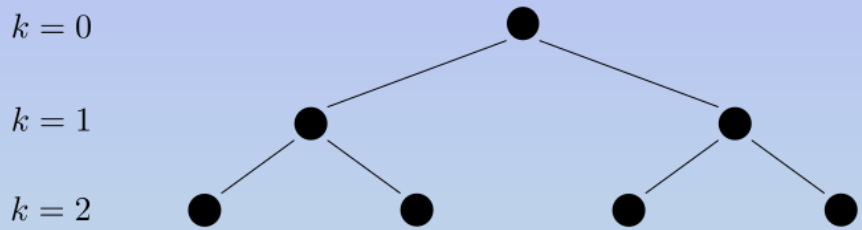
Centered forests





Centered forests

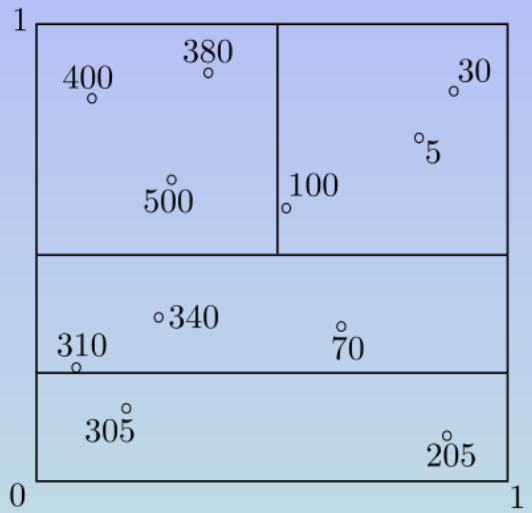


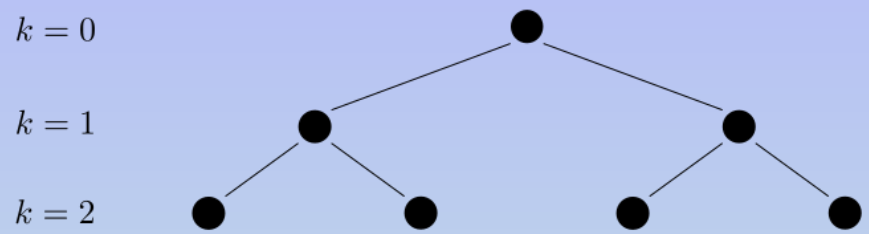


Theorem (Biau [2012])

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

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.

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

$\to$ Trees are not fully developed.

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_j.
- Stop when each cell contains exactly $\texttt{nodesize} = 1$ observation.

Consistency

Theorem [S.(2016)]

Assume that

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

where ε is a centred noise such that $\mathbb{V}[\varepsilon|\mathbf{X} = \mathbf{x}] \leq \sigma^2 < \infty$, $\mathbf{X}$ has a density on $[0,1]^d$ and m is continuous. Then, provided $a_n \to \infty$ and $a_n/n \to 0$, median forests are consistent, i.e.,

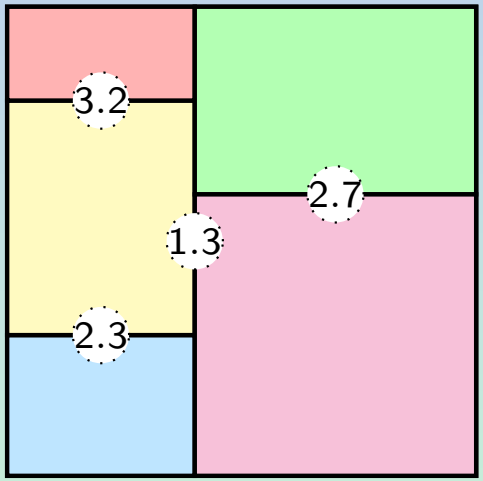
$$\lim_{n\to\infty} \mathbb{E}\left[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\right]^2 = 0.$$

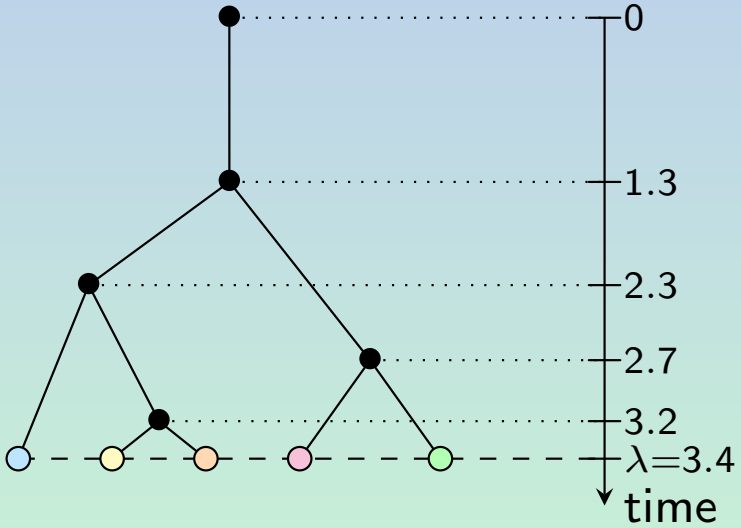
Remarks

- 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.

The Mondrian process (Roy and Teh, 2008)

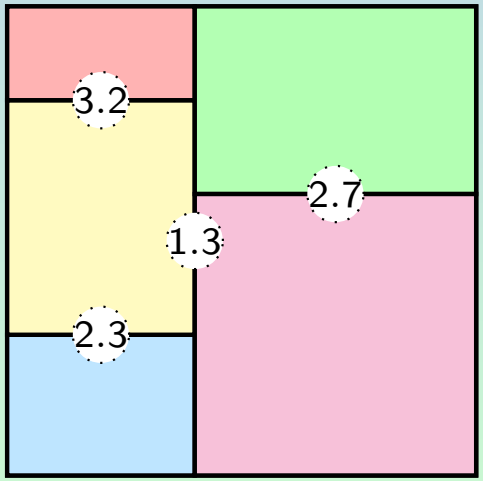
- $\mathrm{MP}(\lambda, C)$: distribution on recursive, axis-aligned partitions of $C = \prod_{j=1}^{d}[a_j, b_j] \subset \mathbf{R}^d$ (= trees).
- $\lambda > 0$ "lifetime" = complexity parameter.

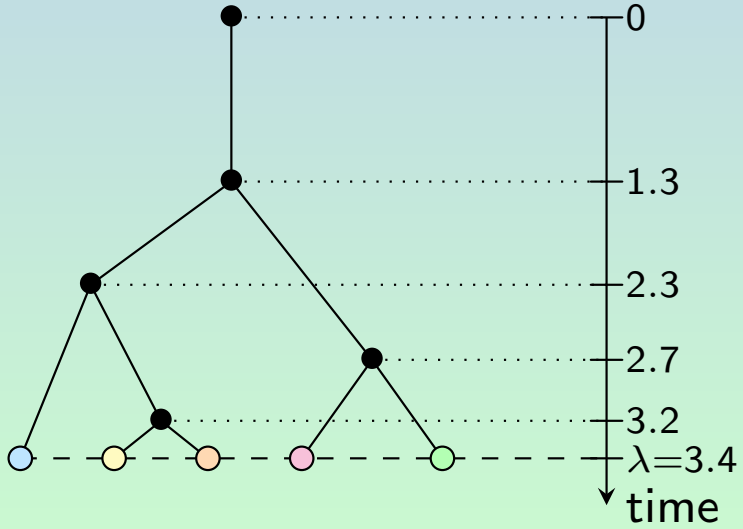




The distribution $\mathrm{MP}(\lambda, C)$

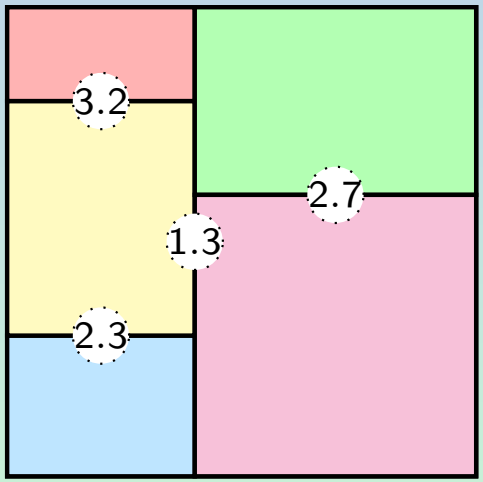
- **Start** with cell C (root), formed at time $\tau_C = 0$.
- **Sample** time till split $E \sim Exp(|C|)$ with $|C| := \sum_{j=1}^{d}(b_j - a_j)$, split coordinate $J \in \{1, \dots, d\}$ with $\mathbb{P}(J = j) = \frac{b_j - a_j}{|A|}$, and split threshold $S_J | J \sim \mathcal{U}([a_J, b_J])$.
- **If** $\tau_C + E \leq \lambda$:
  - Split C in $C_L = \{x \in C : x_J \leq S_J\}$ and $C_R = C \setminus C_L$.
  - 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).

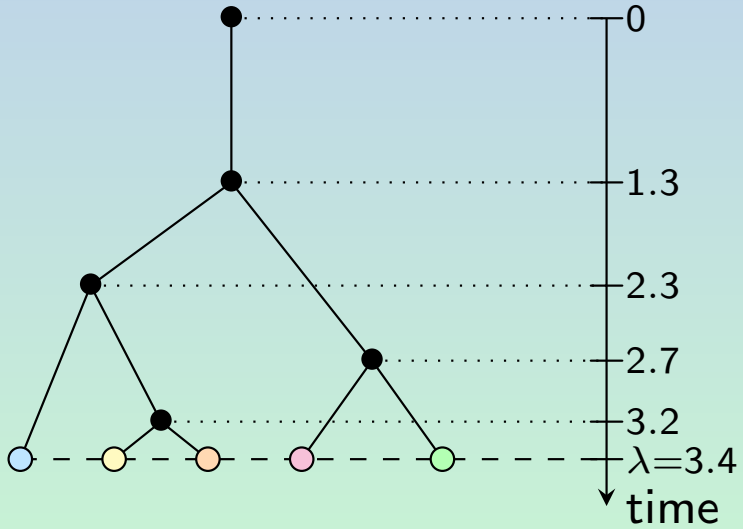




Basic properties

- Mondrian process $(\Pi_\lambda)_{\lambda \in \mathbf{R}^+} \sim \mathrm{MP}(C)$ is a Markov process.
- When $d = 1$, Mondrian partition $\Pi_\lambda \sim \mathrm{MP}(\lambda, [0, 1])$: sub-intervals whose extremities form a Poisson point process of intensity λdx.
- Fundamental restriction property: if $\Pi_\lambda \sim \mathrm{MP}(\lambda, C)$ and $C' \subseteq C$, then $\Pi_\lambda|_{C'} \sim \mathrm{MP}(\lambda, C')$.





Mondrian forests

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

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

Theoretical results

Denote $\widehat{f}_{\lambda,n}^{(M)}$ the (randomized) Mondrian forest estimator with M trees and parameter λ (n = sample size). Assume:

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

Theorem (Mourtada, Gaïffas, S.)

Assume (**H**) *and that* f^* *is* L*-Lipschitz. Then:*

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

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

$$\mathcal{R}(\widehat{f}_{\lambda,n}^{(M)}) = O(n^{-2/(d+2)}), \tag{2}$$

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

"Forest effect" : influence of the number of trees

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

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

Theorem (Mourtada, Gaïffas, S.)

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

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n}^{(M)} - f^*)^2 | X \in [\varepsilon, 1-\varepsilon]^d\big] = O\Big(\frac{1}{M\lambda^2} + \frac{1}{\lambda^4} + \frac{e^{-\lambda\varepsilon}}{\lambda^3} + \frac{(1+\lambda)^d}{n}\Big)$$

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

$$\mathbb{E}\big[(\widehat{f}_{\lambda,n}^{(M)} - f^*)^2 | X \in [\varepsilon, 1-\varepsilon]^d\big] = O(n^{-4/(d+4)})$$

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

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

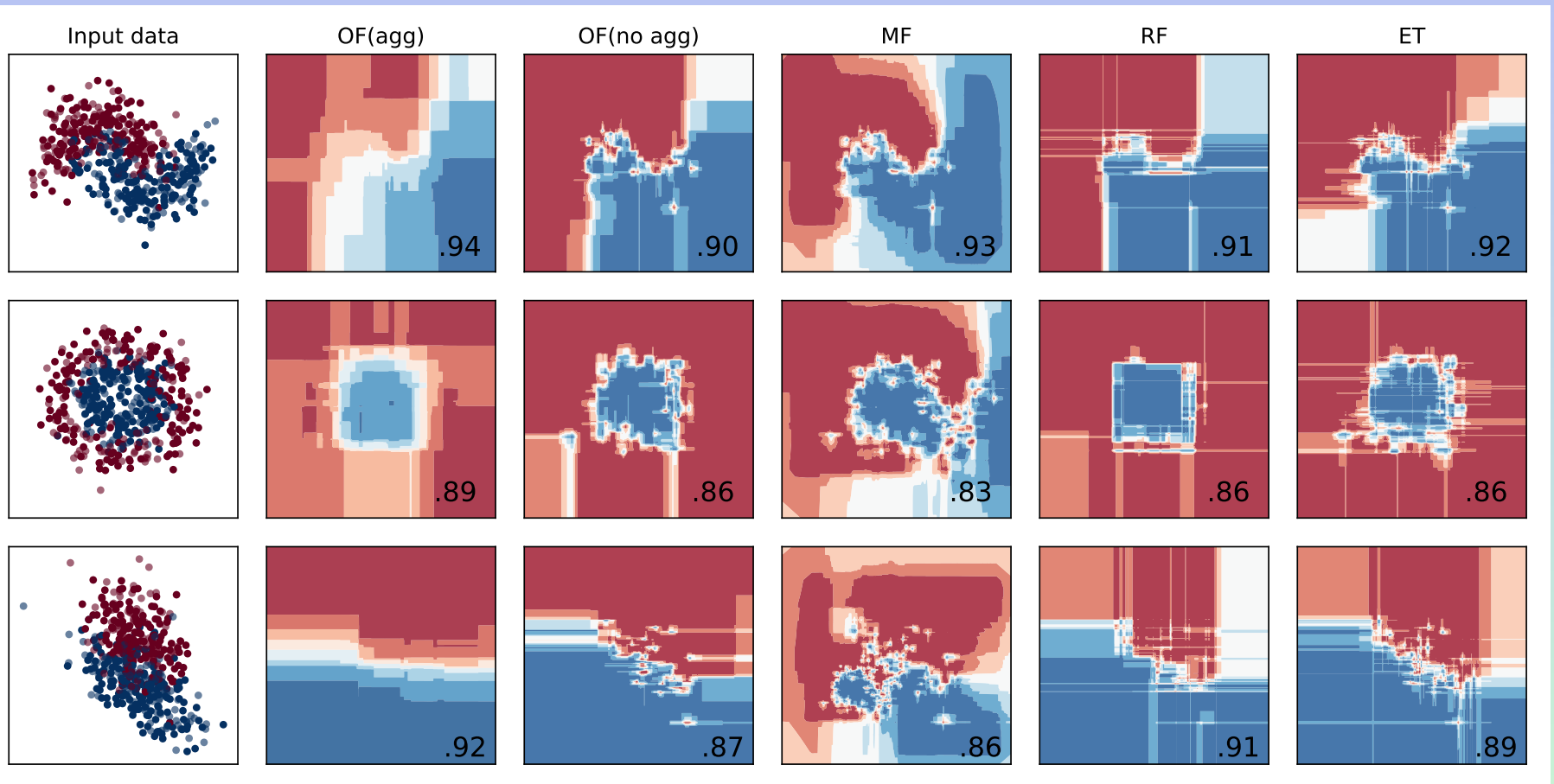
Proof ideas

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

Online implementation and adaptivity to smoothness

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

Experiments



Conclusion

- First optimal rates in arbitrary dimension under nonparametric assumptions for Random forests.
- Influence of the number of trees M: reduction of bias, improved rates for forests in arbitrary dimension.
- Aggregation over trees can be performed efficiently; gives an online algorithm which is parameter-free and competitive with optimal choice of λ ($\Rightarrow$ adaptive to regularity of f^*).
- Minimax rates for Lipschitz / $\mathscr{C}^2$ functions: the best we can hope for Purely Random forests. Further work should consider more refined variants to achieve better adaptivity (e.g. variable selection).

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.

- Mondrian forests: universally consistent. Minimax rates of consistency on both $\mathscr{C}^1$ and $\mathscr{C}^2$.

  $\rightarrow$ Minimax rates on $\mathscr{C}^2$ compared to single Mondrian Trees.

Thank you!

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

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

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

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

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

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

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

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

Jalil Kazemitabar, Arash Amini, Adam Bloniarz, and Ameet S Talwalkar. Variable importance using decision trees. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 426–435. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/6646-variable-importance-using-decision-trees.pdf`.

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.

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

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

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