Analysis of some purely random forests

Sylvain Arlot\textsuperscript{1} (joint work with Robin Genuer\textsuperscript{2})

\textsuperscript{1}Université Paris-Sud
\textsuperscript{2}ISPED, Université Bordeaux 2

TopData Workshop, Banyuls
13 June 2017

arXiv:1407.3939 \hspace{1cm} arXiv:1604.01515
Outline

1. Random forests
2. Purely random forests
3. Toy forests in one dimension
4. Hold-out random forests
Outline

1. Random forests
2. Purely random forests
3. Toy forests in one dimension
4. Hold-out random forests

Analysis of some purely random forests

Sylvain Arlot
Regression: data \((X_1, Y_1), \ldots, (X_n, Y_n)\)
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. $\sim P$)

\[ Y_i = s^*(X_i) + \varepsilon_i \]

with $s^*(X) = \mathbb{E}[Y | X]$ (regression function).
Regression

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

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

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

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

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

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) $D_n$. 

Analysis of some purely random forests

Sylvain Arlot
Regression tree (Breiman et al, 1984)

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

1. Choice of the partition $\mathcal{U}$ (tree structure)
2. For each $\lambda \in \mathcal{U}$ (tree leaf), choice of the estimation $\hat{\beta}_\lambda$ of $s^*(x)$ when $x \in \lambda$.

Here, $\hat{\beta}_\lambda = \overline{Y}_\lambda$ average of the $(Y_i)_{X_i \in \lambda}$.
Random forest (Breiman, 2001)

Definition (Random forest (Breiman, 2001))

\[ \{ \hat{s}_{\Theta_j}, 1 \leq j \leq q \} \] collection of tree predictors, \((\Theta_j)_{1 \leq j \leq q}\) i.i.d. r.v. independent from \(D_n\).

Random forest predictor \(\hat{s}\) obtained by aggregating the tree collection.

\[ \hat{s}(x) = \frac{1}{q} \sum_{j=1}^{q} \hat{s}_{\Theta_j}(x) \]

- 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)
  \[ \Rightarrow \text{resample } D^b_n \]
**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 $D^b_n$  

- Bootstrapping a tree: $\hat{s}^b_{\text{tree}} = \hat{s}_{\text{tree}}(D^b_n)$
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) ⇒ resample $D_n^b$

- Bootstrapping a tree: $\hat{s}_{\text{tree}}^b = \hat{s}_{\text{tree}}(D_n^b)$

- **Bagging**: bootstrap ($q$ independent resamples) then aggregation

$$\hat{s}_{\text{bagging}}(x) = \frac{1}{q} \sum_{j=1}^{q} \hat{s}_{\text{tree}}^{b,j}(x)$$
Definition (RI tree)

In a RI tree, at each node, \texttt{mtry} variables are randomly chosen. Then, the best cut direction is chosen only among the chosen variables.
Random Forest—Random Inputs (Breiman, 2001)

**Definition (RI tree)**

In a RI tree, at each node, \textit{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.
Random Forest - Random Inputs (Breiman, 2001)

Definition (RI tree)

In a RI tree, at each node, \textit{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.

\[ \text{RF-RI} \iff \text{bagging on RI trees} \]
Random Forest-Random Inputs

Random forests

Purely random forests

Toy forests

Hold-out random forests

Conclusion

Analysis of some purely random forests

Sylvain Arlot
Example of application of random forests: Kinect

Depth image $\Rightarrow$ depth comparison features at each pixel

$\Rightarrow$ body part at each pixel $\Rightarrow$ body part positions $\Rightarrow \cdots$

Figures from Shotton et al (2011)
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 \( s^* \)

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 $s^*$


⇒ Here, we consider simplified RF models, for which a precise analysis is possible: purely random forests
Outline

1. Random forests
2. Purely random forests
3. Toy forests in one dimension
4. Hold-out random forests

Analysis of some purely random forests
Purely random forests

Definition (Purely random tree)

\[ \hat{s}_U(x) = \sum_{\lambda \in U} \overline{Y_\lambda}(D_n) \mathbf{1}_{x \in \lambda} \]

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 \( U \) is independent from \( D_n \).

Definition (Purely random forest)

\[ \hat{s}(x) = \frac{1}{q} \sum_{j=1}^{q} \hat{s}_{U_j}(x) \]

with \( U^1, \ldots, U^q \) i.i.d., independent from \( D_n \).
Purely random forests

**Definition (Purely random forest)**

\[
\hat{s}(x) = \frac{1}{q} \sum_{j=1}^{q} \hat{s}_{U_j}(x) = \frac{1}{q} \sum_{j=1}^{q} \sum_{\lambda \in U_j} \overline{Y_\lambda}(D_n) \mathbb{1}_{x \in \lambda}
\]

with \( U^1, \ldots, U^q \) i.i.d., independent from \( D_n \).

Example (“hold-out RF” model): (random) split of the sample into \( D_n \) (used for defining the labels \( \overline{Y_\lambda} \)) and \( D'_n \) (used for building the trees \( U^j = U_{RI}(D'_n^{*j}) \)).
**Definition (Purely random forest)**

\[
\hat{s}(x) = \frac{1}{q} \sum_{j=1}^{q} \hat{s}_{U^j}(x) = \frac{1}{q} \sum_{j=1}^{q} \sum_{\lambda \in U^j} \bar{Y}_\lambda(D_n) \mathbb{1}_{x \in \lambda}
\]

with \( U^1, \ldots, U^q \) i.i.d., independent from \( D_n \).

Example ("hold-out RF" model): (random) split of the sample into \( D_n \) (used for defining the labels \( \bar{Y}_\lambda \)) and \( D'_n \) (used for building the trees \( U^j = U_{RI}(D'_n \ast^j) \)).

⚠️ From now on, \( D_n \) is the sample used for computing the \( \bar{Y}_\lambda(D_n) \), and we assume its size is \( n \).
Purely random forests

\[ \mathbb{U}^1 \downarrow \mathbb{U}^2 \downarrow \cdots \downarrow \cdots \downarrow \mathbb{U}^q \]

Independent from \( D_n \)

Using \( D_n \), with or without resampling

Aggregation

\[ \hat{s}_{\mathbb{U}^1} \rightarrow \hat{s}_{\mathbb{U}^2} \rightarrow \cdots \rightarrow \hat{s}_{\mathbb{U}^q} \rightarrow \hat{s}_{\text{RF-RI}} \]
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)
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)

⇒ 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

$$\hat{s}_{\mathbb{U}}(x) := \sum_{\lambda \in \mathbb{U}} \frac{\overline{Y}_\lambda 1_{x \in \lambda}}{}$$

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

$$\hat{s}_{\mathbb{U}} \in \arg\min_{f \in S_{\mathbb{U}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 \right\}$$

where $S_{\mathbb{U}}$ is the vector space of functions which are constant over each $\lambda \in \mathbb{U}$. 
Risk of a single tree (regressogram)

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

$$\hat{s}_{\mathcal{U}}(x) := \sum_{\lambda \in \mathcal{U}} \overline{Y}_{\lambda} \mathbf{1}_{x \in \lambda}$$

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

$$\hat{s}_{\mathcal{U}} \in \arg\min_{f \in S_{\mathcal{U}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 \right\}$$

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

Define:

$$\tilde{s}_{\mathcal{U}}(x) := \sum_{\lambda \in \mathcal{U}} \beta_{\lambda} \mathbf{1}_{x \in \lambda}$$

where $\beta_{\lambda} := \mathbb{E}[s^*(X) | X \in \lambda]$.

$\Rightarrow \tilde{s}_{\mathcal{U}} \in \arg\min_{f \in S_{\mathcal{U}}} \mathbb{E}\left[ (f(X) - s^*(X))^2 \right]$ and $\tilde{s}_{\mathcal{U}}(x) = \mathbb{E}[\hat{s}_{\mathcal{U}}(x) | \mathcal{U}]$. 

Analysis of some purely random forests

Sylvain Arlot
Risk decomposition: single tree

\[ \mathbb{E}\left[ (\hat{s}_U(X) - s^*(X))^2 \right] \]
\[ = \mathbb{E}\left[ (\tilde{s}_U(X) - s^*(X))^2 \right] + \mathbb{E}\left[ (\hat{s}_U(X) - \tilde{s}_U(X))^2 \right] \]
\[ = \text{Approximation error} + \text{Estimation error} \]
Risk decomposition: single tree

\[
\mathbb{E}\left[ (\hat{s}_U(X) - s^*(X))^2 \right] \\
= \mathbb{E}\left[ (\tilde{s}_U(X) - s^*(X))^2 \right] + \mathbb{E}\left[ (\hat{s}_U(X) - \tilde{s}_U(X))^2 \right] \\
= \text{Approximation error} + \text{Estimation error}
\]

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

\[
\mathbb{E}\left[ (\tilde{s}_U(X) - s^*(X))^2 \right] \propto \frac{1}{K^2}
\]
Risk decomposition: single tree

\[
\mathbb{E}\left[(\hat{s}_U(X) - s^*(X))^2\right] \\
= \mathbb{E}\left[(\hat{s}_U(X) - s^*(X))^2\right] + \mathbb{E}\left[(\hat{s}_U(X) - \tilde{s}_U(X))^2\right] \\
= \text{Approximation error} + \text{Estimation error}
\]

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

\[
\mathbb{E}\left[(\tilde{s}_U(X) - s^*(X))^2\right] \propto \frac{1}{K^2}
\]

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

\[
\mathbb{E}\left[(\hat{s}_U(X) - \tilde{s}_U(X))^2\right] \approx \frac{\sigma^2 K}{n}
\]
Approximation and estimation errors

Approx. error
Estim. error
E[Risk]
Risk decomposition: purely random forest

$$\left( \mathcal{U}^j \right)_{1 \leq j \leq q} \text{ finite partitions, i.i.d. } \sim \mathcal{U}$$

Estimator (forest): $$\tilde{s}_{\mathcal{U}^1 \cdots q}(x) := \frac{1}{q} \sum_{j=1}^{q} \hat{s}_{\mathcal{U}^j}(x)$$

Ideal forest: $$\check{s}_{\mathcal{U}^1 \cdots q}(x) := \frac{1}{q} \sum_{j=1}^{q} \check{s}_{\mathcal{U}^j}(x) = \mathbb{E} \left[ \hat{s}_{\mathcal{U}^1 \cdots q}(x) \mid \mathcal{U}^1 \cdots q \right]$$
Risk decomposition: purely random forest

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

Estimator (forest):
\[
\hat{s}_{U^1\cdots q}(x) := \frac{1}{q} \sum_{j=1}^{q} \hat{s}_{U^j}(x)
\]

Ideal forest:
\[
\tilde{s}_{U^1\cdots q}(x) := \frac{1}{q} \sum_{j=1}^{q} \tilde{s}_{U^j}(x) = \mathbb{E}[\hat{s}_{U^1\cdots q}(x) | U^1\cdots q]
\]

Quadratic risk decomposition (given \(X = x\))

\[
\mathbb{E} \left[ (\hat{s}_{U^1\cdots q}(x) - s^*(x))^2 \right] = \mathbb{E} \left[ (\hat{s}_{U^1\cdots q}(x) - s^*(x))^2 \right]
\]
\[
+ \mathbb{E} \left[ (\hat{s}_{U^1\cdots q}(x) - \tilde{s}_{U^1\cdots q}(x))^2 \right]
\]
Risk decomposition: purely random forest

\[(U^j)_{1 \leq j \leq q}\] finite partitions, i.i.d. \(\sim \mathcal{U}\)

Estimator (forest): \(\hat{s}_{U^1 \ldots U^q}(x) := \frac{1}{q} \sum_{j=1}^q \hat{s}_{U^j}(x)\)

Ideal forest: \(\tilde{s}_{U^1 \ldots U^q}(x) := \frac{1}{q} \sum_{j=1}^q \tilde{s}_{U^j}(x) = \mathbb{E}[\hat{s}_{U^1 \ldots U^q}(x) \mid U^1 \ldots U^q]\)

**Quadratic risk decomposition (given \(X = x\))**

\[
\mathbb{E}\left[ (\hat{s}_{U^1 \ldots U^q}(x) - s^*(x))^2 \right] = \mathbb{E}\left[ (\tilde{s}_{U^1 \ldots U^q}(x) - s^*(x))^2 \right] \\
+ \mathbb{E}\left[ (\hat{s}_{U^1 \ldots U^q}(x) - \tilde{s}_{U^1 \ldots U^q}(x))^2 \right]
\]

Bias term (approximation error):
\[B_{U,q}(x) := \mathbb{E}\left[ (\tilde{s}_{U^1 \ldots U^q}(x) - s^*(x))^2 \right]\]
Bias decomposition (given $X = x$)

$$B_{U,q}(x) = B_{U,\infty}(x) + \frac{V_U(x)}{q}$$

where

$$B_{U,\infty}(x) := \left( \mathbb{E}[\tilde{s}_U(x)] - s^*(x) \right)^2$$

and

$$V_U(x) := \text{var}(\tilde{s}_U(x))$$

$B_{U,\infty}(x)$ is the bias of the infinite forest: $\tilde{s}_{U,\infty}(x) := \mathbb{E}[\tilde{s}_U(x)]$

to be compared with the bias of a single tree

$$B_{U,1}(x) = B_{U,\infty}(x) + V_U(x)$$
Outline

1. Random forests
2. Purely random forests
3. Toy forests in one dimension
4. Hold-out random forests
Toy forests in one dimension

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

\( U \sim U_{k}^{\text{toy}} \) defined by:

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

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:

\[
\tilde{s}_{U,\infty}(x) = (s^* \ast h_k)(x) = \int_0^1 s^*(t) h_k(x - t) \, dt
\]

where

\[
h_k(u) = \begin{cases} 
  k(1 - ku) & \text{if } 0 \leq u \leq \frac{1}{k} \\
  k(1 + ku) & \text{if } -\frac{1}{k} \leq u \leq 0 \\
  0 & \text{if } |u| \geq \frac{1}{k}
\end{cases}
\]
Interpretation of the ideal infinite forest: proof

\[ I_U(x) := \text{the interval of } U \text{ to which } x \text{ belongs} \]

\[ \tilde{s}_U(x) = \frac{1}{|I_U(x)|} \int_{I_U(x)} s^*(t) \, dt \]

If \( x \in \left[ \frac{1}{k}, 1 - \frac{1}{k} \right] \),
\[ I_U(x) = \left[ x + \frac{V_x - 1}{k}, x + \frac{V_x}{k} \right) \]

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

\[
\tilde{s}_\mathbb{U}(x) = \frac{1}{|I_\mathbb{U}(x)|} \int_{I_\mathbb{U}(x)} s^*(t) \, dt
\]

If \( x \in \left[ \frac{1}{k}, 1 - \frac{1}{k} \right] \),

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

where \( V_x \) has uniform distribution over \([0, 1] \).

\[
\tilde{s}_{\mathbb{U}, \infty}(x) = E_\mathbb{U}[\tilde{s}_\mathbb{U}(x)]
\]

\[
= k \int_0^1 s^*(t) \mathbb{P}\left( x + \frac{V_x - 1}{k} \leq t < x + \frac{V_x}{k} \right) \, dt
\]

\[
= k \int_0^1 s^*(t) \mathbb{P}(k(t - x) < V_x \leq k(t - x) + 1) \, dt
\]
Analysis of the approximation error

(H2) \( s^* \) twice differentiable over \((0,1)\) and \( s^{*''} \) bounded

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

\[
\begin{align*}
    s^*(t) - s^*(x) &= s^*(x)(t - x) + \frac{1}{2} s^{*''}(c_{t,x})(t - x)^2 \\
\end{align*}
\]
Analysis of the approximation error

(H2) \( s^* \) twice differentiable over \((0, 1)\) and \( s^{*''} \) bounded

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

\[
s^*(t) - s^*(x) = s^\prime(x)(t - x) + \frac{1}{2} s^{*''}(c_{t,x})(t - x)^2
\]

Therefore,

\[
\tilde{s}_U(x) - s^*(x) = k \int_{x + \frac{V_x - 1}{k}}^{x + \frac{V_x}{k}} (s^*(t) - s^*(x)) \, dt
\]

\[
= k \, s^\prime(x) \int_{x + \frac{V_x - 1}{k}}^{x + \frac{V_x}{k}} (t - x) \, dt + R_1(x)
\]

\[
= \frac{s^\prime(x)}{k} \left( V_x - \frac{1}{2} \right) + R_1(x)
\]

where \( R_1(x) = \frac{k}{2} \int_{x + \frac{V_x - 1}{k}}^{x + \frac{V_x}{k}} s^{*''}(c_{t,x})(t - x)^2 \, dt \)
Analysis of the approximation error

\[
\left( \mathbb{E}_{U} \left[ \tilde{s}_{U}(x) - s^{*}(x) \right] \right)^2 \leq \frac{\Box}{k^4} \quad \nu_{U}(x) \sim \frac{\Box}{k^2}
\]

Proposition (A. & Genuer, 2014)

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

\[
\mathcal{B}_{U_{k}^{\text{toy}}, 1}(x) \sim \frac{\Box}{k^2} \quad \mathcal{B}_{U_{k}^{\text{toy}}, \infty}(x) \leq \frac{\Box}{k^4}
\]

\[
\int_{\frac{1}{k}}^{1 - \frac{1}{k}} \mathcal{B}_{U_{k}^{\text{toy}}, 1}(x) \, dx \sim \frac{\Box}{k^2} \quad \int_{\frac{1}{k}}^{1 - \frac{1}{k}} \mathcal{B}_{U_{k}^{\text{toy}}, \infty}(x) \, dx \leq \frac{\Box}{k^4}
\]
Analysis of the approximation error

\[
\left( \mathbb{E}_U [\tilde{s}_U(x) - s^*(x)] \right)^2 \leq \frac{\Box}{k^4} \quad \forall U(x) \quad k \to +\infty \quad \frac{\Box}{k^2}
\]

Proposition (A. & Genuer, 2014)

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

\[
\mathcal{B}_{U_k}^{\text{toy},1}(x) \quad k \to +\infty \quad \frac{\Box}{k^2} \quad \mathcal{B}_{U_k}^{\text{toy},\infty}(x) \leq \frac{\Box}{k^4}
\]

\[
\int_{\frac{1}{k}}^{1-\frac{1}{k}} \mathcal{B}_{U_k}^{\text{toy},1}(x) \, dx \quad k \to +\infty \quad \frac{\Box}{k^2} \quad \int_{\frac{1}{k}}^{1-\frac{1}{k}} \mathcal{B}_{U_k}^{\text{toy},\infty}(x) \, dx \leq \frac{\Box}{k^4}
\]

Rate \( k^{-4} \) is tight assuming:

(H3) \quad \( s^* \) three times differentiable over \((0,1)\) and \( s^{*'''} \) bounded.
Estimation error

General fact (Jensen's inequality):

\[ \mathbb{E} \left[ (\hat{s}_U, \infty(X) - \tilde{s}_U, \infty(X))^2 \right] \leq \mathbb{E} \left[ (\hat{s}_U(X) - \tilde{s}_U(X))^2 \right] \]
Estimation error

General fact (Jensen’s inequality):

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

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

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

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

(A. & Genuer, 2016)
Summary: risk analysis

\[ \mathbb{E} \left[ (\widehat{s}_{U1\ldots q}(x) - s^*(x))^2 \right] \approx \frac{c_1(s^*, x)}{k^2} + \frac{\sigma^2 k}{n} + \frac{c_2(s^*, x)}{k^4} + \frac{2\sigma^2 k}{3n} \]

\[ \text{where } c_1(s^*, x) = \frac{s''(x)^2}{12} \quad \text{and} \quad c_2(s^*, x) = \frac{s''''(x)^2}{144}. \]

Assumptions:
- \( x \in (0, 1) \) far from boundary
- (H3) \( s^* \) three times differentiable over \( (0, 1) \) and \( s'''' \) bounded
- \( \mathcal{X} \) uniform over \([0, 1]\)
- \( \text{var}(Y|X) = \sigma^2 \)
- no resampling for computing labels
Rates of convergence

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

$$\text{Tree} \geq \Box n^{-2/3}$$

$$\text{Infinite forest} \leq \Box n^{-4/5} \implies \text{minimax } C^2$$
Rates of convergence

Corollary: risk convergence rates (far from boundaries, with $k = k^*_n$ optimal):

$$\text{Tree} \geq \Box n^{-2/3}$$

$$\text{Infinite forest} \leq \Box n^{-4/5} \Rightarrow \text{minimax } C^2$$

Remarks:

- $q \geq \Box (k^*_n)^2$ is sufficient to get an “infinite” forest
Corollary: risk convergence rates (far from boundaries, with $k = k_n^*$ optimal):

\[
\text{Tree} \geq \square n^{-2/3}
\]
\[
\text{Infinite forest} \leq \square n^{-4/5} \implies \text{minimax } C^2
\]

Remarks:
- $q \geq \square (k_n^*)^2$ 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
Outline

1. Random forests
2. Purely random forests
3. Toy forests in one dimension
4. Hold-out random forests
Definition (Biau, 2012)

Split $D_n$ into $D_{n_1}$ and $D_{n_2}$

$\mathbb{U}^1 \downarrow \mathbb{U}^2 \downarrow \cdots \downarrow \mathbb{U}^q$

RI partitions, using $D_{n_1}$

Using $D_{n_2}$, no resampling here

$\hat{s}_{\mathbb{U}^1} \downarrow \hat{s}_{\mathbb{U}^2} \downarrow \cdots \downarrow \hat{s}_{\mathbb{U}^q}$

Aggregation

$\hat{s}_{RF-RI}$

$\Rightarrow$ purely random forest
Numerical experiments: framework

- Data generation:
  \[ X_i \sim \mathcal{U}([0, 1]^d) \quad Y_i = s^*(X_i) + \varepsilon_i \]
  \[ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \sigma^2 = 1/16 \]

\[ s^*: \mathbf{x} \in [0, 1]^d \mapsto \frac{1}{10} \times \left[ 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 \right] . \]

- Data split: \( n_1 = 1280 \quad n_2 = 25600 \)

- Forests definition:
  \( \text{nodesize} = 1 \)
  \( k \in \{ 2^5, 2^6, 2^7, 2^8 \} \)
  \( \text{“Large” forests are made of } q = k \text{ trees.} \)

- Compute integrated approximation/estimation errors
### Numerical experiments: results ($d = 5$)

<table>
<thead>
<tr>
<th></th>
<th>Single tree</th>
<th>Large forest</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No bootstrap</strong></td>
<td>$\frac{0.13}{k^{0.17}} + \frac{1.04\sigma^2 k}{n_2}$</td>
<td>$\frac{0.13}{k^{0.17}} + \frac{1.04\sigma^2 k}{n_2}$</td>
</tr>
<tr>
<td>$mtry = d$</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Bootstrap</strong></td>
<td>$\frac{0.14}{k^{0.17}} + \frac{1.06\sigma^2 k}{n_2}$</td>
<td>$\frac{0.15}{k^{0.29}} + \frac{0.08\sigma^2 k}{n_2}$</td>
</tr>
<tr>
<td>$mtry = d$</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>No bootstrap</strong></td>
<td>$\frac{0.23}{k^{0.19}} + \frac{1.01\sigma^2 k}{n_2}$</td>
<td>$\frac{0.06}{k^{0.31}} + \frac{0.06\sigma^2 k}{n_2}$</td>
</tr>
<tr>
<td>$mtry = \lfloor d/3 \rfloor$</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Bootstrap</strong></td>
<td>$\frac{0.25}{k^{0.20}} + \frac{1.02\sigma^2 k}{n_2}$</td>
<td>$\frac{0.06}{k^{0.34}} + \frac{0.05\sigma^2 k}{n_2}$</td>
</tr>
<tr>
<td>$mtry = \lfloor d/3 \rfloor$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Numerical experiments: results ($d = 10$)

<table>
<thead>
<tr>
<th></th>
<th>Single tree</th>
<th>Large forest</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No bootstrap</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$mtry = d$</td>
<td>$0.11k^{0.12} + \frac{1.03\sigma^2 k}{n_2}$</td>
<td>$0.11k^{0.12} + \frac{1.03\sigma^2 k}{n_2}$</td>
</tr>
<tr>
<td><strong>Bootstrap</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$mtry = d$</td>
<td>$0.11k^{0.11} + \frac{1.05\sigma^2 k}{n_2}$</td>
<td>$0.10k^{0.19} + \frac{0.04\sigma^2 k}{n_2}$</td>
</tr>
<tr>
<td><strong>No bootstrap</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$mtry = \lfloor d/3 \rfloor$</td>
<td>$0.21k^{0.18} + \frac{1.08\sigma^2 k}{n_2}$</td>
<td>$0.08k^{0.25} + \frac{0.04\sigma^2 k}{n_2}$</td>
</tr>
<tr>
<td><strong>Bootstrap</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$mtry = \lfloor d/3 \rfloor$</td>
<td>$0.20k^{0.16} + \frac{1.05\sigma^2 k}{n_2}$</td>
<td>$0.07k^{0.26} + \frac{0.03\sigma^2 k}{n_2}$</td>
</tr>
</tbody>
</table>
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).
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 \( \mathcal{U} \)),

  \[
  \text{tree approx. error } \propto M_2 ^2 \quad \text{infinite forest approx. error } \propto M_2 ^2
  \]

  where \( M_2 \approx \text{average square distance from } x \text{ to the boundary of its cell} (\propto k^{-2} \text{ for toy forests}) \)
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 \( \mathbb{U} \)),

  \[
  \text{tree approx. error } \propto M_2 \quad \text{infinite forest approx. error } \propto M_2^2
  \]

  where \( M_2 \approx \) 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 \( C^2 \) rates)
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 $\mathcal{U}$),

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

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 $C^2$ rates)

- purely uniformly random forests in dimension 1 (split a random cell, chosen with probability equal to its volume): rates similar to toy forests
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 \( \mathbb{U} \)),

  \[
  \text{tree approx. error } \propto M_2 \quad \text{infinite forest approx. error } \propto M_2^2
  \]

  where \( M_2 \approx \text{average square distance from } x \text{ to the boundary of its cell} \ (\propto k^{-2} \text{ for toy forests}) \)

- toy forests in dimension \( d \): approximation error \( \propto k^{-2/d} \) vs. \( k^{-4/d} \) (infinite forest reaches minimax \( C^2 \) 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 \( \alpha = -\log_2 \left( 1 - \frac{1}{2d} \right) \Rightarrow \text{not minimax rates!} \)
Open problems / future work

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

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

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