Is interpolation benign for Random Forest regression?

Erwan Scornet
joint work with Ludovic Arnould (Paris 6) and Claire Boyer (Paris 6)

Outline

Interpolation regimes in ML

Interpolation in random forests

- Non-adaptive RF: centered RF (CRF)
- Non-adaptive RF: KeRF
- Semi-adaptive RF: median RF
- Adaptive RF: Breiman RF

Interpolation regimes in ML

Framework - Nonparametric regression

- Supervised learning: we assume to be given a training set $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ composed of i.i.d. pairs (X_i, Y_i), distributed as the generic pair (X, Y) with $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ (regression).

Framework - Nonparametric regression

- Supervised learning: we assume to be given a training set $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ composed of i.i.d. pairs (X_i, Y_i), distributed as the generic pair (X, Y) with $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ (regression).
- Our goal is to "learn" a predictor f_n, based on the training set $\mathcal{D}_n$, such that

$$\underbrace{f_n(X)}_{\text{prediction on test (unseen) data}} \simeq Y.$$

- Performance measure of a predictor f: $\text{Risk}(f) = \mathbb{E}\left[\left(Y - f(X)\right)^2\right]$
- The minimizer $f^\star$ of the risk is called the Bayes predictor

Framework - Nonparametric regression

- Supervised learning: we assume to be given a training set $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ composed of i.i.d. pairs (X_i, Y_i), distributed as the generic pair (X, Y) with $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ (regression).
- Our goal is to "learn" a predictor f_n, based on the training set $\mathcal{D}_n$, such that

$$\underbrace{f_n(X)}_{\text{prediction on test (unseen) data}} \simeq Y.$$

- Performance measure of a predictor f: $\text{Risk}(f) = \mathbb{E}\left[\left(Y - f(X)\right)^2\right]$
- The minimizer $f^\star$ of the risk is called the Bayes predictor
- Consistency: We say that a predictor f_n is consistent when

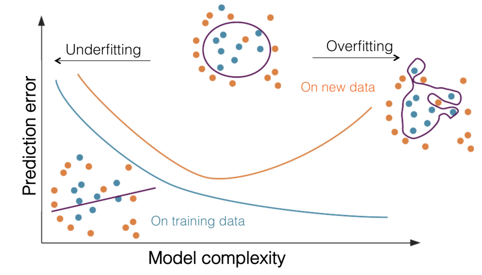
$$\text{Risk}(f_n) \xrightarrow[n \to +\infty]{} \text{Risk}(f^\star).$$

Complexity tuning

- Usually the constructed predictor f_n is constrained to live in a class $\mathcal{F}$ of functions
- Complexity of the model $\equiv$ Size of $\mathcal{F}$
- How to choose it?
  Statistical wisdom: take care of the so-called bias-variance tradeoff

Bias: systematic error, the predictor model is too simple to grasp data complexity

Variance: how much the predictions for a given point vary between different realizations of the model



Going beyond the traditional bias-variance tradeoff

New insights in the parametric world: adding another billion parameters to a neural network improves the predictive performances.

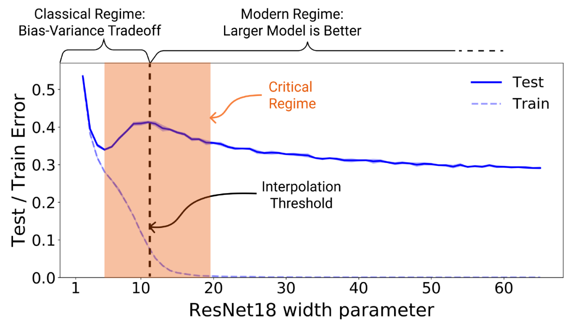


Fig. 1: Nakkiran et al. [2021]

Double descent phenomenon at least well-understood in linear models. [Hastie et al. 2019]

Over-parametrization in neural networks

The risk can be always decomposed as follows

$$\text{Risk} = \text{approximation error} + \text{estimation error} + \text{optimisation error}$$

Why does not overparametrization hurt NN training ?

- approximation error: more parameters, better approx capacities
- optimisation error: more parameters, nicer optimisation space

  [NGuyen et al. 2019, Nguyen 2020]

- estimation error: more parameters, implicit regularisation

  [Deep learning: a statistical viewpoint, Bartlett, Montanari, Rakhlin, 21]

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]

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

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

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

- Local-means estimator: $f(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{\|x - X_i\|}{h}\right)}{\sum_{i=1}^n K\left(\frac{\|x - X_i\|}{h}\right)}$ 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 $K_a(x) = \|x\|^{-a}\mathbb{1}_{\|x\|\leqslant 1}$.

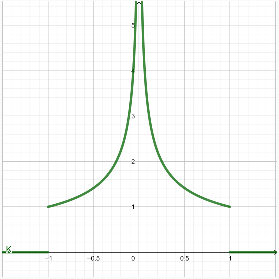


Fig. 2: Singular kernel above for $a = 0.5$

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

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

Predictions of singular kernels

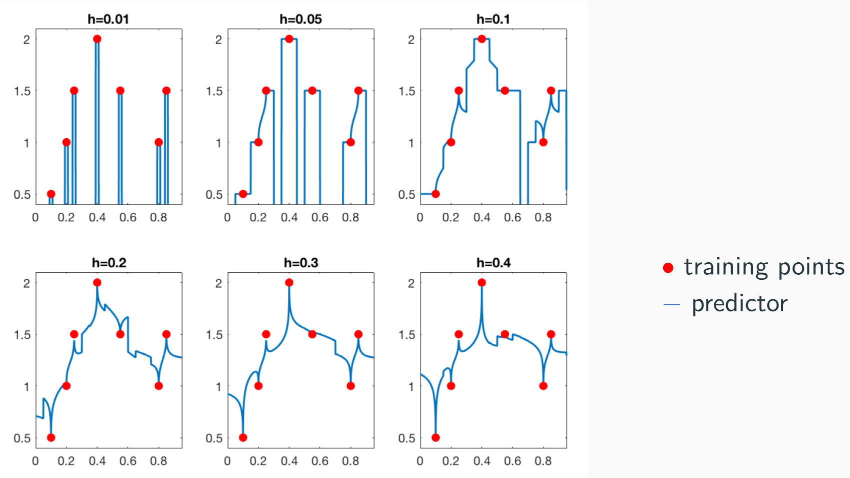


Fig. 3: Interpolation with $K(x) = \|x\|^{-a}\mathbb{1}_{\|x\|\leqslant 1}$ and $a = 0.49$, [Belkin et al., 2019]

Spiked-smooth estimates

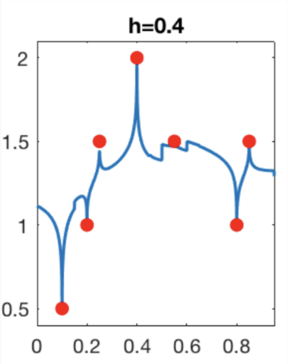


Fig. 4: From [Belkin et al. 2019]

Spiked: the influence of interpolation is very localized around training points.

Smooth: anywhere else, the estimated function remains "smooth".

$$f_n(x) = f^{\text{smooth}}(x) + \Delta^{\text{spiky}}(x)$$

Spiked-smooth estimates

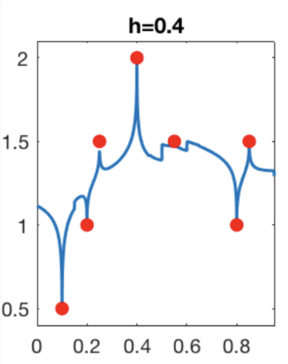


Fig. 4: From [Belkin et al. 2019]

Spiked: the influence of interpolation is very localized around training points.

Smooth: anywhere else, the estimated function remains "smooth".

$$f_n(x) = f^{\text{smooth}}(x) + \Delta^{\text{spiky}}(x)$$

Beyond kernel methods

Can the same be said for random forests?

Interpolation in random forests

Random Forest

Random forest (RF) $f_{M,n}(x) = \dfrac{1}{M}\sum_{m=1}^{M} t_n(x, \theta_m)$

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

Random Forest

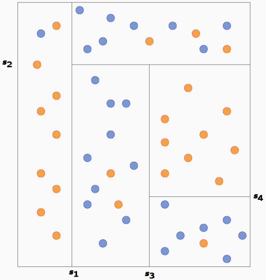
Random forest (RF) $f_{M,n}(x) = \dfrac{1}{M}\sum_{m=1}^{M} t_n(x, \theta_m)$

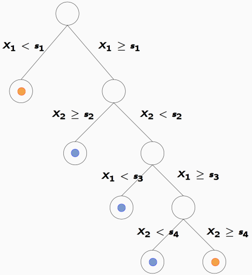
- Non-parametric method
- Based on bagging and random feature selections
- Aggregate the predictions of M trees

Decision Trees (DT)

- DT is a way to partition the input space along coordinates axes
- At each step, the DT finds a feature j and a threshold τ for splitting (usually according to some diversity criterion (entropy, ...))

Decision tree





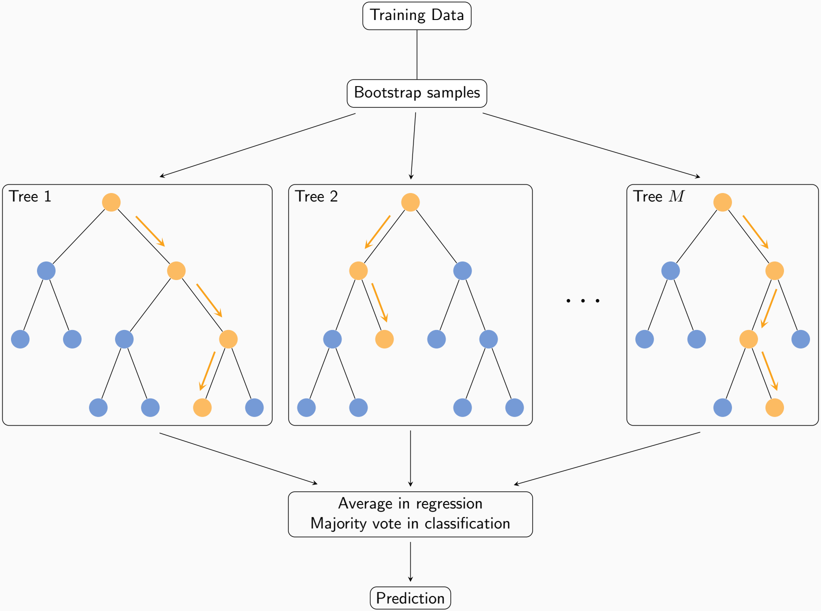
$$t_n(x, \theta) = \sum_{i=1}^{n} Y_i \boxed{\frac{\mathbb{1}_{X_i \in A_n(x,\theta)}}{N_n(x,\theta)}}$$

$\theta \equiv$ randomized cuts

$A_n(x, \theta) \equiv$ leaf containing x

$N_n(x, \theta) \equiv$ number of data points in $A_n(x, \Theta)$

A classical random forest



Research statement

RF are powerful predictors in practice

- Consistency has been proved for several simpler RF models with label-independent splits.
- Most convergence results are based on a control of the tree depth, preventing trees to be fully grown, and thus avoiding interpolation.

Goal

- Is there any random forest model that both interpolate and exhibit consistency properties? In other words,

$$\text{Risk}\left(\text{interpolating RF}\right) \xrightarrow[n \to +\infty]{?} \text{Risk}(f^\star)$$

Research statement

Goal

- Study of the consistency of RF in interpolation regimes in regression

$$\text{Risk}\,(\text{interpolating RF}) \xrightarrow[n \to +\infty]{?} \text{Risk}(f^\star)$$

RF type	Cuts depend on X_i	Cuts depend on Y_i
non-adaptive (centered RF)	✗	✗
semi-adaptive (Median RF)	✓	✗
adaptive (Breiman RF)	✓	✓

Preliminary

- The generative model satisfies

$$Y = f^\star(X) + \varepsilon,$$

with $X \sim \mathcal{U}\left([0,1]^d\right)$ and $\mathbb{E}\left[\varepsilon | X\right] = 0$ almost surely.

- Risk of f_n

$$\text{Risk}(f_n) = \mathbb{E}\left[(f_n(X) - Y)^2\right]$$

- Forest predictor

$$f_{M,n}(x) = \frac{1}{M}\sum_{m=1}^{M} t_n(x, \theta_j)$$

- Infinite forest predictor

$$f_{\infty,n}(x) = \mathbb{E}_\Theta\left[t_n(x, \Theta)\right]$$

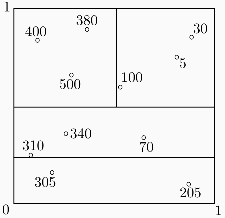
Outline

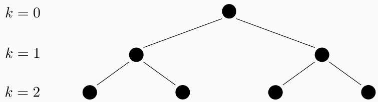
Non-adaptive RF: Centered RF (CRF)

Construction of a centered tree: at each step,

1. a feature is uniformly chosen among all possible d features
2. the split along the chosen feature is made at the center of the current cell

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





Non-adaptive RF: Centered RF (CRF)

Standard CRF

$$f_{M,n}(x, \Theta_M) = \frac{1}{M} \sum_{m=1}^{M} t_n(x, \Theta_m) \qquad f_{\infty,n}(x) = \mathbb{E}_{\Theta}[f_n(x, \Theta)]$$

Theorem [Klusowski, 2021]

The risk of the infinite centered forest $f_{\infty,n}^{\mathrm{CRF}}$ satisfies, for any depth k_n,

$$\mathrm{Risk}(f_{\infty,n}^{\mathrm{CRF}}(X)) - \mathrm{Risk}(f^\star) \leqslant \underbrace{d \sum_{j=1}^{d} \|\partial_j f^\star\|_\infty 2^{2k_n \log(1-1/(2d))}}_{\text{approximation error}}$$

$$+ \underbrace{12\sigma^2 8^d d^{d/2} \frac{2^{k_n}}{n} \frac{1}{k_n^{(d-1)/2}}}_{\text{estimation error}} + \underbrace{B^2 \exp\left(-\frac{n}{2^{k_n+1}}\right)}_{\text{bias related to empty cells}} .$$

Non-adaptive RF: Centered RF (CRF)

Standard CRF

$$f_{M,n}(x, \Theta_M) = \frac{1}{M} \sum_{m=1}^{M} t_n(x, \Theta_m) \qquad f_{\infty,n}(x) = \mathbb{E}_\Theta[f_n(x, \Theta)]$$

Theorem [Klusowski, 2021]

The risk of the infinite centered forest $f_{\infty,n}^{\mathrm{CRF}}$ satisfies, for a depth $k_n = \log_2 n$,

$$\mathrm{Risk}(f_{\infty,n}^{\mathrm{CRF}}(X)) - \mathrm{Risk}(f^\star) \leqslant \underbrace{d \sum_{j=1}^{d} \|\partial_j f^\star\|_\infty n^{2\log(1-1/(2d))}}_{\text{approximation error}}$$

$$+ \underbrace{12\sigma^2 8^d d^{d/2} \frac{1}{(\log_2 n)^{(d-1)/2}}}_{\text{estimation error}} + \underbrace{B^2 \exp\left(-\frac{1}{2}\right)}_{\text{bias related to empty cells}} .$$

Non-adaptive RF: Centered RF (CRF)

Standard CRF

$$f_{M,n}(x, \Theta_M) = \frac{1}{M} \sum_{m=1}^{M} t_n(x, \Theta_m) \qquad f_{\infty,n}(x) = \mathbb{E}_\Theta[f_n(x, \Theta)]$$

Unfortunately...

Proposition [Arnould et al., 2023]

Assume that $\mathbb{E}[f^\star(X)^2] > 0$. Then, in the mean interpolating regime (one point/cell in average, $k = \lfloor \log_2(n) \rfloor$), the CRF $f_{\infty,n}^{\mathrm{CRF}}$ is not consistent.

Non-adaptive RF: Centered RF (CRF)

Addressing the problem of empty cells by not averaging over them!

Void-free CRF

$$f^{\mathrm{VF}}_{M,n}(x, \Theta_M) \propto \sum_{m=1}^{M} t_n(x, \Theta_m) \mathbb{1}_{N_n(x,\Theta_m)>0} \qquad f^{\mathrm{VF}}_{\infty,n}(x) = \mathbb{E}_{\Theta}\left[f_n(x, \Theta) | N_n(x, \Theta) > 0\right]$$

Non-adaptive RF: Centered RF (CRF)

Addressing the problem of empty cells by not averaging over them!

Void-free CRF

$$f^{\mathrm{VF}}_{M,n}(x,\Theta_M) \propto \sum_{m=1}^{M} t_n(x,\Theta_m)\mathbb{1}_{N_n(x,\Theta_m)>0} \qquad f^{\mathrm{VF}}_{\infty,n}(x) = \mathbb{E}_{\Theta}\left[f_n(x,\Theta)|N_n(x,\Theta)>0\right]$$

Proposition [Arnould et al., 2023]

Assume that $f^\star$ has bounded partial derivatives. Then, in the mean interpolating regime ($k = \lfloor \log_2 n \rfloor$), the infinite void-free-CRF $f^{\mathrm{VF}}_{\infty,n}$ is consistent in a noiseless setting ($\sigma = 0$), and, for all $n > 1$,

$$\mathcal{R}\left(f^{\mathrm{VF}}_{\infty,n}(X)\right) \leqslant C_d \left(\frac{n}{\log_2 n}\right)^{2\log_2\left(1-\frac{1}{2d}\right)} + (C_d+2)\, n^{-1/(2\ln 2)},$$

where $C_d = 4d\left(\sum_{j=1}^{d} ||\partial f^\star_j||^2_\infty\right)$.

Centered RF: ideas of proof

Aggregating all cells,

$$\mathrm{Risk}(f_{\infty,n}^{\mathrm{CRF}}(X)) - \mathrm{Risk}(f^\star)$$
$$\geqslant \mathbb{E}\left[f^\star(X)^2 \mathbb{P}\left(N_n(X,\Theta) = 0 | X\right)\right].$$

Aggregating non-empty cells (noiseless setting)

$$\mathrm{Risk}(f_{\infty,n}^{\mathrm{VF}}(X)) - \mathrm{Risk}(f^\star)$$
$$\leqslant \mathrm{bias}^2 + ||f||_\infty^2 \mathbb{P}\left(\forall \Theta, N_n(\Theta, X) = 0\right)$$

CRF vs Void-free CRF

$\mathbb{P}\left(N_n(X,\Theta) = 0\right)$ falling into an empty leaf in a single random tree of the infinite forest.

vs.

$\mathbb{P}_{X,\mathcal{D}_n}\left[\forall \Theta, N_n(X,\Theta) = 0\right]$. falling into empty leaves in all trees of the infinite forest.

Outline

Kernel RF (KeRF)

Still in the mean interpolation regime, one can study KeRF

- to avoid the problem of empty cells
- to control the risk (variance)

Kernel RF (KeRF)

Still in the mean interpolation regime, one can study KeRF

- to avoid the problem of empty cells
- to control the risk (variance)

KeRF

1. grow all centered trees
2. average along all points contained in the leaves in which x falls

$$f_{M,n}^{\mathrm{KeRF}}(x,\Theta) = \frac{\sum_{i=1}^{n} Y_i K_{M,n}(x,X_i)}{\sum_{i=1}^{n} K_{M,n}(x,X_i)} = \frac{\sum_{i=1}^{n} Y_i \sum_{m=1}^{M} \mathbb{1}_{X_i \in A_n(x,\Theta_m)}}{\sum_{i=1}^{n} \sum_{m=1}^{M} \mathbb{1}_{X_i \in A_n(x,\Theta_m)}}$$

Kernel RF (KeRF)

Still in the mean interpolation regime, one can study KeRF

- to avoid the problem of empty cells
- to control the risk (variance)

KeRF

1. grow all centered trees
2. average along all points contained in the leaves in which x falls

$$f_{M,n}^{\mathrm{KeRF}}(x,\Theta) = \frac{\sum_{i=1}^{n} Y_i K_{M,n}(x, X_i)}{\sum_{i=1}^{n} K_{M,n}(x, X_i)} = \frac{\sum_{i=1}^{n} Y_i \sum_{m=1}^{M} \mathbb{1}_{X_i \in A_n(x,\Theta_m)}}{\sum_{i=1}^{n} \sum_{m=1}^{M} \mathbb{1}_{X_i \in A_n(x,\Theta_m)}}$$

Infinite KeRF

$$f_{\infty,n}^{\mathrm{KeRF}}(x) = \frac{\sum_{i=1}^{n} Y_i K_n(x, X_i)}{\sum_{i=1}^{n} K_n(x, X_i)} = \frac{\sum_{i=1}^{n} Y_i \mathbb{P}_{\Theta}\left[X_i \in A_n(x,\Theta)\right]}{\sum_{i=1}^{n} \mathbb{P}_{\Theta}\left[X_i \in A_n(x,\Theta)\right]}$$

KeRF

Theorem [Arnould et al., 2023]

Assume that $f^\star$ is Lipschitz continuous and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Let $d > 5$. Then, in the mean interpolation regime, $k_n = \lfloor \log_2(n) \rfloor$,

$$\text{Risk}(f_{\infty,n}^{\text{KeRF}}) - \text{Risk}(f^\star) \leqslant C_d \log(n)^{-(d-5)/6},$$

with $C_d > 0$ a constant depending on $\sigma, d, \|f^\star\|_\infty$.

KeRF

Theorem [Arnould et al., 2023]

Assume that $f^\star$ is Lipschitz continuous and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Let $d > 5$. Then, in the mean interpolation regime, $k_n = \lfloor \log_2(n) \rfloor$,

$$\mathrm{Risk}(f_{\infty,n}^{\mathrm{KeRF}}) - \mathrm{Risk}(f^\star) \leqslant C_d \log(n)^{-(d-5)/6},$$

with $C_d > 0$ a constant depending on $\sigma, d, \|f^\star\|_\infty$.

Remarks

- In the mean interpolation regime, the infinite KeRF is consistent
- Slow convergence rate
- Almost matching the lower bound $\log(n)^{-d+1}$ for the optimal convergence rate of deep non-adaptive RF [Lin & Jeon, 2006]

Outline

Interpolation regimes in ML

Interpolation in random forests

- Non-adaptive RF: centered RF (CRF)
- Non-adaptive RF: KeRF
- Semi-adaptive RF: median RF
- Adaptive RF: Breiman RF

Towards strict interpolation

- So far, study in the mean interpolation regime only
- To analyze the strict interpolation case, we have to consider semi-adaptive RF

Towards strict interpolation

- So far, study in the mean interpolation regime only
- To analyze the strict interpolation case, we have to consider semi-adaptive RF

Semi-adaptive median RF

1. 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.
2. Median RF: aggregation of median trees

What we know about Median RF

Assumption (H1)

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

What we know about Median RF

Assumption (H1)

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

Theorem [Scornet, 2016]

Grant Assumption **(H1)**. Then, provided $a_n \to \infty$ and $a_n/n \to 0$, the infinite median forest $f_{\infty,n}^{\mathrm{MedRF}}$ is consistent, i.e.,

$$\lim_{n\to\infty} \mathrm{Risk}\left(f_{\infty,n}^{\mathrm{MedRF}}\right) = \mathrm{Risk}(f^\star).$$

What we know about Median RF

Assumption (H1)

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

Theorem [Scornet, 2016]

Grant Assumption **(H1)**. Then, provided $a_n \to \infty$ and $a_n/n \to 0$, the infinite median forest $f_{\infty,n}^{\mathrm{MedRF}}$ is consistent, i.e.,

$$\lim_{n\to\infty} \mathrm{Risk}\left(f_{\infty,n}^{\mathrm{MedRF}}\right) = \mathrm{Risk}(f^\star).$$

Remarks

- First (and only) consistency results for fully grown trees.
- Each tree is not consistent but the forest is, because of subsampling.

What we know about Median RF

Assumption (H1)

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

Theorem [Scornet, 2016]

Grant Assumption **(H1)**. Then, provided $a_n \to \infty$ and $a_n/n \to 0$, the infinite median forest $f_{\infty,n}^{\mathrm{MedRF}}$ is consistent, i.e.,

$$\lim_{n\to\infty} \mathrm{Risk}\left(f_{\infty,n}^{\mathrm{MedRF}}\right) = \mathrm{Risk}(f^\star).$$

Remarks

- First (and only) consistency results for fully grown trees.
- Each tree is not consistent but the forest is, because of subsampling.

Unsatisfying result because forest interpolation only occurs when $a_n = n$.

Semi-adaptive RF: median RF

Theorem [Arnould et al., 2023]

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

$$\mathcal{R}\left(f_{\infty,n}^{\mathrm{MedRF}}\right) \leqslant C_1 d \left(\sum_{\ell=1}^{d} ||\partial_\ell f^\star||_\infty^2\right) \left(1 - \frac{3}{4d}\right)^{\log_2 n} + \sigma^2 C_{2,d} (\log_2 n)^{-(d-1)/2},$$

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

Semi-adaptive RF: median RF

Theorem [Arnould et al., 2023]

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

$$\mathcal{R}\left(f_{\infty,n}^{\mathrm{MedRF}}\right) \leqslant C_1 d \left(\sum_{\ell=1}^{d} \|\partial_\ell f^\star\|_\infty^2\right)\left(1 - \frac{3}{4d}\right)^{\log_2 n} + \sigma^2 C_{2,d} (\log_2 n)^{-(d-1)/2},$$

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

- Interpolating (median) RF are consistent in a noisy setting (first result).
- Slow rate as expected
- 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).

Outline

Adaptive RF: Breiman forests

- Widely used
- Cuts depend on X_i and Y_i

Breiman random forests

- Data sampling : bootstrap
- At each cell, select randomly m_{try} coordinates among $\{1, \ldots, d\}$.
- Choose the split by minimizing the CART-split criterion on the cell along the m_{try} selected coordinates.
- Stop when each cell contains exactly one point.

- Aggregate CART trees

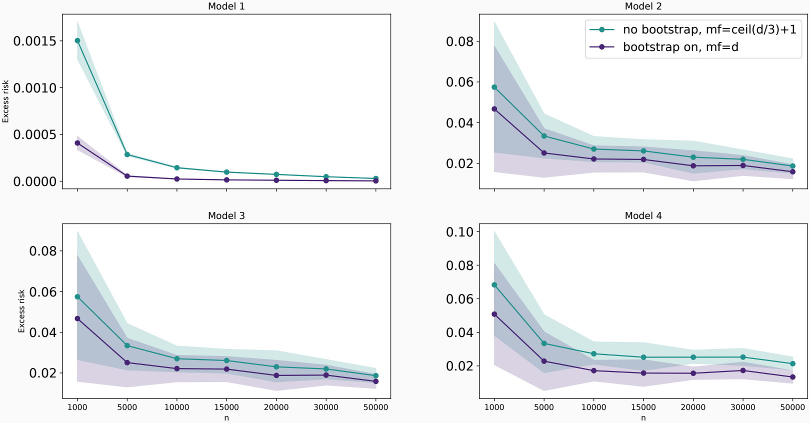
Hard to theoretically analyze (even in non-interpolation regimes)

Numerical XP with interpolating Breiman RF

- Simulated data with 4 different models
- 500 trees per forest, (max-depth$=$ `None`)
- 2 types of forests
  - `max-feature` $= \lceil d/3 \rceil$ + bootstrap off (interpolating)
  - `max-feature` $= d$ + bootstrap on (non-interpolating)

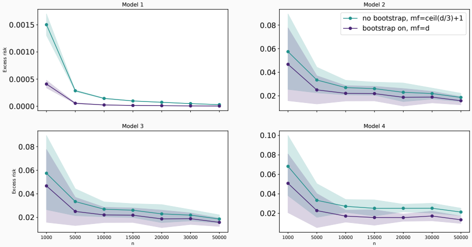
Numerical XP with interpolating Breiman RF

- Simulated data with 4 different models
- 500 trees per forest, (max-depth= `None`)
- 2 types of forests
  - `max-feature` $= \lceil d/3 \rceil$ + bootstrap off (interpolating)
  - `max-feature` $= d$ + bootstrap on (non-interpolating)



Numerical XP with interpolating Breiman RF

- Simulated data with 4 different models
- 500 trees per forest, (max-depth= `None`)
- 2 types of forests
  - `max-feature` $= \lceil d/3 \rceil$ + bootstrap off (interpolating)
  - `max-feature` $= d$ + bootstrap on (non-interpolating)



Conclusion

Interpolating Breiman RF seem to be consistent even in the noisy setting

Breiman RF: how about the interpolation zone?

- ✗ 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: how about the interpolation zone?

- ✗ 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 `max-features` 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

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

- Necessary condition for consistency: $\mathbb{E}\left[\mathrm{vol}(Z_n)\right] \to 0$ as $n \to \infty$
- For most points of the space, more than one point are involved in the prediction of the RF $\rightsquigarrow$ self-averaging property?

Conclusion

- Non-adaptive interpolating RF are not consistent (empty cells)

Conclusion

- Non-adaptive interpolating RF are not consistent (empty cells)
- Adaptive RF: interpolation and consistency become compatible when self-regularisation processes occur
  - Theoretically proved for Median RF
  - Empirical evidence for Breiman RF

Conclusion

- Non-adaptive interpolating RF are not consistent (empty cells)
- Adaptive RF: interpolation and consistency become compatible when self-regularisation processes occur
  - Theoretically proved for Median RF
  - Empirical evidence for Breiman RF
- RF vs kernel methods:
  - Singular Kernel (any bandwidth) versus interpolating RF (large depth)
  - Slow rate of consistency

Conclusion - Thank you!

- Non-adaptive interpolating RF are not consistent (empty cells)
- Adaptive RF: interpolation and consistency become compatible when self-regularisation processes occur
  - Theoretically proved for Median RF
  - Empirical evidence for Breiman RF
- RF vs kernel methods:
  - Singular Kernel (any bandwidth) versus interpolating RF (large depth)
  - Slow rate of consistency

Summary of theoretical contributions

<table>
<tr><th colspan="2" rowspan="3"></th><th colspan="4">Conditions for consistency</th></tr>
<tr><th colspan="2">Regardless of the noise scenario</th><th>In a noisy scenario</th><th></th></tr>
<tr><th>Managing the empty cells issue</th><th>Controlling the bias</th><th>Controlling the variance</th><th>Decreasing volume of the interpolation zone</th></tr>
<tr><td rowspan="3">Mean interpolation regime (non-adaptive RF)</td><td>Centered RF</td><td>✗</td><td>✓</td><td>✓</td><td></td></tr>
<tr><td>Void-free CRF</td><td>✓</td><td>✓</td><td>?</td><td></td></tr>
<tr><td>Centered KeRF</td><td>✓</td><td>✓</td><td>✓</td><td></td></tr>
<tr><td rowspan="2">Exact interpolation (semi-adaptive and adaptive RF)</td><td>Median RF</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td></tr>
<tr><td>Breiman RF</td><td>✓</td><td>?</td><td>?</td><td>✓</td></tr>
</table>

References

Ludovic Arnould, Claire Boyer, and Erwan Scornet. Is interpolation benign for random forest regression? *The 26th International Conference on Artificial Intelligence and Statistics*, 2023.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

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

References II

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12): 124003, 2021.

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

Models in numerical XP

- **Model 1:** $d = 2,\ Y = 2X_1^2 + \exp(-X_2^2)$ (noiseless)
- **Model 2:** $d = 8,\ Y = X_1X_2 + X_3^2 - X_4X_5 + X_6X_7 - X_8^2 + \mathcal{N}(0, 0.5)$
- **Model 3:** $d = 6,\ Y = X_1^2 + X_2^2X_3e^{-|X_4|} + X_5 - X_6 + \mathcal{N}(0, 0.5)$
- **Model 4:** $d = 5$,
  $Y = 1/(1 + \exp(-10(\sum_{i=1}^{d} X_i - 1/2))) + \mathcal{N}(0, 0.05)$