Generalization Error and Out-of-bag Bounds in Random (Uniform) Forests

Saïp Ciss

Generalization Error and Out-of-bag Bounds in Random (Uniform) Forests

Saïp Ciss*

January 27, 2015

Abstract

In the context of ensemble learning, especially for random forests models, the out-of-bag (OOB) procedure, using the training set, produces an estimation of the generalization error. The *OOB error* has the same purpose than the cross-validation error, but comes with very specific points. First, there exists an OOB classifier that leads to the OOB evaluation. Second, the OOB classifier is embedded in the forest classifier. We show in this paper that these two intrinsic properties lead to *produce simple conditions for the test error to be bounded by the OOB error*. Conditions come with the only required and usual assumptions which are the $i.i.d$ one and the existence of first and second order moments. The main interest is that the OOB error is explicitly known, hence one just needs a training set without any other assumption on the model behind the data. As a practical case, we use Random Uniform Forests (Ciss, 2015a), a variant of Random Forests (Breiman, 2001) that inherits of all properties of the latter, to show how OOB bounds apply. We also provide an R package, *randomUniformForest*, allowing to experiment all the arguments described in the paper.

Keywords : Random Uniform Forests, Random Forests, statistical learning, bounds, Out-of-bag error, classification, regression, R package.

*PhD. University Paris Ouest Nanterre La Défense. Modal'X. saip.ciss@wanadoo.fr

1 Introduction

Ensemble learning defines all the methods that use many base learners, then combine them to produce a classifier. It has been shown to be effective (Breiman, 1996, 2001) and two paradigms, with each many variants, constitute the main approaches. The first one leads to construct many independent, or with low correlation, base learners combining their outputs by majority vote (classification) or averaging (regression). The second one builds base learners in a sequential way, using the residuals of one to improve the next one, and the resulted classifier is some function of the base learners. *Bagging* (Breiman, 1996) and *Random Forests* (Breiman, 2001) are the most known models for the first paradigm, while *Gradient Boosting Machines* (Friedman, 2001, 2002) are one of the most effective for the second one, which was originally introduced with *AdaBoost* (Freund and Schapire, 1997), short for adaptative boosting. In this paper, we will focus on *Random Uniform Forests* (Ciss, 2015a), a variant of Random Forests. It shares the same principle : *use a decision tree as a base learner, build many ones with low correlation, then combine them by majority vote or averaging.* To build a random forest like classifier two steps are needed :
- the *bootstrap* one : for each tree, choose at random and with replacement, n observations among n from the training set. This step is not mandatory.
- the *aggregating* one : grow each tree independently from another, without pruning and by selecting at random a subset of variables for each created node, then combine all trees at the end. Since trees are randomized (more or less strongly), the usual combining technique is majority vote (classification) or the average (regression) of the trees outputs.

The bootstrap step can be replaced by a *subsampling* one (choose randomly m observations among the n of the training set, $m < n$), especially for regression. This is the case for Random Uniform Forests. We seek to have little correlation between trees and little variance for each tree. This is essential since, for the most general case of regression, the forest error is bounded by the product of average correlation between trees residuals and the average prediction error of a tree (i.e. average variance between trees residuals), which depends to the variance of the trees. Hence random forests models usually focus on how to obtain low correlation and low variance. Since it is very hard to decrease both in the same time (variance is usually increasing as correlation is decreasing), *Random Uniform Forests are designed to decrease correlation faster than the increase of variance*.

These arguments, correlation and variance, are the main directions from which the whole analysis of random forests is pursued. Low correlation, as a replacement of independence, is a necessary condition for the convergence of random forests. Variance is involved in both correlation and generalization error of the forest classifier.

In section 2, we define the forest classifier for Random Uniform Forests and the OOB classifier. In section 3, we describe the convergence of the prediction error, recalling Breiman's arguments. In section 4, we write the Breiman's bounds and briefly discuss about some cases where they are not adequate. Section 5 is the core of the article and defines conditions to obtain OOB bounds. We discuss about the consequences of the proposed bounds in section 6 and conclude in section 7. Proofs are sent to the appendix.

2 The Random Uniform Forest classifier

Random Uniform Forests are close to Breiman's Random Forests but they come with many differences. The most important ones are the use of *random cut-points* and the *sampling with replacement a set of variables for each candidate node*. Random Uniform Forests are designed to be fairly simple and to *let data speak for themselves* with some kind of global optimization when selecting an optimal node among many random ones.

A Random Uniform Forest is an ensemble of *random uniform decision trees*, which are unpruned and binary random decision trees that use the continuous Uniform distribution to be built. Since the results we are assessing can apply to any ensemble model that uses bootstrap or subsampling and that can build an OOB classifier (see next section), we will omit details of the algorithm and will focus on the general form of the classifier. Note that we use, for convenience, **E** and **Var** as the expectation and variance operators. **I** is the indicator function.

Let us consider $D_n = \{(X_i, Y_i), 1 \leq i \leq n\}$, corresponding to the observations and responses of a training sample, where (X, Y) is a $\mathbb{R}^d \times \mathcal{Y}$-valued random pair, with respect to the *i.i.d.* assumption. Let us suppose, for brevity, that $Y \in \{0, 1\}$ considering, then, the *binary classification* case and when referring on classification in the rest of the paper. The decision rule of a random uniform decision tree is given by

$$g_{\mathcal{P}}(x, A, D_n) = g_{\mathcal{P}}(x) = \begin{cases} 1, \ if \ \sum_{i=1}^{n} \mathbf{I}_{\{X_i \in A, Y_i = 1\}} > \sum_{i=1}^{n} \mathbf{I}_{\{X_i \in A, Y_i = 0\}}, \ x \in A \\ \\ 0, \ otherwise. \end{cases}$$

A is the current terminal and optimal region (node), coming from the recursive partitioning scheme, where falls the observation x,
$g_{\mathcal{P}}$ is the decision rule of the tree for a set $A \in \mathcal{P}$, a partition of the data.
Hence we count, in a terminal node where falls an observation, the number of times each label (defined by Y) appears. The label that has the highest count is the one that is assigned to the observation.

In regression we get

$$g_{\mathcal{P}}(x, A, D_n) = g_{\mathcal{P}}(x) = \frac{1}{\sum_{i=1}^{n} \mathbf{I}_{\{X_i \in A\}}} \sum_{i=1}^{n} Y_i \mathbf{I}_{\{X_i \in A\}}, \ x \in A.$$

The average of all instances of Y, in a terminal node, is taken as the output of the classifier.

The decision rule, $\bar{g}_{\mathcal{P}}^{(B)}$, of the Random Uniform Forest classifier is given by

$$\bar{g}_{\mathcal{P}}^{(B)}(x) = \begin{cases} 1, \ if \ \sum_{b=1}^{B} \mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(x) = 1\}} > \sum_{b=1}^{B} \mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(x) = 0\}} \\ \\ 0, \ otherwise. \end{cases}$$

And for regression :

$$\bar{g}_{\mathcal{P}}^{(B)}(x) = \frac{1}{B} \sum_{b=1}^{B} g_{\mathcal{P}}^{(b)}(x).$$

2.1 The OOB Forest classifier

The Out-of-bag (OOB) informations are the observations that do not participate to the trees growth. For each tree, due to bootstrap or subsampling, some observations (at random) are not chosen and are stored in order to build the OOB classifier, whose decision rule is $\bar{g}^{(B)}_{\mathcal{P},oob}(X)$. *The OOB classifier exists only for the training sample* and use, on average and in the bootstrap case, B' trees, $B' = \lceil \texttt{exp}(-1)B \rceil$, with n observations. Note that the B' trees are not necessary the same for each observation that needs to be evaluated. We have for an observation x and for only D_n,

$$\bar{g}^{(B)}_{\mathcal{P},oob}(x) = \begin{cases} 1, \ if \ \sum_{b=1}^{B} \mathbf{I}_{\left\{g^{(b)}_{\mathcal{P}}(x)=1\right\}} \mathbf{I}_{\left\{b \in G^{-}(x,B)\right\}} > \sum_{b=1}^{B} \mathbf{I}_{\left\{g^{(b)}_{\mathcal{P}}(x)=0\right\}} \mathbf{I}_{\left\{b \in G^{-}(x,B)\right\}} \\ \\ 0, \ otherwise. \end{cases}$$

And, for regression :

$$\bar{g}^{(B)}_{\mathcal{P},oob}(x) = \frac{1}{\sum_{b=1}^{B} \mathbf{I}_{\left\{b \in G^{-}(x,B)\right\}}} \sum_{b=1}^{B} g^{(b)}_{\mathcal{P}}(x) \mathbf{I}_{\left\{b \in G^{-}(x,B)\right\}},$$

where $G^{-}(x, B)$ is the set of trees, among the B, which have never classified x.

In practice, the design of the OOB classifier is the same than the one of the forest classifier, except that the latter use all trees and applies to any test set with respect to the *i.i.d.* assumption. The OOB classifier learns the training set and applies only on the latter. Moreover, the OOB classifier is embedded in the forest classifier. For each tree, some observations that have not been used to grow the tree are now used to evaluate the classifier, as a test set would serve for the forest classifier. Main interests of Out-of-bag evaluation can be summarized with the following lines:
i) the classifier is built during the learning phase, taking only a small computing time in comparison to a k-fold cross-validation,
ii) except for some trees, it is the one that will be used for the test set,
iii) the OOB error is based on all the observations of the training set at once,
iv) one can assess convergence of the classifier by changing the sampling rate, especially for large datasets.

3 Convergence of the prediction error

Let us recall Breiman's arguments, looking first classification. Let us consider mg, the margin function that measures the difference (in frequency) between points (observations) correctly classified and misclassified points. We have

$$mg(X, Y) = \left(\frac{1}{B} \sum_{b=1}^{B} \mathbf{I}_{\{g^{(b)}_{\mathcal{P}}(X)=Y\}} \right) - \left(\frac{1}{B} \sum_{b=1}^{B} \mathbf{I}_{\{g^{(b)}_{\mathcal{P}}(X)\neq Y\}} \right).$$

Let us call (following Breiman's notation) PE^*, the prediction (or generalization) error for B trees, and define $g_{\mathcal{P}}(X) \stackrel{def}{=} g_{\mathcal{P}}(X, \theta)$, where θ is the parameter that translates the

randomness introduced. For the b-th tree, $1 \leq b \leq B$, we define $g_{\mathcal{P}}^{(b)}(X) \overset{def}{=} g_{\mathcal{P}}^{(b)}(X, \theta_b)$. Then

$$PE^* = \mathbf{P}_{\mathbf{X},\mathbf{Y}} \left(mg(X, Y) < 0 \right),$$

and by the *Law of Large Numbers*, if trees are independent and when $B \to \infty$,

$$PE^* \overset{p.s.}{\to} PE = \mathbf{P}_{\mathbf{X},\mathbf{Y}} \left\{ \mathbf{P}_{\theta}(g_{\mathcal{P}}(X, \theta) = Y) - \mathbf{P}_{\theta}(g_{\mathcal{P}}(X, \theta) \neq Y) < \ 0 \right\},$$

(Breiman, 2001, theorem 1.2)
where PE is the *generalization error* of the (infinite) forest.

In regression, we get similar results. We have

$$PE^*(forest) \overset{def}{=} PE^*(\bar{g}_{\mathcal{P}}^{(B)}(X)) = \mathbf{E}_{\mathbf{X},\mathbf{Y}} \left(Y - \frac{1}{B} \sum_{b=1}^{B} g_{\mathcal{P}}^{(b)}(X, \theta_b) \right)^2,$$

and when $B \to \infty$,

$$PE^*(\bar{g}_{\mathcal{P}}^{(B)}(X)) \overset{p.s.}{\to} PE(\mathbf{E}_{\theta} g_{\mathcal{P}}(X, \theta)) = \mathbf{E}_{\mathbf{X},\mathbf{Y}} \left(Y - \mathbf{E}_{\theta} g_{\mathcal{P}}(X, \theta) \right)^2,$$

(Breiman, 2001, theorem 11.1).

As a consequence, *Random (Uniform) Forests do not overfit if trees are independent (in practice, a little dependent) and under the i.i.d. assumption.* One can note that low correlation between trees is easy to achieve in (binary) classification (usually around or less than 0.1) but much harder in regression (usually around 0.3 or more). A second consequence is that one does not need to grow a lot of trees. For a fixed dataset, convergence toward the true prediction error of the model will quickly happen as one is adding trees to the forest (one just needs to measure the difference in OOB error with new trees added). The third one is that in classification one needs first to lower bias while in regression one needs first to reduce correlation. But, the main application of convergence is the ability to just focus on ways to reduce prediction error without the need to further work on statistical consistency. More precisely, the forest will always do *its best*, depending on the hyper-parameters provided. Hence, in practice, the main question will be to know how many observations are needed for such a task, rather than to know if the model can match the lowest possible error. It might, but the latter requires conditions (or assumptions) that are not easily reachable in practice.

4 Bounds

As correlation decreases (required condition), average variance of trees increases and it begins harder to reduce prediction error. Decreasing correlation is easy, since it only requires to increase trees randomness. But then, variance will increase and one has to find ways to not let it move too fast. If one wants to control both correlation and variance, one key is to observe and monitor Breiman's bounds. These are the bounds of Random (Uniform) Forests that ensure that the prediction error (under the *i.i.d.* assumption) will not increase beyond a limit. Most applications of Breiman's bounds are linked with the

OOB classifier, that inherit to the bounds and show if more work (which is also depending on n and on the hyper-parameters) is needed to improve the modeling or if there is no more room for the algorithm.

`Classification`

At first, bounds involve classification and we have two. The first bound is, by the Bienaymé-Tchebychev inequality,

$$PE^* \leq \frac{\mathbf{Var}_{\mathbf{X},\mathbf{Y}}(mr)}{s^2},$$

where

$$mr(X,Y) = \mathbf{P}_\theta(g_{\mathcal{P}}(X,\theta) = Y) - \mathbf{P}_\theta(g_{\mathcal{P}}(X,\theta) \neq Y)$$

is the limit of mg and $s, s > 0$, the *strength* (or margin), is

$$s = \mathbf{E}_{\mathbf{X},\mathbf{Y}}\{mr(X,Y)\}.$$

This first bound states that prediction error would *always* be under a limit which is explicit but unknown (unless one evaluates it using OOB informations). It is the *upper bound of the prediction error* and it can be loose (but useful in case of problems with imbalanced classes or more difficult ones).
The second bound is tighter. We have (Breiman, 2001, theorem 2.3)

$$PE^* \leq \frac{\bar{\rho}(1-s^2)}{s^2},$$

where

$$\bar{\rho} = \frac{\mathbf{E}_{\theta,\theta'}[\rho(\theta,\theta')\sqrt{\mathbf{Var}(\theta)}\sqrt{\mathbf{Var}(\theta')}]}{\mathbf{E}_{\theta,\theta'}[\sqrt{\mathbf{Var}(\theta)}\sqrt{\mathbf{Var}(\theta')}]},$$

$\bar{\rho}$ is the *average (weighted) correlation between trees*,
$\rho(\theta,\theta')$ is the correlation of two trees of random and independently chosen parameters θ and θ',
$\mathbf{Var}(\theta)$ is the variance of the tree over the observations, standing as the variance of the *raw strength*. Let us note the latter rmg. We have

$$rmg(\theta,X,Y) = \mathbf{I}_{\{g_{\mathcal{P}}(X,\theta)=Y\}} - \mathbf{I}_{\{g_{\mathcal{P}}(X,\theta)\neq Y\}},$$

$$\mathbf{Var}(\theta) \stackrel{def}{=} \mathbf{Var}_{\mathbf{X},\mathbf{Y}}(rmg(\theta,X,Y)),$$

and

$$mr(X,Y) = \mathbf{E}_\theta(rmg(\theta,X,Y)).$$

An estimate of $\bar{\rho}$ is, then, given by

$$\hat{\bar{\rho}} = \frac{\sum_{1\leq b<c\leq B}\hat{\rho}_{b,c}(\theta_b,\theta_c)\sqrt{\widehat{\mathbf{Var}}_{\mathbf{X},\mathbf{Y}}(rmg(\theta_b,X,Y))}\sqrt{\widehat{\mathbf{Var}}_{\mathbf{X},\mathbf{Y}}(rmg(\theta_c,X,Y))}}{\sum_{1\leq b<c\leq B}\sqrt{\widehat{\mathbf{Var}}_{\mathbf{X},\mathbf{Y}}(rmg(\theta_b,X,Y))}\sqrt{\widehat{\mathbf{Var}}_{\mathbf{X},\mathbf{Y}}(rmg(\theta_c,X,Y))}}$$

where

$$\widehat{\mathbf{Var}}_{\mathbf{X},\mathbf{Y}}(rmg(\theta,X,Y)) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{I}_{\{g_{\mathcal{P}}(X_i,\theta)=Y_i\}} - \mathbf{I}_{\{g_{\mathcal{P}}(X_i,\theta)\neq Y_i\}} - mg(X_i,Y_i)\right)^2.$$

An estimate of s is given by

$$\hat{s} = \frac{1}{n}\sum_{i=1}^{n} mg(X_i,Y_i) = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\frac{1}{B}\sum_{b=1}^{B}\mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(X_i)=Y_i\}}\right) - \left(\frac{1}{B}\sum_{b=1}^{B}\mathbf{I}_{\{g_{\mathcal{P}}^{(b)}(X_i)\neq Y_i\}}\right)\right\}.$$

Here, we are concerned by correlation and much more by strength, which is connected with bias. In Random Uniform Forests, we first expect to increase the strength by adding more trees. If we try more optimization, strength may also increase but correlation will increase too. Fortunately, we can lower correlation by using more randomization (up to a limit) without not affecting strength too much. we call Breiman's second bound the *expected bound of the prediction error* and one result we are looking for is that it must not be exceeded, thanks to the OOB classifier.

One can note that *the bound might not work in case of imbalanced classes or correlated covariates*. Empirically, we found that for correlated covariates, some issues arise, depending of the number of correlated variables and the dimension of the problem. One solution is to the change the space of representation, but Variable Importance will be lost. A more natural solution is to use first the algorithm without any optimization (e.g. as a purely random forest) and get a benchmark against which more sophisticated options will challenge. In the case of imbalanced classes, the bound does not work because *strength* will focus on the majority class, coming from the intrinsic growth of trees (cases with the majority class will be more favored by their number rather than by their relation with the labels). However, estimating the correct Breiman's bound usually only requires to balance classes, using the appropriate technique.

`Regression`

In regression, Breiman also provides a bound and the *theoretical prediction error of the forest* (generalization error). The key difference with classification resides in the fact that the bound has stronger link with (average) variance of trees residuals.

Suppose that, for all θ, $\mathbf{E}(Y) = \mathbf{E}_{\mathbf{X}}(g_{\mathcal{P}}(X,\theta))$. We have, (Breiman, 2001, theorem 11.2),

$$PE(forest) \leq \bar{\rho} PE(tree),$$

where

$$PE(tree) \stackrel{def}{=} PE(g_{\mathcal{P}}(X,\theta)) = \mathbf{E}_{\theta}\mathbf{E}_{\mathbf{X},\mathbf{Y}}(Y - g_{\mathcal{P}}(X,\theta))^2,$$

is the *average prediction error of a tree* (or the average variance of trees residuals) and

$$\begin{aligned} PE(forest) = PE(\mathbf{E}_{\theta} g_{\mathcal{P}}(X,\theta)) &= \mathbf{E}_{\mathbf{X},\mathbf{Y}}\left(Y - \mathbf{E}_{\theta} g_{\mathcal{P}}(X,\theta)\right)^2 \\ &= \mathbf{E}_{\mathbf{X},\mathbf{Y}}\left\{\mathbf{E}_{\theta}\left(Y - g_{\mathcal{P}}(X,\theta)\right)\right\}^2 \\ &= \mathbf{E}_{\theta}\mathbf{E}_{\theta'}\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left(Y - g_{\mathcal{P}}(X,\theta)\right)\left(Y - g_{\mathcal{P}}(X,\theta')\right). \end{aligned}$$

$\bar{\rho}$ is the *average (weighted) correlation between trees residuals*, defined by

$$\bar{\rho} = \frac{\mathbf{E}_\theta \mathbf{E}_{\theta'} \rho(\theta, \theta') \sqrt{\mathbf{Var}_{\mathbf{X},\mathbf{Y}} \left(Y - g_{\mathcal{P}}(X, \theta)\right)} \sqrt{\mathbf{Var}_{\mathbf{X},\mathbf{Y}}(Y - g_{\mathcal{P}}(X, \theta'))}}{\left(\mathbf{E}_\theta \sqrt{\mathbf{Var}_{\mathbf{X},\mathbf{Y}}(Y - g_{\mathcal{P}}(X, \theta))}\right)^2},$$

where ρ is the correlation between two trees residuals of random and independently chosen parameters θ and θ'.

We also get the *estimate of the theoretical prediction error of the forest*, given by

$$\widehat{PE^*}(forest) = \widehat{PE^*}(\bar{g}_{\mathcal{P}}^{(B)}(X)) = \hat{\bar{\rho}} \left(\frac{1}{B} \sum_{b=1}^{B} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - g_{\mathcal{P}}^{(b)}(X_i)\right)^2} \right)^2,$$

with

$$\hat{\bar{\rho}} = \frac{\sum_{1 \leq b < c \leq B} \hat{\rho}_{b,c}(\theta_b, \theta_c) \sqrt{\widehat{\mathbf{Var}}_{\mathbf{X},\mathbf{Y}}(Y - g_{\mathcal{P}}^{(b)}(X, \theta_b))} \sqrt{\widehat{\mathbf{Var}}_{\mathbf{X},\mathbf{Y}}(Y - g_{\mathcal{P}}^{(c)}(X, \theta_c))}}{\left(\sum_{b=1}^{B} \sqrt{\widehat{\mathbf{Var}}_{\mathbf{X},\mathbf{Y}}(Y - g_{\mathcal{P}}^{(b)}(X, \theta_b))}\right)^2}.$$

In regression we are concerned by both correlation and *average prediction error of trees*. Here, each time one wants to lower variance, correlation increases unless one finds a way to do both reduction. To avoid variance getting high, one strategy is to work more on the dimension, growing large and deep trees and, in many cases, to use post-processing. In regression, the main problem is that randomization leads to a lot of combinations that affect variance but not enough reduce correlation. We can link this with two points addressed by the algorithm:
- due to this correlation, not enough low, a part of the problem is linked with bootstrap which does not generate enough diversity (in terms of different values). That's why it is not used in Random Uniform Forests for regression.
- Due to variance that may remain high, we allow to select a set (or subset) of variables, at each node, *with replacement* in order to increase competition, and thus try to reduce average variance of trees at some cost (expected low) to the correlation.
A way to assess it is to monitor the theoretical prediction error of the forest using the OOB classifier. If it is lower than the OOB prediction error then improvements may be found at the cost of more computation. If not then one has to take care of possible overfitting that is most likely to happen in the regression case, due to higher correlation between trees residuals.

Breiman's bounds provide strong guarantees. However and in practice, one should avoid to use bound, in regression, as an upper one except if correlation and bias are low enough. The most interesting and practical case comes from the fact that all Breiman's bounds and theoretical prediction error of the forest can be estimated using the OOB classifier, allowing first to compare OOB empirical error and OOB estimates of the bounds. While useful, it still does not allow to conclude about the test error, since both measures come from the same training set while we want to assess the test set.

5 Decomposition of the prediction error and OOB bounds

We suppose, here, that *the relation between Y and X is unknown. Only first and second order moments are expected and the i.i.d. assumption.* Hence, we do not look if there is a model behind the data, since the purpose is to know if *there exist bounds that are data-dependent without any knowledge of a best model or hypothesis from this model.*

Then, we can decompose the prediction error in order to find where to look for the control of the latter. For the Random Uniform Forest classifier, the general form of the true prediction error, in regression, is given by

$$\mathbf{E}_{\mathbf{X},\mathbf{Y}} \left(Y - \mathbf{E}_\theta g_{\mathcal{P}}(X,\theta)\right)^2 =$$
$$\mathbf{E}(\epsilon^2) + \left\{\mathbf{E}_{\mathbf{X},\mathbf{Y}} \left[Y - \mathbf{E}_\theta g_{\mathcal{P}}(X,\theta)\right]\right\}^2 + \mathbf{Var}_{\mathbf{X}} \left(\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta)\right) - 2\mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta), Y),$$

with $\epsilon \stackrel{def}{=} Y - \mathbf{E}(Y)$.

One can note that we consider the infinite forest classifier and write the decomposition for all values of the pair (X, Y). Since, in practice, $\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta)$ is unknown, we can derive the relation above by using the forest classifier with the number of trees, B, fixed. For binary classification (and $Y \in \{0, 1\}$), we get

$$\mathbf{P}_{\mathbf{X},\mathbf{Y}} \left(\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y\right) = \mathbf{P}(Y = 1) + \mathbf{P}_{\mathbf{X}}(\bar{g}_{\mathcal{P}}^{(B)}(X) = 1) - 2\mathbf{E}_{\mathbf{X},\mathbf{Y}} \left\{Y \bar{g}_{\mathcal{P}}^{(B)}(X)\right\}, \quad (1)$$

and, simplifying but not too strongly, for regression,

$$\mathbf{E}_{\mathbf{X},\mathbf{Y}}(Y - \bar{g}_{\mathcal{P}}^{(B)}(X))^2 = \mathbf{E}(Y^2) - 2\mathbf{E}_{\mathbf{X},\mathbf{Y}} \left\{Y \bar{g}_{\mathcal{P}}^{(B)}(X)\right\} + \left\{\mathbf{E}_{\mathbf{X}} \left(\bar{g}_{\mathcal{P}}^{(B)}(X)\right)\right\}^2 + \mathbf{Var}_{\mathbf{X}} \left(\bar{g}_{\mathcal{P}}^{(B)}(X)\right). \quad (2)$$

The last step replaces the expectation operator by its empirical counterpart, leading to match the test error. Replacing Breiman's bounds by their OOB counterparts give *OOB bounds*, whose we want to be upper bounds of test error. Hence, the main task is to match the parameters of Y, the latter being expected to not drift.

A- In the classification case, let us write the OOB empirical error and the test error :

$$\overline{PE}_{oob}^{*(B)} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_{\left\{\bar{g}_{\mathcal{P},oob}^{(B)}(X_i) \neq Y_i\right\}},$$
$$\overline{PE}^{*} = \frac{1}{N-n} \sum_{i=n+1}^{N} \mathbf{I}_{\left\{\bar{g}_{\mathcal{P}}^{(B)}(X_i) \neq Y_i\right\}},$$

where $N - n$, $N > n$, is the size of the test sample.

`Proposition 1.`
i) Let us suppose that for any randomly chosen training and test samples, large enough, the first term of the relation (1) sees its empirical counterpart be approximatively equal in both samples.
ii) Consider the OOB classifier and suppose its correlation ρ, with values of Y in the training sample, is approximatively equal with the one of the forest classifier with unknown values of Y in the test sample.

If n, the size of the training sample, and B are large enough and if

$$iii)\ \sqrt{\widehat{\mathbf{Var_X}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X)\right)} - \sqrt{\widehat{\mathbf{Var_X}}\left(\bar{g}_{\mathcal{P},oob}^{(B)}(X)\right)} > \frac{\left(1-\frac{2}{n}\sum_{i=1}^{n}\mathbf{I}_{\{Y_i=1\}}\right)\left(\frac{1}{N-n}\sum_{i=n+1}^{N}\mathbf{I}_{\left\{\bar{g}_{\mathcal{P}}^{(B)}(X_i)=1\right\}} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_{\left\{\bar{g}_{\mathcal{P},oob}^{(B)}(X_i)=1\right\}}\right)}{2\hat{\rho}\sqrt{\widehat{\mathbf{Var}}(Y|D_n)}}$$

then, for any test sample of size, large enough, $N-n$,

$$\overline{PE}^{*} \leq \overline{PE}_{oob}^{*(B)}.$$

The claim states that we can found non-asymptotic conditions where the OOB error will be an upper bound of the test error. It suffices that we have a large enough training sample (in practice test sample will also be large for real world problems) under the *i.i.d.* assumption (required, otherwise the test error could be everywhere). Then the point *i)* will hold. For the point *ii)*, one can note that the OOB classifier is a weaker classifier than the forest one (since it uses less trees) and is a part of the latter. Hence, its correlation with Y in the training sample will usually be lower or close to the correlation of the forest classifier with Y in the test sample. The term "approximatively equal" illustrates the fact that convergence happens, as trees are added, but not monotonically (small oscillations appear).

B- In the regression case, we have

$$\overline{PE}^{*}(\bar{g}_{\mathcal{P},oob}^{(B)}(X,\theta)) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{g}_{\mathcal{P},oob}^{(B)}(X_i)\right)^2,$$

$$\overline{PE}^{*}(\bar{g}_{\mathcal{P}}^{(B)}(X,\theta)) = \frac{1}{N-n}\sum_{i=n+1}^{N}\left(Y_i - \bar{g}_{\mathcal{P}}^{(B)}(X_i)\right)^2.$$

`Proposition 2.`
i) Let us suppose that for any randomly chosen training and test samples, large enough, the empirical counterpart of $\mathbf{Var}(Y)$ *in the test sample is approximatively equal to the one in the training sample.*
ii) Consider the OOB classifier and suppose its correlation ρ*, with values of* Y *in the training sample, is approximatively equal with the one of the forest classifier with unknown values of* Y *in the test sample.*

If n*, the size of the training sample, and* B *are large enough and if*

$$iii) \left| \frac{1}{N-n} \sum_{i=n+1}^{N} \bar{g}_{\mathcal{P}}^{(B)}(X_i) \right| < \left| \frac{1}{n} \sum_{i=1}^{n} \bar{g}_{\mathcal{P},oob}^{(B)}(X_i) \right| \text{ and } \widehat{\mathbf{Var}}_{\mathbf{X}}(\bar{g}_{\mathcal{P}}^{(B)}(X)) < \widehat{\mathbf{Var}}_{\mathbf{X}}(\bar{g}_{\mathcal{P},oob}^{(B)}(X)),$$

$$\begin{aligned} iv) & \left| \left(\frac{1}{N-n} \sum_{i=n+1}^{N} \bar{g}_{\mathcal{P}}^{(B)}(X_i) \right)^2 - \left(\frac{1}{n} \sum_{i=1}^{n} \bar{g}_{\mathcal{P},oob}^{(B)}(X_i) \right)^2 \right| \\ & > 2\hat{\rho} \sqrt{\widehat{\mathbf{Var}}(Y|D_n)} \left(\sqrt{\widehat{\mathbf{Var}_{\mathbf{X}}}(\bar{g}_{\mathcal{P},oob}^{(B)}(X))} - \sqrt{\widehat{\mathbf{Var}_{\mathbf{X}}}(\bar{g}_{\mathcal{P}}^{(B)}(X))} \right) \\ & \quad + \frac{2}{n} \sum_{i=1}^{n} Y_i \left(\frac{1}{n} \sum_{i=1}^{n} \bar{g}_{\mathcal{P},oob}^{(B)}(X_i) - \frac{1}{N-n} \sum_{i=n+1}^{N} \bar{g}_{\mathcal{P}}^{(B)}(X_i) \right), \end{aligned}$$

then for any test sample of size, large enough, $N - n$

$$\overline{PE}^{*}(\bar{g}_{\mathcal{P}}^{(B)}(X)) \leq \overline{PE}^{*}(\bar{g}_{\mathcal{P},oob}^{(B)}(X)).$$

In regression we have to focus on mean and variance of the forest (and OOB) classifier, with respect to each sample, and thus there is more work to do (in classification the sample has more effect on the prediction error, in regression it is the model). Here, while point *i)* is a consequence of the *i.i.d.* assumption, point *ii)* is a convenient assumption with the same purposes than those discussed in classification. Point *iii)* is the main argument for the OOB error to be a bound and relies only on the values of X in training and test samples. But it is not sufficient since in the relation (2), there is a term which depends both from Y and $\bar{g}_{\mathcal{P}}^{(B)}$. A simple assumption lead us to state that its empirical counterpart should be approximatively equal in the training and test samples. Going further, one may want to restrict measures of approximation for the only target variable. Hence point *iv)* ensures that, if itself and point *iii)* apply, *the only reason that leads the OOB error to not be a bound of the test error is the violation of the i.i.d. assumption.*

`Proposition 2` can be alleviated by allowing the empirical variance of the OOB classifier (in the training sample) to be different to the one of the forest classifier in the test sample.

`Proposition 3.`
Suppose the conditions i) and ii) of the proposition 2 hold.
If n, the size of the training sample, and B are large enough and if

$$\left| \frac{1}{N-n} \sum_{i=n+1}^{N} \bar{g}_{\mathcal{P}}^{(B)}(X_i) \right| < \left| \frac{1}{n} \sum_{i=1}^{n} \bar{g}_{\mathcal{P},oob}^{(B)}(X_i) \right|$$

and

$$\begin{aligned}
&\left| \left(\frac{1}{N-n} \sum_{i=n+1}^{N} \bar{g}_{\mathcal{P}}^{(B)}(X_i) \right)^2 - \left(\frac{1}{n} \sum_{i=1}^{n} \bar{g}_{\mathcal{P},oob}^{(B)}(X_i) \right)^2 \right| \\
&> 2\hat{\rho} \sqrt{\widehat{\mathbf{Var}}(Y|D_n)} \left(\sqrt{\widehat{\mathbf{Var}}_{\mathbf{X}}(\bar{g}_{\mathcal{P},oob}^{(B)}(X))} - \sqrt{\widehat{\mathbf{Var}}_{\mathbf{X}}(\bar{g}_{\mathcal{P}}^{(B)}(X))} \right) \\
&\quad + \frac{2}{n} \sum_{i=1}^{n} Y_i \left(\frac{1}{n} \sum_{i=1}^{n} \bar{g}_{\mathcal{P},oob}^{(B)}(X_i) - \frac{1}{N-n} \sum_{i=n+1}^{N} \bar{g}_{\mathcal{P}}^{(B)}(X_i) \right) + \left(\widehat{\mathbf{Var}}_{\mathbf{X}}(\bar{g}_{\mathcal{P}}^{(B)}(X)) - \widehat{\mathbf{Var}}_{\mathbf{X}}(\bar{g}_{\mathcal{P},oob}^{(B)}(X)) \right),
\end{aligned}$$

then for any test sample of size, large enough, $N-n$

$$\overline{PE}^{*}(\bar{g}_{\mathcal{P}}^{(B)}(X)) \leq \overline{PE}^{*}(\bar{g}_{\mathcal{P},oob}^{(B)}(X)).$$

6 Discussion

In practice, one may state that the empirical variance of Y can be very different in the test sample. Then the *i.i.d.* assumption will no longer hold.

`Overriding the i.i.d. case`

Let us consider the regression case and suppose we train a Random Uniform Forest, using an option called *output perturbation sampling*. This allows to perturb Y, increasing its empirical variance, in the training set and changing (randomly and more or less strongly) its mean. Let's illustrate the method below.

For each tree, a new target is sampled from the original one by :
- sampling at random and with replacement all (or a part of) the values of Y in the training sample,
- computing the mean and the variance of the new sample,
- multiplying the variance by a constant *c, c > 1*,
- generating new values of the target, using the Gaussian distribution whose the parameters are the ones above,
- adjusting the new target to have the same definition domain than the original Y.

This procedure has many consequences and we focus only on a few ones. First, the Random Uniform Forest grown with these new target variables (one different per tree) will also converge, because the average correlation between trees residuals will be much

lower than the one between the forest classifier with the original data. Variance will increase, but still slower than the decrease of correlation. Moreover, the test error will be close or even lower (with post-processing) than the one with the original data. However, the main consequence is that *if the i.i.d. assumption does not hold, the conditions provided in the proposition will hold whenever the empirical variance of the test set is lower than the one computed in the training set with the procedure described above.* Since we state that $c > 1$, if conditions of the proposition are fulfilled, then the empirical variance in the test set should drift a lot to prevent the OOB error to be a bound.

In practice, one just needs to compute the forest classifier with the original data and get the forest classifier. Then compute a new forest classifier with the new targets and get the OOB classifier and error. At last, simply embed this OOB classifier with the original forest classifier, using the conditions provided in the proposition(s).

`Estimating the generalization error and getting upper bounds`

One of the greatest interests of Random Forests is the strong link between theory and practice, arising in a simple manner. All theoretical guarantees and measures can be assessed by their OOB empirical counterparts without worrying about overfitting or even optimality. Overfitting could happen as soon as correlation is getting high. While related to the statistical learning theory, we may see optimality as the expression of offered guarantees about what can be done at best, depending on the given data and the model properties. In the case of Random (Uniform) Forests we can get an estimate of PE, the prediction error. Let us suppose that *correlation between trees and bias are low enough.*
In classification we get

$$\overline{PE}^{*(B)}_{oob} \leq \widehat{PE^*}_{oob} = \frac{\hat{\bar{\rho}}_{oob}(1 - \hat{s}^2_{oob})}{\hat{s}^2_{oob}},$$

where $\overline{PE}^{*(B)}_{oob}$ is estimated by computing the misclassification rate from the n observations of the training sample, using $B' = \lceil \texttt{exp}(-1)B \rceil$ trees and $\widehat{PE^*}_{oob}$ is estimated by computing, from the OOB classifier, the average correlation between trees, $\hat{\bar{\rho}}$, and the average strength, $\hat{s}$.
In regression,

$$\overline{PE}^*(\bar{g}^{(B)}_{\mathcal{P},oob}(X)) \leq \widehat{PE}^*(\bar{g}^{(B)}_{\mathcal{P},oob}(X)) = \hat{\bar{\rho}}_{oob} \left(\frac{1}{B'} \sum_{b=1}^{B} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - g^{(b)}_{\mathcal{P}}(X_i) \right)^2 \mathbf{I}_{\left\{ b \in G^-(X_i, B) \right\}}} \right)^2,$$

and

$$\widehat{PE}^*(\bar{g}^{(B)}_{\mathcal{P},oob}(X)) \leq \hat{\bar{\rho}}_{oob} \overline{PE}^*(g_{\mathcal{P},oob}(X, \theta)) = \hat{\bar{\rho}}_{oob} \left\{ \frac{1}{B'} \sum_{b=1}^{B} \left(\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - g^{(b)}_{\mathcal{P}}(X_i) \right)^2 \mathbf{I}_{\left\{ b \in G^-(X_i, B) \right\}} \right) \right\},$$

where, this time, $\hat{\bar{\rho}}_{oob}$ is the OOB estimate of the average correlation between trees residuals,
$\overline{PE}^*(\bar{g}^{(B)}_{\mathcal{P},oob}(X))$ is the OOB estimate of the mean squared error of the forest,
$\widehat{PE}^*(\bar{g}^{(B)}_{\mathcal{P},oob}(X))$ is the OOB estimate of the theoretical prediction error of the forest,
$\overline{PE}^*(g_{\mathcal{P},oob}(X, \theta))$ is the OOB estimate of the average prediction error of a tree.

7 Conclusion

In this paper, we tried to constantly maintain links between theory and practice, assessing the main properties provided by Breiman and deriving simple conditions under which test error is bounded by the OOB error. These conditions lead to obtain guarantees of the effectiveness of Random (Uniform) Forests in real world problems. While ensemble models already have proven their ability to deal with a wide range of datasets, some issues remain about the good properties of Random Forests. From the results presented in the paper, one can observe that (low) correlation between trees is the main ingredient. But, one also observes that correlation is linked with variance, hence having low correlation and high variance should not lower the prediction error. Moreover variance increases as correlation is reduced. Then, having a low prediction error implies to decrease correlation faster than the increase of variance. This is one of the alternatives, the other being to lower variance for a fixed correlation. For both purposes, studying the OOB error is essential since it gives all the arguments (unfortunately as outputs) needed to lower the prediction error : number of trees, variance, correlation and, while discussed only a little here, hyper-parameters. These ones are, in fact, the core (and the most difficult task) that leads to improve the forest, modifying its structure and getting new formulations that should not break the Breiman's paradigm : increase diversity.

8 Proofs

8.1 Proof of Proposition 1

A bias-variance-covariance decomposition in the case of a binary classification, with values in $\{0,1\}$, is associated to the test error and given by :

$$\begin{aligned}
&\mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y\right) \\
&\qquad = \mathbf{Var}(Y) + \left\{\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left(Y - \bar{g}_{\mathcal{P}}^{(B)}(X)\right)\right\}^2 + \mathbf{Var}_{\mathbf{X}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X)\right) - 2\mathbf{Cov}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X), Y\right),
\end{aligned}$$

with

$$\begin{aligned}
\mathbf{Var}(Y) &= \mathbf{P}(Y=1)\mathbf{P}(Y=0), \\
\left\{\mathbf{E}_{\mathbf{X},\mathbf{Y}}(Y - \bar{g}_{\mathcal{P}}^{(B)}(X))\right\}^2 &= \left\{\mathbf{P}(Y=1) - \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\right\}^2, \\
\mathbf{Var}_{\mathbf{X}}(\bar{g}_{\mathcal{P}}^{(B)}(X)) &= \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=0)\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1), \\
\mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(\bar{g}_{\mathcal{P}}^{(B)}(X), Y) &= \mathbf{E}\left\{\left[\bar{g}_{\mathcal{P}}^{(B)}(X) - \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\right][Y - \mathbf{P}(Y=1)]\right\}.
\end{aligned}$$

it follows that

$$\begin{aligned}
&\mathbf{Var}(Y) + \left\{\mathbf{E}_{\mathbf{X},\mathbf{Y}}(Y - \bar{g}_{\mathcal{P}}^{(B)}(X))\right\}^2 \\
&\qquad = \mathbf{P}(Y=1)\mathbf{P}(Y=0) + \left\{\mathbf{P}(Y=1)\right\}^2 - 2\mathbf{P}(Y=1)\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1) + \left\{\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\right\}^2 \\
&\qquad = \mathbf{P}(Y=1)\left(\mathbf{P}(Y=0) + \mathbf{P}(Y=1) - 2\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\right) + \left\{\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\right\}^2 \\
&\qquad = \mathbf{P}(Y=1)\left(1 - 2\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\right) + \left\{\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\right\}^2,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(\bar{g}_{\mathcal{P}}^{(B)}(X), Y) \\
&\quad = \mathbf{E}\left\{Y\bar{g}_{\mathcal{P}}^{(B)}(X)\right\} - \mathbf{P}(Y=1)\mathbf{E}\left\{\bar{g}_{\mathcal{P}}^{(B)}(X)\right\} - \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\mathbf{E}\left\{Y\right\} \\
&\quad\quad + \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\mathbf{P}(Y=1) \\
&\quad = \mathbf{E}\left\{Y\bar{g}_{\mathcal{P}}^{(B)}(X)\right\} - \mathbf{P}(Y=1)\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1).
\end{aligned}$$

One obtains :

$$\begin{aligned}
&\mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y\right) \\
&\quad = \mathbf{P}(Y=1)\left(1 - 2\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\right) + \left\{\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\right\}^2 + \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=0)\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1) \\
&\quad\quad - 2\left(\mathbf{E}\left\{Y\bar{g}_{\mathcal{P}}^{(B)}(X)\right\} - \mathbf{P}(Y=1)\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\right) \\
&\quad = \mathbf{P}(Y=1) + \left\{\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)\right\}^2 + \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=0)\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1) - 2\mathbf{E}\left\{Y\bar{g}_{\mathcal{P}}^{(B)}(X)\right\}.
\end{aligned}$$

Then

$$\mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y\right) = \mathbf{P}(Y=1) + \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1) - 2\mathbf{E}\left\{Y\bar{g}_{\mathcal{P}}^{(B)}(X)\right\},$$

with

$$\begin{aligned}
\mathbf{E}\left\{Y\bar{g}_{\mathcal{P}}^{(B)}(X)\right\} &= \mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(\bar{g}_{\mathcal{P}}^{(B)}(X), Y) + \mathbf{E}(Y)\mathbf{E}(\bar{g}_{\mathcal{P}}^{(B)}(X)) \\
&= \rho\sqrt{\mathbf{Var}(Y)}\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P}}^{(B)}(X))} + \mathbf{E}(Y)\mathbf{E}(\bar{g}_{\mathcal{P}}^{(B)}(X)) \\
&= \rho\sqrt{\mathbf{Var}(Y)}\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P}}^{(B)}(X))} + \mathbf{P}(Y=1)\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1),
\end{aligned}$$

where ρ is the correlation coefficient between $\bar{g}_{\mathcal{P}}^{(B)}(X)$ and Y. We get :

$$\begin{aligned}
&\mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y\right) \\
&\quad = \mathbf{P}(Y=1) + \mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1)(1 - 2\mathbf{P}(Y=1)) - 2\rho\sqrt{\mathbf{Var}(Y)}\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P}}^{(B)}(X))}.
\end{aligned}$$

$\mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y\right) - \mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P},oob}^{(B)}(X) \neq Y\right) \leq 0$ is equivalent to write :

$$\begin{aligned}
&\left(\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1) - \mathbf{P}(\bar{g}_{\mathcal{P},oob}^{(B)}(X)=1)\right)(1 - 2\mathbf{P}\left(Y=1\right)) \\
&\qquad\qquad -2\sqrt{\mathbf{Var}(Y)}\left(\rho\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P}}^{(B)}(X))} - \rho_{oob}\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P},oob}^{(B)}(X))}\right) \leq 0.
\end{aligned}$$

Finally,

$$\begin{aligned}
&\mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X) \neq Y\right) - \mathbf{P}_{\mathbf{X},\mathbf{Y}}\left(\bar{g}_{\mathcal{P},oob}^{(B)}(X) \neq Y\right) \leq 0 \quad if \\
&\rho\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P}}^{(B)}(X))} - \rho_{oob}\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P},oob}^{(B)}(X))} > \frac{(1 - 2\mathbf{P}\left(Y=1\right))\left(\mathbf{P}(\bar{g}_{\mathcal{P}}^{(B)}(X)=1) - \mathbf{P}(\bar{g}_{\mathcal{P},oob}^{(B)}(X)=1)\right)}{2\sqrt{\mathbf{Var}(Y)}}.
\end{aligned}$$

Under the relations i) and ii) of the proposition, we get the relation iii) and for any test set large enough, $\overline{PE}^{*} \leq \overline{PE}_{oob}^{*(B)}$.□

8.2 Proof of Proposition 2

We recall that $\overline{PE}^*(\bar{g}^{(B)}_{\mathcal{P},oob}(X))$ is the OOB prediction error computed on the training sample and corresponds to an estimation of the prediction error (on the test sample) noted $\overline{PE}^*(\bar{g}^{(B)}_{\mathcal{P}}(X))$. The prediction error for a Random (Uniform) Forest is given by:

$$PE(\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta)_{\mathcal{P}}(X)) = \mathbf{E}_{\mathbf{X},\mathbf{Y}}\left(Y - \mathbf{E}_\theta g_{\mathcal{P}}(X,\theta)\right)^2,$$

and its decomposition gives :

$$\begin{aligned}
&PE(\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta)) \\
&\quad = \mathbf{Var}(Y) + \left\{\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left[Y - \mathbf{E}_\theta g_{\mathcal{P}}(X,\theta)\right]\right\}^2 + \mathbf{Var}_{\mathbf{X}}\left(\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta)\right) - 2\mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta), Y) \\
&\quad = \mathbf{E}(Y^2) + \left(\mathbf{E}_{\mathbf{X},\mathbf{Y}}\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta)\right)^2 - 2\mathbf{E}_{\mathbf{X},\mathbf{Y}}\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta)\mathbf{E}(Y) + \mathbf{Var}_{\mathbf{X}}\left(\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta)\right) \\
&\qquad - 2\mathbf{Cov}_{\mathbf{X},\mathbf{Y}}(\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta), Y) \\
&\quad = \mathbf{E}(Y^2) + \left(\mathbf{E}_{\mathbf{X},\mathbf{Y}}\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta)\right)^2 + \mathbf{Var}_{\mathbf{X}}\left(\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta)\right) - 2\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left(\mathbf{E}_\theta g_{\mathcal{P}}(X,\theta)Y\right).
\end{aligned}$$

For the finite forest (with B trees), we get:

$$PE^*(\bar{g}^{(B)}_{\mathcal{P}}(X)) = \mathbf{E}(Y^2) - 2\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left\{Y\bar{g}^{(B)}_{\mathcal{P}}(X)\right\} + \left\{\mathbf{E}_{\mathbf{X}}\left(\bar{g}^{(B)}_{\mathcal{P}}(X)\right)\right\}^2 + \mathbf{Var}_{\mathbf{X}}\left(\bar{g}^{(B)}_{\mathcal{P}}(X)\right).$$

Having $PE^*(\bar{g}^{(B)}_{\mathcal{P}}(X)) - PE^*(\bar{g}^{(B)}_{\mathcal{P},oob}(X)) \leq 0$ is equivalent to write

$$\begin{aligned}
&-2\left(\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left\{Y\bar{g}^{(B)}_{\mathcal{P}}(X)\right\} - \mathbf{E}_{\mathbf{X},\mathbf{Y}}\left\{Y\bar{g}^{(B)}_{\mathcal{P},oob}(X)\right\}\right) + \left(\left\{\mathbf{E}_{\mathbf{X}}\left(\bar{g}^{(B)}_{\mathcal{P}}(X)\right)\right\}^2 - \left\{\mathbf{E}_{\mathbf{X}}\left(\bar{g}^{(B)}_{\mathcal{P},oob}(X)\right)\right\}^2\right) \\
&+ \left(\mathbf{Var}_{\mathbf{X}}\left(\bar{g}^{(B)}_{\mathcal{P}}(X)\right) - \mathbf{Var}_{\mathbf{X}}\left(\bar{g}^{(B)}_{\mathcal{P},oob}(X)\right)\right) \leq 0.
\end{aligned}$$

Under the conditions i) to iii) of the proposition, we need to have

$$-2\left(\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left\{Y\bar{g}^{(B)}_{\mathcal{P}}(X)\right\} - \mathbf{E}_{\mathbf{X},\mathbf{Y}}\left\{Y\bar{g}^{(B)}_{\mathcal{P},oob}(X)\right\}\right) < 0.$$

Since $\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left\{Y\bar{g}^{(B)}_{\mathcal{P}}(X)\right\} = \rho\sqrt{\mathbf{Var}(Y)}\sqrt{\mathbf{Var}(\bar{g}^{(B)}_{\mathcal{P}}(X))} + \mathbf{E}(Y)\mathbf{E}(\bar{g}^{(B)}_{\mathcal{P}}(X))$ we get

$$\begin{aligned}
&2\left(\rho\sqrt{\mathbf{Var}(Y)}\sqrt{\mathbf{Var}(\bar{g}^{(B)}_{\mathcal{P}}(X))} + \mathbf{E}(Y)\mathbf{E}(\bar{g}^{(B)}_{\mathcal{P}}(X))\right) \\
&- 2\left(\rho\sqrt{\mathbf{Var}(Y)}\sqrt{\mathbf{Var}(\bar{g}^{(B)}_{\mathcal{P},oob}(X))} + \mathbf{E}(Y)\mathbf{E}(\bar{g}^{(B)}_{\mathcal{P},oob}(X))\right) \geq 0
\end{aligned}$$

as a sufficient condition. But since we do not want to depend too much on the target variable (whose unconditional distribution would vary) we have to find a stronger condition.
First we write the relation above in terms of the training and test samples under conditions i), ii) and iii).
Second, we bound the relation by looking the empirical counterpart of
$\left|\left\{\mathbf{E}_{\mathbf{X}}\left(\bar{g}^{(B)}_{\mathcal{P}}(X)\right)\right\}^2 - \left\{\mathbf{E}_{\mathbf{X}}\left(\bar{g}^{(B)}_{\mathcal{P},oob}(X)\right)\right\}^2\right|$.

Note that when coming to the estimators, the difference of $\mathbf{E}(Y^2)$ in training and test samples is not 0. That's why conditions are stated to simplify the problem and treat it as it is expected in the *i.i.d.* case. However drift in parameters can be handled (see Discussion), still holding the proposition.
The empirical counterpart of

$$- 2\rho\sqrt{\mathbf{Var}(Y)}\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P}}^{(B)}(X))} - 2\mathbf{E}(Y)\mathbf{E}(\bar{g}_{\mathcal{P}}^{(B)}(X)) + 2\rho\sqrt{\mathbf{Var}(Y)}\sqrt{\mathbf{Var}(\bar{g}_{\mathcal{P},oob}^{(B)}(X))} + 2\mathbf{E}(Y)\mathbf{E}(\bar{g}_{\mathcal{P},oob}^{(B)}(X))$$

is given by

$$I = 2\hat{\rho}\sqrt{\widehat{\mathbf{Var}}(Y|D_n)}\left(\sqrt{\widehat{\mathbf{Var}}_{\mathbf{X}}(\bar{g}_{\mathcal{P},oob}^{(B)}(X))} - \sqrt{\widehat{\mathbf{Var}}_{\mathbf{X}}(\bar{g}_{\mathcal{P}}^{(B)}(X))}\right) + \frac{2}{n}\sum_{i=1}^{n} Y_i\left(\frac{1}{n}\sum_{i=1}^{n}\bar{g}_{\mathcal{P},oob}^{(B)}(X_i) - \frac{1}{N-n}\sum_{i=n+1}^{N}\bar{g}_{\mathcal{P}}^{(B)}(X_i)\right).$$

Hence, it suffices, under conditions i), ii) and iii), that

$$\left|\left(\frac{1}{N-n}\sum_{i=n+1}^{N}\bar{g}_{\mathcal{P}}^{(B)}(X_i)\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\bar{g}_{\mathcal{P},oob}^{(B)}(X_i)\right)^2\right| > I$$

to have $\overline{PE}^*(\bar{g}_{\mathcal{P}}^{(B)}(X)) \leq \overline{PE}^*(\bar{g}_{\mathcal{P},oob}^{(B)}(X))$.□

8.3 Proof of Proposition 3

Recalling that $PE^*(\bar{g}_{\mathcal{P}}^{(B)}(X)) - PE^*(\bar{g}_{\mathcal{P},oob}^{(B)}(X)) \leq 0$ leads to write that

$$- 2\left(\mathbf{E}_{\mathbf{X},\mathbf{Y}}\left\{Y\bar{g}_{\mathcal{P}}^{(B)}(X)\right\} - \mathbf{E}_{\mathbf{X},\mathbf{Y}}\left\{Y\bar{g}_{\mathcal{P},oob}^{(B)}(X)\right\}\right) + \left(\left\{\mathbf{E}_{\mathbf{X}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X)\right)\right\}^2 - \left\{\mathbf{E}_{\mathbf{X}}\left(\bar{g}_{\mathcal{P},oob}^{(B)}(X)\right)\right\}^2\right) + \left(\mathbf{Var}_{\mathbf{X}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X)\right) - \mathbf{Var}_{\mathbf{X}}\left(\bar{g}_{\mathcal{P},oob}^{(B)}(X)\right)\right) \leq 0.$$

With no assumption of the order between $\mathbf{Var}_{\mathbf{X}}\left(\bar{g}_{\mathcal{P},oob}^{(B)}(X)\right)$ and $\mathbf{Var}_{\mathbf{X}}\left(\bar{g}_{\mathcal{P}}^{(B)}(X)\right)$, it suffices (following the proof of proposition 2) to fulfill the conditions of the proposition to get the result.□

References

Biau, G., 2012. Analysis of a Random Forests Model. *The Journal of Machine Learning Research* 13, 1063-1095.

Biau, G., Devroye, L., Lugosi, G., 2008. Consistency of random forests and other averaging classifiers. *The Journal of Machine Learning Research* 9, 2015-2033.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C., 1984. *Classification and Regression Trees*. New York: Chapman and Hall.

Breiman, L., 1996. Bagging predictors. *Machine learning* 24, 123-140.

Breiman, L., 1996. Heuristics of instability and stabilization in model selection. *The annals of statistics* 24, 2350-2383

Breiman, L., 2001. Random forests. *Machine learning* 45, 5-32.

Breiman, L., 2001. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science* 16, 199-231

Ciss, S., 2014. *randomUniformForest: random Uniform Forests for Classification, Regression and Unsupervised Learning*.
R package version 1.1.2, http://CRAN.R-project.org/package=randomUniformForest.

Ciss, S., 2015. Random Uniform Forests. hal-01104340.

Devroye, L., Györfi, L., Lugosi, G., 1996. *A probabilistic theory of pattern recognition*. New York: Springer.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29, 1189-1232.

Friedman, J.H., 2002. Stochastic gradient boosting. *Computational Statistics and Data Analysis* 38, 367-378.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. *Machine Learning* 63, 3-42.

Hastie, T., Tibshirani, R., Friedman, J.J.H., 2001. *The elements of statistical learning*. New York: Springer.

Ho, T.K., 1998. The random subspace method for constructing decision forests.*IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 832-844.

Lin, Y., Jeon, Y., 2002. Random Forests and Adaptive Nearest Neighbors. *Journal of the American Statistical Association* 101-474.

Scornet, E., Biau, G., Vert, J. P., 2014. Consistency of Random Forests. *arXiv* preprint arXiv:1405.2881.

Shannon, C.E., 1949. *The Mathematical Theory of Communication.* University of Illinois Press.

Vapnik, V.N., 1995. *The nature of statistical learning theory.* Springer-Verlag New York.