Variance reduction in purely random forests

Robin Genuer$^{a,b,*}$

$^a$University of Bordeaux, ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Bordeaux, F-33000, France; $^b$INSERM, ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Bordeaux, F-33000, France

(Received 21 December 2010; final version received 16 March 2012)

Random forests (RFs), introduced by Leo Breiman in 2001, are a very effective statistical method. The complex mechanism of the method makes theoretical analysis difficult. Therefore, simplified versions of RF, called purely RFs (PRF), which can be theoretically handled more easily, have been considered. In this paper, we study the variance of such forests. First, we show a general upper bound which emphasises the fact that a forest reduces the variance. We then introduce a simple variant of PRFs, that we call purely uniformly RFs. For this variant and in the context of regression problems with a one-dimensional predictor space, we show that both random trees and RFs reach minimax rate of convergence. In addition, we prove that compared with random trees, RFs improve accuracy by reducing the estimator variance by a factor of three-fourths.

Keywords: random forests; nonparametric regression; rates of convergence; randomisation; ensemble methods

1. Introduction

Random forests (RFs), introduced in Breiman (2001), are a very effective statistical method. They have outstanding performances in several situations for both regression and classification problems. Mathematical understanding of these good performances remains quite unknown. As defined in Breiman (2001), a RF is a collection of tree-predictors \( \{ h(x, \Theta_l), 1 \leq l \leq q \} \), where \( (\Theta_l)_{1 \leq l \leq q} \) are i.i.d. random vectors and a RF predictor is obtained by aggregating this collection of trees. In addition to consistency results, one of the main theoretical challenges is to explain why a RF improves so much the performance of a single tree.

Breiman (2001) introduced a specific instance of RF, called RFs-random inputs (RFRI), which has been adopted in many fields as a reference method. Indeed, RFRI are simple to use, and are efficiently coded in the popular R-package randomForest (Liaw and Wiener 2002). They are effective for a predictive goal and they can also be used for variable selection (see e.g. Díaz-Uriarte and Alvarez de Andrés 2006; Genuer, Poggi and Tuleau 2010).

*Email: robin.genuer@isped.u-bordeaux2.fr

ISSN 1048-5252 print/ISSN 1029-0311 online
© American Statistical Association and Taylor & Francis 2012
http://dx.doi.org/10.1080/10485252.2012.677843
http://www.tandfonline.com
However, RFRI are very difficult to handle theoretically. This is why people are interested in simplified versions, called purely random forests (PRF). The main difference is that in PRF, the splits of tree nodes are randomly drawn independent of the learning sample, while in the random RFRI, the splits are optimised using the learning sample. This independence between splits and learning sample makes mathematical analysis easier. Cutler and Zhao (2001) introduced Perfect Random Tree Ensemble, an algorithm which builds some PRFs, and illustrated its good performance on benchmark data sets. More recently, Biau, Devroye and Lugosi (2008) showed that both purely random trees (PRTs) and PRFs are universally consistent.

In this paper, we essentially focus on the variance of PRFs. First, we show a general upper bound on the forest variance, which emphasises the fact that a forest does reduce the variance. This upper bound relies on a quantity depending only on the distribution of the splits of the trees. Roughly, if this quantity is small compared with the number of splits of the trees, then the forest variance is small compared with the tree variance. Let us mention here the very insightful analysis of Biau (2012), who also prove a result of variance reduction (for another kind of random forest).

Then, our paper offers to examine a simple variant of PRFs. We call it purely uniformly RFs (PURF) and we analyse its risk, only in a regression framework with a one-dimensional predictor space. The main goal is to emphasise the gain of using a forest instead of a tree. The results obtained for PURF are twofold: first, we show that both purely uniformly random trees (PURT) and forests risks reach minimax rate of convergence on the Lipschitz functions class; second, we show that forests improve the variance term by a factor of three-fourths while not increasing the bias.

The paper is organised as follows. Section 2 presents the model. Section 3 presents a general upper bound for PRFs variance. Section 4 gives some risk bounds for PURT and forests. In Section 5, we lead a simulation study illustrating our results and compare PURF and RFRI. Section 6 concludes the paper, while proofs are collected in Section 7.

2. Framework

The framework we consider all along the paper is the classical random design regression framework.

More precisely, consider a learning set \( \mathcal{L}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \) made of \( n \) i.i.d. observations of a vector (\( X, Y \)) from an unknown distribution. \( Y \) is real-valued since we are in a regression framework. \( X \in \mathcal{X} \) is a measurable space (e.g. \( \mathcal{X} = \mathbb{R}^d \) with any \( d \geq 1 \)). We consider the following statistical model:

\[
Y_i = s(X_i) + \varepsilon_i \quad \text{for } i = 1, \ldots, n.
\]

\( s : \mathcal{X} \to \mathbb{R} \) is the unknown regression function and the goal is to estimate \( s \). Finally, we suppose that \((\varepsilon_1, \ldots, \varepsilon_n)\) are i.i.d. observations of \( \varepsilon \) with value in \( \mathbb{R} \), independent of \( \mathcal{L}_n \), with \( \mathbb{E}[\varepsilon] = 0 \) and \( \text{Var}(\varepsilon) = \sigma^2 < +\infty \).

This paper aims at comparing performances in estimating \( s \) using a single random tree or a RF. As a result, we emphasise a variance reduction brought by the forest.

3. A general upper bound for PRFs variance

We begin by giving a general analysis of the PRFs variance.

3.1. PRT definition

Let us first mention that, all along the paper, we make a slight language abuse. Indeed, we refer to random tree, the tree himself (as a graph), the corresponding partition of \( \mathcal{X} \), as well as the corresponding estimator.
Thus, the main difference between RFs and PRFs is that in the purely random case, partitions of the input space $\mathcal{X}$ are drawn randomly, independent of $L_n$. We recall that in classical RFs, partitions of the input space are most of the time obtained by random perturbations of a partitioning scheme where splits are calculated using $L_n$. Hence, for classical RFs, the random perturbations are independent of $L_n$, but the partition is not.

We denote by $\mathbb{U}$ a random partition of $\mathcal{X}$ in $k$ cells with distribution $U$. $k$ is a natural integer which will depend on the number of observations $n$.

A PRT, associated with $\mathbb{U}$, is defined for $x \in \mathcal{X}$ as

$$\tilde{s}_U(x) = \sum_{\lambda \in \mathbb{U}} \hat{\beta}_\lambda 1_{x \in \lambda},$$

where

$$\hat{\beta}_\lambda = \frac{1}{\sharp \{i : X_i \in \lambda\}} \sum_{i : X_i \in \lambda} Y_i$$

with $\sharp \mathcal{E}$ denoting the cardinality of the set $\mathcal{E}$.

**Remark 1**  We mention that if $\sharp \{i : X_i \in \lambda\} = 0$, we set $\hat{\beta}_\lambda = 0$.

In addition, let us define, for $x \in \mathcal{X}$:

$$\tilde{s}_U(x) = \sum_{\lambda \in \mathbb{U}} \beta_\lambda 1_{x \in \lambda},$$

where

$$\beta_\lambda = \mathbb{E}[Y \mid X \in \lambda].$$

Conditionally on $\mathbb{U}$, $\tilde{s}_U$ is the best approximation of $s$ among all the regressograms based on $\mathbb{U}$, but of course it depends on the unknown distribution of $(X, Y)$.

With these notations, we can write a bias-variance decomposition of the quadratic risk of $\tilde{s}_U$ as follows:

$$\mathbb{E}[(\tilde{s}_U(X) - s(X))^2] = \mathbb{E}[(\tilde{s}_U(X) - \tilde{s}_U(X))^2] + \mathbb{E}[(\tilde{s}_U(X) - s(X))^2].$$

$$= \text{variance term} + \text{bias term}$$

(2)

To clarify these variance and bias terms, we emphasise that for a given partition $u$ and a given $x$, we have

$$\mathbb{E}[\tilde{s}_u(x)] = \tilde{s}_u(x)$$

so $\mathbb{E}[(\tilde{s}_u(x) - \tilde{s}_u(x))^2]$ is the variance of the estimator $\tilde{s}_u(x)$ and $\mathbb{E}[(\tilde{s}_u(x) - s(x))^2]$ is its bias. We then integrate with respect to (w.r.t) $X$ and $\mathbb{U}$ to get decomposition (2).

**3.2. PRF definition**

A RF is the aggregation of a collection of random trees. Therefore, in the context of PRFs, the principle is to generate several PRT by drawing several random partitions, and to aggregate them.

Let $\mathbb{V}_q = (\mathbb{U}_1, \ldots, \mathbb{U}_q)$ be $q$ i.i.d. random partitions with distribution $\mathcal{U}$.
A PRF, associated with $\mathcal{V}_q$, is defined for $x \in [0, 1]$ as follows:

$$\hat{s}_{\mathcal{V}_q}(x) = \frac{1}{q} \sum_{l=1}^{q} \hat{s}_U(x).$$

Let us define, for $x \in [0, 1]$:

$$\tilde{s}_{\mathcal{V}_q}(x) = \frac{1}{q} \sum_{l=1}^{q} \tilde{s}_U(x).$$

Again, we have a bias-variance decomposition of the quadratic risk of $\hat{s}_{\mathcal{V}_q}$, given by

$$\mathbb{E}[(\hat{s}_{\mathcal{V}_q}(X) - s(X))^2] = \mathbb{E}[(\tilde{s}_{\mathcal{V}_q}(X) - \tilde{s}_{\mathcal{V}_q}(X))^2] + \mathbb{E}[(\tilde{s}_{\mathcal{V}_q}(X) - s(X))^2].$$

$$= \text{variance term} + \text{bias term} \quad (3)$$

### 3.3. PRT variance

We start to deal with the variance term of Decomposition (2). First, we work conditionally on $\mathcal{U}$, then the problem reduces to the case of a regressogram on a deterministic partition, and we can apply the following proposition which comes from Arlot (2008).

**Proposition 3.1** Conditionally on $\mathcal{U}$, the variance term of Decomposition (2) satisfies:

$$\mathbb{E}[(\hat{s}_U(X) - \tilde{s}_U(X))^2 | \mathcal{U}] = \frac{1}{n} \sum_{\lambda \in \mathcal{U}} (1 + \delta_{n,p_\lambda})(\sigma^2 + \sigma_\lambda^2),$$

$$\quad \text{where}$$

- $p_\lambda = \mathbb{P}(X \in \lambda)$,
- $\sigma_\lambda^2 = \mathbb{E}[(s(X) - \tilde{s}_U(X))^2 | X \in \lambda]$,
- $\delta_{n,p_\lambda} \xrightarrow{n \to +\infty} 0$.

We now integrate Equation (4) w.r.t. $\mathcal{U}$, and we get the following equality:

$$\mathbb{E}[(\hat{s}_U(X) - \tilde{s}_U(X))^2] = \frac{1}{n} \sum_{\lambda \in \mathcal{U}} (\sigma^2 + \sigma^2 \mathbb{E}[\delta_{n,p_\lambda}] + \mathbb{E}[\sigma_\lambda^2] + \mathbb{E}[\sigma_\lambda^2 \delta_{n,p_\lambda}]).$$

We will see in Section 4.1 that in a specific case, the three last terms of Equality (5) are negligible compared with the constant term $\sigma^2$. Hence in this case, the variance of a tree is equivalent to $\sigma^2 k/n$.

We claim, that if $\mathcal{X}$ is bounded, say $\mathcal{X} = [0, 1]^d$, any reasonable distribution of $\mathcal{U}$ may lead to a variance of $\sigma^2 k/n$, as soon as $k \xrightarrow{n \to +\infty} +\infty$ and $(k/n) \xrightarrow{n \to +\infty} 0$. These last conditions on $k$ and $n$ are in fact what is required by Biau et al. (2008) to get the universal consistency of some PRF.

### 3.4. PRF variance upper bound

We now study the variance term of Decomposition (3). We begin to show that when letting the number of trees $q$ grow to infinity, the variance of a PURF is close to the covariance between two PURT.
Indeed, since \( \bar{s}_{V,q}(x) = (1/q) \sum_{t=1}^{q} \hat{s}_{U,t}(x) \), the variance term satisfies:

\[
\mathbb{E}[(\bar{s}_{V,q}(X) - \hat{s}_{V,q}(X))^2] = \frac{1}{q} \sum_{t=1}^{q} \mathbb{E}[(\hat{s}_{U,t}(X) - \hat{s}_{U,t}(X))^2] \\
+ \frac{1}{q} \sum_{t \neq q} \mathbb{E}[(\hat{s}_{U,t}(X) - \hat{s}_{U,t}(X))(\hat{s}_{U,q}(X) - \hat{s}_{U,q}(X))] \\
= \frac{1}{q} \mathbb{E}[(\hat{s}_{U,1}(X) - \hat{s}_{U,1}(X))^2] \\
+ \frac{q(q-1)}{q^2} \mathbb{E}[(\hat{s}_{U,1}(X) - \hat{s}_{U,1}(X))(\hat{s}_{U,2}(X) - \hat{s}_{U,2}(X))],
\]

where the last equality comes from the fact that the \( (\hat{s}_{U,t}(X) - \hat{s}_{U,t}(X))_{1 \leq t \leq q} \) have the same distribution.

Now, if we let \( q \) grow to infinity, we get

\[
\mathbb{E}[(\bar{s}_{V,q}(X) - \hat{s}_{V,q}(X))^2] = \mathbb{E}[(\hat{s}_{U,1}(X) - \hat{s}_{U,1}(X))(\hat{s}_{U,2}(X) - \hat{s}_{U,2}(X))] \left(1 + \frac{o}{q} \rightarrow +\infty \right). \quad (6)
\]

The next step is to upper bound the covariance between two PURT

\[
\mathbb{E}[(\hat{s}_{U,1}(X) - \hat{s}_{U,1}(X))(\hat{s}_{U,2}(X) - \hat{s}_{U,2}(X))].
\]

Let us denote by \( U^1 = \{\lambda_1, \ldots, \lambda_k\} \) and \( U^2 = \{\lambda_{1}', \ldots, \lambda_{k}'\} \) the cells of the partitions respectively associated with the trees \( \hat{s}_{U,1} \) and \( \hat{s}_{U,2} \).

Finally, we state the following theorem, which gives a general upper bound for PRF variance:

**Theorem 3.2**

\[
\mathbb{E}[(\bar{s}(X) - \bar{s}(X))^2] \leq \frac{1}{n} \mathbb{E} \left[ \sum_{t=1}^{N_U} (1 + \delta_{n,\hat{p}_t}) (\sigma_t^2 + \Sigma_t^2) \right] \left(1 + \frac{o}{q} \rightarrow +\infty \right),
\]

where

\[
N_U = k - \sum_{1 \leq r < r' \leq k} \sum_{1 \leq s \leq k} \mathbbm{1}_{\lambda_r, \lambda_r' \subset \lambda_s, \lambda_s' \subset \lambda_r}.
\]

\( \Sigma_t^2 \) denotes a sum of terms \( \mathbb{E}[(\hat{s}_{U,t}(X) - s(X))(\hat{s}_{U,t}(X) - s(X)) | X \in \lambda_t'] \) for several consecutive values of \( t' \) and with \( l = 1 \) or 2.

\( \hat{p}_t \) denotes for some \( j \in \{1, \ldots, k\} \) either \( p_{\lambda_j} \) or \( p_{\lambda_j'} \) depending on the relative positions between the \( (\lambda_1, \ldots, \lambda_k) \) and the \( (\lambda_1', \ldots, \lambda_k') \).

Theorem 3.2 is to be compared with Equality (5) and tells us that the variance of a PRF is upper bounded by a sum of terms of the same kind than those appearing in the variance of a PRT. The actual gain comes from the number of such terms in the sum. Indeed \( N_U \leq k \), and the larger the double sum in the r.h.s of (7), the smaller the \( N_U \).

We stress that quantities appearing in \( N_U \) only depend of the two partitions \( U^1, U^2 \). And our results mean that if \( U^1 \) and \( U^2 \) are different enough, the covariance between \( \hat{s}_{U,1} \) and \( \hat{s}_{U,2} \) will be smaller than a single tree variance.

In Section 4.4, we will see that in a special case, \( \mathbb{E}[N_U] \) equals \( 3k/4 \). This allows to claim that the forest reduces variance by a factor of \( 3/4 \) for this case.
4. Risk bounds for PURFs

We now give a detailed analysis of a specific variant of PRF, in a context of a one-dimensional predictor space. Hence, in this section, we assume $\mathcal{X} = [0, 1]$.

The principle of PURT is that we draw $k$ uniform random variables, which form the partition of the input space $[0, 1]$. Then we build a regressogram on this partition, that we call a tree.

Note that, unlike PRFs or RFRI, the tree structure of individual predictors is not obvious. This comes from the fact that in PURT, the partition is not obtained in a recursive manner. Nevertheless, we keep the vocabulary of trees and forests to distinguish individual predictors from aggregated ones.

More precisely, let $U = (U_1, \ldots, U_k)$ be $k$ i.i.d. random variables of uniform distribution on $[0, 1]$.

4.1. PURT variance

Therefore, using the fact that, in our case, $U$ is made of $k$ i.i.d. random variables of uniform distribution on $[0, 1]$, we deduce from Equation (5) the following proposition:

**Corollary 4.1** If $k \to+\infty$, $(k/n) \to 0$, $\mu > 0$ and $s$ is C-Lipschitz, the variance of a PURT satisfies:

$$\mathbb{E}[(\hat{s}_U(X) - \tilde{s}_U(X))^2] = \frac{\sigma^2(k + 1)}{n} + \frac{k}{n} \to 0,$$

(8)

where the notation $\frac{k}{n} \to 0$ denotes a function $f(n)$ such as $(f(n)/k/n) \to 0$.

Details of the proof of Corollary 4.1 can be found in Section 7.2.

The first two hypotheses of Corollary 4.1 $(k \to+\infty, k/n \to 0)$ are the same natural conditions found by Biau et al. (2008) for consistency of PRF. They guarantee that the number of splits of the tree must grow to infinity but slower than the number of samples.

4.2. PURT bias

We now turn to the bias term of Decomposition (2). Direct calculations (see Section 7.3 for details) lead to the following upper bound for the bias term of a PURT:

**Proposition 4.2** If $\mu$ is bounded by $M > 0$ and $s$ is C-Lipschitz, the bias of a PURT is upper bounded by

$$\mathbb{E}[(\tilde{s}_U(X) - s(X))^2] \leq \frac{6MC^2}{(k + 1)^2}.$$  

(9)

4.3. Risk bounds for PURT

Putting together (8) and (9) leads to the following risk bound for a PURT.

**Proposition 4.3** If $k \to+\infty$, $(k/n) \to 0$, $0 < \mu \leq M$ and $s$ is C-Lipschitz, the risk of a PURT satisfies:

$$\mathbb{E}[(\hat{s}_U(X) - s(X))^2] \leq \frac{\sigma^2(k + 1)}{n} + \frac{6MC^2}{(k + 1)^2} + \frac{k}{n} \to 0.$$  

(10)
The balance between the two first terms of the right-hand side (r.h.s.) of (10) leads to take 
\((k + 1) = n^{1/3}\), and gives the following upper bound for the risk of a PURT.

**Corollary 4.4** Under the assumptions of Theorem 4.3,

\[
\mathbb{E}[(\hat{s}_U(X) - s(X))^2] \leq Kn^{-2/3} + o\left(\frac{n^{-2/3}}{n \to +\infty}\right),
\]

where \(K\) is a positive constant.

Therefore, a PURT reaches the minimax rate of convergence associated with the class of Lipschitz functions (see e.g. Ibragimov and Khasminskii 1981).

Let us now analyse PURFs.

### 4.4. PURF variance

From Theorem 3.2, we deduce the following proposition:

**Proposition 4.5** If \(k \xrightarrow{n \to +\infty} +\infty\), \((k/n) \xrightarrow{n \to +\infty} 0\), \(\mu > 0\), \(s\) is C-Lipschitz and \(q \xrightarrow{n \to +\infty} +\infty\), the variance of a PURF satisfies the following upper bound:

\[
\mathbb{E}[(\hat{s}(X) - \tilde{s}(X))^2] \leq \frac{3 \sigma^2 (k + 1)}{4n} + o\left(\frac{k}{n} \xrightarrow{n \to +\infty}\right).
\]

(11)

We give details of the proof of Proposition 4.5 in Section 7.4.

Proposition 4.5 is to be compared with Corollary 4.1 and tells us that the variance of a PURF is upper bounded by three-fourths times the variance of a PURT. Hence, the rate of decay (in terms of power of \(n\)) of the PURF variance is the same as the PURT variance, and the actual gain appears in the multiplicative constant.

Let us, finally, comment on the hypotheses of Proposition 4.5. First, note that the hypotheses on \(k\) and \(n\) are the same as in Corollary 4.1, which allows a fair comparison between the two results. Finally, the other hypotheses (\(\mu > 0\), \(s\) is C-Lipschitz) are the same as in Corollary 4.1 and help to control negligible terms.

### 4.5. PURF bias

We now deal with the bias term of Decomposition (3). A convex inequality gives that the bias of a forest is not larger than the bias of a single tree:

\[
\mathbb{E}[(\hat{s}(X) - s(X))^2] \leq \frac{1}{q} \sum_{l=1}^{q} \mathbb{E}[(\hat{s}_U(X) - s(X))^2]
= \mathbb{E}[(\tilde{s}_U(X) - s(X))^2].
\]

So from Proposition 4.2, we deduce that:

**Proposition 4.6** If \(\mu\) is bounded by \(M > 0\) and \(s\) is C-Lipschitz, the bias of a PURF satisfies the same inequality as (9), that is:

\[
\mathbb{E}[(\hat{s}(X) - s(X))^2] \leq \frac{6MC^2}{(k + 1)^2}.
\]

(12)
4.6. Risk bounds for PURF

Putting together (11) and (12) leads to the following risk bound for a PURF.

**Proposition 4.7** If \( k \xrightarrow{n \to +\infty} +\infty \), \( (k/n) \xrightarrow{n \to +\infty} 0 \), \( 0 < \mu \leq M \), \( s \) is \( C \)-Lipschitz and \( q \xrightarrow{n \to +\infty} +\infty \), the risk of a PURF satisfies:

\[
\mathbb{E}[(\hat{s}(X) - s(X))^2] \leq \frac{3}{4} \frac{\sigma^2(k+1)}{n} + \frac{6MC^2}{(k+1)^2} + o\left(\frac{k}{n}\right)
\]

Again, taking \((k+1) = n^{1/3}\) gives the upper bound for the risk:

**Corollary 4.8** Under the assumptions of Theorem 4.7,

\[
\mathbb{E}[(\hat{s}(X) - s(X))^2] \leq Kn^{-2/3} + o\left(n^{-2/3}\right),
\]

where \( K \) is a positive constant.

So, a PURF reaches the minimax rate of convergence for \( C \)-Lipschitz functions.

Secondly, as the variance of a PURF is systematically reduced compared with a PURT and the bias of a PURF is not larger than the one of a PURT, the risk of a PURF is actually lower.

5. Simulations

In this section, we explain simulation experiments aiming at illustrating the results of Section 4 and compare PURF with RFRI.

5.1. Experiments

Experimentations were performed on four simulated data sets. We keep the framework of Model (1). In addition, we assume that \( \varepsilon \sim \mathcal{N}(0, \frac{1}{4}) \), and we take for \( s \) the following functions:

- **sinus data** \( s : [0, 1] \to \mathbb{R}, x \mapsto \sin(2\pi x) \),
- **square data** \( s : [0, 1] \to \mathbb{R}, x \mapsto x^2 \),
- **abs data** \( s : [0, 1] \to \mathbb{R}, x \mapsto |x - \frac{1}{2}| \),
- **stump data** \( s : [0, 1] \to \mathbb{R}, x \mapsto \mathbb{1}_{x \leq 1/2} - \mathbb{1}_{x > 1/2} \).

We also take several values for the number of data:

\( n \in \{100, 500, 1000, 5000, 10,000\} \). And for each value of \( n \), we fix \( k = \lfloor n^{1/3} \rfloor \) and \( q = \lfloor k/n \rfloor \) (the choice of \( k \) is motivated by discussions on PURT and PURF risks in Sections 4.3 and 4.6; the choice of \( q \) is sufficient condition for Equation (6) to hold).

We are interested in comparing PURT and PURF on one hand, and random trees-random inputs (RTRI) and RFRI on the other hand. Finally, we confront PURF and RFRI.

RFRI are extensively used in practice and present very good performances. We stress that they do not belong to the PRFs family, because partitions associated with each trees of RFRI are optimised using the learning sample. Hence, the results of the paper do not apply to them, but we still include RFRI in our simulation study to compare their behaviour with the PURF one.

For each function \( s \) and each value of \( n \), we simulate 20 data sets, on which we run 20 forests (or trees) and evaluate performances on 50 additional points.
5.1.1. Results

The results of experiments on sinus data are summarised in Figure 1. In the top-left graph, we plot the ratio PURFvariance/PURTvariance as a function of $n$. The ratios PURFbias/PURTbias and PURFrisk/PURTrisk are respectively plotted in the top-right and the bottom-left graphs. Finally, we plot in the bottom-right graph the ratio RFRI risk/RTRI risk: the solid line corresponds to the RFRI algorithm with default values, that is when each individual tree is grown to its maximal size (the rule being that we do not cut a node containing less than 5 data). The dashed line corresponds to the RFRI(k) algorithm, where each tree is grown until it has $k$ terminal nodes. We choose this variant to give a fair comparison between PURF and RFRI, that is when trees, associated with both methods, construct partitions in $k$ sets.

Figure 2–4 are obtained in the same way as Figure 1 and present results respectively for square data, abs data and stump data.
Finally, we give in Figure 5, the estimated risks of RFRI, RFRI(k) and PURF on the four simulated data sets.

5.2. Comments and discussions

We see that graphs of Figures 1–3 are very similar. We summarise the results by the following comments:

- the ratio PURF variance/PURT variance remains almost constant (when \( n \) is growing) around the value 0.4. This confirms that PURF and PURT variances are of the same order of magnitude, and shows that PURF effectively reduces the variance, here by a factor of less than \( \frac{3}{4} \). So our upper bound with factor \( \frac{3}{4} \) seems to be improvable.
Figure 5. Estimated risks of RFRI, RFRI(k) and PURF for the four simulated data sets.

- the ratio PURFbias/PURTbias decreases when $n$ is growing. This suggests that PURF seems to reduce the order of magnitude of the bias.
- as a consequence of the precedent remark, PURF seems to improve the rate of convergence of the risk (we see that the ratio PURFrisk/PURTrisk decreases as $n$ increases).
- the ratios between forest and tree risks for RFRI and RFRI(k) are roughly constant, again around 0.4. This suggests that RFRI reduces the risk, by a factor less than $\frac{1}{2}$.

Results for stump data in Figure 4 are quite different. Indeed, we again have a constant variance ratio (around a value a little bit less than 0.4), but here both bias ratio and risk ratio remain constant too (say around 0.4). So for stump data, PURF reduces the bias, but only by a constant. As a consequence, rates of convergence of PURF and PURT seem to be the same in this case.

The ratio RFRIrisk/RTRIrisk remains almost constant around 0.4 (may be with a slow decrease for the variant RFRI(k)).

These differences are natural because the regression function of stump data is discontinuous, whereas it is more regular in the other data sets (Lipschitz for abs data and indefinitely derivable of sinus data and square data).

As an intermediary conclusion, we can say that, in our simulation experiments, forests always improve the risk, compared with trees. More precisely, we observe that the risk is always at least two times smaller for a forest. Moreover, the improvement brought by forests can even affect the rate of convergence, especially for the PURF method when the regression function is regular.

Let us now give some comments about Figure 5. For sinus data, square data and abs data, we see that RFRI gives the worst performances, its risk being constant as $n$ becomes larger. RFRI(k) significantly improves the performance, with a risk converging to 0. And finally PURF is even better than RFRI(k).

We do not manage to explain the constant behaviour of RFRI risk. But the comparison between RFRI and RFRI(k) is interesting, because it shows a case where using fully grown trees in a forest performs worse than using relatively small trees. However, we stress that this phenomenon is likely due to the fact that we deal with one-dimensional input data. Indeed, RFRI with fully grown trees have shown very good performances many times when the dimension of input data is large (and potentially very large), see e.g. Breiman (2001) and Goldstein, Hubbard, Cutler, and Barcellos (2010). This point surely deserves a more intensive study.
The situation of stump data is again very different from other data sets. Here, PURF performs the worst (even if it manages to give a reasonable risk for very large values of \(n\)), RFRI risk is again constant, and RFRI(k) significantly reduces the risk. This suggests that RFRI(k) is much better than PURF for non-regular functions. Our explanation is that since RFRI(k) optimises the partitions associated with trees using the learning sample, it can better track some discontinuity of the regression function than PURF, in which partitions are chosen independently of the learning sample. In addition, even if PURF is better than RFRI(k) in the three other data sets, RFRI(k) is still competitive, whereas for stump data PURF is significantly worse than RFRI(k). Hence, when estimating regression with unknown regularity, RFRI(k) seems to be the better choice among the three methods we compared in this study.

6. Conclusion

In the context of PRFs, we give a general upper bound showing that the variance of a forest can actually be smaller than the variance of a tree.

We also emphasise, for a very simple version of RFs, the actual gain of using a RF instead of using a single random tree. First, we showed that both trees and forests reach the minimax rate of convergence. Then, we manage to highlight a reduction of the variance of a forest, compared with the variance of a tree. This is, in this specific context, a proof of the well-known conjecture for RFs: ‘a RF, by aggregating several random trees, reduces variance and leaves the bias unchanged’ which can be found for example in Hastie, Tibshirani and Friedman (2009).

Furthermore, our simulation study indicates that there is room for improvement, because a forest seems to be able to reduce bias as well. In addition, for sufficiently regular regression function, a forest could reduce the order of magnitude of the bias.

An interesting open problem would be to generalise this result, which could handle more complex versions of RFs and relax the hypotheses we made here. Obviously, a more ambitious goal would be to give some precise insights explaining the outstanding performances of RFRI (especially when the dimension of input data is large).

7. Proofs

7.1. Proof of Theorem 3.2

Before entering into details of the proof of Theorem 3.2, we recall that in the proof of Proposition 3.1 (which can be found in Arlot 2008), calculations lead to the following equality:

\[
E[(\hat{s}_U(X) - \tilde{s}_U(X))^2 | U] = \sum_{\lambda \in U} p_{\lambda} E \left[ \frac{1}{n\hat{p}_\lambda} \right] (\sigma^2 + \sigma^2_{\lambda}),
\]  

(13)

where \(\hat{p}_\lambda = \mathbb{1}(i : X_i \in U)/n\).

Then, an estimation of \(p_{\lambda} E[1/n\hat{p}_\lambda]\) gives the expression \((1/n)(1 + \delta_{n,p_\lambda})\) in Proposition 3.1.

We note

\[
\text{Var}_\lambda = p_{\lambda} E \left[ \frac{1}{n\hat{p}_\lambda} \right] (\sigma^2 + \sigma^2_{\lambda}),
\]  

(14)

a generic term of the sum in the r.h.s. of (13).

We now address the proof of Theorem 3.2. We begin by introducing some notations

In the sequel, we denote the covariance between two PURT by

\[
\mathbb{C}(\hat{s}_U, \hat{s}_U') = E[(\hat{s}_U(X) - \tilde{s}_U(X))(\hat{s}_U'(X) - \tilde{s}_U'(X))].
\]
Let us consider $\mathbb{U}^1 = (\lambda^1_1, \ldots, \lambda^1_k)$ and $\mathbb{U}^2 = (\lambda^2_1, \ldots, \lambda^2_k)$ the partitions, respectively, associated with $\hat{s}_{\mathbb{U}^1}$ and $\hat{s}_{\mathbb{U}^2}$.

Then we denote by $\mathbb{W}$ the intersection of the partitions $\mathbb{U}^1$ and $\mathbb{U}^2$. That is, all sets of $\mathbb{W}$ are obtained in intersecting one set of $\mathbb{U}^1$ and another set of $\mathbb{U}^2$. And we note $(\mu_1, \ldots, \mu_m)$ the sets of the partition $\mathbb{W}$, where $m$ is the number of sets of $\mathbb{W}$. We have $2k - 1 \leq m \leq k^2$.

The term $(\hat{s}_{\mathbb{U}^1}(X) - \hat{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \hat{s}_{\mathbb{U}^2}(X))$ equals, by definition, to

$$\left(\sum_{r=1}^{k} (\hat{\beta}_{\mu_1, \lambda_r^1} - \hat{\beta}_{\lambda_r^1}) \mathbb{I}_{X \in \lambda_r^1}\right) \left(\sum_{s=1}^{k} (\hat{\beta}_{\mu_1, \lambda_r^2} - \hat{\beta}_{\lambda_r^2}) \mathbb{I}_{X \in \lambda_r^2}\right) = \sum_{t=1}^{m} (\hat{\beta}_{\mu_1, \lambda_t^1} - \hat{\beta}_{\mu_1, \lambda_t^2}) (\hat{\beta}_{\mu_1, \lambda_t^2} - \hat{\beta}_{\mu_1, \lambda_t^1}) \mathbb{I}_{X \in \mu_t},$$

where

$\hat{\beta}_{\mu_1, \lambda_t^1} = \hat{\beta}_{\lambda_t^1}$ and $\hat{\beta}_{\mu_1, \lambda_t^2} = \beta_{\lambda_t^2}$ if $\mu_t \subset \lambda_t^1$,

$\hat{\beta}_{\mu_1, \lambda_t^2} = \beta_{\lambda_t^2}$ and $\hat{\beta}_{\mu_1, \lambda_t^1} = \beta_{\lambda_t^1}$ if $\mu_t \subset \lambda_t^2$.

Now, let us give some details for the first term of (15), denoted by $S_1(X)$. Without loss of generality, we suppose that $\mu_1 = \lambda_1^1 \cap \lambda_1^2 \neq \emptyset$. So,

$$S_1(X) = (\hat{s}_{\mathbb{U}^1}(X) - \hat{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \hat{s}_{\mathbb{U}^2}(X)) \mathbb{I}_{X \in \mu_1} = (\hat{\beta}_{\lambda_1^1} - \hat{\beta}_{\lambda_1^2})(\hat{\beta}_{\lambda_1^2} - \beta_{\lambda_1^1}) \mathbb{I}_{X \in \mu_1} = \left(\frac{1}{n\hat{p}_{\lambda_1^1}} \sum_{i : X_i \in \lambda_1^1} (Y_i - \hat{\beta}_{\lambda_1^1})\right) \left(\frac{1}{n\hat{p}_{\lambda_1^2}} \sum_{i : X_i \in \lambda_1^2} (Y_i - \hat{\beta}_{\lambda_1^2})\right) \mathbb{I}_{X \in \mu_1} = \frac{1}{n\hat{p}_{\lambda_1^1} n\hat{p}_{\lambda_1^2}} \sum_{i : X_i \in \lambda_1^1 \cap \lambda_1^2} (Y_i - \hat{\beta}_{\lambda_1^1})(Y_i - \hat{\beta}_{\lambda_1^2}) \mathbb{I}_{X \in \mu_1}$$

If we denote by $\mathbb{E}^{\lambda_{1,2}}[\cdot]$ the conditional expectation $\mathbb{E}[\cdot | (\mathbb{I}_{X_i \in \lambda_1^1})_{1 \leq i \leq n}, (\mathbb{I}_{X_i \in \lambda_1^2})_{1 \leq i \leq n}]$, we have:

$$\mathbb{E}[S_1(X) | \mathbb{U}^1, \mathbb{U}^2] = q_1 \mathbb{E}\left[\frac{1}{n\hat{p}_{\lambda_1^1} n\hat{p}_{\lambda_1^2}} \sum_{i : X_i \in \lambda_1^1 \cap \lambda_1^2} \mathbb{E}^{\lambda_{1,2}}[(Y_i - \hat{\beta}_{\lambda_1^1})(Y_i - \hat{\beta}_{\lambda_1^2})] \mathbb{I}_{X \in \mu_1} | \mathbb{U}^1, \mathbb{U}^2\right],$$

where $q_1 = P(X \in \mu_1)$ but

$$i^1 \neq i^2 \implies \mathbb{E}^{\lambda_{1,2}}[(Y_i^1 - \hat{\beta}_{\lambda_1^1})(Y_i^2 - \hat{\beta}_{\lambda_1^2})] = 0$$

because $Y_i^1$ and $Y_i^2$ are independent. Hence:

$$\mathbb{E}[S_1(X) | \mathbb{U}^1, \mathbb{U}^2] = q_1 \mathbb{E}\left[\frac{1}{n\hat{p}_{\lambda_1^1} n\hat{p}_{\lambda_1^2}} \sum_{i : X_i \in \mu_1} \mathbb{E}^{\lambda_{1}}[(Y_i - \hat{\beta}_{\lambda_1^1})(Y_i - \hat{\beta}_{\lambda_1^2})] \mathbb{I}_{X_i \in \lambda_1^1} | \mathbb{U}^1, \mathbb{U}^2\right]$$

$$= q_1 \mathbb{E}\left[\frac{1}{n\hat{p}_{\lambda_1^1} n\hat{p}_{\lambda_1^2}} \sum_{i : X_i \in \mu_1} \mathbb{E}[(Y_i - \hat{\beta}_{\lambda_1^1})(Y_i - \hat{\beta}_{\lambda_1^2}) | X_i \in \mu_1] \mathbb{I}_{X_i \in \lambda_1^1} | \mathbb{U}^1, \mathbb{U}^2\right],$$

where $\mathbb{E}^{\lambda_{1}}[\cdot]$ denotes the conditional expectation $\mathbb{E}[\cdot | (\mathbb{I}_{X_i \in \lambda_1^1})_{1 \leq i \leq n}]$. 


Now, as
\[ \mathbb{E}[(Y_i - \beta_{\lambda_i})(Y_i - \beta_{\lambda_i}) | X_i \in \mu_1] = \mathbb{E}[(Y - \beta_{\lambda_i})(Y - \beta_{\lambda_i}) | X \in \mu_1] \]
for all \( i \), and
\[ \mathbb{E}[(Y - \beta_{\lambda_i})(Y - \beta_{\lambda_i}) | X \in \mu_1] = \sigma^2 + \rho^2_{\mu_1}, \]
where
\[ \rho^2_{\mu_1} = \mathbb{E}[(s(X) - \tilde{s}_{U1}(X))(s(X) - \tilde{s}_{U2}(X)) | X \in \mu_1] \]
we get
\[ \mathbb{E}[S_1(X) | U^1, U^2] = q_1 \mathbb{E} \left[ \frac{1}{\hat{n}_\lambda} \right] (\sigma^2 + \rho^2_{\mu_1}). \]  
If we suppose in addition that \( \mu_2 = \lambda^1_2 \cap \lambda^2_1 \), and that \( \mu_1 \cup \mu_2 = \lambda^1_1 \) we similarly get for the second term of (15):
\[ \mathbb{E}[S_2(X) | U^1, U^2] = \mathbb{E}[(\tilde{s}_{U1}(X) - \tilde{s}_{U1}(X))(\tilde{s}_{U2}(X) - \tilde{s}_{U2}(X)) \mathbb{I}_{X \in \mu_2} | U^1, U^2] \]
\[ = q_2 \mathbb{E} \left[ \frac{n\hat{q}}{n\hat{p}_\lambda} \right] (\sigma^2 + \rho^2_{\mu_2}). \]  
where
\[ q_2 = P(X \in \mu_2), \]
\[ n\hat{q}_2 = E[i : X_i \in \mu_2], \]
\[ \rho^2_{\mu_2} = \mathbb{E}[(s(X) - \tilde{s}_{U1}(X))(s(X) - \tilde{s}_{U2}(X)) | X \in \mu_2]. \]
Since \( \mu_2 \) is included in \( \lambda^1_2 \), we have \( \hat{q}_2 \leq \hat{p}_\lambda \), so:
\[ \mathbb{E}[S_2(X) | U^1, U^2] \leq q_2 \mathbb{E} \left[ \frac{1}{n\hat{p}_\lambda} \right] (\sigma^2 + \rho^2_{\mu_2}). \]  
Finally, by summing the two terms \( S_1(X) \) and \( S_2(X) \), we deduce that
\[ \mathbb{E}[S_1(X) + S_2(X) | U^1, U^2] \leq p\lambda \mathbb{E} \left[ \frac{1}{n\hat{p}_\lambda} \right] (\sigma^2 + \rho^2_{\mu_1} + \rho^2_{\mu_2}). \]  
In conclusion, we succeeded to bound the sum of the first two terms of (15) by an expression very close to \( \text{Var}_\lambda \) (defined in (14)). The only difference comes from the fact that instead of \( \lambda^2_\hat{k} \) we have \( \rho^2_{\mu_1} + \rho^2_{\mu_2} \).

We can easily generalise this fact by proving the following lemma.

We denote by \( S_r(X) \) the \( r \)th term of (15), i.e. \( S_r(X) = (\tilde{s}_{U1}(X) - \tilde{s}_{U1}(X))(\tilde{s}_{U2}(X) - \tilde{s}_{U2}(X)) \mathbb{I}_{X \in \mu_r} \).

**Lemma 7.1** Let \( r \) be in \( \{1, \ldots, k\} \) and denote by \( K_r \) a subset of \( \{1, \ldots, m\} \) such that
\[ \bigcup_{i \in K_r} \mu_i = \lambda^1_r \]  
then
\[ \mathbb{E} \left[ \sum_{i \in K_r} S_r(X) | U^1, U^2 \right] \leq p\lambda \mathbb{E} \left[ \frac{1}{n\hat{p}_\lambda} \right] (\sigma^2 + \Sigma^2_r), \]
where \( \Sigma^2_r = \sum_{i \in K_r} \rho^2_{\mu_i} \).
Indeed for all \( t \in K_r \),
\[
\mathbb{E}[S_t(X) | \cup_1^t, \cup_2^t] \leq q_t \mathbb{E} \left[ \frac{1}{n\hat{p}_{\lambda^t}} \right] (\sigma^2 + \rho^2_{\mu_t}),
\]
where
\[
q_t = P(X \in \mu_t),
\]
\[
\rho^2_{\mu_t} = \mathbb{E}[(s(X) - \tilde{s}_{\cup_1^t}(X))(s(X) - \tilde{s}_{\cup_2^t}(X)) | X \in \mu_t].
\]
Thus,
\[
\mathbb{E} \left[ \sum_{t \in K_r} S_t(X) | \cup_1^t, \cup_2^t \right] \leq P \left( X \in \bigcup_{t \in K_r} \mu_t \right) \mathbb{E} \left[ \frac{1}{n\hat{p}_{\lambda^t}} \right] (\sigma^2 + \Sigma^2_r).
\]
From Relation (15) we have \( \bigcup_{t \in K_r} \mu_t = \lambda^1_r \), which concludes the proof of Lemma 7.1.

Therefore, we can upper bound the initial sum (15) of \( m \) terms (we recall that \( 2k - 1 \leq m \leq k^2 \)) by a sum of \( k \) terms of the same order as \( \text{Var}_{\lambda} \) only involving sets of the partition \( \cup^1 \). At this stage, we get an upper bound for the variance of a forest which is of the same order as the variance of a tree. But we can do better. With similar arguments, we can prove the following lemma:

**Lemma 7.2** If there exist \( r \neq r' \) and \( s \) such as
\[
\lambda^1_r \subset \lambda^2_s, \quad \lambda^1_{r'} \subset \lambda^2_s
\]
the expression
\[
\mathbb{E}[\hat{\delta}_{\cup^1}(X) - \tilde{\delta}_{\cup^1}(X))(\hat{\delta}_{\cup^2}(X) - \tilde{\delta}_{\cup^2}(X))\mathbb{1}_{\lambda^1_s, \cup_1^t, \cup_2^t} | \cup^1, \cup^2]
\]
is upper bounded by
\[
p_{\lambda^1_s} \mathbb{E} \left[ \frac{1}{n\hat{p}_{\lambda^t}} \right] (\sigma^2 + \Sigma^2_s),
\]
where \( \Sigma^2_s = \rho^2_t + \rho^2_{r'} \), with \( t \) and \( r' \) are such that \( \lambda^1_t = \mu_t \) and \( \lambda^1_{r'} = \mu_{r'} \)

Indeed,
\[
\mathbb{E}[\hat{\delta}_{\cup^1}(X) - \tilde{\delta}_{\cup^1}(X))(\hat{\delta}_{\cup^2}(X) - \tilde{\delta}_{\cup^2}(X))\mathbb{1}_{\lambda^1_s} | \cup^1, \cup^2]
\]
\[
\leq p_{\lambda^1_s} \mathbb{E} \left[ \frac{1}{n\hat{p}_{\lambda^t}} \right] (\sigma^2 + \rho^2_t)
\]
and
\[
\mathbb{E}[\hat{\delta}_{\cup^1}(X) - \tilde{\delta}_{\cup^1}(X))(\hat{\delta}_{\cup^2}(X) - \tilde{\delta}_{\cup^2}(X))\mathbb{1}_{\lambda^1_{r'}} | \cup^1, \cup^2]
\]
\[
\leq p_{\lambda^1_{r'}} \mathbb{E} \left[ \frac{1}{n\hat{p}_{\lambda_{r'}}} \right] (\sigma^2 + \rho^2_{r'}).
\]
Finally, since \( p_{\lambda^1_s} + p_{\lambda^1_{r'}} \leq p_{\lambda^1_s} \), \( \rho^2_t \leq \rho^2_t + \rho^2_{r'} \) and \( \rho^2_{r'} \leq \rho^2_t + \rho^2_{r'} \), the result is obtained by summing the two terms.

As in Proposition 3.1, we replace all \( p_{\lambda^1_s} \mathbb{E}[1/n\hat{p}_{\lambda^t}] \) by their estimates \((1 + \delta_n p_{\lambda^t})\).
By repeatedly applying this lemma for all sets, we can upper bound
\[
\mathbb{E}[(\tilde{s}_{U}(X) - \tilde{s}_{U}(X))(\hat{s}_{U}(X) - \tilde{s}_{U}(X)) | U^1, U^2]
\]
by a sum of \(N_{l^d}\) terms of the form \((1 + \delta_{n,p_j^d})(\sigma^2 + \Sigma_i^2)\), where \(\tilde{p}_{k^d}\) denotes for some \(j \in \{1, \ldots, k\}\) either \(p_{k^d}\) or \(p_{k^d}\) depending on the fact that we are in the situation of Lemma 7.1 or Lemma 7.2. And \(N_{l^d} = k - M_{l^d}\) with
\[
M_{l^d} = \sum_{1 \leq r < r' \leq k} \sum_{1 \leq s \leq k} 1_{\lambda_r^d = \lambda_s^d, \lambda_r^d < \lambda_s^d}.
\]
We conclude the proof of Theorem 3.2 by integrating with respect to \((U_1, U_2)\).

7.2. Proof of Corollary 4.1

Here, we have \(U = \{0, U(1), [U(1), U(2)],[U(1), U(2), \ldots, [U(1), 1]\) where \((U(1), U(2), \ldots, U(\delta))\) denotes the ordered statistics of \((U_1, U_2, \ldots, U_k)\). With these notations, Equation (5) becomes
\[
\mathbb{E}[(\tilde{s}_{U}(X) - \tilde{s}_U(X))^2] = \frac{1}{n} \sum_{j=0}^{k} \left( \sigma^2 + \sigma^2 \mathbb{E}[\delta_{n,p_j}] + \mathbb{E}[\sigma_j^2] + \mathbb{E}[\sigma_j^2 \delta_{n,p_j}] \right),
\]
where
\[
p_j = P(U_{(j)} \leq X < U_{(j+1)}),
\]
\[
\sigma_j^2 = \mathbb{E}[(s(X) - \tilde{s}_{U}(X))^2 | U_{(j)} \leq X < U_{(j+1)}].
\]
And we must prove that the three last terms in the sum of Equation (16) are negligible compared with the constant term \(\sigma^2\).

Let us fix \(0 \leq j \leq k\). As it can be found e.g. in David and Nagaraja (2003, Chapter 6), the probability density function of \(U_{(j+1)} - U_{(j)}\) is the function \(t \in [0, 1] \mapsto k(1 - t)^{k-1}\).

- For the second term \(\mathbb{E}[\delta_{n,p_j}]\):
  from Arlot (2008) we have \(\delta_{n,p_j} \leq \kappa_3 (np_j)^{-1/4}\), where \(\kappa_3\) is a positive constant. So,
  \[
  \mathbb{E}[\delta_{n,p_j}] \leq \kappa_3 \mathbb{E}[(np_j)^{-1/4}] = \frac{\kappa_3}{n^{-1/4}} \mathbb{E}[p_j^{-1/4}] \leq \frac{\kappa_3}{(an)^{-1/4}} \mathbb{E}[(U_{(j+1)} - U_{(j)})^{-1/4}] \leq \frac{\kappa_4}{\alpha^{-1/4}} \left( \frac{k}{n} \right)^{1/4},
  \]
  where \(\alpha = \min_{[0,1]} \mu > 0\) and \(\kappa_4\) is another positive constant.

  Since \((k/n) \xrightarrow[n \to +\infty]{} 0\) the last upper bound tends to 0 as \(n\) tends to infinity.

- For the third term \(\mathbb{E}[\sigma_j^2]\):
  \[
  \sigma_j^2 = \mathbb{E}[(s(X) - \tilde{s}_{U}(X))^2 | U_{(j)} \leq X < U_{(j+1)}] \leq C^2 (U_{(j+1)} - U_{(j)})^2 \text{ because } s \text{ is } C\text{-Lipschitz}
  \]
  So,
  \[
  \mathbb{E}[\sigma_j^2] \leq C^2 \mathbb{E}[(U_{(j+1)} - U_{(j)})^2] = C^2 \frac{2}{(k+1)(k+2)}
  \]
  which tends to 0 as \(k\) tends to infinity.
For the last term, the following inequality is sufficient to conclude:
\[ \mathbb{E}[\sigma_j^2 \delta_{n,j}] \leq C^2 \mathbb{E}[\delta_{n,j}], \] because \( U_{(j+1)} - U_{(j)} \leq 1. \)

### 7.3. Proof of Proposition 4.2

We keep the notations of Section 7.2. In addition, we define \( \beta_j \) for any \( 0 \leq j \leq k \) by
\[ \beta_j = \mathbb{E}[Y \mid U_{(j)} < X \leq U_{(j+1)}]. \]

Function \( s \) is supposed to be \( C \)-Lipschitz, so
\[
\mathbb{E}[(\tilde{s}_1(U)(X) - s(X))^2] = \mathbb{E}\left[ \left( \sum_{j=0}^{k} (s(X) - \beta_j) \mathbb{I}_{U_{(j)} < X \leq U_{(j+1)}} \right)^2 \right]
\]
\[= \mathbb{E}\left[ \sum_{j=0}^{k} (s(X) - \beta_j)^2 \mathbb{I}_{U_{(j)} < X \leq U_{(j+1)}} \right] \leq \mathbb{E}\left[ \sum_{j=0}^{k} C^2 (U_{(j+1)} - U_{(j)})^2 \mathbb{I}_{U_{(j)} < X \leq U_{(j+1)}} \right]
\]
\[= C^2 \mathbb{E}\left[ \sum_{j=0}^{k} (U_{(j+1)} - U_{(j)})^2 \mathbb{P}(U_{(j)} < X \leq U_{(j+1)}) \right] \leq C^2 \mathbb{E}\left[ \sum_{j=0}^{k} M(U_{(j+1)} - U_{(j)})^3 \right] \]
because \( \mu \) is bounded by \( M \)
\[= MC^2 \sum_{j=0}^{k} \mathbb{E}[(U_{(j+1)} - U_{(j)})^3] \]
\[= MC^2 \frac{6}{(k+2)(k+3)} \leq \frac{6MC^2}{(k+1)^2}. \]

### 7.4. Proof of Proposition 4.5

Using the fact that we explicitly know the distribution of \( \mathbb{U} \), we deduce from Theorem 3.2 the following corollary.

**Corollary 7.3** If \( k \xrightarrow{n \to +\infty} +\infty, (k/n) \xrightarrow{n \to +\infty} 0, \mu > 0 \) and \( s \) is \( C \)-Lipschitz, we have,
\[
\mathbb{E}[(\tilde{s}_1(U)(X) - \tilde{s}_1(U)(X))(\tilde{s}_2(U)(X) - \tilde{s}_2(U)(X))] \leq \frac{\sigma^2 \mathbb{E}[N_U]}{n} + \frac{\sigma}{n \to +\infty} \left( \frac{k}{n} \right)
\]
\[\leq \frac{3}{4} \sigma^2 (k+1) + \frac{\sigma}{n \to +\infty} \left( \frac{k}{n} \right). \]
Because of the simple draws of random partitions, the number \( M_{1d} \) is explicitly computable (we know the distribution of the two ordered statistics) and it is shown to be equivalent to \( \frac{1}{4} (k + 1) \) as \( k \) tends to \(+\infty\) (see Lemma 7.4).

As in Corollary 4.1, we have to prove that all terms of the sum are negligible compared with the constant one \( \sigma^2 \). To deal with the fact that the number of terms in the sum is now random, we use the following inequality:

\[
\mathbb{E} \left[ \sum_{t=0}^{N_{1t}} (\sigma^2 \delta_{n,p_t} + (\Sigma_t^{d,1,2})^2 + (\Sigma_t^{d,1,2})^2 \delta_{n,p_t}) \right] 
\leq \sum_{t=0}^{k} (\mathbb{E}[\sigma^2 \delta_{n,p_t}] + \mathbb{E}[(\Sigma_t^{d,1,2})^2] + \mathbb{E}[(\Sigma_t^{d,1,2})^2 \delta_{n,p_t}]).
\]

These quantities are of the same kind as the three last terms in the sum of Equation (17). So with the same techniques we get that

\[
\frac{1}{n} \mathbb{E} \left[ \sum_{t=0}^{N_{1t}} (\sigma^2 \delta_{n,p_t} + (\Sigma_t^{d,1,2})^2 + (\Sigma_t^{d,1,2})^2 \delta_{n,p_t}) \right] = o \left( \frac{k}{n} \right).
\]

So, we have

\[
\mathbb{E}[(\tilde{\delta}_{12}(X) - \tilde{\delta}_{12}(X))(\tilde{\delta}_{12}(X) - \tilde{\delta}_{12}(X))] \leq \frac{\sigma^2 \mathbb{E}[N_{1d}]}{n} + o \left( \frac{k}{n} \right).
\]

Finally, the following technical result allows to conclude the proof of Corollary 7.3, and thus, using Equality (6), the proof of Proposition 4.5.

**Lemma 7.4**

\[
\mathbb{E}[M_{1d}] = \frac{(k - 2)(k - 3)}{2(2k - 1)} \left( 1 + \frac{4}{(k + 1)(k - 3)} \right).
\]

**Hence,**

\[
\mathbb{E}[M_{1d}] = \frac{k + 1}{4} + o \left( \frac{k}{k+\infty} \right).
\]

We then obtain that

\[
\mathbb{E}[N_{1d}] = \frac{3}{4} (k + 1) + o \left( \frac{k}{k + \infty} \right).
\]

Let us prove Lemma 7.4.

\[
\mathbb{E}[M_{1d}] = \sum_{r=1}^{k-2} \sum_{s=1}^{k-1} \mathbb{P}(U_{(s)}^2 < U_{(r)}^1 < U_{(r+1)}^1 < U_{(r+2)}^1 < U_{(s+1)}^2).
\]

As we know the distribution of ordered statistics (see e.g. David and Nagaraja 2003, Section 2.2), we can compute the following probability:

\[
\mathbb{P}(U_{(s)}^2 < U_{(r)}^1 < U_{(r+1)}^1 < U_{(r+2)}^1 < U_{(s+1)}^2) = \mathbb{P}(U_{(s)}^2 < U_{(r)}^1 \ and \ U_{(r+2)}^1 < U_{(s+1)}^2)
\]

\[
= \sum_{j=r+2}^{k} \sum_{i=0}^{r-1} \frac{k!}{i!(j-i)!(k-j)!} \mathbb{E}[(U_{(s)}^1)^i(U_{(s+1)}^2)^j - U_{(s)}^1)^i(1 - U_{(s+1)}^2)^j]
\]

\[
= \sum_{j=r+2}^{k} \sum_{i=0}^{r-1} \frac{k!}{i!(k-j)!(s-1)!(s+1)!} \frac{(i + s - 1)!(2k - (j + s) - 1)!}{(2k)!}
\]
So,

\[
\mathbb{E}[M_{t,k}] = \frac{(k!)^2}{(2k)!} \sum_{r=1}^{k-2} \sum_{s=1}^{k-1} \left( \sum_{i=0}^{r-1} \binom{i + (s - 1)}{i} \right) \left( \sum_{j=r+2}^{k} \binom{k-j+k-(s+1)}{k-j} \right)
\]

\[
= \frac{(k!)^2}{(2k)!} \sum_{r=1}^{k-2} \sum_{s=1}^{k-1} \binom{r-1}{r-1} \binom{k-r-2}{k-r-2} \binom{2k-r-2-s}{t+1} \binom{k-3-t}{k-3-r} \binom{2k-3}{k-3}
\]

(by elementary properties of binomial coefficients, see e.g. Graham, Knuth and Patashnik 1989, p. 160)

\[
= \frac{k-2}{4(2k-1)} \sum_{t=0}^{2k-5} \sum_{r=t-k+2}^{t} \binom{t+1}{r} \binom{2k-3-(t+1)}{k-3-r} \binom{2k-3}{k-3}
\]

(by defining \( t = r + s \))

\[
= \frac{k-2}{4(2k-1)} \sum_{t=0}^{2k-5} \left[ F_{\mathcal{H}(2k-3,t+1,k-3)}(t) - F_{\mathcal{H}(2k-3,t+1,k-3)}(t-k+1) \right]
\]

(where \( F_{\mathcal{H}(N,m,n)} \) denotes the cumulative distribution function of the hyper-geometric distribution)

\[
= \frac{k-2}{4(2k-1)} \sum_{t=0}^{k-3} F_{\mathcal{H}(2k-3,t+1,k-3)}(t)
\]

\[
= \frac{k-2}{2(2k-1)} \left[ \sum_{t=0}^{k-4} \left( 1 - \binom{t+1}{t+1} \binom{2k-3-(t+1)}{k-3-(t+1)} \frac{2k-3}{k-3} \right) + 1 \right]
\]

\[
= \frac{k-2}{2(2k-1)} \left( k - 3 + \frac{4}{k+1} \right).
\]

**Acknowledgements**

The author would like to thank Pascal Massart for very valuable discussions, Jean-Michel Poggi and Nicolas Verzelen for interesting remarks, and Lucile Debrosse for her precious help with English language. The author also thank two anonymous referees who help to significantly improve this paper.

**References**

Arlot, S. (2008), ‘V-Fold Cross-Validation Improved: V-Fold Penalization’, preprint, arXiv:0802.0566v2.

Biau, G. (2012), ‘Analysis of a random forests model’, *Journal of Machine Learning Research*, in press.

Biau, G., Devroye, L., and Lugosi, G. (2008), ‘Consistency of Random Forests and other Averaging Classifiers’, *Journal of Machine Learning Research*, 9, 2039–2057.

Breiman, L. (2001), ‘Random Forests’, *Machine Learning*, 45, 5–32.

Cutler, A., and Zhao, G. (2001), ‘Pert – Perfect Random Tree Ensembles’, *Computing Science and Statistics*, 33, 490–497.

David H.A., and Nagaraja H.N. (2003), *Order Statistics*, Wiley Series in Probability and Statistics, New York: Wiley.

Díaz-Uriarte, R., and Alvarez de Andrés, S. (2006), ‘Gene Selection and Classification of Microarray Data using Random Forest’, *BMC Bioinformatics*, 7, 3.

Genuer, R., Poggi, J.-M., and Tuleau, C. (2010), ‘Variable Selection using Random Forests’, *Pattern Recognition Letters*, 31, 2225–2236.
Goldstein, B., Hubbard, A., Cutler, A., and Barcellos, L. (2010), ‘An Application of Random Forests to a Genome-Wide Association Dataset: Methodological Considerations & New Findings’, *BMC Genetics*, 11(1), 49.

Graham, R.L., Knuth, D.E., and Patashnik, O. (1989), *Concrete Mathematics*, Reading, MA: Addison-Wesley.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning* (2nd ed.), New York: Springer-Verlag.

Ibragimov, I.A., and Khasminskii, R.Z. (1981), *Statistical Estimation: Asymptotic Theory*, New York: Springer-Verlag.

Liaw, A., and Wiener, M. (2002), ‘Classification and Regression by randomForest’, *R News*, 2(3), 18–22.