# Risk bounds for purely uniformly random forests

Robin Genuer

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *Risk bounds for purely uniformly random forests*

Robin Genuer

**N° 7318**

Juin 2010

Thème COG


ISSN 0249-6399 ISRN INRIA/RR--7318--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY - ÎLE-DE-FRANCE

# Risk bounds for purely uniformly random forests

Robin Genuer * †

Thème COG — Systèmes cognitifs
Équipes-Projets SELECT

Rapport de recherche n° 7318 — Juin 2010 — 19 pages

**Abstract:** 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, a simplified version of random forests, called purely random forests, which can be theoretically handled more easily, has been considered. In this paper we introduce a variant of this kind of random forests, that we call purely uniformly random forests. 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.

**Key-words:** RANDOM FORESTS, NON-PARAMETRIC REGRESSION, RATE OF CONVERGENCE, RANDOMIZATION.

* Univ Paris-Sud, Laboratoire de Mathématique, UMR 8628, Orsay F-91405
† Inria Saclay Ile-de-France

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université
4, rue Jacques Monod, 91893 ORSAY Cedex
Téléphone : +33 1 72 92 59 00

# Bornes de risque pour les forêts purement uniformément aléatoires.

**Résumé :** Introduites par Leo Breiman en 2001, les forêts aléatoires sont une méthode statistique très performante. D'un point de vue théorique, leur analyse est difficile, du fait de la complexité de l'algorithme. Pour expliquer ces performances, des versions de forêts aléatoires simplifiées, et donc plus faciles à analyser, ont été introduites. Ces versions ont été appelées forêts purement aléatoires. Dans cet article, nous introduisons une autre version simplifiée, que nous appelons forêts purement uniformément aléatoires. Dans un contexte de régression, avec une seule variable explicative, nous montrons que les arbres aléatoires ainsi que les forêts aléatoires atteignent la vitesse de convergence minimax. De plus, nous prouvons que les forêts aléatoires améliorent les performances des arbres aléatoires, en réduisant la variance des estimateurs associés d'un facteur de trois quarts.

**Mots-clés :** Forêts aléatoires, Régression non-paramétrique, Vitesse de convergence, Randomisation.

# 1 Introduction

Random forests (RF), introduced by Leo Breiman in 2001 [3], 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 by Leo Breiman, 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.

In [3], Leo Breiman introduced a specific instance of random forest, called random forests-RI, which has been adopted in many fields as a reference method. Indeed, random forests-RI are simple to use, and are efficiently coded in the popular R-package `randomForest` [11]. They are effective for a predictive goal and they can also be used for variable selection (see e.g. [6], [7]).

However, forests-RI 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 forests-RI, the splits are optimized using the learning sample. This independence between splits and learning sample makes mathematical analysis easier. In [4], Cutler and Zhao 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. [2] showed that both purely random trees and purely random forests are universally consistent.

Our paper offers to examine another simple variant of random forests, which can be put in the so-called purely random forests family. We call it *purely uniformly random forests* 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 of this paper 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 and Section 4 give some risk bounds for purely uniformly random trees and purely uniformly random forests respectively. Section 5 concludes the paper, while proofs are collected in Section 6.

# 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. We consider the following

statistical model:

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

$s$ is the unknown regression function and the goal is to estimate $s$. We make the following assumptions on model (1):

- $X \in [0,1]$ *with continuous density function* $\mu$*;*
- $(\varepsilon_1, \dots, \varepsilon_n)$ *are i.i.d. observations of* $\varepsilon$*, independent of* $\mathcal{L}_n$*, with* $\mathbb{E}[\varepsilon] = 0$ *and where* $\mathrm{Var}(\varepsilon) = \sigma^2$ *is assumed to be known.*

Note that we deal only with a one-dimensional predictor space.

This paper aims at comparing performances in estimating $s$ using a single random tree and a random forest of a special kind, described in the next section.

# 3 Risk bounds for Purely Uniformly Random Trees

## 3.1 Tree definition

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-RI, 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.

Let us 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 $[0,1]$, as well as the corresponding estimator.

More precisely, let $\mathbb{U} = (U_1, \dots, U_k)$ be $k$ i.i.d. random variables of uniform distribution on $[0,1]$, where $k$ is a natural integer which will depend on the number of observations $n$.

A Purely Uniformly Random Tree (PURT), associated with $\mathbb{U}$, is defined for $x \in [0,1]$ as:

$$\hat{s}_{\mathbb{U}}(x) = \sum_{j=0}^{k} \hat{\beta}_j \mathbb{1}_{U_{(j)} < x \leqslant U_{(j+1)}}$$

where

$$\hat{\beta}_j = \frac{1}{\sharp\{i : \, U_{(j)} < X_i \leqslant U_{(j+1)}\}} \sum_{i: \, U_{(j)} < X_i \leqslant U_{(j+1)}} Y_i$$

and $(U_{(1)}, \dots, U_{(k)})$ is the ordered statistics of $(U_1, \dots, U_k)$ and $U_{(0)} = 0$, $U_{(k+1)} = 1$. $\sharp\,\mathcal{E}$ denotes the cardinality of the set $\mathcal{E}$.

**Remark 1** *Let us mention that if* $\sharp\{i : \, U_{(j)} < X_i \leqslant U_{(j+1)}\} = 0$*, we set* $\hat{\beta}_j = 0$*. However as we will see in Section 3.2, our assumptions on* $k$ *and* $n$ *will make the probability of observing such an event tend to* $0$*.*

In addition, let us define, for $x \in [0,1]$:

$$\tilde{s}_{\mathbb{U}}(x) = \sum_{j=0}^{k} \beta_j \mathbb{1}_{U_{(j)} < x \leqslant U_{(j+1)}}$$

where

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

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:

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

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 Variance of a tree

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 [1].

**Proposition 1** *Conditionally on $\mathbb{U}$, the variance term of decomposition (2) satisfies:*

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2 \,|\, \mathbb{U}] = \frac{1}{n} \sum_{j=0}^{k} (1 + \delta_{n,p_j})(\sigma^2 + (\sigma_j^d)^2) \tag{3}$$

*where*

- $p_j = \mathrm{P}(U_{(j)} < X \leqslant U_{(j+1)})$,
- $(\sigma_j^d)^2 = \mathbb{E}[(s(X) - \tilde{s}_{\mathbb{U}}(X))^2 \,|\, U_{(j)} < X \leqslant U_{(j+1)}]$,
- $\delta_{n,p} \xrightarrow[np \to +\infty]{} 0$.

■

We now integrate equation (3) w.r.t. $\mathbb{U}$, and we get the following equality:

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2] = \frac{1}{n}\sum_{j=0}^{k}\left(\sigma^2 + \sigma^2\mathbb{E}[\delta_{n,p_j}] + \mathbb{E}[(\sigma_j^d)^2] + \mathbb{E}[(\sigma_j^d)^2\delta_{n,p_j}]\right) \quad (4)$$

Let us stress that equation (4) is general, since it does not depend on the distribution of $\mathbb{U}$. Hence, it can be used for any random partition distributions.

Finally, 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 (4) the following proposition:

**Proposition 2** *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 PUR Tree satisfies:*

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2] = \frac{\sigma^2(k+1)}{n} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right) \quad (5)$$

*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 Proposition 2 can be found in Section 6.1.

The first two hypotheses of Proposition 2 ($k \xrightarrow[n\to+\infty]{} +\infty$, $\frac{k}{n} \xrightarrow[n\to+\infty]{} 0$) are the same natural conditions found by Biau et al. [2] 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.

## 3.3 Bias of a tree

We now turn to the bias term of decomposition (2). Direct calculations (see Section 6.2 for details) lead to the following upper bound for the bias term of a PURT:

**Proposition 3** *If* $\mu$ *is bounded by* $M > 0$ *and* $s$ *is* $C$*-Lipschitz, the bias of a PURT is upper bounded by:*

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

■

## 3.4 Risk bounds for a tree

Putting together (5) and (6) leads to the following risk bound for a PURT.

**Theorem 1** *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:*

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - s(X))^2] \leqslant \frac{\sigma^2(k+1)}{n} + \frac{6MC^2}{(k+1)^2} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right) \quad (7)$$

■

The balance between the two first terms of the right hand side (r.h.s.) of (7) leads to take $(k+1) = n^{1/3}$, and gives the following upper bound for the risk of a PURT.

**Corollary 1** *Under the assumptions of Theorem 1,*

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

*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 [10]).

Let us now analyze purely uniformly random forests. As a result, we emphasize an improvement given by a forest compared to a single tree.

# 4 Risk bounds for Purely Uniformly Random Forests

## 4.1 Forest definition

A random forest is the aggregation of a collection of random trees. So, in the context of Purely Uniformly Random Forests (PURF), the principle is to generate several PUR Trees by drawing several random partitions given by uniform random variables, and to aggregate them.

Let $\mathbb{V} = (\mathbb{U}^1, \ldots, \mathbb{U}^q)$ be $q$ i.i.d. random vectors of the same distribution as $\mathbb{U}$ (defined in Section 3.1). That is for $l = 1, \ldots, q, \quad \mathbb{U}^l = (U_1^l, \ldots, U_k^l)$ where the $(U_j^l)_{1\leqslant j\leqslant k}$ are i.i.d. random variables of uniform distribution on $[0,1]$.

A PURF, associated with $\mathbb{V}$, is defined for $x \in [0,1]$ as follows:

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

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

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

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

$$\begin{aligned}\mathbb{E}[(\hat{s}(X) - s(X))^2] &= \mathbb{E}[(\hat{s}(X) - \tilde{s}(X))^2] + \mathbb{E}[(\tilde{s}(X) - s(X))^2] \\ &= \text{variance term} + \text{bias term}\end{aligned} \tag{8}$$

## 4.2 Variance of a forest

We first deal with the variance term of decomposition (8). 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 $\hat{s}(x) = \frac{1}{q}\sum_{l=1}^{q} \hat{s}_{\mathbb{U}^l}(x)$, the variance term satisfies:

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

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

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

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

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

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

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

(it is detailed in Section 6.3) and it leads to the following theorem, which gives the behavior of the variance of a PURF:

**Theorem 2** *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:*

$$\mathbb{E}[(\hat{s}(X) - \tilde{s}(X))^2] \leqslant \frac{3}{4}\frac{\sigma^2(k+1)}{n} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right) \tag{9}$$

■

Theorem 2 is to be compared with Proposition 2 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.

We mention that, as in the analysis of the variance of a tree (see equation (4)), we derive, in the proof of Theorem 2, a general statement (see equation (13) in Section 6.3), which does not depend on the distribution of the partition defining the random trees.

Let us, finally, comment the hypotheses of Theorem 2. First, note that the hypotheses on $k$ and $n$ are the same as in Proposition 2, which allows a fair comparison between the two results. Second, the hypothesis on $q$ allows to ensure that the upper bound on the covariance (given by Corollary 3 in Section 6.3) leads to the same upper bound for the variance of the forest. Finally, the other hypotheses ($\mu > 0$, $s$ is $C$-Lipschitz) are the same as in Proposition 2 and help to control negligible terms.

## 4.3 Bias of a forest

We now deal with the bias term of decomposition (8). 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 3, we deduce that:

**Proposition 4** *If $\mu$ is bounded by $M > 0$ and $s$ is $C$-Lipschitz, the bias of a PURF satisfies the same inequality as (6), that is:*

$$\mathbb{E}[(\tilde{s}(X) - s(X))^2] \leqslant \frac{6MC^2}{(k+1)^2} \tag{10}$$

■

## 4.4 Risk bounds for a forest

Putting together (9) and (10) leads to the following risk bound for a PURF.

**Theorem 3** *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 2** *Under the assumptions of Theorem 3,*

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

*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 Conclusion

We 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. [9].

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-RI.

# 6 Proofs

## 6.1 Proof of Proposition 2

We must show that the three last terms in the sum of equation (4) are negligible compared to the constant term $\sigma^2$.
Let us fix $0 \leqslant j \leqslant k$. As it can be found e.g. in Chapter 6 of [5], 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 [1] we have $\delta_{n,p_j} \leqslant \kappa_3 (np_j)^{-1/4}$, where $\kappa_3$ is a positive constant. So,

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

  where $m = \min_{[0,1]} \mu$ and $\kappa_4$ 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 $\mathbb{E}[(\sigma_j^d)^2]$:

  $$\begin{aligned}(\sigma_j^d)^2 &= \mathbb{E}[(s(X) - \tilde{s}_{\mathbb{U}}(X))^2 \,|\, U_{(j)} < X \leqslant U_{(j+1)}] \\ &\leqslant C^2 (U_{(j+1)} - U_{(j)})^2 \quad \text{because } s \text{ is } C-\text{Lipschitz}\end{aligned}.$$

  So, $\mathbb{E}[\sigma_j^d)^2] \leqslant C^2 \mathbb{E}[(U_{(j+1)} - U_{(j)})^2] = C^2 \frac{2}{(k+1)(k+2)}$ which tends to 0 as $k$ tends to infinity.

- For the last term, the following inequality is sufficient to conclude:
  $\mathbb{E}[(\sigma_j^d)^2 \delta_{n,p_j}] \leqslant C^2 \mathbb{E}[\delta_{n,p_j}]$, because $U_{(j+1)} - U_{(j)} \leqslant 1$.

## 6.2 Proof of Proposition 3

Function $s$ is supposed to be $C$-Lipschitz, so

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

## 6.3 Proof of Theorem 2

Before entering into details of the proof of Theorem 2, we recall that in the proof of Proposition 1 (which can be found in [1]), calculations lead to the following equality:

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2 \mid \mathbb{U}] = \sum_{j=0}^{k} p_j\mathbb{E}\Big[\frac{1}{n\hat{p}_j}\Big](\sigma^2 + (\sigma_j^d)^2) \tag{11}$$

where $\hat{p}_j = \dfrac{\sharp\{i :\ U_{(j)} < X_i \leqslant U_{(j+1)}\}}{n}$.

Then, an estimation of $p_j\mathbb{E}\Big[\dfrac{1}{n\hat{p}_j}\Big]$ gives the expression $\dfrac{1}{n}(1 + \delta_{n,p_j})$ in Proposition 1.

We note

$$Var_j = p_j\mathbb{E}\Big[\frac{1}{n\hat{p}_j}\Big](\sigma^2 + (\sigma_j^d)^2) \tag{12}$$

a generic term of the sum in the r.h.s. of (11).

We now address the proof of Theorem 2. We begin by introducing some notations and establish an intermediate result. The following proposition is not only useful to prove Theorem 2, but has its own interest. Indeed, it gives a

general upper bound (to be compared to equation (3)) which does not depend on the distribution of random partitions defining the trees.

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 $\mathbb{U}^1 = (U_1^1, \ldots, U_k^1)$ and $\mathbb{U}^2 = (U_1^2, \ldots, U_k^2)$ two sequences of i.i.d. uniform random variables, with respective ordered statistics $(U_{(1)}^1, \ldots, U_{(k)}^1)$ and $(U_{(1)}^2, \ldots, U_{(k)}^2)$.

Then we denote by $(V_{(1)}, \ldots, V_{(2k)})$ the ordered statistics of the complete vector $(U_1^1, \ldots, U_k^1, U_1^2, \ldots, U_k^2)$, $V_{(0)} = 0$ and $V_{(2k+1)} = 1$.

$(\Sigma_t^{d,1,2})^2$ denotes a sum of terms $\mathbb{E}[(\tilde{s}_{\mathbb{U}^1}(X) - s(X))(\tilde{s}_{\mathbb{U}^2}(X) - s(X)) \,|\, V_{(t')} < X \leqslant V_{(t'+1)}]$ for several consecutive values of $t'$.

Finally $\tilde{p}_t$ denotes for some $j \in \{0, \ldots, k\}$ either $p_j^1$ or $p_j^2$ depending on the relative positions between the $(U_1^1, \ldots, U_k^1)$ and the $(U_1^2, \ldots, U_k^2)$ in $(V_{(1)}, \ldots, V_{(2k)})$ (see details below).

**Proposition 5** *The covariance between two PURT satisfies the following upper bound:*

$$\mathbb{C}(\hat{s}_{\mathbb{U}^1}, \hat{s}_{\mathbb{U}^2}) \leqslant \frac{1}{n}\mathbb{E}\left[\sum_{t=0}^{N_{1,2}}(1 + \delta_{n,\tilde{p}_t})(\sigma^2 + (\Sigma_t^{d,1,2})^2)\right] \tag{13}$$

*where* $N_{1,2} = k + 1 - \sum_{r=1}^{k-2}\sum_{s=1}^{k-1} \mathbb{1}_{U_{(s)}^2 < U_{(r)}^1 < U_{(r+1)}^1 < U_{(r+2)}^1 < U_{(s+1)}^2}$ .

■

**Remark 2** *The gain in variance for a PURF comes from the fact that the number of terms in the sum of equation (13) is smaller than* $k + 1$*. Indeed, it is* $k + 1 - M_{1,2}$ *where* $M_{1,2}$ *is the number of times that* 3 *consecutive ordered statistics of* $\mathbb{U}^1$ *are included in* 2 *consecutive ordered statistics of* $\mathbb{U}^2$*.*

We now prove inequality (13) of Proposition 5. 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:

$$\begin{aligned}&\left(\sum_{r=0}^{k}(\hat{\beta}_r^1 - \beta_r^1)\mathbb{1}_{U_{(r)}^1 < X \leqslant U_{(r+1)}^1}\right)\left(\sum_{s=0}^{k}(\hat{\beta}_s^2 - \beta_s^2)\mathbb{1}_{U_{(s)}^2 < x \leqslant U_{(s+1)}^2}\right)\\ &= \sum_{t=0}^{2k}(\hat{\beta}_{t,r}^1 - \beta_{t,r}^1)(\hat{\beta}_{t,s}^2 - \beta_{t,s}^2)\mathbb{1}_{V_{(t)} < X \leqslant V_{(t+1)}}\end{aligned} \tag{14}$$

where $(V_{(1)}, \ldots, V_{(2k)})$ is the ordered statistics of the vector $(U_1^1, \ldots, U_k^1, U_1^2, \ldots, U_k^2)$, $V_{(0)} = 0$, $V_{(2k+1)} = 1$, and

$$\begin{cases} \hat{\beta}_{t,r}^1 = \hat{\beta}_r^1 \text{ and } \beta_{t,r}^1 = \beta_r^1, & \text{if } ]V_{(t)}, V_{(t+1)}] \subset ]U_{(r)}^1, U_{(r+1)}^1] \\ \hat{\beta}_{t,s}^2 = \hat{\beta}_s^2 \text{ and } \beta_{t,s}^2 = \beta_s^2, & \text{if } ]V_{(t)}, V_{(t+1)}] \subset ]U_{(s)}^2, U_{(s+1)}^2] \end{cases}$$

For $l = 1, 2$ and $j = 0, \ldots, k$, we define $\hat{p}_j^l = \dfrac{\sharp\{i :\ U_{(j)}^l < X_i \leqslant U_{(j+1)}^l\}}{n}$.

Now, let us give some details for the first term of (14), denoted by $S_1(X)$. Without loss of generality, we suppose that $V_{(1)} = U^1_{(1)}$ (i.e. $U^1_{(1)} < U^2_{(1)}$). So,

$$\begin{aligned}
S_1(X) &= (\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \tilde{s}_{\mathbb{U}^2}(X))\mathbb{1}_{0<X\leqslant U^1_{(1)}} \\
&= (\hat{\beta}^1_1 - \beta^1_1)(\hat{\beta}^2_1 - \beta^2_1)\mathbb{1}_{0<X\leqslant U^1_{(1)}} \\
&= \Big(\frac{1}{n\hat{p}^1_1}\sum_{i:\ 0<X_i\leqslant U^1_{(1)}}(Y_i - \beta^1_1)\Big)\Big(\frac{1}{n\hat{p}^2_1}\sum_{i:\ 0<X_i\leqslant U^2_{(1)}}(Y_i - \beta^2_1)\Big)\mathbb{1}_{0<X\leqslant U^1_{(1)}} \\
&= \frac{1}{n\hat{p}^1_1 n\hat{p}^2_1}\sum_{i^1:\ 0<X_{i^1}\leqslant U^1_{(1)}\, i^2:\ 0<X_{i^2}\leqslant U^2_{(1)}}(Y_{i^1} - \beta^1_1)(Y_{i^2} - \beta^2_1)\mathbb{1}_{0<X\leqslant U^1_{(1)}}
\end{aligned}$$

If we denote by $\mathbb{E}^{\Lambda_{1,2}}[.]$ the conditional expectation $\mathbb{E}[.\,|\,(\mathbb{1}_{0<X_{i^1}\leqslant U^1_{(1)}})_{1\leqslant i^1\leqslant n}, (\mathbb{1}_{0<X_{i^2}\leqslant U^2_{(1)}})_{1\leqslant i^2\leqslant n}]$, we have:

$$\begin{aligned}
&\mathbb{E}[S_1(X)\,|\,\mathbb{U}^1,\mathbb{U}^2] \\
&= \mathbb{E}\Big[p^1_1\mathbb{E}\Big[\frac{1}{n\hat{p}^1_1 n\hat{p}^2_1}\sum_{i^1:\ 0<X_{i^1}\leqslant U^1_{(1)}\, i^2:\ 0<X_{i^2}\leqslant U^2_{(1)}}\mathbb{E}^{\Lambda_{1,2}}[(Y_{i^1} - \beta^1_1)(Y_{i^2} - \beta^2_1)]\Big]\,\Big|\,\mathbb{U}^1,\mathbb{U}^2\Big]
\end{aligned}$$

but

$$i^1 \neq i^2 \quad\Longrightarrow\quad \mathbb{E}^{\Lambda_{1,2}}[(Y_{i^1} - \beta^1_1)(Y_{i^2} - \beta^2_1)] = 0$$

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

$$\begin{aligned}
&\mathbb{E}[S_1(X)\,|\,\mathbb{U}^1,\mathbb{U}^2] \\
&= \mathbb{E}\Big[p^1_1\mathbb{E}\Big[\frac{1}{n\hat{p}^1_1 n\hat{p}^2_1}\sum_{i:\ 0<X_i\leqslant U^1_{(1)}}\mathbb{E}^{\Lambda_1}[(Y_i - \beta^1_1)(Y_i - \beta^2_1)]\Big]\,\Big|\,\mathbb{U}^1,\mathbb{U}^2\Big] \\
&= \mathbb{E}\Big[p^1_1\mathbb{E}\Big[\frac{1}{n\hat{p}^1_1 n\hat{p}^2_1}\sum_{i:\ 0<X_i\leqslant U^1_{(1)}}\mathbb{E}[(Y_i - \beta^1_1)(Y_i - \beta^2_1)\,|\,0<X_i\leqslant U^1_{(1)}]\Big]\,\Big|\,\mathbb{U}^1,\mathbb{U}^2\Big]
\end{aligned}$$

where $\mathbb{E}^{\Lambda_1}[.]$ denotes the conditional expectation $\mathbb{E}[.\,|\,(\mathbb{1}_{0<X_i\leqslant U^1_{(1)}})_{1\leqslant i\leqslant n}]$.

Now, as

$$\mathbb{E}[(Y_i - \beta^1_1)(Y_i - \beta^2_1)\,|\,0<X_i\leqslant U^1_{(1)}] = \mathbb{E}[(Y - \beta^1_1)(Y - \beta^2_1)\,|\,0<X\leqslant U^1_{(1)}]$$

for all $i$, and

$$\mathbb{E}[(Y - \beta^1_1)(Y - \beta^2_1)\,|\,0<X\leqslant U^1_{(1)}] = \sigma^2 + (\sigma^{d,1,2}_0)^2$$

where

$$(\sigma^{d,1,2}_0)^2 = \mathbb{E}[(s(X) - \tilde{s}_{\mathbb{U}^1}(X))(s(X) - \tilde{s}_{\mathbb{U}^2}(X))\,|\,0<X\leqslant V_{(1)}]$$

we get

$$\mathbb{E}[S_1(X)\,|\,\mathbb{U}^1,\mathbb{U}^2] = p^1_1\mathbb{E}\Big[\frac{1}{n\hat{p}^2_1}\Big](\sigma^2 + (\sigma^{d,1,2}_0)^2)\ .$$

If we suppose in addition that $V_{(2)} = U^2_{(1)}$, we similarly get for the second term of (14):

$$\begin{aligned}&\mathbb{E}[S_2(X) \,|\, \mathbb{U}^1, \mathbb{U}^2] \\ &= \mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \tilde{s}_{\mathbb{U}^2}(X))\mathbb{1}_{U^1_{(1)} < X \leqslant U^2_{(1)}} \,|\, \mathbb{U}^1, \mathbb{U}^2] \\ &= q_2 \mathbb{E}\left[\frac{n\hat{q}_2}{n\hat{p}^1_2 n\hat{p}^2_1}(\sigma^2 + (\sigma_1^{d,1,2})^2)\right]\end{aligned}$$

where

$$q_2 = \mathrm{P}(V_{(1)} < X \leqslant V_{(2)}) = \mathrm{P}(U^1_{(1)} < X \leqslant U^2_{(1)})$$
$$n\hat{q}_2 = \sharp\{i : V_{(1)} < X_i \leqslant V_{(2)}\}$$

and

$$(\sigma_1^{d,1,2})^2 = \mathbb{E}[(s(X) - \tilde{s}_{\mathbb{U}^1}(X))(s(X) - \tilde{s}_{\mathbb{U}^2}(X)) \,|\, V_{(1)} < X \leqslant V_{(2)}] \,.$$

Since $]V_{(1)}, V_{(2)}]$ is included in $]U^1_{(1)}, U^1_{(2)}]$, we have $\hat{q}_2 \leqslant \hat{p}^1_2$, so:

$$\mathbb{E}[S_2(X) \,|\, \mathbb{U}^1, \mathbb{U}^2] \leqslant q_2 \mathbb{E}\Big[\frac{1}{n\hat{p}^2_1}\Big](\sigma^2 + (\sigma_1^{d,1,2})^2) \,.$$

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

$$\mathbb{E}[S_1(X) + S_2(X) \,|\, \mathbb{U}^1, \mathbb{U}^2] \leqslant p^2_1 \mathbb{E}\Big[\frac{1}{n\hat{p}^2_1}\Big](\sigma^2 + (\sigma_0^{d,1,2})^2 + (\sigma_1^{d,1,2})^2)$$

In conclusion, we succeeded to bound the sum of the first two terms of (14) by an expression very close to $Var_j$ (defined in (12)). The only difference comes from the fact that instead of $(\sigma^d_j)^2$ we have $(\sigma_0^{d,1,2})^2 + (\sigma_1^{d,1,2})^2$. But as we saw in proof of Proposition 2, these terms are negligible, so $p^2_1 \mathbb{E}\Big[\frac{1}{n\hat{p}^2_1}\Big](\sigma^2 + (\sigma_0^{d,1,2})^2 + (\sigma_1^{d,1,2})^2)$ is of the same order than $Var_j$.

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

We denote by $S_j(X)$ the $j$-th term of (14), i.e. $S_j(X) = (\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \tilde{s}_{\mathbb{U}^2}(X))\mathbb{1}_{V_{(j)} < X \leqslant V_{(j+1)}}$.

**Lemma 1** *Let $r$ be in $\{0, \ldots, k\}$ and denote by $t$, $t'$ the integers such that*

$$U^1_{(r)} = V_{(t)} < V_{(t'+1)} = U^1_{(r+1)} \tag{15}$$

*then*

$$\mathbb{E}\Big[\sum_{j=t}^{t'} S_j(X) \,|\, \mathbb{U}^1, \mathbb{U}^2\Big] \leqslant p^1_r \mathbb{E}\Big[\frac{1}{n\hat{p}^1_r}\Big](\sigma^2 + (\Sigma_r^{d,1,2})^2)$$

*where* $(\Sigma_r^{d,1,2})^2 = \sum_{j=t}^{t'} (\sigma_j^{d,1,2})^2$.

■

Indeed for all $j \in \{t, t+1, \ldots, t'\}$,

$$\mathbb{E}[S_j(X) \,|\, \mathbb{U}^1, \mathbb{U}^2] \leqslant q_j \mathbb{E}\Big[\frac{1}{n\hat{p}_r^1}\Big](\sigma^2 + (\sigma_j^{d,1,2})^2)$$

where

$$q_j = \mathrm{P}(V_{(j)} < X \leqslant V_{(j+1)})$$

and

$$(\sigma_j^{d,1,2})^2 = \mathbb{E}[(s(X) - \tilde{s}_{\mathbb{U}^1}(X))(s(X) - \tilde{s}_{\mathbb{U}^2}(X)) \,|\, V_{(j)} < X \leqslant V_{(j+1)}] \,.$$

Thus,

$$\mathbb{E}\Big[\sum_{j=t}^{t'} S_j(X) \,|\, \mathbb{U}^1, \mathbb{U}^2\Big] \leqslant \mathrm{P}(V_{(t)} < X \leqslant V_{(t'+1)}) \; \mathbb{E}\Big[\frac{1}{n\hat{p}_r^1}\Big](\sigma^2 + (\Sigma_r^{d,1,2})^2) \,.$$

From relation (15) we have $\mathrm{P}(V_{(t)} < X \leqslant V_{(t'+1)}) = p_r^1$, which concludes the proof of Lemma 1.

Therefore, we can upper bound the initial sum (14) of $2k+1$ terms by a sum of $k+1$ terms of the same order as $Var_j$ only involving intervals 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 2** *If there exist $r$ and $s$ such as*

$$U_{(s)}^2 < U_{(r)}^1 < U_{(r+1)}^1 < U_{(r+2)}^1 < U_{(s+1)}^2$$

*the expression*

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

*is upper bounded by*

$$p_s^2 \mathbb{E}\Big[\frac{1}{n\hat{p}_s^2}\Big](\sigma^2 + (\Sigma_s^{d,1,2})^2) \,.$$

*where* $(\Sigma_s^{d,1,2})^2 = (\sigma_{r+s}^{d,1,2})^2 + (\sigma_{r+s+1}^{d,1,2})^2$.

■

Indeed,

$$\begin{aligned}
&\mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \tilde{s}_{\mathbb{U}^2}(X))\mathbb{1}_{U_{(r)}^1 < X \leqslant U_{(r+1)}^1} \,|\, \mathbb{U}^1, \mathbb{U}^2] \\
&\leqslant p_r^1 \mathbb{E}\Big[\frac{1}{n\hat{p}_s^2}\Big](\sigma^2 + (\sigma_{r+s}^{d,1,2})^2)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \tilde{s}_{\mathbb{U}^2}(X))\mathbb{1}_{U_{(r+1)}^1 < X \leqslant U_{(r+2)}^1} \,|\, \mathbb{U}^1, \mathbb{U}^2] \\
&\leqslant p_{r+1}^1 \mathbb{E}\Big[\frac{1}{n\hat{p}_s^2}\Big](\sigma^2 + (\sigma_{r+s+1}^{d,1,2})^2) \,.
\end{aligned}$$

Finally, since $p^1_r + p^1_{r+1} \leqslant p^2_s$, $(\sigma^{d,1,2}_{r+s})^2 \leqslant (\sigma^{d,1,2}_{r+s})^2 + (\sigma^{d,1,2}_{r+s+1})^2$ and $(\sigma^{d,1,2}_{r+s+1})^2 \leqslant (\sigma^{d,1,2}_{r+s})^2 + (\sigma^{d,1,2}_{r+s+1})^2$, the result is obtained by summing the two terms.

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

By repeatedly applying this lemma for all intervals, 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 $N_{1,2}$ terms of the form $(1+\delta_{n,\tilde{p}_t})(\sigma^2 + (\Sigma^{d,1,2}_t)^2)$, where $\tilde{p}_t$ denotes for some $j \in \{0,\dots,k\}$ either $p^1_j$ or $p^2_j$ depending on the fact that we are in the situation of Lemma 1 or Lemma 2, $N_{1,2} = k+1-M_{1,2}$ and

$$M_{1,2} = \sum_{r=1}^{k-2}\sum_{s=1}^{k-1} \mathbb{1}_{U^2_{(s)} < U^1_{(r)} < U^1_{(r+1)} < U^1_{(r+2)} < U^2_{(s+1)}} \ .$$

This concludes the proof of Proposition 5. Now, using the fact that we deal with uniform partitions, we manage to prove the following corollary.

**Corollary 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,*

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

■

Because of the simple draws of random partitions, the number $M_{1,2}$ 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 3 below). As in Proposition 2, we have to prove that all terms of the sum are negligible compared to 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 simple inequality:

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=0}^{N_{1,2}} (\sigma^2\delta_{n,p_t} + (\Sigma^{d,1,2}_t)^2 + (\Sigma^{d,1,2}_t)^2\delta_{n,p_t})\right] \\ &\leqslant \sum_{t=0}^{k}\Big(\mathbb{E}[\sigma^2\delta_{n,p_t}] + \mathbb{E}[(\Sigma^{d,1,2}_t)^2] + \mathbb{E}[(\Sigma^{d,1,2}_t)^2\delta_{n,p_t}]\Big) \ .\end{aligned}$$

These quantities are of the same kind as the three last terms in the sum of equation 4. So with the same techniques we get that

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

So, we have

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

Finally, the following technical result allows to conclude the proof of Corollary 3, and thus the proof of Theorem2.

**Lemma 3**

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

*Hence,*

$$\mathbb{E}[M_{1,2}] = \frac{k+1}{4} + \underset{k\to+\infty}{\mathrm{o}}(k) .$$

■

We then obtain that

$$\mathbb{E}[N_{1,2}] = \frac{3}{4}(k+1) + \underset{k\to+\infty}{\mathrm{o}}(k) .$$

Let us demonstrate lemma 3.

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

As we know the distribution of ordered statistics (see e.g. Section 2.2 of [5]), we can compute the following probability:

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

So,

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

(by elementary properties of binomial coefficients (see e.g. [8] p.160))

$$= \frac{k-2}{4(2k-1)} \sum_{t=0}^{2k-5} \sum_{r=t-k+2}^{t} \frac{\left( \begin{array}{c} t+1 \\ r \end{array} \right) \left( \begin{array}{c} 2k-3-(t+1) \\ k-3-r \end{array} \right)}{\left( \begin{array}{c} 2k-3 \\ k-3 \end{array} \right)}$$

(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 hypergeometric distribution)

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

$$= \frac{k-2}{2(2k-1)} \left[ \sum_{t=0}^{k-4} \left( 1 - \frac{\left( \begin{array}{c} t+1 \\ t+1 \end{array} \right) \left( \begin{array}{c} 2k-3-(t+1) \\ k-3-(t+1) \end{array} \right)}{\left( \begin{array}{c} 2k-3 \\ k-3 \end{array} \right)} \right) + 1 \right]$$

$$= \frac{k-2}{2(2k-1)} \Big( k - 3 + \frac{4}{k+1} \Big).$$

# References


[1] Arlot, S.. *V-fold cross-validation improved: V-fold penalization*. Preprint. arXiv : 0802.0566v2 (2008)

[2] Biau, G., Devroye, L., Lugosi, G.. *Consistency of random forests and other averaging classifiers*. Journal of Machine Learning Research. 9, 2039-2057 (2008)

[3] Breiman, L.. *Random Forests*. Machine Learning. 45, 5-32 (2001)

[4] Cutler, A., Zhao, G.. *Pert - Perfect random tree ensembles*. Computing Science and Statistics. 33, 490-497 (2001)

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

[6] Díaz-Uriarte, R., Alvarez de Andrés, S.. *Gene Selection and classification of microarray data using random forest*. BMC Bioinformatics. 7, 3 (2006)

[7] Genuer, R., Poggi, J.-M. and Tuleau, C. *Variable selection using random forests*. Pattern Recognition Letters (in press) doi:10.1016/j.patrec.2010.03.014 (2010)

[8] Graham R.L., Knuth D.E., Patashnik O.. *Concrete mathematics*. Addison-Wesley (1989)

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

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

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

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399