Analyse du risque de forêts purement aléatoires
L'intérêt de la diversité dans les forêts

Sylvain Arlot[1] (collaboration avec Robin Genuer[2])

[1]Université Paris-Sud

[2]ISPED, Université Bordeaux 2

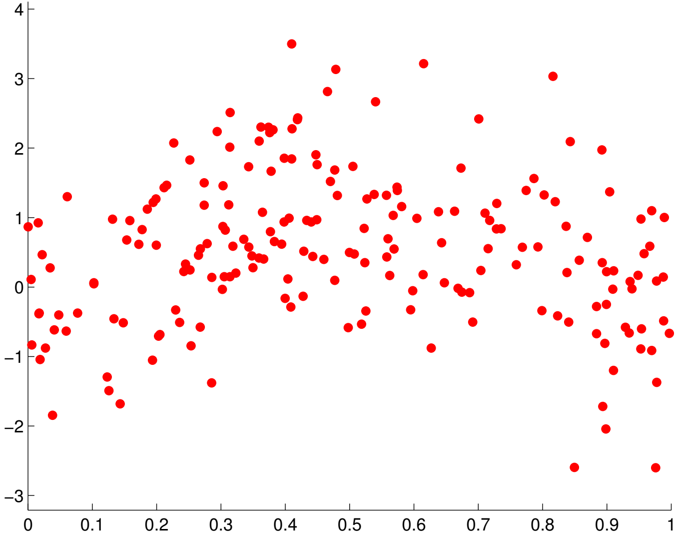
Journée "Promenade en forêt aléatoire", IHP
11 Janvier 2018

Références: arXiv:1407.3939 arXiv:1604.01515

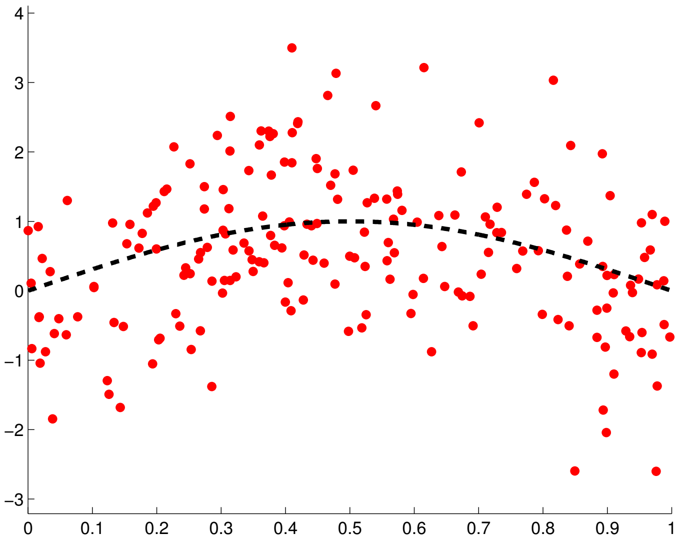
Outline

Outline

Regression: data $(X_1, Y_1), \dots, (X_n, Y_n)$



Goal: find the signal (denoising)



Regression

- Data D_n : $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ (i.i.d. $\sim P$)

$$Y_i = s^\star(X_i) + \varepsilon_i$$

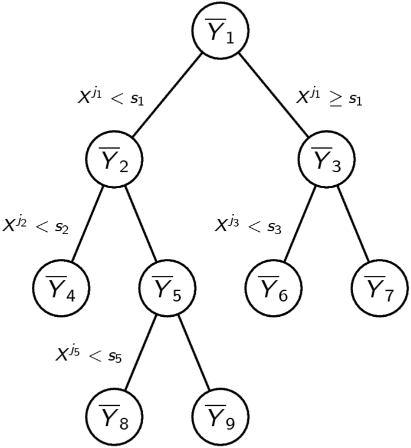
with $s^\star(X) = \mathbb{E}[Y \mid X]$ (regression function).

- Goal: learn f measurable function $\mathcal{X} \to \mathbb{R}$ s.t. the quadratic risk

$$\mathbb{E}_{(X,Y)\sim P}\left[\left(f(X) - s^\star(X)\right)^2\right]$$

is minimal.

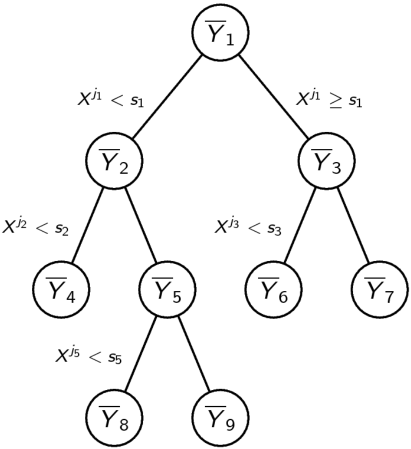
Regression tree (Breiman et al, 1984)



Tree: piecewise-constant predictor, obtained by partitioning recursively $\mathbb{R}^d$.

Restriction: splits parallel to the axes.

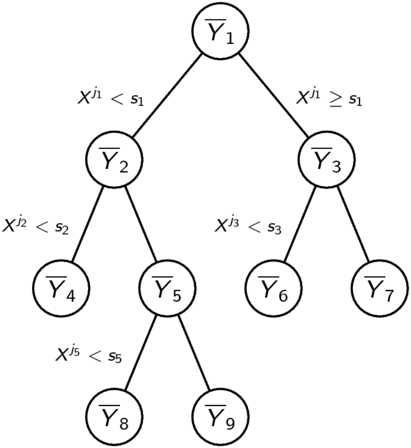
Regression tree (Breiman et al, 1984)



Tree: piecewise-constant predictor, obtained by partitioning recursively $\mathbb{R}^d$.

1. Choice of the partition $\mathbb{U}$ (tree structure)
   Usually, at each step, one looks for the best split of the data into two groups (minimize sum of within-group variances) D_n.

Regression tree (Breiman et al, 1984)



Tree: piecewise-constant predictor, obtained by partitioning recursively $\mathbb{R}^d$.

1. Choice of the partition $\mathbb{U}$ (tree structure)
2. For each $\lambda \in \mathbb{U}$ (tree leaf), choice of the estimation $\widehat{\beta}_\lambda$ of $s^\star(x)$ when $x \in \lambda$.
   Here, $\widehat{\beta}_\lambda = \overline{Y}_\lambda$ average of the $(Y_i)_{X_i \in \lambda}$.

Random forest (Breiman, 2001): general definition

Definition (Random forest (Breiman, 2001))

$\left\{ \widehat{s}_{\Theta_j}, 1 \leqslant j \leqslant q \right\}$ collection of tree predictors, $(\Theta_j)_{1 \leqslant j \leqslant q}$ i.i.d. r.v. independent from D_n.

Random forest predictor $\widehat{s}$ obtained by aggregating the tree collection.

$$\widehat{s}(x) = \frac{1}{q} \sum_{j=1}^{q} \widehat{s}_{\Theta_j}(x)$$

- ensemble method (Dietterich, 1999, 2000)
- powerful statistical learning algorithm, for both classification and regression.

Bagging ("bootstrap aggregating")

- Bootstrap (Efron, 1979): draw n i.i.d. r.v., uniform over $\{(X_i, Y_i) \,/\, i = 1, \ldots, n\}$ (sampling with replacement) $\Rightarrow$ resample D_n^b
- Bootstrapping a tree: $\widehat{s}^{b}_{\mathrm{tree}} = \widehat{s}_{\mathrm{tree}}(D_n^b)$
- Bagging: bootstrap (q independent resamples) then aggregation

$$\widehat{s}_{\mathrm{bagging}}(x) = \frac{1}{q} \sum_{j=1}^{q} \widehat{s}^{b,j}_{\mathrm{tree}}(x)$$

Random Forest-Random Inputs (Breiman, 2001)

Definition (RI tree)

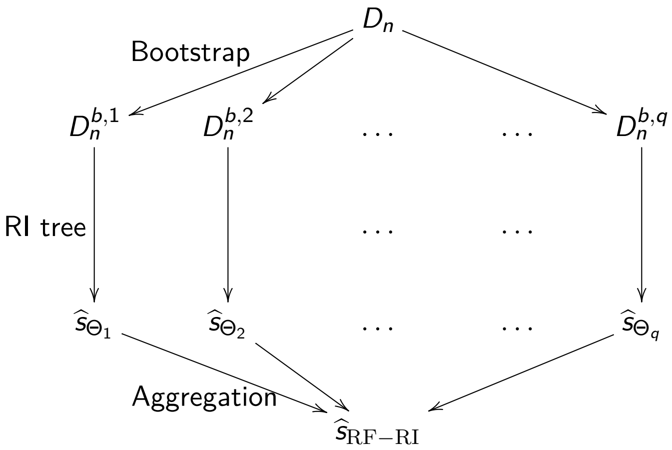
In a RI tree, at each node, `mtry` variables are randomly chosen. Then, the best cut direction is chosen only among the chosen variables.

Definition (Random forest RI)

A random forest RI (RF-RI) is obtained by aggregating RI trees built on independent bootstrap resamples.

$$\text{RF-RI} \quad \Leftrightarrow \quad \text{bagging on RI trees}$$

Random Forest-Random Inputs



Theoretical results on RF-RI

- Few theoretical results on Breiman's original RF-RI
- Most results:
  - focus on a specific part of the algorithm (resampling, split criterion),
  - modify the algorithm (eg, subsampling instead of resampling)
  - make strong assumptions on $s^\star$
- References (see survey paper by Biau and Scornet, 2016): Mentch & Hooker (2014), Scornet, Biau & Vert (2015), Wager & Athey (2015), ...

Theoretical results on RF-RI

- Few theoretical results on Breiman's original RF-RI
- Most results:
  - focus on a specific part of the algorithm (resampling, split criterion),
  - modify the algorithm (eg, subsampling instead of resampling)
  - make strong assumptions on $s^\star$
- References (see survey paper by Biau and Scornet, 2016): Mentch & Hooker (2014), Scornet, Biau & Vert (2015), Wager & Athey (2015), ...

$\Rightarrow$ Here, we consider simplified RF models, for which a precise analysis is possible: purely random forests

Outline

Purely random forests

Definition (Purely random tree)

$$\widehat{s}_{\mathbb{U}}(x) = \sum_{\lambda \in \mathbb{U}} \overline{Y_\lambda}(D_n) \mathbb{1}_{x \in \lambda}$$

where $\overline{Y_\lambda}(D_n)$ is the average of $(Y_i)_{X_i \in \lambda \, , \, (X_i, Y_i) \in D_n}$ and the partition $\mathbb{U}$ is independent from D_n.

Definition (Purely random forest)

$$\widehat{s}(x) = \frac{1}{q} \sum_{j=1}^{q} \widehat{s}_{\mathbb{U}^j}(x)$$

with $\mathbb{U}^1, \ldots, \mathbb{U}^q$ i.i.d., independent from D_n.

Purely random forests

Definition (Purely random forest)

$$\widehat{s}(x) = \frac{1}{q} \sum_{j=1}^{q} \widehat{s}_{\mathbb{U}^j}(x) = \frac{1}{q} \sum_{j=1}^{q} \sum_{\lambda \in \mathbb{U}^j} \overline{Y_\lambda}(D_n) \mathbb{1}_{x \in \lambda}$$

with $\mathbb{U}^1, \ldots, \mathbb{U}^q$ i.i.d., independent from D_n.

Example ("hold-out RF" model): use some extra data D'_n for building the trees: $\mathbb{U}^j = \mathbb{U}_{\mathrm{RI}}(D_n^{\prime \star j})$) (can be done by splitting the sample into two subsamples D_n and D'_n).

Purely random forests

Definition (Purely random forest)

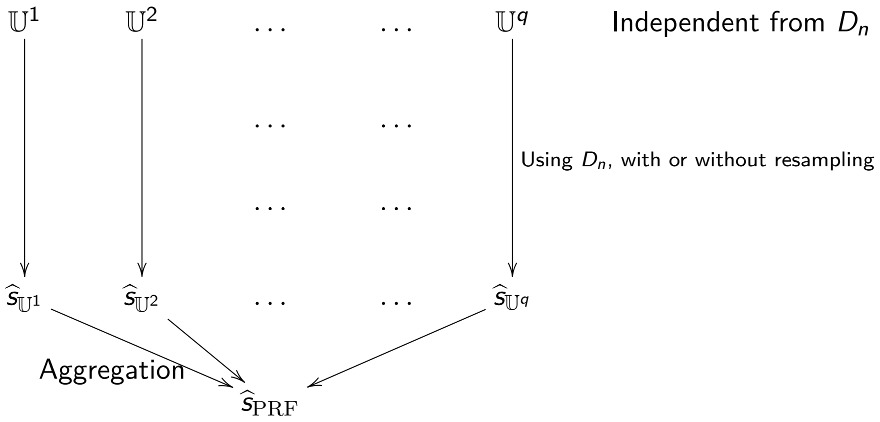
$$\widehat{s}(x) = \frac{1}{q} \sum_{j=1}^{q} \widehat{s}_{\mathbb{U}^j}(x) = \frac{1}{q} \sum_{j=1}^{q} \sum_{\lambda \in \mathbb{U}^j} \overline{Y_\lambda}(D_n) \mathbb{1}_{x \in \lambda}$$

with $\mathbb{U}^1, \ldots, \mathbb{U}^q$ i.i.d., independent from D_n.

Example ("hold-out RF" model): use some extra data D_n' for building the trees: $\mathbb{U}^j = \mathbb{U}_{\mathrm{RI}}(D_n^{\prime \star j})$ (can be done by splitting the sample into two subsamples D_n and D_n').

⚠ From now on, D_n is the sample used for computing the $\overline{Y_\lambda}(D_n)$, and we assume its size is n.

Purely random forests



Purely random forests: theory

- Consistency: Biau, Devroye & Lugosi (2008), Scornet (2014)

- Rates of convergence: Breiman (2004), Biau (2012)
- Some adaptivity to dimension reduction (sparse framework): Biau (2012)

- Forests decrease the estimation error (Biau, 2012; Genuer, 2012)

$\Rightarrow$ What about approximation error?
Almost the same for a forest and a tree?

Risk of a single tree (regressogram)

Given the partition $\mathbb{U}$, regressogram estimator

$$\widehat{s}_{\mathbb{U}}(x) := \sum_{\lambda \in \mathbb{U}} \overline{Y_\lambda} \mathbb{1}_{x \in \lambda}$$

where $\overline{Y_\lambda}$ is the average of $(Y_i)_{X_i \in \lambda}$.

$$\widehat{s}_{\mathbb{U}} \in \operatorname*{argmin}_{f \in S_{\mathbb{U}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \bigl(Y_i - f(X_i) \bigr)^2 \right\}$$

where $S_{\mathbb{U}}$ is the vector space of functions which are constant over each $\lambda \in \mathbb{U}$.

Define:

$$\tilde{s}_{\mathbb{U}}(x) := \sum_{\lambda \in \mathbb{U}} \beta_\lambda \mathbb{1}_{x \in \lambda} \qquad \text{where } \beta_\lambda := \mathbb{E}[s^\star(X) \mid X \in \lambda] \enspace .$$

$\Rightarrow \tilde{s}_{\mathbb{U}} \in \operatorname{argmin}_{f \in S_{\mathbb{U}}} \mathbb{E}\Bigl[\bigl(f(X) - s^\star(X) \bigr)^2 \Bigr]$ and $\tilde{s}_{\mathbb{U}}(x) = \mathbb{E}\bigl[\widehat{s}_{\mathbb{U}}(x) \mid \mathbb{U} \bigr]$

Risk decomposition: single tree

$$\mathbb{E}\left[\left(\widehat{s}_{\mathbb{U}}(X) - s^\star(X)\right)^2\right]$$
$$= \mathbb{E}\left[\left(\tilde{s}_{\mathbb{U}}(X) - s^\star(X)\right)^2\right] + \mathbb{E}\left[\left(\widehat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X)\right)^2\right]$$
$$= \text{Approximation error} + \text{Estimation error}$$

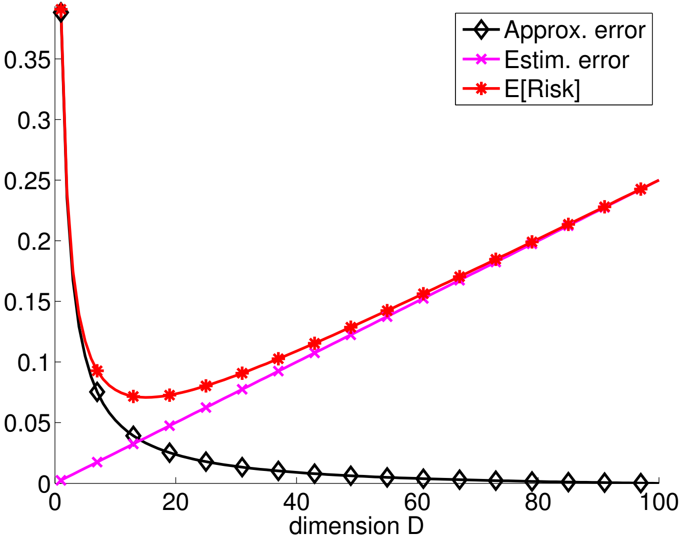
If $s^\star$ is smooth, $X \sim \mathcal{U}([0,1])$ and $\mathbb{U}$ regular partition into D pieces, then

$$\mathbb{E}\left[\left(\tilde{s}_{\mathbb{U}}(X) - s^\star(X)\right)^2\right] \propto \frac{1}{D^2}$$

If $\operatorname{var}(Y \mid X) = \sigma^2$ does not depend on X, then

$$\mathbb{E}\left[\left(\widehat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X)\right)^2\right] \approx \frac{\sigma^2 D}{n}$$

Approximation and estimation errors



Risk decomposition: purely random forest

$(\mathbb{U}^j)_{1 \leqslant j \leqslant q}$ finite partitions, i.i.d. $\sim \mathcal{U}$

Estimator (forest): $\widehat{s}_{\mathbb{U}^{1\cdots q}}(x) := \dfrac{1}{q} \sum_{j=1}^{q} \widehat{s}_{\mathbb{U}^j}(x)$

Ideal forest: $\tilde{s}_{\mathbb{U}^{1\cdots q}}(x) := \dfrac{1}{q} \sum_{j=1}^{q} \tilde{s}_{\mathbb{U}^j}(x) = \mathbb{E}\left[\widehat{s}_{\mathbb{U}^{1\cdots q}}(x) \mid \mathbb{U}^{1\cdots q}\right]$

Quadratic risk decomposition (given $X = x$)

$$\mathbb{E}\left[\left(\widehat{s}_{\mathbb{U}^{1\cdots q}}(x) - s^\star(x)\right)^2\right] = \mathbb{E}\left[\left(\tilde{s}_{\mathbb{U}^{1\cdots q}}(x) - s^\star(x)\right)^2\right] + \mathbb{E}\left[\left(\widehat{s}_{\mathbb{U}^{1\cdots q}}(x) - \tilde{s}_{\mathbb{U}^{1\cdots q}}(x)\right)^2\right]$$

Approximation error: $\mathcal{B}_{\mathcal{U},q}(x) := \mathbb{E}\left[\left(\tilde{s}_{\mathbb{U}^{1\cdots q}}(x) - s^\star(x)\right)^2\right]$

Bias decomposition (given $X = x$)

$$\mathcal{B}_{\mathcal{U},q}(x) = \mathcal{B}_{\mathcal{U},\infty}(x) + \frac{\mathcal{V}_{\mathcal{U}}(x)}{q}$$

$$\text{where} \quad \mathcal{B}_{\mathcal{U},\infty}(x) := \Big(\mathbb{E}\big[\tilde{s}_{\mathbb{U}}(x)\big] - s^{\star}(x)\Big)^2$$

$$\text{and} \quad \mathcal{V}_{\mathcal{U}}(x) := \operatorname{var}\big(\tilde{s}_{\mathbb{U}}(x)\big)$$

$\mathcal{B}_{\mathcal{U},\infty}(x)$ is the approx. error of the infinite forest: $\tilde{s}_{\mathbb{U},\infty}(x) := \mathbb{E}\big[\tilde{s}_{\mathbb{U}}(x)\big]$

to be compared with the approximation error of a single tree

$$\mathcal{B}_{\mathcal{U},1}(x) = \mathcal{B}_{\mathcal{U},\infty}(x) + \mathcal{V}_{\mathcal{U}}(x)$$

Outline

Toy forests in one dimension

Assume: $\mathcal{X} = [0,1)$ and X uniform over $[0,1)$

$\mathbb{U} \sim \mathcal{U}_k^{\mathtt{toy}}$ defined by:

$$\mathbb{U} = \left\{ \left[0, \frac{1-T}{k} \right) , \left[\frac{1-T}{k}, \frac{2-T}{k} \right) , \ldots , \left[\frac{k-T}{k}, 1 \right) \right\}$$

where T has uniform distribution over $[0,1]$.

Interpretation of the ideal infinite forest

Proposition (A. & Genuer, 2014)

For any $x \in \left[\frac{1}{k}, 1 - \frac{1}{k}\right]$, *the ideal infinite forest at* x *satisfies:*

$$\tilde{s}_{\mathbb{U},\infty}(x) = (s^\star * h_k)(x) = \int_0^1 s^\star(t) h_k(x - t) \,\mathrm{d}t$$

where

$$h_k(u) = \begin{cases} k(1 - ku) & \text{if } 0 \leqslant u \leqslant \frac{1}{k} \\ k(1 + ku) & \text{if } -\frac{1}{k} \leqslant u \leqslant 0 \\ 0 & \text{if } |u| \geqslant \frac{1}{k} \end{cases}$$

Interpretation of the ideal infinite forest: proof

$I_{\mathbb{U}}(x) :=$ the interval of $\mathbb{U}$ to which x belongs

$$\tilde{s}_{\mathbb{U}}(x) = \frac{1}{|I_{\mathbb{U}}(x)|} \int_{I_{\mathbb{U}}(x)} s^{\star}(t) \, \mathrm{d}t$$

If $x \in \left[\frac{1}{k}, 1 - \frac{1}{k}\right]$, $\quad I_{\mathbb{U}}(x) = \left[x + \frac{V_x - 1}{k}, x + \frac{V_x}{k}\right)$

where V_x has uniform distribution over $[0, 1]$.

Interpretation of the ideal infinite forest: proof

$I_{\mathbb{U}}(x) :=$ the interval of $\mathbb{U}$ to which x belongs

$$\tilde{s}_{\mathbb{U}}(x) = \frac{1}{|I_{\mathbb{U}}(x)|} \int_{I_{\mathbb{U}}(x)} s^{\star}(t) \, \mathrm{d}t$$

If $x \in \left[\frac{1}{k}, 1 - \frac{1}{k}\right]$, $\quad I_{\mathbb{U}}(x) = \left[x + \frac{V_x - 1}{k}, x + \frac{V_x}{k}\right)$

where V_x has uniform distribution over $[0, 1]$.

$$\begin{aligned}
\tilde{s}_{\mathbb{U},\infty}(x) &= \mathbb{E}_{\mathbb{U}}\left[\tilde{s}_{\mathbb{U}}(x)\right] \\
&= k \int_0^1 s^{\star}(t) \, \mathbb{P}\left(x + \frac{V_x - 1}{k} \leqslant t < x + \frac{V_x}{k}\right) \mathrm{d}t \\
&= k \int_0^1 s^{\star}(t) \, \mathbb{P}\left(k(t - x) < V_x \leqslant k(t - x) + 1\right) \mathrm{d}t
\end{aligned}$$

Analysis of the approximation error

(H2) $s^\star$ twice differentiable over $(0,1)$ and $s^{\star\prime\prime}$ bounded

Taylor-Lagrange formula: for every $t \in (0,1)$, some $c_{t,x} \in (0,1)$ exists such that

$$s^\star(t) - s^\star(x) = s^{\star\prime}(x)(t-x) + \frac{1}{2} s^{\star\prime\prime}(c_{t,x})(t-x)^2$$

Therefore,

$$\begin{aligned}\tilde{s}_{\mathbb{U}}(x) - s^\star(x) &= k \int_{x+\frac{V_x-1}{k}}^{x+\frac{V_x}{k}} \left(s^\star(t) - s^\star(x)\right) \mathrm{d}t \\ &= k\, s^{\star\prime}(x) \int_{x+\frac{V_x-1}{k}}^{x+\frac{V_x}{k}} (t-x) \,\mathrm{d}t + R_1(x) \\ &= \frac{s^{\star\prime}(x)}{k} \left(V_x - \frac{1}{2} \right) + R_1(x)\end{aligned}$$

where $R_1(x) = \frac{k}{2} \int_{x+\frac{V_x-1}{k}}^{x+\frac{V_x}{k}} s^{\star\prime\prime}(c_{t,x})(t-x)^2 \,\mathrm{d}t$

Analysis of the approximation error

$$\left(\mathbb{E}_{\mathbb{U}}\left[\tilde{s}_{\mathbb{U}}(x) - s^{\star}(x)\right]\right)^2 \leqslant \frac{\square}{k^4} \qquad \mathcal{V}_{\mathcal{U}}(x) \underset{k\to+\infty}{\sim} \frac{\square}{k^2}$$

Proposition (A. & Genuer, 2014)

Assuming (H2), for every $x \in \left[\frac{1}{k}, 1 - \frac{1}{k}\right]$,

$$\mathcal{B}_{\mathcal{U}_k^{\mathrm{toy}},1}(x) \underset{k\to+\infty}{\sim} \frac{\square}{k^2} \qquad \mathcal{B}_{\mathcal{U}_k^{\mathrm{toy}},\infty}(x) \leqslant \frac{\square}{k^4}$$

$$\int_{\frac{1}{k}}^{1-\frac{1}{k}} \mathcal{B}_{\mathcal{U}_k^{\mathrm{toy}},1}(x)\,\mathrm{d}x \underset{k\to+\infty}{\sim} \frac{\square}{k^2} \qquad \int_{\frac{1}{k}}^{1-\frac{1}{k}} \mathcal{B}_{\mathcal{U}_k^{\mathrm{toy}},\infty}(x)\,\mathrm{d}x \leqslant \frac{\square}{k^4}$$

Rate k^{-4} is tight assuming:

(H3) $s^{\star}$ three times differentiable over $(0, 1)$ and $s^{\star\prime\prime\prime}$ bounded

Estimation error

General fact (Jensen's inequality):

$$\mathbb{E}\left[\left(\widehat{s}_{\mathbb{U},\infty}(X) - \tilde{s}_{\mathbb{U},\infty}(X)\right)^2\right] \leqslant \mathbb{E}\left[\left(\widehat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X)\right)^2\right]$$

Estimation error

General fact (Jensen's inequality):

$$\mathbb{E}\left[\left(\widehat{s}_{\mathbb{U},\infty}(X) - \tilde{s}_{\mathbb{U},\infty}(X)\right)^2\right] \leqslant \mathbb{E}\left[\left(\widehat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X)\right)^2\right]$$

For the toy forest, without any resampling for computing labels and assuming that $\operatorname{var}(Y|X) = \sigma^2$:

$$\mathbb{E}\left[\left(\widehat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X)\right)^2\right] \approx \frac{\sigma^2 k}{n}$$

$$\mathbb{E}\left[\left(\widehat{s}_{\mathbb{U},\infty}(X) - \tilde{s}_{\mathbb{U},\infty}(X)\right)^2\right] \approx \frac{2}{3}\frac{\sigma^2 k}{n}$$

(A. & Genuer, 2016)

Summary: risk analysis

	Single tree	Infinite forest
	$(q = 1)$	$(q = \infty)$
$\mathbb{E}\left[\left(\widehat{s}_{\mathbb{U}^{1\cdots q}}(x) - s^\star(x)\right)^2\right] \approx$	$\dfrac{c_1(s^\star, x)}{k^2} + \dfrac{\sigma^2 k}{n}$	$\dfrac{c_2(s^\star, x)}{k^4} + \dfrac{2\sigma^2 k}{3n}$

where $c_1(s^\star, x) = \dfrac{s^{\star\prime}(x)^2}{12}$ and $c_2(s^\star, x) = \dfrac{s^{\star\prime\prime}(x)^2}{144}$.

Assumptions:

- $x \in (0, 1)$ far from boundary
- (H3) $s^\star$ three times differentiable over $(0, 1)$ and $s^{\star\prime\prime\prime}$ bounded
- $\mathcal{X}$ uniform over $[0, 1]$
- $\operatorname{var}(Y | X) = \sigma^2$
- no resampling for computing labels

Rates of convergence

Corollary: risk convergence rates (far from boundaries, with $k = k_n^\star$ optimal):

$$\text{Tree} \geqslant \square \, n^{-2/3}$$

$$\text{Infinite forest} \leqslant \square \, n^{-4/5} \quad \Rightarrow \quad \text{minimax } \mathcal{C}^2$$

Remarks:

- $q \geqslant \square \, (k_n^\star)^2$ is sufficient to get an "infinite" forest
- with subsampling a out of n for computing labels: estimation error of a single tree $\frac{\sigma^2 k}{a}$ instead of $\frac{\sigma^2 k}{n}$; no change for infinite forest

Conclusion

- Forests improve the order of magnitude of the approximation error, compared to a single tree
- Estimation error seems to change only by a constant factor (at least for toy forests);
  not contradictory with literature: here, we fix k; different picture if `nodesize` is fixed (+subsampling)
- Randomization:
  randomization of labels seems to have no impact;
  strong impact of randomization of partitions (hold-out RF: both bootstrap and `mtry`)

Approximation error: generalization

- General result on the approximation error under (H2)/(H3): e.g., roughly, if x is centered in its cell (on average over $\mathbb{U}$),

  tree approx. error $\propto \mathcal{M}_2$ infinite forest approx. error $\propto \mathcal{M}_2^2$

  where $\mathcal{M}_2 \approx$ average square distance from x to the boundary of its cell ($\propto k^{-2}$ for toy forests)

Approximation error: generalization

- General result on the approximation error under (H2)/(H3): e.g., roughly, if x is centered in its cell (on average over $\mathbb{U}$),

  tree approx. error $\propto \mathcal{M}_2$ infinite forest approx. error $\propto \mathcal{M}_2^2$

  where $\mathcal{M}_2 \approx$ average square distance from x to the boundary of its cell ($\propto k^{-2}$ for toy forests)

- toy forests in dimension d: approximation error $\propto k^{-2/d}$ vs. $k^{-4/d}$ (infinite forest reaches minimax $\mathcal{C}^2$ rates)
- purely uniformly random forests in dimension 1 (split a random cell, chosen with probability equal to its volume): rates similar to toy forests
- balanced purely random forests (full binary tree, uniform splits) in dimension d: $k^{-\alpha}$ (tree) vs. $k^{-2\alpha}$ (forest) where $\alpha = -\log_2\left(1 - \frac{1}{2d}\right) \Rightarrow$ not minimax rates!

Open problems / future work

- Theory on approximation error of hold-out RF?
  $\Rightarrow$ understand the typical shape of the cell that contains x, for a RI tree
  (x centered on average? square distance to boundary?)

- Theory on estimation error of other models (beyond toy)?
  of hold-out RF?

- Extensive numerical experiments? (other functions $s^\star$, ...)