Minimax rates for Mondrian trees and forests

Jaouad Mourtada Stéphane Gaïffas Erwan Scornet

ÉCOLE POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

Motivation

Random Forests (RF) classifiers, starting in early 2000s with [2]:

- Widely used in practice;
- Often achieve **state-of-the-art results**;
- But theory is **underdeveloped**: some **consistency** results, rarely **rates of convergence** [1].

Mondrian Forests is an **online** RF algorithm [3] based on the **Mondrian Process** [5], a distribution on **tree partitions** of $[0,1]^d$.

- **Computationally attractive** (can be updated easily);
- Good accuracy/complexity tradeoff;
- **Lack theoretical guarantees/analysis**;
- Depends on **complexity parameter** $\lambda \in \mathbf{R}^+$: how to **tune it**?

Our contributions

- Analysis of the **statistical properties** of Mondrian Forests:
- Universal **consistency** of the procedure;
- **Minimax nonparametric rates** for a **proper tuning** of λ, in **arbitrary dimension** d.
  First minimax optimal rates for a RF method in arbitrary dimension $d \geqslant 1$ (case $d = 1$ was done in [1]).
- **Improved rates** for large enough Mondrian **forests** over single **trees** under some conditions.

Setting

Regression (same for classification):

- Sample $\mathcal{D}_n : (X_1, Y_1), \ldots, (X_n, Y_n) \in [0,1]^d \times \mathbf{R}$, i.i.d distributed as (X, Y). μ distribution of X, $f^*(x) = \mathbb{E}[Y|X = x]$ (unknown) true **regression function**.
- **Goal**: using the sample $\mathcal{D}_n$, output a (possibly randomized) regression function $\widehat{f}_n : [0,1]^d \to \mathbf{R}$ such that as $n \to \infty$

$$R(\widehat{f}_n) = \mathbb{E}[(\widehat{f}_n(X) - f^*(X))^2] \to 0\,.$$

- **Online algorithm**: new points (X_i, Y_i) arrive sequentially, estimators are updated on the fly.

References

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

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

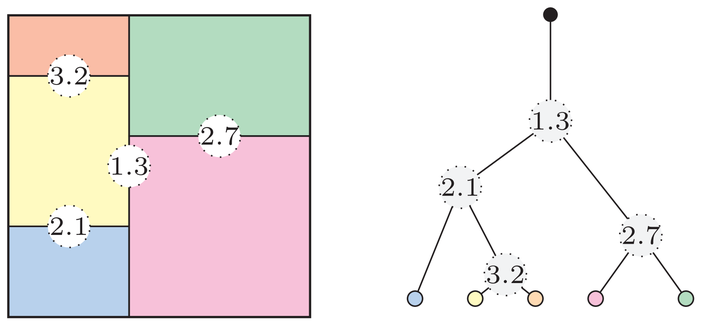
[3] B. Lakshminarayanan, D. M. Roy, and Y. W. Teh. Mondrian forests: Efficient online random forests. In *NIPS*, pages 3140–3148, 2014.

[4] J. Mourtada, S. Gaïffas, and E. Scornet. Universal consistency and minimax rates for online Mondrian forests. In *NIPS*, pages 3759–3768, 2017.

[5] D. M. Roy and Y. W. Teh. The Mondrian process. In *NIPS*, pages 1377–1384, 2009.

The Mondrian process

Distribution $\mathsf{MP}(\lambda, C)$ on **tree partitions** (kd-trees) of rectangular box $C = \prod_{j=1}^d [a_j, b_j] \subseteq \mathbf{R}^d$ [5]. $\lambda \in \mathbf{R}^+$: **lifetime parameter**, guides the **complexity** of partitions.



$\mathsf{Mondrian}(\lambda, C)$: Samples $\Pi_\lambda \sim \mathsf{MP}(\lambda, C)$

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

Mondrian Forests

- Sample independent partitions $\Pi_\lambda^{(1)}, \ldots, \Pi_\lambda^{(M)} \sim \mathsf{MP}(\lambda, [0,1]^d)$.
- On each leaf cell C of $\Pi_\lambda^{(k)}$, constant prediction = average of the Y_i s.t. $X_i \in C$;
- **Mondrian Forest estimator** $\widehat{f}_{\lambda,n}^{(M)}$: average the predictions of trees $k = 1, \ldots, M$.
- **Online** implementation [3] using the **properties of the Mondrian process**, with an increasing λ_n [4].

Universal consistency

Theorem 1. *Assume that $\mathbb{E}[Y^2] < \infty$. Let $\lambda_n \to \infty$ such that $\lambda_n^d / n \to 0$. Then, Mondrian forests with lifetime λ_n are **consistent**: $R(\widehat{f}_{\lambda_n,n}^{(M)}) \to 0$.*

Remark 1. **No rate of convergence**, but true under virtually **no assumption** on the regression function f^*.

Minimax nonparametric rates

Theorem 2. *Assume that $f^* : [0,1]^d \to \mathbf{R}$ is **Lipschitz**. Then, Mondrian Forests with lifetime $\lambda_n \asymp n^{1/(d+2)}$ satisfy*

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

*which is the **optimal convergence rate** under this hypothesis.*

Remark 2. True for **any number of trees** M, in particular for single trees ($M = 1$).

"Forest effect"

Theorem 3. *Assume that f^* **is of class** $\mathscr{C}^2$, and that X has a positive, Lipschitz density on $[0,1]^d$. Then, for every $\varepsilon > 0$, for $\lambda := \lambda_n \asymp n^{1/(d+4)}$ and $M := M_n \gtrsim n^{2/(d+4)}$,*

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

*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 3. In this variant of RF, **averaging reduces bias (approximation error)**, and **not variance**. Common intuition on other variants of RF is the opposite.

Proof ideas

Bias-variance decomposition: **approximation error** + **estimation error**. **Difficulty**: in dimension $d \geqslant 2$, tree partitions have a **recursive structure**, not straightforward to control precisely (dependence on previous splits…).

- Controlling first the combinatorial tree structure, then the geometry of the partition is **suboptimal** ($\Rightarrow$ suboptimal rates).
- Mondrian processes have appealing **restriction properties** that enable to directly control the induced partition.

Key Lemmas

Tight control of **local** and **global** properties of tree partitions.

Lemma 1. *Let $D_\lambda(x)$ be the **diameter of the cell** containing $x \in [0,1]^d$ in a Mondrian $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$. For every $\delta > 0$, we have*

$$\mathbb{P}(D_\lambda(x) \geqslant \delta) \leqslant d \left(1 + \frac{\lambda\delta}{\sqrt{d}}\right) \exp\left(-\frac{\lambda\delta}{\sqrt{d}}\right).$$

Lemma 2. *If $\Pi_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$, the **number of splits** K_λ in Π_λ satisfies: $\mathbb{E}[K_\lambda] = (1+\lambda)^d$.*

Online and adaptive MF

- **Parameter-free** algorithm competitive with the **best choice** of λ_n through efficient **aggregation** ($\Rightarrow$ **adaptive** rates).
- Relies on the **extension properties** of the Mondrian process + efficient aggregation procedure on trees to obtain an **online algorithm**.

Open problems

- Mondrian Forests satisfy the **best guarantees we can hope for "purely random forests"**: trees grown **independently of data**.
- What **advantage** can **more sophisticated variants** demonstrably provide ? E.g., selection of important variables ?

Contact information

jaouad.mourtada@polytechnique.edu