HAL
archives-ouvertes.fr

Variance reduction in purely random forests

Robin Genuer

▶ **To cite this version:**

Robin Genuer. Variance reduction in purely random forests. Journal of Nonparametric Statistics, American Statistical Association, 2012, 2, pp.18 - 562. <10.1007/978-1-4899-0027-2>. <hal-01590513>

HAL Id: hal-01590513
https://hal.inria.fr/hal-01590513

Submitted on 19 Sep 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY - Attribution 4.0 International License

Journal of Nonparametric Statistics
Vol. 00, No. 00, Month 2011, 1–23

RESEARCH ARTICLE

Variance reduction in purely random forests

Robin Genuer[a,b] *

[a] *Univ. 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*

(*28 November 2011*)

Random forests, 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 random forests, called purely random forests, 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 emphasizes the fact that a forest reduces the variance. We then introduce a simple variant of purely random forests, that we call purely uniformly random forests. For this variant and in the context of regression problems with a one-dimensional predictor space, we show that both random trees and random forests reach minimax rate of convergence. In addition, we prove that compared to random trees, random forests improve accuracy by reducing the estimator variance by a factor of three fourths.

Keywords: Random Forests, Non-parametric regression, Rates of convergence, Randomization, Ensemble methods

* Corresponding author. Email: Robin.Genuer@isped.u-bordeaux2.fr

ISSN: 1048-5252 print/ISSN 1029-0311 online
© 2011 Taylor & Francis
DOI: 10.1080/1048525YYxxxxxxxx
http://www.informaworld.com

1. Introduction

Random forests (RF), introduced in Breiman (2001), are a very effective statistical method. They give outstanding performances in a lot of situations for both regression and classification problems. Mathematical understanding of these good performances remains quite unknown. As defined in Breiman (2001), a random forest is a collection of tree-predictors $\{h(x, \Theta_l), 1 \leqslant l \leqslant q\}$, where $(\Theta_l)_{1 \leqslant l \leqslant q}$ are i.i.d. random vectors, and a random forest 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 random forest improves so much the performance of a single tree.

Breiman (2001) introduced a specific instance of random forest, called random forests-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 et al. 2010).

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 *independently* of the learning sample; while in random RFRI, the splits are optimized using the learning sample. This independence between splits and learning sample makes mathematical analysis easier. Cutler and Zhao (2001) introduced PERT (Perfect Random Tree Ensemble), an algorithm which builds some purely random forests, and illustrated its good performance on benchmark datasets. More recently Biau et al. (2008) showed that both purely random trees and purely random forests are universally consistent.

In this paper, we essentially focus on the variance of purely random forests. First, we show a general upper bound on the forest variance, which emphasizes 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 to the number of splits of the trees, then the forest variance is small compared to the tree variance.

Then, our paper offers to examine a simple variant of purely random forests. We call it *purely uniformly random forests* (PURF) and we analyze its risk, only in a regression framework with a one-dimensional predictor space. The main goal is to emphasize 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 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 organized as follows. Section 2 presents the model. Section 3 presents a general upper bound for purely random forests variance. Section 4 gives some risk bounds for purely uniformly random trees and forests. In Section 5, we lead a simulation study illustrating our results and comparing 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. $X \in \mathcal{X}$ a mesurable space (e.g. $\mathcal{X} = \mathbb{R}^d$ with any $d \geq 1$). We consider the following statistical model:

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

$s : \mathcal{X} \to \mathbb{R}$ 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 ε with value in $\mathbb{R}$, independent of $\mathcal{L}_n$, with $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 < +\infty$.

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

3. A general upper bound for purely random forests variance

We begin by giving a general analysis of the purely random forests variance.

3.1. *Purely Random Tree 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 random forests and purely random forests is that in the purely random case, partitions of the input space $\mathcal{X}$ are drawn randomly, independently of $\mathcal{L}_n$. We recall that in classical random forests, partitions of the input space are most of the time obtained by random perturbations of a partitioning scheme where splits are calculated using $\mathcal{L}_n$. Hence, for classical random forests, the random perturbations are independent of $\mathcal{L}_n$, but the partition is not.

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

A Purely Random Tree (PRT), associated with $\mathbb{U}$, is defined for $x \in \mathcal{X}$ as:

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

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 $x \in \mathcal{X}$:

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

where

$$\beta_\lambda = \mathbb{E}[Y \,|\, X \in \lambda] \; .$$

Conditionally on $\mathbb{U}$, $\tilde{s}_{\mathbb{U}}$ 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 $\hat{s}_{\mathbb{U}}$ as follows:

$$\begin{array}{rcccl} \mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - s(X))^2] & = & \mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2] & + & \mathbb{E}[(\tilde{s}_{\mathbb{U}}(X) - s(X))^2] \\ & = & \text{variance term} & + & \text{bias term} \end{array} \tag{2}$$

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

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

so $\mathbb{E}[(\hat{s}_u(x) - \tilde{s}_u(x))^2]$ is the variance of the estimator $\hat{s}_u(x)$ 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. *Purely Random Forest definition*

A random forest is the aggregation of a collection of random trees. So, in the context of Purely Random Forests (PRF), 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 $\mathbb{V}_q$, is defined for $x \in [0,1]$ as follows:

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

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

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

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

$$\begin{array}{rcccl} \mathbb{E}[(\hat{s}_{\mathbb{V}_q}(X) - s(X))^2] & = & \mathbb{E}[(\hat{s}_{\mathbb{V}_q}(X) - \tilde{s}_{\mathbb{V}_q}(X))^2] & + & \mathbb{E}[(\tilde{s}_{\mathbb{V}_q}(X) - s(X))^2] \\ & = & \text{variance term} & + & \text{bias term} \end{array} \tag{3}$$

3.3. *PRT Variance*

We start to deal with the variance term of Decomposition (2). First, we work conditionally on $\mathbb{U}$, 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* $\mathbb{U}$*, the variance term of Decomposition (2) satisfies:*

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2 \,|\, \mathbb{U}] = \frac{1}{n}\sum_{\lambda \in \mathbb{U}} (1 + \delta_{n,p_\lambda})(\sigma^2 + \sigma_\lambda^2) \tag{4}$$

where

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

■

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

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2] = \frac{1}{n}\sum_{\lambda \in \mathbb{U}} \left(\sigma^2 + \sigma^2 \mathbb{E}[\delta_{n,p_\lambda}] + \mathbb{E}[\sigma_\lambda^2] + \mathbb{E}[\sigma_\lambda^2 \delta_{n,p_\lambda}]\right) \tag{5}$$

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

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

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 $\hat{s}_{\mathbb{V}_q}(x) = \dfrac{1}{q}\sum_{l=1}^{q} \hat{s}_{\mathbb{U}^l}(x)$, the variance term satisfies:

$$\mathbb{E}[(\hat{s}_{\mathbb{V}_q}(X) - \tilde{s}_{\mathbb{V}_q}(X))^2] = \frac{1}{q^2}\sum_{l=1}^{q} \mathbb{E}[(\hat{s}_{\mathbb{U}^l}(X) - \tilde{s}_{\mathbb{U}^l}(X))^2] + \frac{1}{q^2}\sum_{l \neq q} \mathbb{E}[(\hat{s}_{\mathbb{U}^l}(X) - \tilde{s}_{\mathbb{U}^l}(X))(\hat{s}_{\mathbb{U}^m}(X) - \tilde{s}_{\mathbb{U}^m}(X))]$$

$$= \frac{1}{q}\mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))^2] + \frac{q(q-1)}{q^2}\mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \tilde{s}_{\mathbb{U}^2}(X))]$$

where the last equality comes from the fact that the $(\hat{s}_{\mathbb{U}^l}(X) - \tilde{s}_{\mathbb{U}^l}(X))_{1\leqslant l\leqslant q}$have the same distribution.

Now, if we let q grow to infinity, we get:

$$\mathbb{E}[(\hat{s}_{\mathbb{V}_q}(X) - \tilde{s}_{\mathbb{V}_q}(X))^2] = \mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \tilde{s}_{\mathbb{U}^2}(X))]\big(1 + \underset{q\to+\infty}{\mathrm{o}}(1)\big) \quad (6)$$

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

$$\mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \tilde{s}_{\mathbb{U}^2}(X))]$$

Let us note $\mathbb{U}^1 = \{\lambda_1^1, \ldots, \lambda_k^1\}$ and $\mathbb{U}^2 = \{\lambda_1^2, \ldots, \lambda_k^2\}$ the cells of the partitions respectively associated to the trees $\hat{s}_{\mathbb{U}^1}$ and $\hat{s}_{\mathbb{U}^2}$.

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

Theorem 3.2 :

$$\mathbb{E}[(\hat{s}(X) - \tilde{s}(X))^2] \leqslant \frac{1}{n}\mathbb{E}\left[\sum_{t=1}^{N_{\mathcal{U}}}(1 + \delta_{n,\tilde{p}_t})(\sigma^2 + \Sigma_t^2)\right]\big(1 + \underset{q\to+\infty}{\mathrm{o}}(1)\big)$$

where

$$N_{\mathcal{U}} = k - \sum_{1\leqslant r<r'\leqslant k}\sum_{1\leqslant s\leqslant k}\mathbb{1}_{\lambda_r^1\subset\lambda_s^2\,,\,\lambda_{r'}^1\subset\lambda_s^2} \quad (7)$$

Σ_t^2 *denotes a sum of terms* $\mathbb{E}[(\tilde{s}_{\mathbb{U}^1}(X) - s(X))(\tilde{s}_{\mathbb{U}^2}(X) - s(X))\,|\,X \in \lambda_{t'}^l]$ *for several consecutive values of* t' *and with* $l = 1$ *or* 2 *.*
$\tilde{p}_t$ *denotes for some* $j \in \{1, \ldots, k\}$ *either* $p_{\lambda_j^1}$ *or* $p_{\lambda_j^2}$ *depending on the relative positions between the* $(\lambda_1^1, \ldots, \lambda_k^1)$ *and the* $(\lambda_1^2, \ldots, \lambda_k^2)$*.*

■

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 $N_{\mathcal{U}} \leqslant k$, and the larger the double sum in the r.h.s of (7) is, the smaller $N_{\mathcal{U}}$ is.

We stress that quantities appearing in $N_{\mathcal{U}}$ only depends of the two partitions $\mathbb{U}^1, \mathbb{U}^2$. And our results means that if $\mathbb{U}^1$ and $\mathbb{U}^2$ are *different* enough, the covariance between $\hat{s}_{\mathbb{U}^1}$ and $\hat{s}_{\mathbb{U}^2}$ will be smaller than a single tree variance.

In Section 4.4, we will see that in a special case, $\mathbb{E}[N_{\mathcal{U}}]$ equals $\frac{3k}{4}$. This allows to claim that the forest reduces variance by a factor of $\frac{3}{4}$, for this case.

4. Risk bounds for Purely Uniformly Random Forests

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

The principle of Purely Uniformly Random Trees (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 purely random forests or random forests-random inputs, 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 $\mathbb{U} = (U_1, \ldots, U_k)$ be k i.i.d. random variables of uniform distribution on $[0,1]$.

4.1. *PURT variance*

So, using the fact that, in our case, $\mathbb{U}$ 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* $k \xrightarrow[n\to+\infty]{} +\infty$, $\frac{k}{n} \xrightarrow[n\to+\infty]{} 0$, $\mu > 0$ *and* s *is* C*-Lipschitz, the variance of a PURT satisfies:*

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2] = \frac{\sigma^2(k+1)}{n} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right) \tag{8}$$

where the notation $\underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right)$ *denotes a function* $f(n)$ *such as* $\frac{f(n)}{k/n} \xrightarrow[n\to+\infty]{} 0$.

■

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

The first two hypotheses of Corollary 4.1 ($k \xrightarrow[n\to+\infty]{} +\infty$, $\frac{k}{n} \xrightarrow[n\to+\infty]{} 0$) 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* μ *is bounded by* $M > 0$ *and* s *is* C*-Lipschitz, the bias of a PUR Tree is upper bounded by:*

$$\mathbb{E}[(\tilde{s}_{\mathbb{U}}(X) - s(X))^2] \leqslant \frac{6MC^2}{(k+1)^2} \tag{9}$$

■

4.3. *Risk bounds for PURT*

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

Proposition 4.3: *If* $k \xrightarrow[n\to+\infty]{} +\infty$, $\frac{k}{n} \xrightarrow[n\to+\infty]{} 0$, $0 < \mu \leqslant M$ *and* s *is* C*-Lipschitz, the risk of a PURT satisfies:*

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - s(X))^2] \leqslant \frac{\sigma^2(k+1)}{n} + \frac{6MC^2}{(k+1)^2} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right) \tag{10}$$

■

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,*

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - s(X))^2] \leqslant Kn^{-2/3} + \underset{n\to+\infty}{\mathrm{o}}(n^{-2/3})$$

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 analyze purely uniformly random forests.

4.4. *PURF variance*

From Theorem 3.2, we deduce the following proposition:

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

$$\mathbb{E}[(\hat{s}(X) - \tilde{s}(X))^2] \leqslant \frac{3}{4}\frac{\sigma^2(k+1)}{n} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right) \tag{11}$$

■

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 PUR Forest is upper bounded by three fourths times the variance of a PUR Tree. So, the rate of decay (in terms of power of n) of the PUR Forest variance is the same as the PUR Tree variance, and the actual gain appears in the multiplicative constant.

Let us, finally, comment 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:

$$\begin{aligned}\mathbb{E}[(\tilde{s}(X) - s(X))^2] &\leqslant \frac{1}{q}\sum_{l=1}^{q}\mathbb{E}[(\tilde{s}_{\mathbb{U}^l}(X) - s(X))^2] \\ &= \mathbb{E}[(\tilde{s}_{\mathbb{U}^1}(X) - s(X))^2] \,.\end{aligned}$$

So from Proposition 4.2, we deduce that:

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

$$\mathbb{E}[(\tilde{s}(X) - s(X))^2] \leqslant \frac{6MC^2}{(k+1)^2} \tag{12}$$

■

4.6. *Risk bounds for PURF*

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

Proposition 4.7: *If $k \xrightarrow[n\to+\infty]{} +\infty$, $\frac{k}{n} \xrightarrow[n\to+\infty]{} 0$, $0 < \mu \leqslant M$, s is C-Lipschitz and $q \xrightarrow[n\to+\infty]{} +\infty$, the risk of a PURF satisfies:*

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

■

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

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

$$\mathbb{E}[(\hat{s}(X) - s(X))^2] \leqslant Kn^{-2/3} + \underset{n\to+\infty}{\mathrm{o}}(n^{-2/3})$$

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 PUR Forest is systematically reduced compared to a PUR Tree and the bias of a PUR Forest is not larger than the one of a PUR Tree, the risk of a PUR Forest is actually lower.

5. Simulations

In this section, we lead simulation experiments aiming at illustrating results of Section 4 and comparing PURF with RFRI.

5.1. *Experiments*

Experimentations are done on four simulated datasets. We keep the framework of Model (1). In addition, we assume that $\varepsilon \sim \mathcal{N}(0, 1/4)$, 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` : $s : [0,1] \to \mathbb{R},\ x \mapsto |x - 1/2|$
- `stump data` : $s : [0,1] \to \mathbb{R},\ x \mapsto \mathbb{1}_{x \leqslant 1/2} - \mathbb{1}_{x > 1/2}$

We also take several values for the number of data :
$n \in \{100, 500, 1000, 5000, 10000\}$. 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 Purely Uniformly Random Trees (PURT) and Purely Uniformly Random Forests (PURF) on one hand, and Random Trees-Random Inputs (RTRI) and Random Forests-Random Inputs (RFRI) on the other hand. Finally, we confront PURF and RFRI.

RFRI are extensively used in practive and present very good performances. We stress that they do not belong to the Purely Random Forests family, because partitions associated to each trees of RFRI are optimized using the learning sample. So, 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 datasets, 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 summarized in Figure 1.

In the top-left graph, we plot the ratio PURFvariance/PURTvariance as a function of n. The ratio PURFbias/PURTbias and PURFrisk/PURTrisk are respectively plot 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 induvidual 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, 3, 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 datasets.

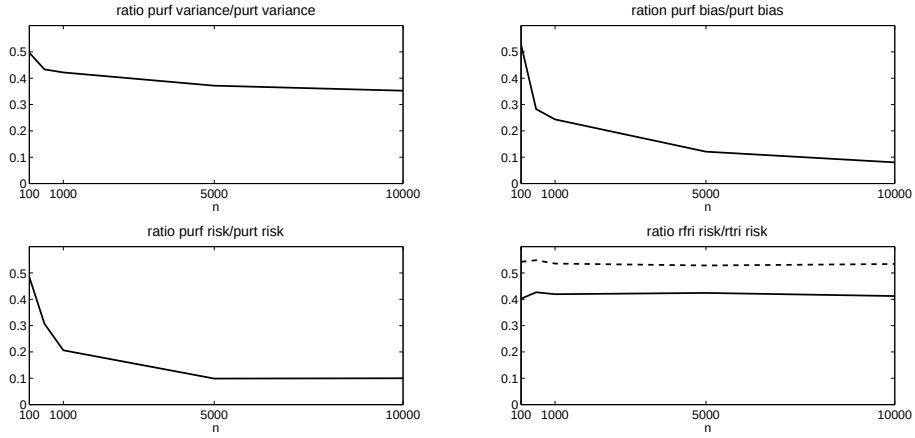

Figure 1. Comparisons between forests and trees for `sinus data`

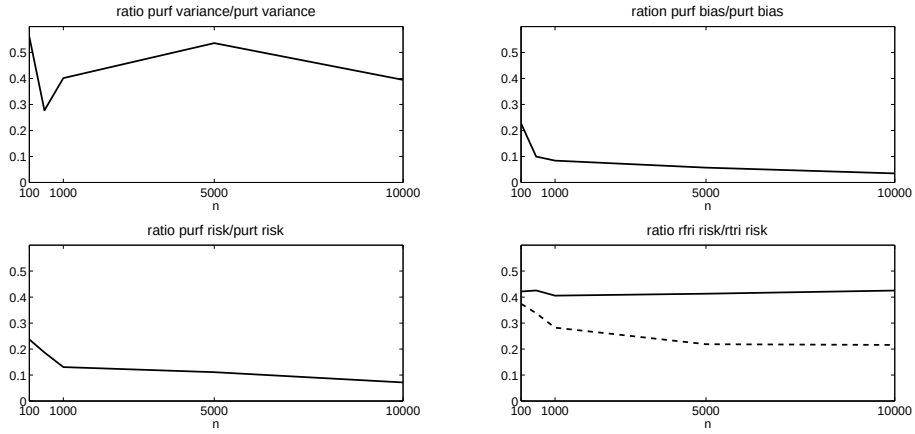

Figure 2. Comparisons between forests and trees for `square data`

5.2. *Comments and discussions*

We see that graphs of Figures 1, 2 and 3 are very similar. We summarize the results by the following comments:

- the ratio PURFvariance/PURTvariance 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 1/2. So our upper bound with factor 3/4 seems to be improvable.
- 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 conver-

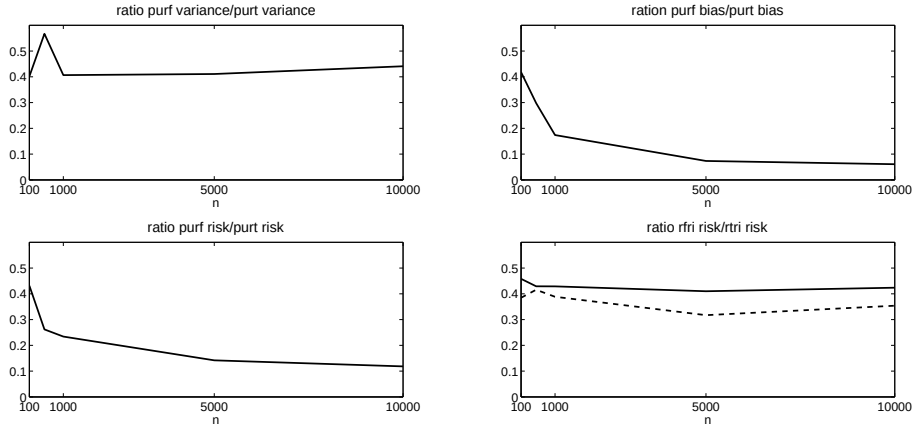

Figure 3. Comparisons between forests and trees for `abs data`

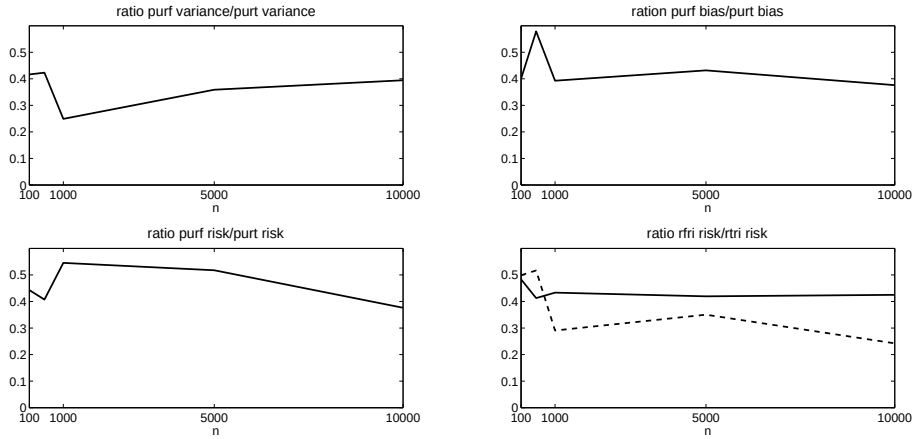

Figure 4. Comparisons between forests and trees for `stump data`

gence of the risk (we see that the ratio PURFrisk/PURTrisk decreases as n increases).

- the ratio 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 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 (maybe with a slow decrease for the variant RFRI(k)).

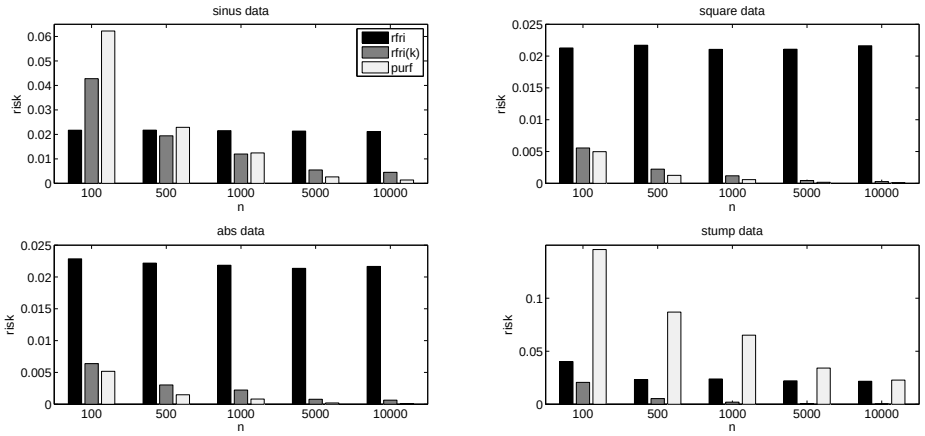

Figure 5. Estimated risks of RFRI, RFRI(k) and PURF for the four simulated datasets

These differences are natural because regression function of `stump data` is discontinuous, whereas it is more regular in the other datasets (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 improves the risk, compared to 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 beetwen 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); Goldstein et al. (2010). This point surely deserves a more intensive study.

The situation of `stump data` is again very different from other datasets. Here, PURF performs the worst (even if it manage to give a raisonable 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 si that since RFRI(k) optimizes the partitions associated to trees using the learning sample, it can better track some discontinuity of the regression function than PURF, in which partitions are chosen indepently of the learning sample. In addition, even if PURF is better thant RFRI(k) in the three other datasets, 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 compare in this study.

6. Conclusion

In the context of purely random forests, 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 emphasize, for a very simple version of random forests, the actual gain of using a random forest 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 to the variance of a tree. This is, in this specific context, a proof of the well-known conjecture for random forests: "a random forest, by aggregating several random trees, reduces variance and leaves the bias unchanged" which can be found for example in Hastie et al. (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 generalize this result, which could handle more complex versions of random forests and relax the hypotheses we made here. Obviously, a more ambitious goal would be to give some precise insights explaining the outstanding performances of random forests-random inputs (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:

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2 \,|\, \mathbb{U}] = \sum_{\lambda \in \mathbb{U}} p_\lambda \mathbb{E}\Big[\frac{1}{n\hat{p}_\lambda}\Big](\sigma^2 + \sigma_\lambda^2) \tag{13}$$

where $\hat{p}_\lambda = \dfrac{\sharp\{i :\ X_i \in \mathbb{U}\}}{n}$.

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

$$Var_\lambda = p_\lambda \mathbb{E}\Big[\frac{1}{n\hat{p}_\lambda}\Big](\sigma^2 + \sigma_\lambda^2) \tag{14}$$

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:

$$\mathbb{C}(\hat{s}_{\mathbb{U}^1}, \hat{s}_{\mathbb{U}^2}) = \mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \tilde{s}_{\mathbb{U}^2}(X))]$$

Let us consider $\mathbb{U}^1 = (\lambda_1^1, \ldots, \lambda_k^1)$ and $\mathbb{U}^2 = (\lambda_1^2, \ldots, \lambda_k^2)$ the partitions respectively associated to $\hat{s}_{\mathbb{U}^1}$ 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 $2k - 1 \leqslant m \leqslant k^2$.

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

$$\begin{aligned}
&\left(\sum_{r=1}^{k} (\hat{\beta}_{\lambda_r^1} - \beta_{\lambda_r^1}) \mathbb{1}_{X \in \lambda_r^1}\right) \left(\sum_{s=1}^{k} (\hat{\beta}_{\lambda_s^2} - \beta_{\lambda_s^2}) \mathbb{1}_{X \in \lambda_s^2}\right) \\
&= \sum_{t=1}^{m} (\hat{\beta}_{\mu_t, \lambda_r^1} - \beta_{\mu_t, \lambda_r^1})(\hat{\beta}_{\mu_t, \lambda_s^2} - \beta_{\mu_t, \lambda_s^2}) \mathbb{1}_{X \in \mu_t} \qquad (15)
\end{aligned}$$

where
$\begin{cases} \hat{\beta}_{\mu_t, \lambda_r^1} = \hat{\beta}_{\lambda_r^1} \text{ and } \beta_{\mu_t, \lambda_r^1} = \beta_{\lambda_r^1}, \text{ if } \mu_t \subset \lambda_r^1 \\ \hat{\beta}_{\mu_t, \lambda_s^2} = \hat{\beta}_{\lambda_s^2} \text{ and } \beta_{\mu_t, \lambda_s^2} = \beta_{\lambda_s^2}, \text{ if } \mu_t \subset \lambda_s^2 \end{cases}$

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,

$$\begin{aligned}
S_1(X) &= (\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \tilde{s}_{\mathbb{U}^2}(X)) \mathbb{1}_{X \in \mu_1} \\
&= (\hat{\beta}_{\lambda_1^1} - \beta_{\lambda_1^1})(\hat{\beta}_{\lambda_1^2} - \beta_{\lambda_1^2}) \mathbb{1}_{X \in \mu_1} \\
&= \Big(\frac{1}{n\hat{p}_{\lambda_1^1}} \sum_{i:\, X_i \in \lambda_1^1} (Y_i - \beta_{\lambda_1^1})\Big)\Big(\frac{1}{n\hat{p}_{\lambda_1^2}} \sum_{i:\, X_i \in \lambda_1^2} (Y_i - \beta_{\lambda_1^2})\Big) \mathbb{1}_{X \in \mu_1} \\
&= \frac{1}{n\hat{p}_{\lambda_1^1} n\hat{p}_{\lambda_1^2}} \sum_{i^1:\, X_{i^1} \in \lambda_1^1} \sum_{i^2:\, X_{i^2} \in \lambda_1^2} (Y_{i^1} - \beta_{\lambda_1^1})(Y_{i^2} - \beta_{\lambda_1^2}) \mathbb{1}_{X \in \mu_1}
\end{aligned}$$

If we denote by $\mathbb{E}^{\Lambda_{1,2}}[.]$ the conditional expectation $\mathbb{E}[. \,|\, (\mathbb{1}_{X_{i^1} \in \lambda_1^1})_{1 \leqslant i^1 \leqslant n}, (\mathbb{1}_{X_{i^2} \in \lambda_1^2})_{1 \leqslant i^2 \leqslant n}]$, we have:

$$\begin{aligned}
&\mathbb{E}[S_1(X) \,|\, \mathbb{U}^1, \mathbb{U}^2] \\
&= q_1 \mathbb{E}\Big[\frac{1}{n\hat{p}_{\lambda_1^1} n\hat{p}_{\lambda_1^2}} \sum_{i^1:\, X_{i^1} \in \lambda_1^1} \sum_{i^2:\, X_{i^2} \in \lambda_1^2} \mathbb{E}^{\Lambda_{1,2}}[(Y_{i^1} - \beta_{\lambda_1^1})(Y_{i^2} - \beta_{\lambda_1^2})] \,\Big|\, \mathbb{U}^1, \mathbb{U}^2\Big]
\end{aligned}$$

where $q_1 = \mathrm{P}(X \in \mu_1)$

but

$$i^1 \neq i^2 \quad \Longrightarrow \quad \mathbb{E}^{\Lambda_{1,2}}[(Y_{i^1} - \beta_{\lambda_1^1})(Y_{i^2} - \beta_{\lambda_1^2})] = 0$$

because Y_{i^1} and Y_{i^2} are independent. Hence:

$$\begin{aligned}
&\mathbb{E}[S_1(X) \,|\, \mathbb{U}^1, \mathbb{U}^2] \\
&= q_1 \mathbb{E}\Big[\frac{1}{n\hat{p}_{\lambda_1^1} n\hat{p}_{\lambda_1^2}} \sum_{i:\, X_i \in \mu_1} \mathbb{E}^{\Lambda_1}[(Y_i - \beta_{\lambda_1^1})(Y_i - \beta_{\lambda_1^2})] \,\Big|\, \mathbb{U}^1, \mathbb{U}^2\Big]
\end{aligned}$$

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

where $\mathbb{E}^{\Lambda_1}[.]$ denotes the conditional expectation $\mathbb{E}[. \,|\, (\mathbb{1}_{X_i \in \lambda_1^1})_{1 \leqslant i \leqslant n}]$.

Now, as

$$\mathbb{E}[(Y_i - \beta_{\lambda_1^1})(Y_i - \beta_{\lambda_1^2}) \,|\, X_i \in \mu_1] = \mathbb{E}[(Y - \beta_{\lambda_1^1})(Y - \beta_{\lambda_1^2}) \,|\, X \in \mu_1]$$

for all i, and

$$\mathbb{E}[(Y - \beta_{\lambda_1^1})(Y - \beta_{\lambda_1^2}) \,|\, X \in \mu_1] = \sigma^2 + \rho^2_{\mu_1}$$

where

$$\rho^2_{\mu_1} = \mathbb{E}[(s(X) - \tilde{s}_{\mathbb{U}^1}(X))(s(X) - \tilde{s}_{\mathbb{U}^2}(X)) \,|\, X \in \mu_1]$$

we get

$$\mathbb{E}[S_1(X) \,|\, \mathbb{U}^1, \mathbb{U}^2] = q_1 \mathbb{E}\Big[\frac{1}{n\hat{p}_{\lambda_1^2}}\Big](\sigma^2 + \rho^2_{\mu_1}) \,.$$

If we suppose in addition that $\mu_2 = \lambda_2^1 \cap \lambda_1^2$, and that $\mu_1 \cup \mu_2 = \lambda_1^2$ we similarly get for the second term of (15):

$$\begin{aligned}
&\mathbb{E}[S_2(X) \,|\, \mathbb{U}^1, \mathbb{U}^2] \\
&= \mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \tilde{s}_{\mathbb{U}^2}(X))\mathbb{1}_{X \in \mu_2} \,|\, \mathbb{U}^1, \mathbb{U}^2] \\
&= q_2 \mathbb{E}\left[\frac{n\hat{q}_2}{n\hat{p}_{\lambda_2^1} n\hat{p}_{\lambda_1^2}}\right](\sigma^2 + \rho^2_{\mu_2})
\end{aligned}$$

where

$$\begin{aligned}
q_2 &= \mathrm{P}(X \in \mu_2) \\
n\hat{q}_2 &= \sharp\{i : X_i \in \mu_2\} \\
\rho^2_{\mu_2} &= \mathbb{E}[(s(X) - \tilde{s}_{\mathbb{U}^1}(X))(s(X) - \tilde{s}_{\mathbb{U}^2}(X)) \,|\, X \in \mu_2] \,.
\end{aligned}$$

Since μ_2 is included in λ_2^1, we have $\hat{q}_2 \leqslant \hat{p}_{\lambda_2^1}$, so:

$$\mathbb{E}[S_2(X) \,|\, \mathbb{U}^1, \mathbb{U}^2] \leqslant q_2 \mathbb{E}\Big[\frac{1}{n\hat{p}_{\lambda_1^2}}\Big](\sigma^2 + \rho^2_{\mu_2}) \,.$$

Finally, by summing the two terms $S_1(X)$ and $S_2(X)$, we deduce that

$$\mathbb{E}[S_1(X) + S_2(X) \,|\, \mathbb{U}^1, \mathbb{U}^2] \leqslant p_{\lambda_1^2} \mathbb{E}\Big[\frac{1}{n\hat{p}_{\lambda_1^2}}\Big](\sigma^2 + \rho^2_{\mu_1} + \rho^2_{\mu_2})$$

In conclusion, we succeeded to bound the sum of the first two terms of (15) by an expression very close to Var_λ (defined in (14)). The only difference comes from the fact that instead of σ^2_λ we have $\rho^2_{\mu_1} + \rho^2_{\mu_2}$.

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

We denote by $S_t(X)$ the t-th term of (15), i.e. $S_t(X) = (\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \tilde{s}_{\mathbb{U}^2}(X))\mathbb{1}_{X\in\mu_t}$.

Lemma 7.1: *Let r be in $\{1,\dots,k\}$ and denote by K_r a subset of $\{1,\dots,m\}$ such that*

$$\bigcup_{t\in K_r} \mu_t = \lambda_r^1 \tag{16}$$

then

$$\mathbb{E}\Big[\sum_{t\in K_r} S_t(X) \,|\, \mathbb{U}^1, \mathbb{U}^2\Big] \leqslant p_{\lambda_r^1}\mathbb{E}\Big[\frac{1}{n\hat{p}_{\lambda_r^1}}\Big](\sigma^2 + \Sigma_r^2)$$

where $\Sigma_r^2 = \sum_{t\in K_r} \rho_{\mu_t}^2$.

■

Indeed for all $t \in K_r$,

$$\mathbb{E}[S_t(X) \,|\, \mathbb{U}^1, \mathbb{U}^2] \leqslant q_t\mathbb{E}\Big[\frac{1}{n\hat{p}_{\lambda_r^1}}\Big](\sigma^2 + \rho_{\mu_t}^2)$$

where

$$\begin{aligned} q_t &= \mathrm{P}(X \in \mu_t) \\ \rho_{\mu_t}^2 &= \mathbb{E}[(s(X) - \tilde{s}_{\mathbb{U}^1}(X))(s(X) - \tilde{s}_{\mathbb{U}^2}(X)) \,|\, X \in \mu_t]\,. \end{aligned}$$

Thus,

$$\mathbb{E}\Big[\sum_{t\in K_r} S_t(X) \,|\, \mathbb{U}^1, \mathbb{U}^2\Big] \leqslant \mathrm{P}(X \in \bigcup_{t\in K_r} \mu_t)\ \mathbb{E}\Big[\frac{1}{n\hat{p}_{\lambda_r^1}}\Big](\sigma^2 + \Sigma_r^2)\,.$$

From Relation (16) we have $\bigcup_{t\in K_r} \mu_t = \lambda_r^1$, which concludes the proof of Lemma 7.1.

Therefore, we can upper bound the initial sum (15) of m terms (we recall that $2k-1 \leqslant m \leqslant k^2$) by a sum of k terms of the same order as Var_λ only involving sets of the partition $\mathbb{U}^1$. 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*

$$\lambda_r^1 \subset \lambda_s^2\ ,\ \lambda_{r'}^1 \subset \lambda_s^2$$

the expression

$$\mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \tilde{s}_{\mathbb{U}^2}(X))\mathbb{1}_{X\in\lambda_r^1\cup\lambda_{r'}^1} \,|\, \mathbb{U}^1, \mathbb{U}^2]$$

is upper bounded by

$$p_{\lambda_s^2}\mathbb{E}\Big[\frac{1}{n\hat{p}_{\lambda_s^2}}\Big](\sigma^2+\Sigma_s^2)\ .$$

where $\Sigma_s^2=\rho_t^2+\rho_{t'}^2$*, with* t *and* t' *are such that* $\lambda_r^1=\mu_t$ *and* $\lambda_{r'}^1=\mu_{t'}$

■

Indeed,

$$\begin{aligned}&\mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X)-\tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X)-\tilde{s}_{\mathbb{U}^2}(X))\mathbb{1}_{X\in\lambda_r^1}\,|\,\mathbb{U}^1,\mathbb{U}^2]\\&\leqslant p_{\lambda_r^1}\mathbb{E}\Big[\frac{1}{n\hat{p}_{\lambda_s^2}}\Big](\sigma^2+\rho_t^2)\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X)-\tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X)-\tilde{s}_{\mathbb{U}^2}(X))\mathbb{1}_{X\in\lambda_{r'}^1}\,|\,\mathbb{U}^1,\mathbb{U}^2]\\&\leqslant p_{\lambda_{r'}^1}\mathbb{E}\Big[\frac{1}{n\hat{p}_{\lambda_s^2}}\Big](\sigma^2+\rho_{t'}^2)\ .\end{aligned}$$

Finally, since $p_{\lambda_r^1}+p_{\lambda_{r'}^1}\leqslant p_{\lambda_s^2}$, $\rho_t^2\leqslant\rho_t^2+\rho_{t'}^2$ and $\rho_{t'}^2\leqslant\rho_t^2+\rho_{t'}^2$, the result is obtained by summing the two terms.

As in Proposition 3.1, we replace all $p_{\lambda_j}^l\,\mathbb{E}\Big[\frac{1}{n\hat{p}_{\lambda_j^l}}\Big]$ by their estimates $(1+\delta_{n,p_{\lambda_j^l}})$.

By repeatedly applying this lemma for all sets, we can upper bound

$$\mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X)-\tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X)-\tilde{s}_{\mathbb{U}^2}(X))\,|\,\mathbb{U}^1,\mathbb{U}^2]$$

by a sum of $N_{\mathcal{U}}$ terms of the form $(1+\delta_{n,\tilde{p}_{\lambda_t^l}})(\sigma^2+\Sigma_t^2)$, where $\tilde{p}_{\lambda_t^l}$ denotes for some $j\in\{1,\dots,k\}$ either $p_{\lambda_j^1}$ or $p_{\lambda_j^2}$ depending on the fact that we are in the situation of Lemma 7.1 or Lemma 7.2. And $N_{\mathcal{U}}=k-M_{\mathcal{U}}$ with

$$M_{\mathcal{U}}=\sum_{1\leqslant r<r'\leqslant k}\ \sum_{1\leqslant s\leqslant k}\mathbb{1}_{\lambda_r^1\subset\lambda_s^2\,,\,\lambda_{r'}^1\subset\lambda_s^2}\ .$$

We conclude the proof of Theorem 3.2 by integrating with respect to $(\mathbb{U}_1,\mathbb{U}_2)$.

7.2. *Proof of Corollary 4.1*

Here, we have $\mathbb{U}=\{[0,U_{(1)}[,[U_{(1)},U_{(2)}[,\dots,[U_{(k)},1[\}$ where $(U_{(1)},U_{(2)},\dots,U_{(k)})$ denotes the ordered statistics of $(U_1,U_2,\dots,U_k)$. With these notations, Equation (5) becomes:

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X)-\tilde{s}_{\mathbb{U}}(X))^2]=\frac{1}{n}\sum_{j=0}^{k}\big(\sigma^2+\sigma^2\mathbb{E}[\delta_{n,p_j}]+\mathbb{E}[\sigma_j^2]+\mathbb{E}[\sigma_j^2\delta_{n,p_j}]\big)\qquad(17)$$

where

$$p_j = \mathrm{P}(U_{(j)} \leqslant X < U_{(j+1)})$$
$$\sigma_j^2 = \mathbb{E}[(s(X) - \tilde{s}_{\mathbb{U}}(X))^2 \,|\, U_{(j)} \leqslant X < U_{(j+1)}]$$

And we must prove that the three last terms in the sum of Equation (17) are negligible compared to the constant term σ^2.
Let us fix $0 \leqslant j \leqslant k$. As it can be found e.g. in David and Nagaraja (2003, chap. 6), the probability density function of $U_{(j+1)} - U_{(j)}$ is the function $t \in [0,1] \longmapsto k(1-t)^{k-1}$.

- For the second term $\mathbb{E}[\delta_{n,p_j}]$:
  from Arlot (2008) we have $\delta_{n,p_j} \leqslant \kappa_3 (np_j)^{-1/4}$, where κ_3 is a positive constant. So,

$$\begin{aligned}
\mathbb{E}[\delta_{n,p_j}] &\leqslant \kappa_3 \mathbb{E}[(np_j)^{-1/4}] \\
&= \frac{\kappa_3}{n^{-1/4}} \mathbb{E}[p_j^{-1/4}] \\
&\leqslant \frac{\kappa_3}{(\alpha n)^{-1/4}} \mathbb{E}[(U_{(j+1)} - U_{(j)})^{-1/4}] \\
&\leqslant \frac{\kappa_4}{\alpha^{-1/4}} \left(\frac{k}{n}\right)^{1/4}
\end{aligned}$$

  where $\alpha = \min_{[0,1]} \mu > 0$ and κ_4 is another positive constant.
  Since $\frac{k}{n} \xrightarrow[n \to +\infty]{} 0$ the last upper bound tends to 0 as n tends to infinity.
- For the third term $\mathbb{E}[\sigma_j^2]$:
  $\sigma_j^2 = \mathbb{E}[(s(X) - \tilde{s}_{\mathbb{U}}(X))^2 \,|\, U_{(j)} < X \leqslant U_{(j+1)}]$
  $\leqslant C^2 (U_{(j+1)} - U_{(j)})^2$ because s is $C-$Lipschitz.
  So, $\mathbb{E}[\sigma_j^2] \leqslant C^2 \mathbb{E}[(U_{(j+1)} - U_{(j)})^2] = C^2 \frac{2}{(k+1)(k+2)}$ which tends to 0 as k tends to infinity.
- For the last term, the following inequality is sufficient to conclude:
  $\mathbb{E}[\sigma_j^2 \delta_{n,p_j}] \leqslant C^2 \mathbb{E}[\delta_{n,p_j}]$, because $U_{(j+1)} - U_{(j)} \leqslant 1$.

7.3. *Proof of Proposition 4.2*

We keep the notations of Section 7.2. In addition, we define β_j for any $0 \leqslant j \leqslant k$ by:

$$\beta_j = \mathbb{E}[Y \,|\, U_{(j)} < X \leqslant U_{(j+1)}]$$

Function s is supposed to be C-Lipschitz, so

$$\mathbb{E}[(\tilde{s}_{\mathbb{U}}(X) - s(X))^2] = \mathbb{E}[(\sum_{j=0}^{k} (s(X) - \beta_j) \mathbb{1}_{U_{(j)} < X \leqslant U_{(j+1)}})^2]$$

$$
\begin{aligned}
&= \mathbb{E}[\sum_{j=0}^{k}(s(X)-\beta_j)^2\ \mathbb{1}_{U_{(j)}<X\leqslant U_{(j+1)}}] \\
&\leqslant \mathbb{E}[\sum_{j=0}^{k} C^2(U_{(j+1)}-U_{(j)})^2\ \mathbb{1}_{U_{(j)}<X\leqslant U_{(j+1)}}] \\
&= C^2\mathbb{E}[\sum_{j=0}^{k}(U_{(j+1)}-U_{(j)})^2\ \mathrm{P}(U_{(j)}<X\leqslant U_{(j+1)})] \\
&\leqslant C^2\mathbb{E}[\sum_{j=0}^{k} M(U_{(j+1)}-U_{(j)})^3] \\
&\qquad\qquad \text{because } \mu \text{ is bounded by } M \\
&= MC^2\sum_{j=0}^{k}\mathbb{E}[(U_{(j+1)}-U_{(j)})^3] \\
&= MC^2\frac{6}{(k+2)(k+3)} \\
&\leqslant \frac{6MC^2}{(k+1)^2}\ .
\end{aligned}
$$

7.4. *Proof of Proposition 4.5*

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

Corollary 7.3: *If* $k \xrightarrow[n\to+\infty]{} +\infty$, $\dfrac{k}{n} \xrightarrow[n\to+\infty]{} 0$, $\mu > 0$ *and* s *is* C*-Lipschitz, we have,*

$$
\begin{aligned}
\mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X)-\tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X)-\tilde{s}_{\mathbb{U}^2}(X))] &\leqslant \frac{\sigma^2\mathbb{E}[N_{\mathcal{U}}]}{n} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right) \\
&\leqslant \frac{3}{4}\frac{\sigma^2(k+1)}{n} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right)\ .
\end{aligned}
$$

■

Because of the simple draws of random partitions, the number $M_{\mathcal{U}}$ is explicitly computable (we know the distribution of the two ordered statistics) and it is shown to be equivalent to $\dfrac{1}{4}(k+1)$ as k tends to $+\infty$ (see Lemma 7.4 below).
As in Corollary 4.1, we have to prove that all terms of the sum are negligible compared to the constant one σ^2. To deal with the fact that the number of terms in the sum is now random, we use the following inequality:

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

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

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

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

So, we have

$$\mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \tilde{s}_{\mathbb{U}^2}(X))] \leqslant \frac{\sigma^2 \mathbb{E}[N_{\mathcal{U}}]}{n} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right) .$$

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:

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

Hence,

$$\mathbb{E}[M_{\mathcal{U}}] = \frac{k+1}{4} + \underset{k\to+\infty}{\mathrm{o}}(k) .$$

■

We then obtain that

$$\mathbb{E}[N_{\mathcal{U}}] = \frac{3}{4}(k+1) + \underset{k\to+\infty}{\mathrm{o}}(k) .$$

Let us prove lemma 7.4.

$$\mathbb{E}[M_{\mathcal{U}}] = \sum_{r=1}^{k-2}\sum_{s=1}^{k-1} \mathrm{P}(U_{(s)}^2 < U_{(r)}^1 < U_{(r+1)}^1 < U_{(r+2)}^1 < U_{(s+1)}^2)$$

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

$$\begin{aligned}
&\mathrm{P}(U_{(s)}^2 < U_{(r)}^1 < U_{(r+1)}^1 < U_{(r+2)}^1 < U_{(s+1)}^2) \\
&= \mathrm{P}(U_{(s)}^2 < U_{(r)}^1 \text{ and } U_{(r+2)}^1 < U_{(s+1)}^2) \\
&= \sum_{j=r+2}^{k}\sum_{i=0}^{r-1} \frac{k!}{i!(j-i)!(k-j)!}\mathbb{E}[(U_{(s)}^2)^i (U_{(s+1)}^2 - U_{(s)}^2)^{j-i}(1 - U_{(s+1)}^2)^{k-j}]
\end{aligned}$$

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

So,

$$\begin{aligned}
&\mathbb{E}[M_{\mathcal{U}}] \\
&= \frac{(k!)^2}{(2k)!} \sum_{r=1}^{k-2} \sum_{s=1}^{k-1} \left(\sum_{i=0}^{r-1} \binom{i+(s-1)}{i} \right) \left(\sum_{j=r+2}^{k} \binom{k-j+k-(s+1)}{k-j} \right) \\
&= \frac{(k!)^2}{(2k)!} \sum_{r=1}^{k-2} \sum_{s=1}^{k-1} \binom{r-1+s}{r-1} \binom{2k-r-2-s}{k-r-2}
\end{aligned}$$

(by elementary properties of binomial coefficients, see e.g. Graham et al. 1989, p.160)

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

(by defining $t = r + s$)

$$= \frac{k-2}{4(2k-1)} \sum_{t=0}^{2k-5} [\mathrm{F}_{\mathcal{H}(2k-3,t+1,k-3)}(t) - \mathrm{F}_{\mathcal{H}(2k-3,t+1,k-3)}(t-k+1)]$$

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

$$\begin{aligned}
&= \frac{k-2}{4(2k-1)} 2 \sum_{t=0}^{k-3} \mathrm{F}_{\mathcal{H}(2k-3,t+1,k-3)}(t) \\
&= \frac{k-2}{2(2k-1)} \left[\sum_{t=0}^{k-4} \left(1 - \frac{\binom{t+1}{t+1} \binom{2k-3-(t+1)}{k-3-(t+1)}}{\binom{2k-3}{k-3}} \right) + 1 \right] \\
&= \frac{k-2}{2(2k-1)} \Big(k - 3 + \frac{4}{k+1} \Big).
\end{aligned}$$

References

Arlot, S. (2008), 'V-fold cross-validation improved: V-fold penalization,' Preprint, arXiv:0802.0566v2.

Biau, G., Devroye, L., 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., Zhao, G. (2001), 'Pert - Perfect random tree ensembles,' *Computing Science and Statistics*, 33:490-497.

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

Díaz-Uriarte, R., 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., Patashnik O. (1989), *Concrete mathematics*. Addison-Wesley.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning*. Second edition. Springer.

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

Liaw, A., Wiener, M. (2002), 'Classification and Regression by randomForest,' *R News*, 2(3):18-22.