Consistency and minimax rates for online Mondrian Forests

Jaouad Mourtada Stéphane Gaïffas Erwan Scornet

Motivation

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

- Are widely used in practice;
- Often achieve **state-of-the-art results**;
- But the theory is **underdeveloped**: some **consistency** results (often under some restrictions on the joint distribution), rarely **rates of convergence** [1].

Mondrian Forests is an **online** RF algorithm [3] based on the **Mondrian Process** [4], 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

- Amending the procedure to increase $\lambda = \lambda_n$ in a streaming setting: otherwise **inconsistent**;
- Analysis of the **statistical properties** of Mondrian Forests:
- Universal **consistency** of the amended procedure;
- **Much better:** in fact, **proper tuning** of λ_n leads to **minimax nonparametric rates**, in **arbitrary dimension** d.
  First minimax optimal rates for a RF method when $d \geqslant 2$ (case $d = 1$ was done by [1]).

Setting

Classification (same for regression):

- Samples $\mathcal{D}_n : (X_1, Y_1), \dots, (X_n, Y_n) \in [0,1]^d \times \{0,1\}$, i.i.d., same distribution as (X, Y). μ distribution of X, $\eta(x) = \mathbb{P}(Y = 1 \,|\, X = x)$ conditional class probability.
- **Goal**: using the samples $\mathcal{D}_n$, output a (possibly randomized) classification rule $g_n : [0,1]^d \to \{0,1\}$ such that as $n \to \infty$

$$L(g_n) := \mathbb{P}(g_n(X) = Y) \to L^* := \inf_{g:[0,1]^d \to \{0,1\}} \mathbb{P}(g(X) \neq Y)$$

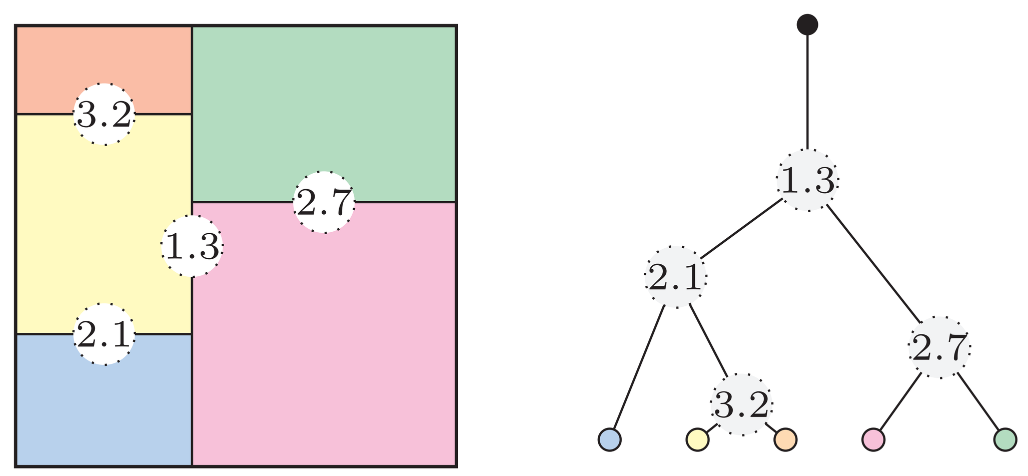
- **Online algorithm**: new points (X_t, Y_t) arrive sequentially, classifiers are updated on the fly.

Contact information

Mail: `jaouad.mourtada@polytechnique.edu`

The Mondrian process

A distribution $\mathsf{MP}(\lambda, C)$ on **tree partitions** (kd-trees) of the rectangular box $C \subseteq \mathbf{R}^d$ [4]. $\lambda \in \mathbf{R}^+$ is the **lifetime parameter** which guides the **complexity** of the partitions.



`Mondrian`(λ, C): Samples $M_\lambda \sim \mathsf{MP}(\lambda, C)$

Start with the root cell C, formed at time $\tau_C = 0$.
for $A = \prod_{j=1}^d [a_j, b_j]$ a leaf of current partition formed at τ_A **do**
  Sample $E_A \sim \mathsf{Exp}(|A|)$, where $|A| := \sum_{j=1}^d (b_j - a_j)$.
  if $\tau_A + E_A \leqslant \lambda$ **then**
    Draw a **split dimension** $J \in \{1, \dots, d\}$, with $\mathbb{P}(J = j) = (b_j - a_j)/|A|$, and a **split threshold** $s_J \sim \mathcal{U}([a_J, b_J])$
    Split A at (J, s_J) with children A_L, A_R formed at $\tau_A + E_A$.
    Add A_L, A_R to the leaves of the partition.
  end if
  Remove A from the remaining leaves.
end for

Mondrian Forests

At step $t \geqslant 1$, given $\mathcal{D}_t$, we have K independent randomized decision trees $g_t^k(\cdot)$, $1 \leqslant k \leqslant K$, $1 \leqslant t \leqslant n$. As new sample point (X_{t+1}, Y_{t+1}) arrives:

- Efficiently update the **structure** of the trees, using the **properties of the Mondrian process**. Update in space domain (original MF) *vs* time domain (ours). Can be **combined**.

Original MF [3]	Our **modified** MF
upd. **data range**: $C_{t+1} \supset C_t$	upd. **lifetime**: $\lambda_{t+1} > \lambda_t$
$\mathsf{MP}(\lambda, C_t) \to \mathsf{MP}(\lambda, C_{t+1})$	$\mathsf{MP}(\lambda_t, C) \to \mathsf{MP}(\lambda_{t+1}, C)$

- **Update** leaf labels given (X_t, Y_t).

Universal consistency

Theorem 1 (Consistency). *Assume that $\lambda_n \to \infty$ and $\lambda_n^d / n \to 0$. Then, under **no restriction** on the distribution of (X, Y), Mondrian Forests with lifetimes sequence (λ_n) are **consistent**: $L(g_n) \to L^*$.*

Minimax nonparametric rates

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

$$L(g_n) - L^* = o(n^{-1/(d+2)})$$

*which is the **optimal convergence rate** under this hypothesis [5].*

Proof ideas

Bias-variance decomposition for forests [1]. **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 $M_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$. 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 $M_\lambda \sim \mathsf{MP}(\lambda, [0,1]^d)$, the **number of splits** K_λ in M_λ satisfies: $\mathbb{E}[K_\lambda] \leqslant \big(e(1+\lambda)\big)^d$.*

Subsequent work

- **"Forest effect"** [1]: in practice, **Forests outperform single trees**. Can be explained theoretically: if η is **smooth**, Forests exhibit **improved rates** (by smoothing predictions).
- **Parameter-free** algorithm competitive with the **best choice** of λ_n through efficient **aggregation** ($\Rightarrow$ **adaptive** rates).

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

[5] Y. Yang. Minimax nonparametric classification. I. Rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, 1999.