Asymptotic Study of Ensemble Methods for Imbalanced Classification

M. Mayala, **E. Scornet**, C. Tillier, O. Wintenberger

Vannes, October 2025

Outline

1. Random Forests construction
2. U-statistics and link with RF
3. Asymptotic analysis of Infinite Centered RF
4. Numerical experiments

Summary

Statistical framework

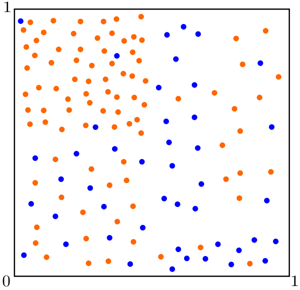
Consider $Z_i := (X_i, Y_i)$ i.i.d. copies of the pair (X, Y)

- **Input variable** $X \in \mathcal{X} = [0,1]^d$
- **Output variable** $Y \in \{0,1\}$.
- **Regression function**: $\mu(\mathrm{x}) = \mathbb{P}\left(Y = 1 | \mathrm{X} = \mathrm{x}\right)$.

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

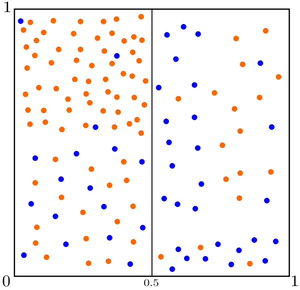
- Non-parametric method
- Based on bagging and random feature selections
- Aggregate the predictions of M trees

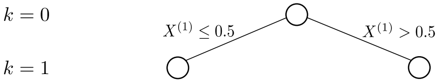
Construction of Decision trees - classification



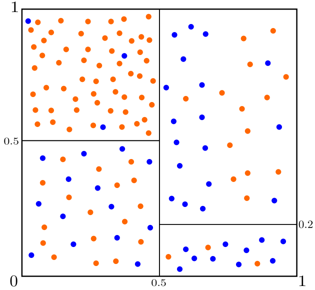
$k = 0$

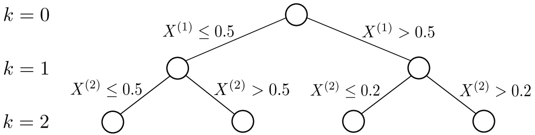
Construction of Decision trees - classification



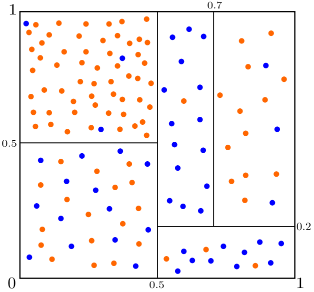


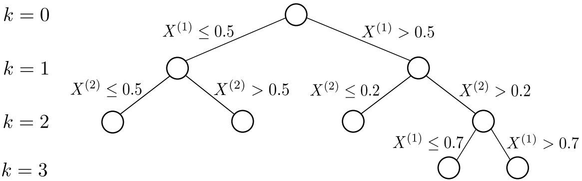
Construction of Decision trees - classification



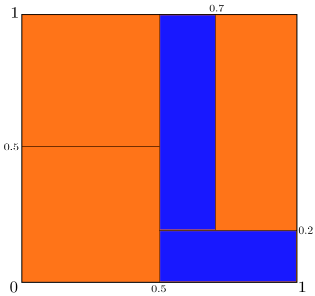


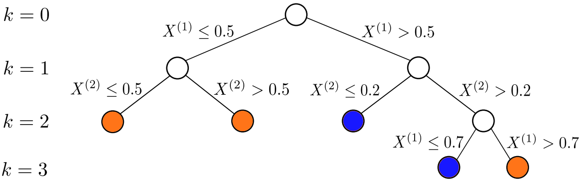
Construction of Decision trees - classification



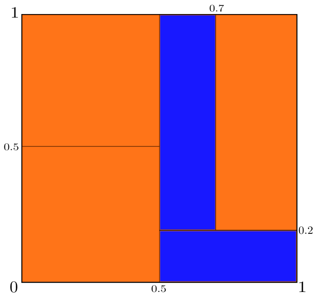


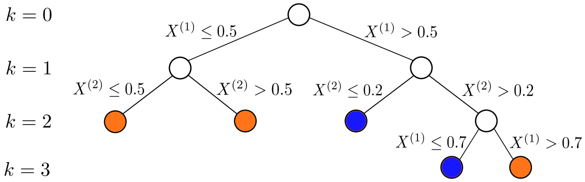
Construction of Decision trees - classification





Construction of Decision trees - classification





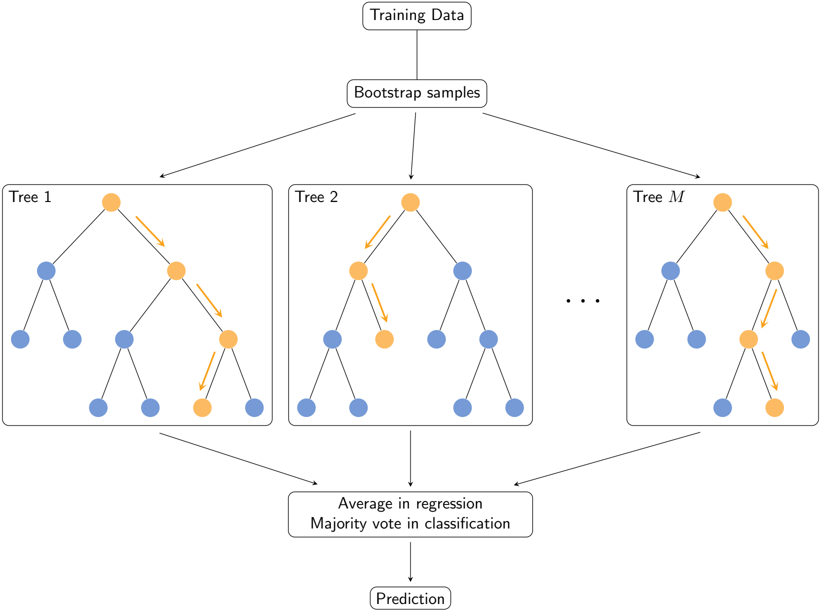
Decrease in impurity for a split (j, s):

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

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

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

A classical random forest



Toward infinite forests

Let $Z_S := (Z_{i_1}, \dots, Z_{i_s})$ be a training subsample of size s.

- The tree prediction at a point $x \in \mathcal{X}$ is

$$T^s(x; \Theta; Z_S) := \frac{\sum_{i \in S} Y_i \mathbb{1}\{X_i \in L_\Theta(x)\}}{N_{L_\Theta(x)}(X_S)},$$

  - $L_\Theta(x)$ is the leaf of the tree built with randomness Θ
  - $N_{L_\Theta(x)}(X_S) := \sum_{i \in S} \mathbb{1}\{X_i \in L_\Theta(x)\}$, number of observations

- Finite forest with M trees

$$\widehat{\mu}_{M,s}(x; Z_n) := \frac{1}{M} \sum_{m=1}^{M} T^s(x; \Theta_m; Z_{S_m}).$$

- Infinite forest

$$\widehat{\mu}_{M,s}(x; Z_n) \underset{M \to \infty}{\longrightarrow} \underbrace{\mathbb{E}[T^s(x; \Theta; Z_{\mathbb{S}}) | Z_{I_n}]}_{\widehat{\mu}^s(x)}$$

Theory of RF (Breiman, 2001): literature review

- **Simplified RF versions**, whose construction is **independent of the dataset**.
  (Biau et al.; Biau, 2012; Genuer, 2012; Arlot and Genuer, 2014; Scornet, 2016; 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.; Biau, 2012; Genuer, 2012; Arlot and Genuer, 2014; Scornet, 2016; 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.; Biau, 2012; Genuer, 2012; Arlot and Genuer, 2014; Scornet, 2016; 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, 2025)

Summary

Subsampling Infinite Random Forests (IRF)

Subsampling IRF

$$\widehat{\mu}^s(\mathrm{x}) := \binom{n}{s}^{-1} \sum_{S \subset \{1,\ldots,n\}, |S|=s} \mathbb{E}[T^s(\mathrm{x}; \Theta; \mathrm{Z}_S) \mid \mathrm{Z}_S].$$

[1] (Lee, 2019)

Subsampling Infinite Random Forests (IRF)

Subsampling IRF

$$\widehat{\mu}^s(\mathrm{x}) := \binom{n}{s}^{-1} \sum_{S \subset \{1,\dots,n\}, |S|=s} \mathbb{E}[T^s(\mathrm{x}; \Theta; \mathrm{Z}_S) \mid \mathrm{Z}_S].$$

Connection to U-statistics[1]

Assume we are given i.i.d. $Z_1, \dots, Z_n$ and we want to estimate $\mu = \mathbb{E}[h(Z_1, \dots, Z_s)]$. Then the unbiased estimator with minimal variance is the U-statistics defined as

$$U_n = \frac{1}{\binom{n}{s}} \sum_i h(Z_{i_1}, \dots, Z_{i_s}).$$

[1] (Lee, 2019)

$$\widehat{\mu}^s(\mathrm{x}) = \underbrace{\mathbb{E}[T^s(\mathrm{x};\Theta;\mathrm{Z}_S)] + \frac{s}{n}\sum_{i=1}^{n} T_1^s(\mathrm{x};Z_i)} + \sum_{r=2}^{s}\binom{s}{r}\widehat{\mu}^s_{n,r}(\mathrm{x}).$$

- First two terms: **Hàjek projection**[2] $\mathring{\widehat{\mu}}(\mathrm{x})$
- $T_1^s = \mathbb{E}[T^s(\mathrm{x};\Theta;\mathrm{Z}_S) \mid Z_1]$
- $\widehat{\mu}^s_{n,r}(\mathrm{x})$ **uncorrelated** terms

[2] der Vaart (2000)
[3] https://www.stat.berkeley.edu/~pitman/s205f02/lecture10.pdf

Technical tools - Hoeffding decomposition

$$\widehat{\mu}^s(\mathrm{x}) = \underbrace{\mathbb{E}[T^s(\mathrm{x};\Theta;\mathrm{Z}_S)] + \frac{s}{n}\sum_{i=1}^{n} T_1^s(\mathrm{x};Z_i)} + \sum_{r=2}^{s}\binom{s}{r}\widehat{\mu}^s_{n,r}(\mathrm{x}).$$

- First two terms: **Hàjek projection**[2] $\mathring{\widehat{\mu}}(\mathrm{x})$
- $T_1^s = \mathbb{E}[T^s(\mathrm{x};\Theta;\mathrm{Z}_S) \mid Z_1]$
- $\widehat{\mu}^s_{n,r}(\mathrm{x})$ **uncorrelated** terms

Sketch of proof

- Prove that $\mathrm{Var}[\widehat{\mu}^s(\mathrm{x})] \sim \mathrm{Var}[\mathring{\widehat{\mu}}(\mathrm{x})]$.
- CLT for $\mathring{\widehat{\mu}}(\mathrm{x}) \Rightarrow$ CLT for $\widehat{\mu}^s(\mathrm{x})$
- Use Lindeberg condition for triangular arrays to obtain a CLT on $\mathring{\widehat{\mu}}(\mathrm{x})$ [3]

[2] der Vaart (2000)
[3] https://www.stat.berkeley.edu/~pitman/s205f02/lecture10.pdf

CLT for subsampling IRF

(H1) The individual trees $T^s(\mathrm{x};\Theta;\mathrm{Z}_S)$ are bounded a.s. and satisfy $nV_1^s \to \infty$, as $n \to \infty$, where

$$V_1^s = \mathrm{Var}[\mathbb{E}[T^s(\mathrm{x};\Theta;\mathrm{Z}_S) \mid Z_1]].$$

Theorem 1 (Mayala, Wintenberger, Tillier and Dombry '24)

Let $\widehat{\mu}^s(\mathrm{x})$ be the subsampling IRF at point $\mathrm{x} \in \mathcal{X}$. Under **(H1)**,

$$\sqrt{\frac{n}{s^2 V_1^s}}(\widehat{\mu}^s(\mathrm{x}) - \mathbb{E}[\widehat{\mu}^s(\mathrm{x})]) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)\,.$$

CLT for subsampling IRF

(H1) The individual trees $T^s(\mathrm{x}; \Theta; Z_S)$ are bounded a.s. and satisfy $nV_1^s \to \infty$, as $n \to \infty$, where

$$V_1^s = \mathrm{Var}[\mathbb{E}[T^s(\mathrm{x}; \Theta; Z_S) \mid Z_1]].$$

(H2)

$$\sqrt{\frac{n}{s^2 V_1^s}} \left|\mathbb{E}[\widehat{\mu}^s(\mathrm{x})] - \mu(\mathrm{x})\right| \to 0.$$

Theorem 1 (Mayala, Wintenberger, Tillier and Dombry '24)

Let $\widehat{\mu}^s(\mathrm{x})$ be the subsampling IRF at point $\mathrm{x} \in \mathcal{X}$. Under **(H1)** and **(H2)**,

$$\sqrt{\frac{n}{s^2 V_1^s}} (\widehat{\mu}^s(\mathrm{x}) - \mu(x)) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1).$$

Literature review on CLT for RF

First CLT on random forests by Mentch and Hooker (2016) with

- Different asymptotics in the number of trees
- Valid for any subsample size $s = o(\sqrt{n})$
- But require $\lim_{n,s\to\infty} V_1^s \neq 0$, where

$$V_1^s = \mathrm{Var}[\mathbb{E}[T^s(\mathrm{x}; \Theta; \mathrm{Z}_S) \mid Z_1]].$$

- Not true for fully-grown trees used in RF

Extension to subsampling with replacement (and variance estimation) by Zhou et al. (2021).

Literature review on CLT for RF

Wager and Athey (2018) establish a CLT for trees such that :

- (honest tree) dataset is split in two parts (building the tree/estimating the mean in each leaf)
- (α-regular) leaves at least a fraction α of samples in each child
- (split randomization) positive probability of splitting each variable

For such trees, Wager and Athey (2018) establish a CLT

- Valid for any subsample size $s \simeq n^{\beta}$, for all $\beta \in (\eta, 1)$ where η depends on the tree structure
- Centered at the true value of the regression function $\mu(x)$.
- But with non-explicit convergence rates and variance

Peng et al. (2022) established CLT for various random forests, whose trees are built independently of the labels of the data.

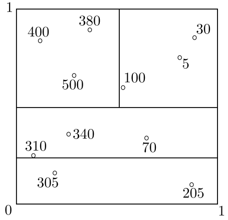
All results are built on Hájek projections.

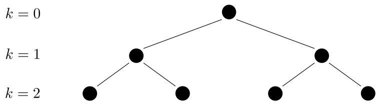
Summary

Centered Random Forests (CRF)

Construction of a centered tree (Breiman, 2004; Biau, 2012):

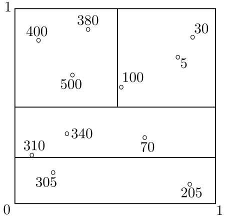
- Select s among n observations without replacement,
- Start at the root $[0,1]^d$, and at each node,
  1. a feature is uniformly chosen among all possible d features
  2. Along this feature, split is made at the center of the cell
- Stop when each cell has been split exactly k times.

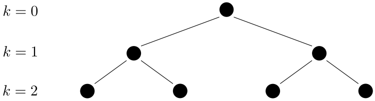




If the new point x falls into an empty cell, the tree arbitrarily predicts 0.

Infinite Centered Random Forests (ICRF)





Let $\mathrm{Z}_S := (Z_{i_1}, \ldots, Z_{i_s})$ be a training subsample of size s. The centered tree prediction at $\mathrm{x} \in [0,1]^d$ is

$$T^s(\mathrm{x}; \Theta; \mathrm{Z}_S) := \frac{\sum_{i \in S} Y_i \mathbb{1}\{X_i \in L_\Theta(\mathrm{x})\}}{N_{L_\Theta(\mathrm{x})}(\mathrm{X}_S)}.$$

Infinite CRF

$$\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x}) := \binom{n}{s}^{-1} \sum_{S \subset \{1,\ldots,n\}, |S|=s} \mathbb{E}[T^s(\mathrm{x}; \Theta; \mathrm{Z}_S) \mid \mathrm{Z}_S].$$

Assumptions

Infinite CRF

$$\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x}) := \binom{n}{s}^{-1} \sum_{S \subset \{1,\dots,n\}, |S|=s} \mathbb{E}[T^s(\mathrm{x}; \Theta; \mathrm{Z}_S) \mid \mathrm{Z}_S].$$

(H0) **Covariate Condition:** X is uniformly distributed on $\mathcal{X}$.

(H1) **Smoothness Condition:** The regression function μ is a L-lipschitz function with respect to the max norm.

(G1) **Tree complexity Condition:** The subsample size s and the tree depth k tend to infinity and satisfy $s/(k2^k) \to \infty$, as $n \to \infty$.

CLT for subsampling ICRF

Theorem (Mayala, Scornet, Tillier and Wintenberger 2025)

Let $d \geqslant 2$ and $\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x})$ be the ICRF estimator at point x. Assume **(H0)**, **(H1)** and **(G1)** hold and

$$\frac{n2^k}{s^2 k^{(d-1)/2}} \to \infty, \tag{1}$$

as $n, s, k \to \infty$. Then, for all $\mathrm{x} \in [0,1]^d$, we have

$$\sqrt{\frac{nk^{(d-1)/2}}{2^k}} \left(\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x}) - \mathbb{E}[\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x})] \right) \xrightarrow[n \to \infty]{d} \mathcal{N}(0, C(d)\mu(\mathrm{x})(1-\mu(\mathrm{x}))).$$

CLT for subsampling ICRF

Theorem (Mayala, Scornet, Tillier and Wintenberger 2025)

Let $d \geqslant 2$ and $\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x})$ be the ICRF estimator at point x. Assume **(H0)**, **(H1)** and **(G1)** hold and

$$\frac{n2^k}{s^2 k^{(d-1)/2}} \to \infty,$$

as $n, s, k \to \infty$. Then, for all $\mathrm{x} \in [0,1]^d$, we have

$$\sqrt{\frac{nk^{(d-1)/2}}{2^k}} \left(\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x}) - \mathbb{E}[\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x})] \right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, C(d)\mu(\mathrm{x})(1-\mu(\mathrm{x}))) .$$

- First CLT for random forests, with an explicit convergence rate, asymptotic variance, and condition on tree complexity.

CLT for subsampling ICRF

Theorem (Mayala, Scornet, Tillier and Wintenberger 2025)

Let $d \geqslant 2$ and $\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x})$ be the ICRF estimator at point x. Assume **(H0)**, **(H1)** and **(G1)** hold and

$$\frac{n2^k}{s^2 k^{(d-1)/2}} \to \infty,$$

as $n, s, k \to \infty$. Then, for all $\mathrm{x} \in [0,1]^d$, we have

$$\sqrt{\frac{nk^{(d-1)/2}}{2^k}} \left(\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x}) - \mathbb{E}[\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x})]\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, C(d)\mu(\mathrm{x})(1-\mu(\mathrm{x}))).$$

- (1) imposes that $s = o(n)$, since $2^k \leqslant s$ to avoid empty cells. Usual condition in the literature (see e.g. Wager and Athey, 2018; Peng et al., 2022)
- This result holds for trees close to bootstrapped ($s = n$) fully grown trees ($k = \log_2(s)$) but rate of convergence is very slow for such trees

CLT for subsampling ICRF

Theorem (Mayala, Scornet, Tillier and Wintenberger 2025)

Let $d \geqslant 2$ and $\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x})$ be the ICRF estimator at point x. Assume **(H0)**, **(H1)** and **(G1)** hold and

$$\frac{n2^k}{s^2 k^{(d-1)/2}} \to \infty,$$

as $n, s, k \to \infty$. Then, for all $\mathrm{x} \in [0,1]^d$, we have

$$\sqrt{\frac{nk^{(d-1)/2}}{2^k}} \left(\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x}) - \mathbb{E}[\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x})]\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, C(d)\mu(\mathrm{x})(1-\mu(\mathrm{x}))).$$

- Beware! The CLT is not centered at the true value of the regression function $\mu(x)$

CLT for subsampling ICRF

Theorem (Mayala, Scornet, Tillier and Wintenberger 2025)

Let $d \geqslant 2$ and $\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x})$ be the ICRF estimator at point x. Assume **(H0)**, **(H1)** and **(G1)** hold and

$$\frac{n2^k}{s^2 k^{(d-1)/2}} \to \infty, \quad \text{and} \quad 2^k k^{-\frac{d-1}{2}} n^{-\frac{d \log 2}{1+d \log 2}} \to \infty,$$

as $n, s, k \to \infty$. Then, for all $\mathrm{x} \in [0,1]^d$, we have

$$\sqrt{\frac{nk^{(d-1)/2}}{2^k}} \left(\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x}) - \mu(\mathrm{x}) \right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, C(d)\mu(\mathrm{x})(1-\mu(\mathrm{x}))) .$$

CLT for subsampling ICRF

Theorem (Mayala, Scornet, Tillier and Wintenberger 2025)

Let $d \geqslant 2$ and $\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x})$ be the ICRF estimator at point x. Assume **(H0)**, **(H1)** and **(G1)** hold and

$$\frac{n2^k}{s^2 k^{(d-1)/2}} \to \infty, \quad \text{and} \quad 2^k k^{-\frac{d-1}{2}} n^{-\frac{d \log 2}{1+d \log 2}} \to \infty,$$

as $n, s, k \to \infty$. Then, for all $\mathrm{x} \in [0,1]^d$, we have

$$\sqrt{\frac{nk^{(d-1)/2}}{2^k}} \left(\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x}) - \mu(\mathrm{x}) \right) \xrightarrow[n \to \infty]{d} \mathcal{N}(0, C(d)\mu(\mathrm{x})(1-\mu(\mathrm{x}))) .$$

- First CLT for random forests with an asymptotically unbiased estimator and explicit constants and conditions on tree structure.
- Comes at the price of an additional assumption

CLT for subsampling ICRF

Theorem (Mayala, Scornet, Tillier and Wintenberger 2025)

Let $d \geqslant 2$ and $\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x})$ be the ICRF estimator at point x. Assume **(H0)**, **(H1)** and **(G1)** hold and

$$\frac{n2^k}{s^2 k^{(d-1)/2}} \to \infty, \quad \text{and} \quad 2^k k^{-\frac{d-1}{2}} n^{-\frac{d \log 2}{1+d \log 2}} \to \infty,$$

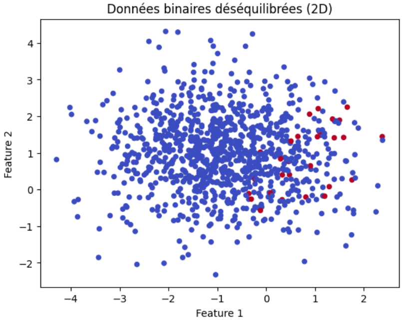
as $n, s, k \to \infty$. Then, for all $\mathrm{x} \in [0,1]^d$, we have

$$\sqrt{\frac{nk^{(d-1)/2}}{2^k}} \left(\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x}) - \mu(\mathrm{x}) \right) \xrightarrow[n \to \infty]{d} \mathcal{N}(0, C(d)\mu(\mathrm{x})(1-\mu(\mathrm{x}))).$$

- The first (resp. second) condition imposes that s is not too large (resp. too small). Let $s = n^\alpha$ and $2^k = n^\beta$. These two conditions can be rewritten as

$$\frac{d \log 2}{1 + d \log 2} < \alpha < 1 \quad \text{and} \quad \max\left(\frac{d \log 2}{1 + d \log 2}, 2\alpha - 1 \right) < \beta < \alpha.$$

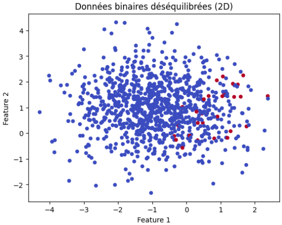
Imbalanced learning



Imbalance setting: $\mathbb{P}[Y = 1] \ll \mathbb{P}[Y = 0]$

- Learning algorithms may struggle in such settings
- Tendency to only predict the majority class
- Recall of the minority class may be very low
- Need for designing specific methods

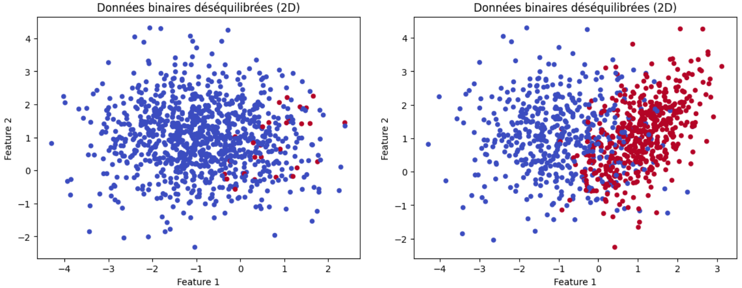
Imbalanced learning



Rebalancing strategies (Ramyachitra and Manikandan, 2014; Krawczyk, 2016):

- Undersample the majority class / oversample the minority class
- Assign weights to samples (King and Zeng, 2001)
- Change the loss (focal loss Lin et al., 2017)
- Generate synthetic data
  - SMOTE-like strategies (Chawla et al., 2002)
  - GAN (Xu et al., 2019), Diffusion (Jolicoeur-Martineau et al.)

Rebalancing framework



Now, assume that we are given a i.i.d. rebalanced training set $Z'_n := (Z'_1, \ldots, Z'_n)$, with $Z'_i = (X'_i, Y'_i)$ such that

$$\begin{cases} \mu'(\mathrm{x}) := \mathbb{P}(Y' = 1 | X' = \mathrm{x}),\ \mathrm{x} \in \mathcal{X}, \\ \mathbb{P}(Y' = 1) = p', \\ \mathbb{P}(X' \in \cdot \mid Y' = j) = \mathbb{P}(X \in \cdot \mid Y = j),\ j = \{0, 1\}. \end{cases}$$

Rebalanced ICRF

Given $\mathrm{Z}'_S := (Z'_{i_1}, \ldots, Z'_{i_s})$, with $S = \{i_1, \ldots, i_s\} \subset I$ and $|S| = s$, centered tree evaluated at point x trained on the s-subsample Z'_S that takes the form

$$T^s(\mathrm{x}; \Theta; \mathrm{Z}'_S) := \frac{\sum_{i \in S} Y'_i \mathbb{1}\{X'_i \in L_\Theta(\mathrm{x})\}}{N_{L_\Theta(\mathrm{x})}(\mathrm{X}'_S)},$$

Rebalanced ICRF

$$\widehat{\mu}^{\mathsf{ICRF}}_{\mathrm{RB},s}(\mathrm{x}) := \binom{n}{s}^{-1} \sum_{S \subset \{1,\ldots,n\}, |S| = s} \mathbb{E}[T^s(\mathrm{x}; \Theta; \mathrm{Z}'_S) \mid \mathrm{Z}'_S].$$

CLT for rebalanced forest

Theorem 2 (Mayala, Scornet, Tillier and Wintenberger, 2025)

Let $d \geqslant 2$ and $\widehat{\mu}^{\mathsf{ICRF}}_{\mathrm{RB},s}(\mathrm{x})$ be the Rebalanced ICRF estimator at point x. Assume **(H0)**, **(H1)** and **(G1)** hold and

$$\frac{n2^k}{s^2 k^{(d-1)/2}} \to \infty, \quad \text{and} \quad 2^k k^{-\frac{d-1}{2}} n^{-\frac{d \log 2}{1+d \log 2}} \to \infty,$$

as $n, s, k \to \infty$. Then, for all $\mathrm{x} \in [0,1]^d$ we have

$$\sqrt{\frac{n}{s^2 V'_{1,s}}}\left(\widehat{\mu}^{\mathsf{ICRF}}_{\mathrm{RB},s}(\mathrm{x}) - \mu'(\mathrm{x})\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)\,,$$

where

$$\frac{c'_1(\mathrm{x})2^k}{s^2 k^{(d-1)/2}} \leqslant \frac{V'_{1,s}}{\mu'(\mathrm{x})(1-\mu'(\mathrm{x}))} \leqslant \frac{c'_2(\mathrm{x})2^k}{s^2 k^{(d-1)/2}}.$$

CLT for rebalanced forest

Theorem 2 (Mayala, Scornet, Tillier and Wintenberger, 2025)

Let $d \geqslant 2$ and $\widehat{\mu}^{\mathsf{ICRF}}_{\mathrm{RB},s}(\mathrm{x})$ be the Rebalanced ICRF estimator at point x. Assume **(H0)**, **(H1)** and **(G1)** hold and

$$\frac{n2^k}{s^2 k^{(d-1)/2}} \to \infty, \quad \text{and} \quad 2^k k^{-\frac{d-1}{2}} n^{-\frac{d \log 2}{1+d \log 2}} \to \infty,$$

as $n, s, k \to \infty$. Then, for all $\mathrm{x} \in [0,1]^d$ we have

$$\sqrt{\frac{n}{s^2 V'_{1,s}}} \left(\widehat{\mu}^{\mathsf{ICRF}}_{\mathrm{RB},s}(\mathrm{x}) - \mu'(\mathrm{x})\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)\,,$$

where

$$\frac{c'_1(\mathrm{x})2^k}{s^2 k^{(d-1)/2}} \leqslant \frac{V'_{1,s}}{\mu'(\mathrm{x})(1-\mu'(\mathrm{x}))} \leqslant \frac{c'_2(\mathrm{x})2^k}{s^2 k^{(d-1)/2}}.$$

- Due to the rebalancing step, X' is not uniform on $[0,1]^d$ anymore: exact constant/rates cannot be derived anymore but the asymptotic remains the same.

CLT for rebalanced forest

Theorem 2 (Mayala, Scornet, Tillier and Wintenberger, 2025)

Let $d \geqslant 2$ and $\widehat{\mu}^{\mathrm{ICRF}}_{\mathrm{RB},s}(\mathrm{x})$ be the Rebalanced ICRF estimator at point x. Assume **(H0)**, **(H1)** and **(G1)** hold and

$$\frac{n2^k}{s^2 k^{(d-1)/2}} \to \infty, \quad \text{and} \quad 2^k k^{-\frac{d-1}{2}} n^{-\frac{d \log 2}{1+d \log 2}} \to \infty,$$

as $n, s, k \to \infty$. Then, for all $\mathrm{x} \in [0,1]^d$ we have

$$\sqrt{\frac{n}{s^2 V'_{1,s}}} \left(\widehat{\mu}^{\mathrm{ICRF}}_{\mathrm{RB},s}(\mathrm{x}) - \mu'(\mathrm{x})\right) \xrightarrow[n \to \infty]{d} \mathcal{N}(0,1)\,,$$

where

$$\frac{c'_1(\mathrm{x}) 2^k}{s^2 k^{(d-1)/2}} \leqslant \frac{V'_{1,s}}{\mu'(\mathrm{x})(1-\mu'(\mathrm{x}))} \leqslant \frac{c'_2(\mathrm{x}) 2^k}{s^2 k^{(d-1)/2}}.$$

- There is a bias: due to rebalancing, the CLT is not centered at $\mu(\mathrm{x})$ but at $\mu'(\mathrm{x})$!

Debiasing rebalanced forest

Due to the rebalancing approach, we have

$$\begin{cases} \mu'(\mathrm{x}) := \mathbb{P}(Y' = 1 | X' = \mathrm{x}),\ \mathrm{x} \in \mathcal{X}, \\ \mathbb{P}(Y' = 1) = p', \\ \mathbb{P}(X' \in \cdot \mid Y' = j) = \mathbb{P}(X \in \cdot \mid Y = j),\ j = \{0, 1\}. \end{cases}$$

Using standard calculation, we obtain

$$\mu(\mathrm{x}) = \frac{p(1-p')\mu'(\mathrm{x})}{p'(1-p)(1-\mu'(\mathrm{x})) + (1-p')p\mu'(\mathrm{x})}.$$

Thanks to the function

$$g(u) = \frac{p(1-p')u}{p'(1-p)(1-u) + (1-p')pu},$$

one can debias the RB-ICRF estimate using $\mu(\mathrm{x}) = g(\mu'(\mathrm{x}))$.

Importance sampling ICRF

$$\widehat{\mu}^{\mathrm{ICRF}}_{\mathrm{IS},s}(\mathrm{x}) := \frac{n_1(1-p')\widehat{\mu}^{\mathrm{ICRF}}_{\mathrm{RB},s}(\mathrm{x})}{p' n_0(1-\widehat{\mu}^{\mathrm{ICRF}}_{\mathrm{RB},s}(\mathrm{x})) + n_1(1-p')\widehat{\mu}^{\mathrm{ICRF}}_{\mathrm{RB},s}(\mathrm{x})}.$$

CLT for the Importance Sampling ICRF

$$\widehat{\mu}^{\mathrm{ICRF}}_{\mathrm{IS},s}(\mathrm{x}) := \frac{n_1(1-p')\widehat{\mu}^{\mathrm{ICRF}}_{\mathrm{RB},s}(\mathrm{x})}{p' n_0(1-\widehat{\mu}^{\mathrm{ICRF}}_{\mathrm{RB},s}(\mathrm{x})) + n_1(1-p')\widehat{\mu}^{\mathrm{ICRF}}_{\mathrm{RB},s}(\mathrm{x})}.$$

Corollary (Mayala, Scornet, Tillier and Wintenberger, 2025)

Let $d \geqslant 2$ and $\widehat{\mu}^{\mathrm{ICRF}}_{\mathrm{IS},s}(\mathrm{x})$ be the importance sampling ICRF estimator. Let $p \neq 0$, $p' \neq 1$ and assume **(H0)**, **(H1)** and **(G1)** hold and

$$\frac{n2^k}{s^2 k^{(d-1)/2}} \to \infty, \quad \text{and} \quad 2^k k^{-\frac{d-1}{2}} n^{-\frac{d \log 2}{1+d \log 2}} \to \infty,$$

as $n, s, k \to \infty$. Then, for all $\mathrm{x} \in [0,1]^d$, we have

$$\frac{1}{g'(\mu'(\mathrm{x}))}\sqrt{\frac{n}{s^2 V'_{1,s}}}\left(\widehat{\mu}^{\mathrm{ICRF}}_{\mathrm{IS},s}(\mathrm{x}) - \mu(\mathrm{x})\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)\,.$$

Variance comparison: highly imbalanced setting

Aim: comparing asymptotic variances of ICRF and IS-ICRF in high imbalanced settings ($p \to 0$).

Variance comparison: highly imbalanced setting

Aim: comparing asymptotic variances of ICRF and IS-ICRF in high imbalanced settings ($p \to 0$).

Problem - Assumptions

- $p \to 0$
- $f_X(x) = pf_{X|Y=1}(x) + (1-p)f_{X|Y=0}(x)$ is the uniform density

$\Rightarrow$ Conditional distributions $f_{X|Y=0,1}$ must change when $p \to 0$.

Variance comparison: highly imbalanced setting

Aim: comparing asymptotic variances of ICRF and IS-ICRF in high imbalanced settings ($p \to 0$).

New framework (H3)

We fix $f_{X|Y=0}$ and $f_{X|Y=0}$ such that

- Both are L-Lipschitz
- $0 < b_1 \leqslant f_{X|Y=0}(\cdot), f_{X|Y=1}(\cdot) \leqslant b_2 < \infty$
- There exists p'' such that

$$p'' f_{X|Y=1} + (1 - p'') f_{X|Y=0}$$

  is the uniform density on $[0, 1]^d$.

Variance comparison: highly imbalanced setting

Aim: comparing asymptotic variances of ICRF and IS-ICRF in high imbalanced settings ($p \to 0$).

New framework (H3)

We fix $f_{X|Y=0}$ and $f_{X|Y=0}$ such that

- Both are L-Lipschitz
- $0 < b_1 \leqslant f_{X|Y=0}(\cdot), f_{X|Y=1}(\cdot) \leqslant b_2 < \infty$
- There exists p'' such that

$$p'' f_{X|Y=1} + (1 - p'')f_{X|Y=0}$$

is the uniform density on $[0, 1]^d$.

(G1) Tree complexity Condition: The subsample size s and the tree depth k tend to infinity and satisfy $s/(k2^k) \to \infty$, as $n \to \infty$.

Variance comparison for fixed $p \ll 1$

Corollary (Mayala, Scornet, Tillier and Wintenberger, 2025)

Let $d \geqslant 2$, $p \neq 0$ and $p' \neq 1$. Grant **(H3)** and **(G1)**. Assume that

$$\frac{n2^k}{s^2 k^{(d-1)/2}} \to \infty, \quad \text{and} \quad 2^k k^{-\frac{d-1}{2}} n^{-\frac{d \log 2}{1+d \log 2}} \to \infty,$$

as $n, s, k \to \infty$. Then, for all $\mathrm{x} \in [0,1]^d$, we have

$$\sqrt{\frac{n}{s^2 V_{1,s}}} (\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x}) - \mu(\mathrm{x})) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1) \, ,$$

and

$$\frac{1}{g'(\mu'(\mathrm{x}))} \sqrt{\frac{n}{s^2 V'_{1,s}}} (\widehat{\mu}_{\mathrm{IS},s}^{\mathrm{ICRF}}(\mathrm{x}) - \mu(\mathrm{x})) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1) \, .$$

Thus, for all k large enough,

$$\frac{V'_{1,s}}{V_{1,s}} g'(\mu'(\mathrm{x}))^2 = O(p).$$

Summary

Simulation setting

(Sim. setting) n i.i.d. pairs (X_i, Y_i) distributed as (X, Y), with $X \sim U([0,1]^2)$ and

$$\mathbb{P}(Y = 1 | X = x) = \mu(x) = \frac{1}{1 + \exp(-(\beta_0 + 3x_1 + 2x_2))},$$

where β_0 is such that $\mathbb{P}(Y = 1) = 0.1$.
For each (n, α, β), we repeat $B = 1000$ times:

1. Generate a dataset with n observations **(Sim. setting)**.
2. A RF[4] is trained with defaut parameters and $s = n^\alpha$, $\texttt{max.depth} = \beta \log_2 n$.
3. The forest prediction $\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x})$ is evaluated at $\mathrm{x} = (0.7, 0.7)$.

We use these predictions to estimate

$$\mathbb{E}[\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x})] - \mu(\mathrm{x}) \quad \text{and} \quad \log\left(\mathbb{E}\left[\left(\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x}) - \mathbb{E}[\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x})]\right)^2\right]\right).$$

[4] R package `ranger` (see Wright and Ziegler, 2017)

Theoretical results

According to our theoretical analysis, if CLT were to hold in L^2, we would obtain

$$\log(\mathbb{E}[(\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x}) - \mathbb{E}[\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x})])^2]) \\ \sim -(1-\beta)\log n - \frac{d-1}{2}\log\log n + C_{1,d,\beta}(\mathrm{x})$$

with

$$C_{1,d,\beta}(\mathrm{x}) = \log(C(d)\mu(\mathrm{x})(1-\mu(\mathrm{x}))) - \frac{d-1}{2}\log\beta - \frac{d-1}{2}\log\log 2.$$

Theoretical results

According to our theoretical analysis, if CLT were to hold in L^2, we would obtain

$$\log(\mathbb{E}[(\widehat{\mu}_s^{\text{ICRF}}(\mathrm{x}) - \mathbb{E}[\widehat{\mu}_s^{\text{ICRF}}(\mathrm{x})])^2]) \sim -(1-\beta)\log n - \frac{d-1}{2}\log\log n + C_{1,d,\beta}(\mathrm{x})$$

with

$$C_{1,d,\beta}(\mathrm{x}) = \log(C(d)\mu(\mathrm{x})(1-\mu(\mathrm{x}))) - \frac{d-1}{2}\log\beta - \frac{d-1}{2}\log\log 2.$$

Thus, $\log(\mathbb{E}[(\widehat{\mu}_s^{\text{ICRF}}(\mathrm{x}) - \mathbb{E}[\widehat{\mu}_s^{\text{ICRF}}(\mathrm{x})])^2])$

- is approximately linear in $\log n$
- with lower slopes for larger values of β
- depends on β but not on the subsample size.

Rates of convergence - variance

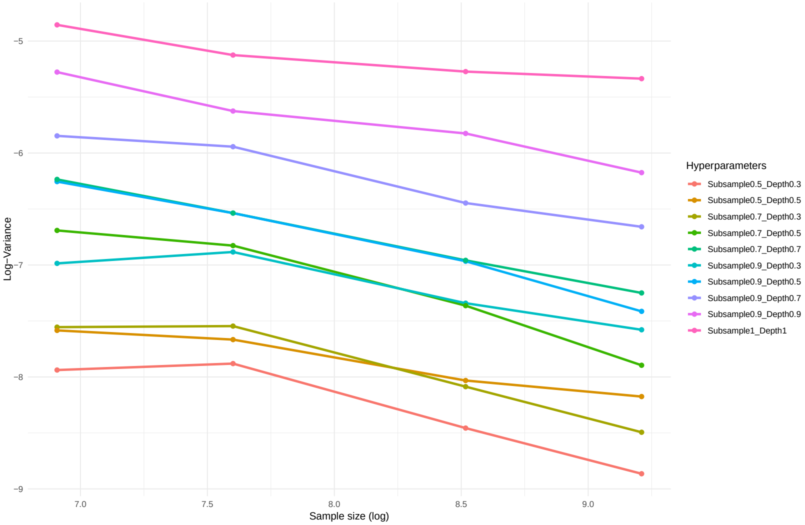


Figure: Log-variance of the classic RF

Rates of convergence - variance

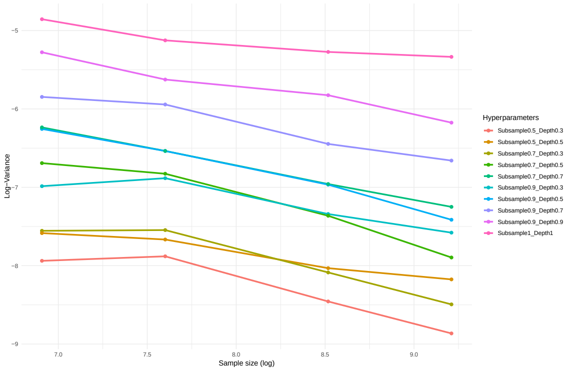


Figure: Log-variance of the classic RF

Theory tells us that the log-variance

- is approximately linear in $\log n$ ✓
- with lower slopes for larger values of β ✓
- depends on β but not on the subsample size $\simeq$

Bias

According to our theoretical results, the bias satisfies

$$\left(\frac{\beta}{\log 2}\right)^{(d-1)/2} n^{1-\beta}(\log n)^{(d-1)/2}\left(\mathbb{E}[\widehat{\mu}_s^{\mathsf{ICRF}}(\mathsf{x})] - \mu(\mathsf{x})\right) \to 0.$$

Thus,

$$\left(\mathbb{E}[\widehat{\mu}_s^{\mathsf{ICRF}}(\mathsf{x})] - \mu(\mathsf{x})\right) = o\left(n^{\beta-1}(\log n)^{-(d-1)/2}\right)$$

Rates of convergence - bias

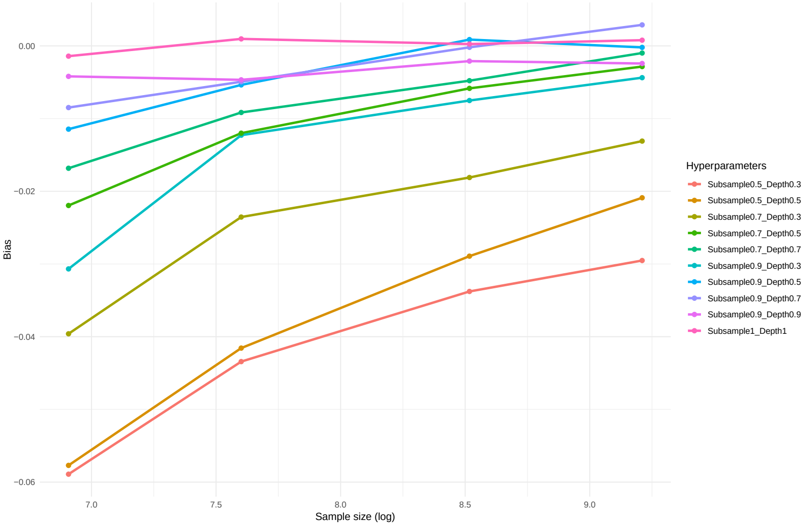


Figure: Bias of the classic RF

Rates of convergence - bias

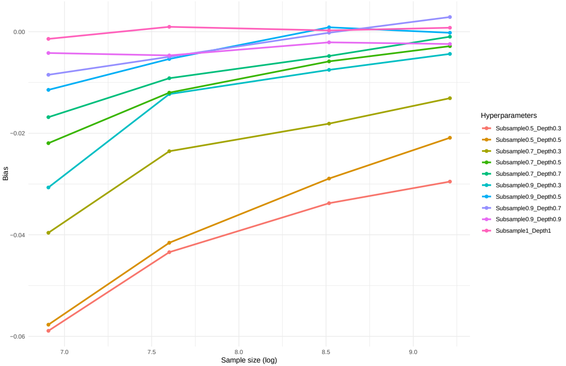


Figure: Bias of the classic RF

- Negative bias: majority of 0, RF pred. shifted toward 0.
- All biases tends to zero - Expected since tree depth increases ($k = \beta \log_2 n$).

Numerical Illustrations

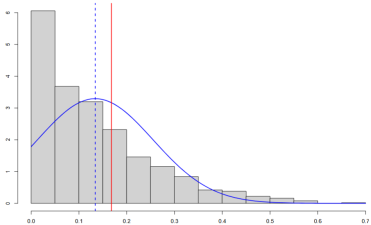


Figure: RF

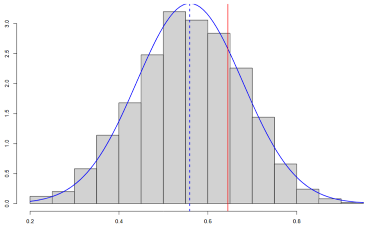


Figure: RB-RF

Histograms of predictions for each estimator with $p' = 0.5$, $n = 100$ and $B = 1000$ replicates.The empirical variances are: 0.121 (RF), 0.119 (RB-RF).

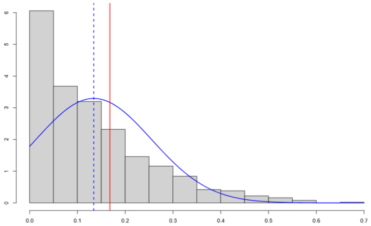


Figure: RF

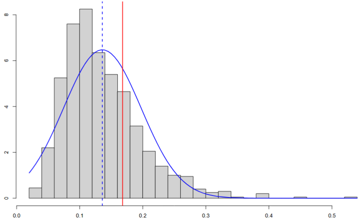


Figure: IS-RF

Histograms of predictions for each estimator with $p' = 0.5$, $n = 100$ and $B = 1000$ replicates.The empirical variances are: 0.121 (RF), 0.061 (IS-RF).

Conclusion

- We establish a CLT for centered random forest under the assumption that covariates are uniformly distributed on $[0,1]^d$.
    - Convergence rate and asymptotic variance are made explicit
    - Assumptions on tree structure (subsampling rate, tree depth)
    - First CLT on random forests with explicit rate of convergence and assumptions on tree structure

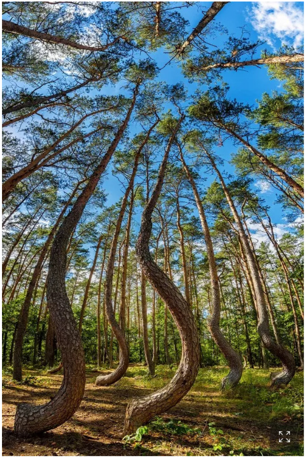
Conclusion

- We establish a CLT for centered random forest under the assumption that covariates are uniformly distributed on $[0, 1]^d$.
  - Convergence rate and asymptotic variance are made explicit
  - Assumptions on tree structure (subsampling rate, tree depth)
  - First CLT on random forests with explicit rate of convergence and assumptions on tree structure
- We analyze imbalanced learning problems
  - CLT for rebalanced forest, with non explicit constant as the new covariate distribution is not uniform
  - CLT is not centered at the correct value - bias of rebalancing strategies
  - We correct this bias and establish a CLT for the IS estimate
  - In a high imbalanced framework, $V_{IS} \ll V_{RF}$

Conclusion

- We establish a CLT for centered random forest under the assumption that covariates are uniformly distributed on $[0,1]^d$.
  - Convergence rate and asymptotic variance are made explicit
  - Assumptions on tree structure (subsampling rate, tree depth)
  - First CLT on random forests with explicit rate of convergence and assumptions on tree structure
- We analyze imbalanced learning problems
  - CLT for rebalanced forest, with non explicit constant as the new covariate distribution is not uniform
  - CLT is not centered at the correct value - bias of rebalancing strategies
  - We correct this bias and establish a CLT for the IS estimate
  - In a high imbalanced framework, $V_{IS} \ll V_{RF}$
- Numerical experiments show that our findings on centered forest can be partially extended to Breiman's random forests.

Thank you for your attention!



Asymptotic Normality of Infinite Centered Random Forests - Application to Imbalanced Classification, M. Mayala, E. Scornet, C. Tillier and O. Wintenberger

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 volume=9 number=9 year=2008 G. Lugosi, journal=JMLR. Consistency of random forests and other averaging classifiers.

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. 2004.

N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 2002.

References II

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.

A.W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

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

A. Jolicoeur-Martineau, K. Fatras, and T. Kachman. Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees. In *AISTAT*.

G. King and L. Zeng. Logistic regression in rare events data. *Political analysis*, 2001.

J. Klusowski. Sharp analysis of a simple model for random forests. In *AISTAT*, 2021.

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

B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in artificial intelligence*, 2016.

References III

A J Lee. *U-statistics: Theory and Practice*. Routledge, 2019.

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

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *IEEE international conference on computer vision*, 2017.

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, *NeurIPS*, pages 431–439, 2013.

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

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

W. Peng, T. Coleman, and L. Mentch. Rates of convergence for random forests via generalized u-statistics. *Electronic Journal of Statistics*, 2022.

D. Ramyachitra and P. Manikandan. Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)*, 2014.

References IV

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

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: A review. 2025.

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.

S. Wager, T. Hastie, and B. Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 2014.

M.N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of statistical software*, 2017.

References V

L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. Modeling tabular data using conditional gan. *NeurIPS*, 2019.

Z. Zhou, L. Mentch, and G. Hooker. V-statistics and variance estimation. *JMLR*, 2021.

Differences between uniform and non-uniform cases

Considering the centered forest with uniform covariates, we have

$$p_{k,\Theta}(\mathrm{x}) = \mathbb{P}(X \in L_\Theta(\mathrm{x})|\Theta) = 2^{-k}.$$

However, when the distribution is not uniform, we prove that

$$\begin{aligned} p'_{k,\Theta}(\mathrm{x}) &= \mathbb{P}(X' \in L_\Theta(\mathrm{x})|\Theta) \\ &= \frac{c'(\mathrm{x})}{2^k} \left(1 + \alpha'(\mathrm{x})\varepsilon'_\Theta(\mathrm{x})\mathrm{Diam}(L_\Theta(\mathrm{x}))\right) \qquad \text{a.s.} \end{aligned}$$

- Random variable and not a deterministic quantity
- Depends on x

Asymptotic constant

In our theoretical result for uniform covariates

$$\sqrt{\frac{nk^{(d-1)/2}}{2^k}}\left(\widehat{\mu}_s^{\mathrm{ICRF}}(\mathrm{x}) - \mu(\mathrm{x})\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, C(d)\mu(\mathrm{x})(1-\mu(\mathrm{x}))) .$$

with

$$C(d) = \frac{2\Gamma(d-1)}{(\log 2)^{d-1}\Gamma((d-1)/2)}\mathbb{E}\left[\left(\frac{\|(\mathrm{N}-\overline{\mathrm{N}}\mathbb{1})\|_2}{\|(\mathrm{N}-\overline{\mathrm{N}}\mathbb{1})\|_1}\right)^{d-1}\right] ,$$

where $\mathrm{N} = (N_1, \ldots, N_d)$ with $N_1, \ldots, N_d$ independent $\mathcal{N}(0,1)$ and $\overline{\mathrm{N}} = (1/d)\sum_{j=1}^{d} N_j$.

This comes from a new control of the quantity: when $k \to \infty$,

$$\mathbb{E}\Big[\mathbb{P}(X_1 \in L_\Theta(\mathrm{x})|X_1)^2\Big] \sim \frac{C(d)}{2^k k^{(d-1)/2}}.$$