Journal of Machine Learning Research 17 (2016) 1-38 Submitted 5/15; Revised 7/16; Published 11/16

Wavelet decompositions of Random Forests - smoothness analysis, sparse approximation and applications

Oren Elisha *School of Mathematical Sciences*
University of Tel-Aviv
and GE Global Research
Israel

Shai Dekel *School of Mathematical Sciences*
University of Tel-Aviv
and GE Global Research
Israel

Editors: Lawrence Carin

Abstract

In this paper we introduce, in the setting of machine learning, a generalization of wavelet analysis which is a popular approach to low dimensional structured signal analysis. The wavelet decomposition of a Random Forest provides a sparse approximation of any regression or classification high dimensional function at various levels of detail, with a concrete ordering of the Random Forest nodes: from 'significant' elements to nodes capturing only 'insignificant' noise. Motivated by function space theory, we use the wavelet decomposition to compute numerically a 'weak-type' smoothness index that captures the complexity of the underlying function. As we show through extensive experimentation, this sparse representation facilitates a variety of applications such as improved regression for difficult datasets, a novel approach to feature importance, resilience to noisy or irrelevant features, compression of ensembles, etc.

Keywords: Random Forest, Wavelets, Besov spaces, adaptive approximation, feature importance.

1. Introduction

Our work brings together Function Space theory, Harmonic Analysis and Machine Learning for the analysis of high dimensional big data. In the field of (low-dimensional) signal processing, there is a complete theory that models structured datasets (e.g audio, images, video) as functions in certain Besov spaces (DeVore 1998), (DeVore et. al. 1992). When representing the signal using time-frequency localized dictionaries, this theory characterizes

the performance of adaptive approximation and is used in a variety of applications, such as denoising, compression, feature extraction, etc. using very simple algorithms.

The first contribution of this work is a construction of wavelet decomposition of Random Forests (Breiman 2001), (Biau and Scornet 2016), (Denil et. al. 2014). Wavelets (Daubechies 1992), (Mallat 2009) and geometric wavelets (Dekel and Leviatan 2005), (Alani et. al 2007), (Dekel and Gershtansky 2012), are a powerful yet simple tool for constructing sparse representations of 'complex' functions. The Random Forest (RF) (Biau and Scornet 2016), (Criminisi et. al. 2011), (Hastie et. al. 2009) introduced by Breiman (Breiman 2001), (Breiman 1996), is a very effective machine learning method that can be considered as a way to overcome the 'greedy' nature and high variance of a single decision tree. When combined, the wavelet decomposition of the RF unravels the sparsity of the underlying function and establishes an order of the RF nodes from 'important' components to 'negligible' noise. Therefore, the method provides a better understanding of any constructed RF. This helps to avoid over-fitting in certain scenarios (e.g. small number of trees), to remove noise or provide compression. Our approach could also be considered as an alternative method for pruning of ensembles (Chen et. al. 2009), (Kulkarni and Sinha 2012), (Yang et. al. 2012), (Joly et. al. 2012) where the most important decision nodes of a huge and complex ensemble of models can be quickly and efficiently extracted. Thus, instead of controlling complexity by restricting trees' depth or node size, one controls complexity through adaptive wavelet approximation.

Our second contribution is to generalize the function space characterization of adaptive algorithms (DeVore 1998), (Devore and Lorentz 1993), to a typical machine learning setup. Using the wavelet decomposition of a RF, we can actually numerically compute a 'weak-type' smoothness index of the underlying regression or classification function overcoming noise. We prove the first part of the characterization and demonstrate, using several examples, the correspondence between the smoothness of the underlying function and properties such as compression.

Applying a 'wavelet-type' machinery for learning tasks, using 'Treelets', was introduced by (Lee et. al. 2008). Treelets provide a decomposition of the domain into localized basis functions that enable a sparse representation of smooth signals. This method performs a bottom-up construction in the feature space, where at each step, a local PCA among two correlated variables generates a new node in a tree. Our method is different, since for supervised learning tasks, the response variable should be used during the construction of the adaptive representation. Also, our work significantly improves upon the 'wavelet-type' construction of (Gavish et. al. 2010). First, since our wavelet decomposition is built on the solid foundations of RFs, it leverages on the well-known fact that over-complete representations/ensembles outperform the critical sampled representations/single decision trees

in problems such as regression, estimation, etc. Secondly, from the theoretical perspective, the Lipschitz space analysis of (Gavish et. al. 2010) is generalized by our Besov space analysis, which is the right mathematical setup for adaptive approximation using wavelets.

The paper is organized as follows: In Section 2 we review Random Forests. In Section 3 we present our main wavelet construction and list some of its key properties. In section 4 we present some theoretical aspects of function space theory and its connection to sparsity. This characterization quantifies the sparsity of the data with respect to the response variable. In Section 5 we review how a novel form of Variable Importance (VI) is computed using our approach. Section 6 provides extensive experimental results that demonstrate the applicative added value of our method in terms of regression, classification, compression and variable importance quantification.

2. Overview of Random Forests

We begin with an overview of single trees. In statistics and machine learning (Breiman et. al. 1984), (Alpaydin 2004), (Biau and Scornet 2016), (Denil et. al. 2014), (Hastie et. al. 2009) the construction is called a Decision Tree or the Classification and Regression Tree (CART) while in image processing and computer graphics (Radha et. al. 1996), (Salembier and Garrido 2000) it is coined as the Binary Space Partition (BSP) tree. We are given a real-valued function $f \in L_2(\Omega_0)$ or a discrete dataset $\{x_i \in \Omega_0, f(x_i)\}_{i\in I}$, in some convex bounded domain $\Omega_0 \subset \mathbb{R}^n$. The goal is to find an efficient representation of the underlying function, overcoming the complexity, geometry and possibly non-smooth nature of the function values. To this end, we subdivide the initial domain Ω_0 into two subdomains, e.g. by intersecting it with a hyper-plane. The subdivision is performed to minimize a given cost function. This subdivision process then continues recursively on the subdomains until some stopping criterion is met, which in turn, determines the leaves of the tree. We now describe one instance of the cost function which is related to minimizing variance. At each stage of the subdivision process, at a certain node of the tree, the algorithm finds, for the convex domain $\Omega \subset \mathbb{R}^n$ associated with the node:

(i) A partition by an hyper-plane into two convex subdomains Ω', Ω'' (see Figure 1),

(ii) Two multivariate polynomials $Q_{\Omega'}, Q_{\Omega''} \in \Pi_{r-1}(\mathbb{R}^n)$, of fixed (typically low) total degree r $-$ 1.

The subdomains and the polynomials are chosen to minimize the following quantity

$$\|f - Q_{\Omega'}\|^p_{L_p(\Omega')} + \|f - Q_{\Omega''}\|^p_{L_p(\Omega'')}, \qquad \Omega' \cup \Omega'' = \Omega. \tag{1}$$

Here, for $1 \leq p < \infty$, we used the definition

$$\|g\|_{L_p(\tilde{\Omega})} := \left(\int_{\tilde{\Omega}} |g(x)|^p \, dx \right)^{1/p},$$

If the dataset is discrete, consisting of feature vectors $x_i \in \mathbb{R}^n, i \in I$, with response values $f(x_i)$, then a discrete functional is minimized

$$\sum_{x_i \in \Omega'} |f(x_i) - Q_{\Omega'}(x_i)|^p + \sum_{x_i \in \Omega''} |f(x_i) - Q_{\Omega''}(x_i)|^p, \quad \Omega' \cup \Omega'' = \Omega. \tag{2}$$

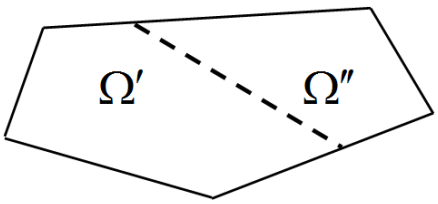


Figure 1: Illustration of a subdivision by an hyperplane of a parent domain Ω into two children Ω', Ω''.

Observe that for any given subdividing hyperplane, the approximating polynomials in (2) can be uniquely determined for $p = 2$ by least square minimization (see (Avery)) for a survey of local polynomial regression). For the order $r = 1$, the approximating polynomials are nothing but the mean of the function values over each of the subdomains

$$Q_{\Omega'}(x) = C_{\Omega'} = \frac{1}{\#\{x_i \in \Omega'\}} \sum_{x_i \in \Omega'} f(x_i), \quad Q_{\Omega''}(x) = C_{\Omega''} = \frac{1}{\#\{x_i \in \Omega''\}} \sum_{x_i \in \Omega''} f(x_i). \tag{3}$$

In many applications of decision trees, the high-dimensionality of the data does not allow to search through all possible subdivisions. As in our experimental results, one may restrict the subdivisions to the class of hyperplanes aligned with the main axes. In contrast, there are cases where one would like to consider more advanced form of subdivisions, where they take certain hyper-surface form, such as conic-sections. Our paradigm of wavelet decompositions can support in principle all of these forms.

Random Forest (RF) is a popular machine learning tool that collects decision trees into an ensemble model (Breiman 2001), (Bernard et. al 2012), (Biau 2012), (Biau and Scornet 2016). The trees are constructed independently in a diverse fashion and prediction is done by a voting mechanism among all trees. A key element (Breiman 2001), is that large diversity between the trees reduces the ensemble's variance. There are many RFs variations

that differ in the way randomness is injected into the model, e.g bagging, random feature subset selection and the partition criterion (Boulesteix et. al. 2012), (Criminisi et. al. 2011), (Hastie et. al. 2009). Our wavelet decomposition paradigm is applicable to most of the RF versions known from the literature.

Bagging (Breiman 1996) is a method that produces partial replicates of the training data for each tree. A typical approach is to randomly select for each tree a certain percentage of the training set (e.g. 80%) or to randomly select samples with repetitions (Hastie et. al. 2009). From an approximation theoretical perspective, this form of RF allows to create an over-complete representation (Christensen 2002) of the underlying function that overcomes the 'greedy' nature of a single tree .

Additional methods to inject randomness can be achieved at the node partitioning level. For each node, we may restrict the partition criteria to a small random subset of the parameter values (hyper-parameter). A typical selection is to search for a partition from a random subset of $\sqrt{n}$ features (Breiman 2001). This technique is also useful for reducing the amount of computations when searching the appropriate partition for each node. Bagging and random feature selections are not mutually exclusive and could be used together.

For $j = 1, ..., J$, one creates a decision tree $\mathcal{T}_j$, based on a subset of the data, X^j. One then provides a weight (score) w_j to the tree $\mathcal{T}_j$, based on the estimated performance of the tree. In the supervised learning, one typically uses the remaining data points $x_i \notin X^j$ to evaluate the performance of $\mathcal{T}_j$. We note that for any point $x \in \Omega_0$, the approximation associated with the j^{th} tree, denoted by $\tilde{f}_j(x)$, is computed by finding the leaf $\Omega \in \mathcal{T}_j$ in which x is contained and then evaluating $\tilde{f}_j(x_i) := Q_\Omega(x)$, where Q_Ω is the corresponding polynomial associated with the decision node Ω. One then assigns a weight $w_j > 0$ to each tree $\mathcal{T}_j$, such that $\sum_{j=1}^{J} w_j = 1$. For simplicity, we will mostly consider in this paper the choice of uniform weights $w_j = 1/J$. One then assigns a value to any point $x \in \Omega_0$ by

$$\tilde{f}(x) = \sum_{j=1}^{J} w_j \tilde{f}_j(x).$$

Typically, in classification problems, the response variables does not have a numeric value, but rather are labeled by one of L classes. In this scenario, each input training point $x_i \in \mathbb{R}^n$ is assigned with a class $Cl(x_i)$. To convert the problem to the 'functional' setting described above one assigns to each class Cl the value of a node on the regular simplex consisting of L vertices in $\mathbb{R}^{L-1}$ (all with equal pairwise distances). Thus, we may assume that the input data is in the form

$$\{x_i, Cl(x_i)\}_{i \in I} \in \left(\mathbb{R}^n, \mathbb{R}^{L-1}\right).$$

In this case, if we choose approximation using constants ($r = 1$), then the calculated mean over any subdomain Ω is in fact a point $\vec{E}_\Omega \in \mathbb{R}^{L-1}$, inside the simplex. Obviously, any value inside the multidimensional simplex, can be mapped back to a class, along with an estimated certainty level, by calculating the closest vertex of the simplex to it. As will become obvious, these mappings can be applied to any wavelet approximation of functions receiving multidimensional values in the simplex.

3. Wavelet decomposition of a random forest

In some applications, there is a need to understand which nodes of a forest encapsulate more information than the others. Furthermore, in the presence of noise, one popular approach is to limit the levels of the tree, so as not to over-fit and contaminate the decisions by noise. Following the classic paradigm of nonlinear approximation using wavelets (Daubechies 1992), (DeVore 1998), (Mallat 2009) and the geometric function space theory presented in (Dekel and Leviatan 2005), (Karaivanov and Petrushev 2003), we present a construction of a wavelet decomposition of a forest. Some aspects of the theoretical justification for the construction are covered in the next section. Let Ω' be a child of Ω in a tree $\mathcal{T}$, i.e. $\Omega' \subset \Omega$ and Ω' was created by a partition of Ω as in Figure 1. Denote by $\mathbf{1}_{\Omega'}$, the indicator function over the child domain Ω', i.e. $\mathbf{1}_{\Omega'}(x) = 1$, if $x \in \Omega'$ and $\mathbf{1}_{\Omega'}(x) = 0$, if $x \notin \Omega'$. We use the polynomial approximations $Q_{\Omega'}, Q_\Omega \in \Pi_{r-1}(\mathbb{R}^n)$, computed by the local minimization (1) and define

$$\psi_{\Omega'} := \psi_{\Omega'}(f) := \mathbf{1}_{\Omega'}(Q_{\Omega'} - Q_\Omega), \tag{4}$$

as the **geometric wavelet** associated with the subdomain Ω' and the function f, or the given discrete dataset $\{x_i, f(x_i)\}_{i \in I}$. Each wavelet $\psi_{\Omega'}$, is a 'local difference' component that belongs to the detail space between two levels in the tree, a 'low resolution' level associated with Ω and a 'high resolution' level associated with Ω'. Also, the wavelets (4) have the 'zero moments' property, i.e., if the response variable is sampled from a polynomial of degree $r-1$ over Ω, then our local scheme will compute $Q_{\Omega'}(x) = Q_\Omega(x) = f(x)$, $\forall x \in \Omega$, and therefore $\psi_{\Omega'} = 0$.

Under certain mild conditions on the tree $\mathcal{T}$ and the function f, we have by the nature of the wavelets, the 'telescopic' sum of differences

$$f = \sum_{\Omega \in \mathcal{T}} \psi_\Omega, \qquad \psi_{\Omega_0} := Q_{\Omega_0}. \tag{5}$$

For example, (5) holds in L_p-sense, $1 \leq p < \infty$, if $f \in L_p(\Omega_0)$ and for any $x \in \Omega_0$ and series of domains $\Omega_l \in \mathcal{T}$, each on a level l with $x \in \Omega_l$, we have that $\lim_{l \to \infty} \operatorname{diam}(\Omega_l) = 0$.

In the setting of a real-valued function, the norm of a wavelet is computed by

$$\|\psi_{\Omega'}\|_2^2 = \int_{\Omega'} \left(Q_{\Omega'}(x) - Q_{\Omega}(x)\right)^2 dx,$$

and in the discrete case by,

$$\|\psi_{\Omega'}\|_2^2 = \sum_{x_i \in \Omega'} \left|Q_{\Omega'}(x_i) - Q_{\Omega}(x_i)\right|^2, \tag{6}$$

where Ω' is a child of Ω.

Observe that for $r = 1$, the subdivision process for partitioning a node by minimizing (1) is equivalent to maximizing the sum of squared norms of the wavelets that are formed in that partition

Lemma 1 *For any partition $\Omega = \Omega' \cup \Omega''$ denote*

$$V_\Omega := \sum_{x_i \in \Omega'} |f(x_i) - C_{\Omega'}|^2 + \sum_{x_i \in \Omega''} |f(x_i) - C_{\Omega''}|^2,$$

where $C_{\Omega'}, C_{\Omega''}$ are defined in (3) and

$$W_\Omega := \|\psi_{\Omega'}\|_2^2 + \|\psi_{\Omega''}\|_2^2.$$

Then, the minimization (2) of V_Ω is equivalent to maximization of W_Ω over all choices of subdomains Ω', Ω'', $\Omega = \Omega' \cup \Omega''$ and constants $C_{\Omega'}, C_{\Omega''}$.

Proof See Appendix.

Recall that our approach is to convert classification problems into a 'functional' setting by assigning the L class labels to vertices of a simplex in $\mathbb{R}^{L-1}$. In such cases of multi-valued functions, choosing $r = 1$, the wavelet $\psi_{\Omega'} : \mathbb{R}^n \to \mathbb{R}^{L-1}$ is

$$\psi_{\Omega'} = \mathbf{1}_{\Omega'} \left(\vec{E}_{\Omega'} - \vec{E}_{\Omega}\right),$$

and its norm is given by

$$\|\psi_{\Omega'}\|_2^2 = \sum_{x_i \in \Omega'} \left\|\vec{E}_{\Omega'} - \vec{E}_{\Omega}\right\|_{l_2}^2 = \left\|\vec{E}_{\Omega'} - \vec{E}_{\Omega}\right\|_{l_2}^2 \# \left\{x_i \in \Omega'\right\}., \tag{7}$$

where for $\vec{v} \in \mathbb{R}^{L-1}$,$\|\vec{v}\|_{l_2} := \sqrt{\sum_{i=1}^{L-1} v_i^2}$.

Using any given weights assigned to the trees, we obtain a wavelet representation of the entire RF

$$\tilde{f}(x) = \sum_{j=1}^{J} \sum_{\Omega \in \mathcal{T}_j} w_j \psi_\Omega(x). \tag{8}$$

The theory (see Theorem 4 below) tells us that sparse approximation is achieved by ordering the wavelet components based on their norm

$$w_{j(\Omega_{k_1})} \left\| \psi_{\Omega_{k_1}} \right\|_2 \geq w_{j(\Omega_{k_2})} \left\| \psi_{\Omega_{k_2}} \right\|_2 \geq w_{j(\Omega_{k_3})} \left\| \psi_{\Omega_{k_3}} \right\|_2 \cdots \tag{9}$$

with the notation $\Omega \in \mathcal{T}_j \Rightarrow j(\Omega) = j$. Thus, the adaptive M-term approximation of a RF is

$$f_M(x) := \sum_{m=1}^{M} w_{j(\Omega_{k_m})} \psi_{\Omega_{k_m}}(x). \tag{10}$$

Observe that, contrary to existing tree pruning techniques, where each tree is pruned separately, the above approximation process applies a 'global' pruning strategy where the significant components can come from any node of any of the trees at any level. For simplicity, one could choose $w_j = 1/J$, and obtain

$$f_M(x) = \frac{1}{J} \sum_{m=1}^{M} \psi_{\Omega_{k_m}}(x). \tag{11}$$

Figure 2 below depicts an M-term (11) selected from an RF ensemble. The red colored nodes illustrate the selection of the M wavelets with the highest norm values from the entire forest. Observe that they can be selected from any tree at any level, with no connectivity restrictions.

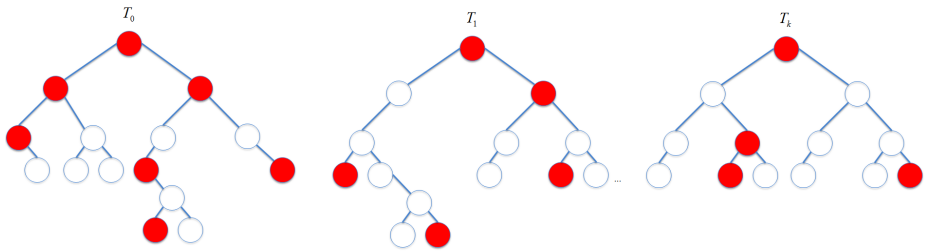


Figure 2: Selection of an M-term approximation from the entire forest.

Figure 3 depicts how the parameter M is selected for the challenging "Red Wine Quality" dataset from the UCI repository (UCI repository). The generation of 10 decision trees on the training set creates approximately 3500 wavelets. The parameter M is then selected by minimization of the approximation error on a validation set. In contrast with other pruning

methods (Loh 2011), using (9), the wavelet approximation method may select significant components from any tree and any level in the forest. By this method, one does not need to predetermine the maximal depth of the trees and over-fitting is controlled by the selection of significant wavelet components.

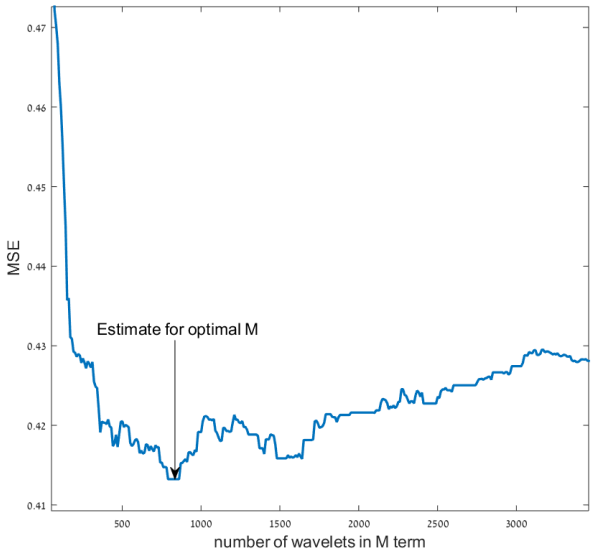


Figure 3: : "Red Wine Quality" dataset - Numeric computation of M for optimal regression.

In a similar manner to certain successful applications in signal processing (e.g. coefficient quantization in the image compression standard JPEG), one may replace the selection of the parameter M in (11), with a threshold parameter $\varepsilon > 0$, chosen suitably for the problem (see for example Section 6.2). One then creates a wavelet approximation using all wavelet terms with norm (6) greater than ε.

In some cases, as presented in (Strobl et. al. 2006) explanatory attributes may be non-descriptive and even noisy, leading to the creation of problematic nodes in the decision trees. Nevertheless, in these cases, the corresponding wavelet norms are controlled and these nodes can be pruned out of the sparse representation (11). The following example demonstrates exactly this, that with high probability, the wavelets associated with the correct variables have relatively higher norms than wavelets associated with non-descriptive variables. Hence the wavelet based criterion will choose, with high probability the correct variable.

Example 1 *Let* $\{y_i\}_{i=1}^m$, *where* $y_i \sim Ber(1/2)$ *i.i.d. and* $\{x_i\}_{i=1}^m \subset [0,1]^n$, $x_i = (x_{i1}, \ldots, x_{ik}, \ldots, x_{in}) \in \mathbb{R}^n$ *with* $x_{ik} = y_i$ *and* x_{ij}, $j \neq k$, *uniformly distributed in* $[0,1]$. *Then, for a subdivision along the* j*th axis,* $[0,1]^n = \Omega' \cup \Omega''$, *and given* $\delta \in (0,1)$, *w.p.* $\geq 1 - \delta$,

1. *If $j \neq k$, then* $\|\psi_{\Omega'}\|_2^2, \|\psi_{\Omega''}\|_2^2 \leq 2\log(2/\delta)$,

2. *If $j = k$ and the subdivision minimizes (1), then*

$$\|\psi_{\Omega'}\|_2^2, \|\psi_{\Omega''}\|_2^2 \geq \left(\frac{m}{2} - \sqrt{\frac{\log(2/\delta)}{2m}}\right)^3 \Big/ m^2.$$

Proof See Appendix.

4. 'Weak-Type' Smoothness and Sparse Representations of the response variable

In this section, we generalize to unstructured and possibly high dimensional datasets, a theoretical framework that has been applied in the context of signal processing, where the data is well structured and of low dimension (DeVore 1998), (DeVore et. al. 1992). The 'sparsity' of a function in some representation is an important property that provides a robust computational framework (Elad 2010(@). Approximation Theory relates the sparsity of a function to its Besov smoothness index and supports cases where the function is not even continuous. Our motivation is to provide additional tools that can be used in the context of machine learning to associate a Besov-index, which is roughly a 'complexity' score, to the underlying function of a dataset. As the theory below and the experimental results show, this index correlates well with the performance of RFs and wavelet decompositions of RFs.

For a function $f \in L_\tau(\Omega)$, $0 < \tau \leq \infty$, $h \in \mathbb{R}^n$ and $r \in \mathbb{N}$, we recall the r-th order difference operator

$$\Delta_h^r(f,x) := \Delta_h^r(f,\Omega,x) := \begin{cases} \sum_{k=0}^{r} (-1)^{r+k} \binom{r}{k} f(x+kh) & [x, x+rh] \subset \Omega, \\ 0 & \text{otherwise,} \end{cases}$$

where $[x,y]$ denotes the line segment connecting any two points $x, y \in \mathbb{R}^n$. The ***modulus of smoothness of order*** r over Ω is defined by

$$\omega_r(f,t)_\tau := \sup_{|h|\leq t} \|\Delta_h^r(f,\Omega,\cdot)\|_{L_\tau(\Omega)}, \quad t > 0,$$

where for $h \in \mathbb{R}^n$, $|h|$ denotes the norm of h. We also denote

$$\omega_r(f,\Omega)_\tau := \omega_r\left(f, \frac{diam(\Omega)}{r}\right)_\tau.$$

Definition 2 *For $0 < p < \infty$ and $\alpha > 0$, we set $\tau = \tau(\alpha, p)$, to be $1/\tau := \alpha + 1/p$. For a given function $f \in L_p(\Omega_0)$, $\Omega_0 \subset \mathbb{R}^n$, and tree $\mathcal{T}$, we define the associated B-space smoothness in $\mathcal{B}_\tau^{\alpha,r}(\mathcal{T})$, $r \in \mathrm{N}$, by*

$$|f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{T})} := \left(\sum_{\Omega \in \mathcal{T}} \left(|\Omega|^{-\alpha} \omega_r (f, \Omega)_\tau \right)^\tau \right)^{1/\tau}, \tag{12}$$

where, $|\Omega|$ denotes the volume of Ω.

We now show that a 'well clustered' function is in fact infinitely smooth in the right adaptively chosen Besov space.

Lemma 3 *Let $f(x) = \sum_{k=}^{K} P_k(x) \mathbf{1}_{B_k}(x)$, where each $B_k \subset \Omega_0$ is a box with sides parallel to the main axes and $P_k \in \Pi_{r-1}$. We further assume that $B_k \cap B_j = \emptyset$, whenever $j \neq k$. Then, there exists an adaptive tree partition $\mathcal{T}$, such that $f \in B_\alpha^{\alpha,r}(\mathcal{T})$, for any $\alpha > 0$.*

Proof See Appendix.

For a given forest $\mathcal{F} = \{\mathcal{T}_j\}_{j=1}^J$ and weights $w_j = 1/J$, the α Besov semi-norm associated with the forest is

$$|f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{F})} := \frac{1}{J} \left(\sum_{j=1}^{J} |f|^\tau_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{T}_j)} \right)^{1/\tau}. \tag{13}$$

The Besov index of f is determined by the maximal index α for which (13) is finite. The above definition generalizes the classical function space theory of Besov spaces, where the tree partitions are non-adaptive. That is, classical Besov spaces may be defined by the special case of partitioning into dyadic cubes, each time using n levels of the tree.

Remark An active research area of approximation theory is the characterization of more geometrically adaptive approximation algorithms by generalizations of the classic 'isotropic' Besov space to more 'geometric' Besov-type spaces (Dahmen et. al. 2001), (Dekel and Leviatan 2005), (Karaivanov and Petrushev 2003). It is known that different geometric approximation schemes are characterized by different flavors of Besov-type smoothness. In this work, for example, we assume all trees are created using partitions along the main n axes. This restriction may lead in general to potentially lower Besov smoothness of the underlying function and the sparsity of the wavelet representation. Yet, the theoretical definitions and results of this paper can also apply to more generalized schemes where for example the tree partitions are by arbitrary hyper-planes. In such a case, the smoothness index of a given function may increase.

Next, for a given tree $\mathcal{T}$ and parameter $0 < \tau < p$ we denote the τ-strength of the tree by

$$N_\tau(f, \mathcal{T}) = \left(\sum_{\Omega \in \mathcal{T}} \|\psi_\Omega\|_p^\tau \right)^{1/\tau}. \tag{14}$$

Observe that

$$\lim_{\tau \to 0} N_\tau(f, \mathcal{T})^\tau = \{\#\Omega \in \mathcal{T} : \psi_\Omega \neq 0\}.$$

Let us further denote the τ-strength of a forest $\mathcal{F}$, by

$$\begin{aligned} N_\tau(f, \mathcal{F}) &:= \frac{1}{J} \left(\sum_{j=1}^{J} \sum_{\Omega \in \mathcal{T}_j} \|\psi_\Omega\|_p^\tau \right)^{1/\tau} \\ &= \frac{1}{J} \left(\sum_{j=1}^{J} N_\tau(f, \mathcal{T}_j)^\tau \right)^{1/\tau}. \end{aligned}$$

In the setting of a single tree constructed to represent a real-valued function, under mild conditions on the partitions (see remark after (5) and condition (17)) , the theory of (Dekel and Leviatan 2005) proves the equivalence

$$|f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{T})} \sim N_\tau(f, \mathcal{T}). \tag{15}$$

This implies that there are constants $0 < C_1 < C_2 < \infty$, that depend on parameters such as α, p, n, r and ρ in condition (17) below, such that

$$C_1 |f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{T})} \leq N_\tau(f, \mathcal{T}) \leq C_2 |f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{T})}.$$

Therefore, we also have for the forest model

$$|f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{F})} \sim N_\tau(f, \mathcal{F}). \tag{16}$$

We now present a "Jackson-type estimate" for the degree of the adaptive wavelet forest approximation. Its proof is in the Appendix.

Theorem 4 *Let $\mathcal{F} = \{\mathcal{T}_j\}_{j=1}^J$ be a forest. Assume there exists a constant $0 < \rho < 1$, such that for any domain $\Omega \in \mathcal{F}$ on a level l and any domain $\Omega' \in \mathcal{F}$, on the level $l+1$, with $\Omega \cap \Omega' \neq \emptyset$, we have*

$$|\Omega'| \leq \rho |\Omega|, \tag{17}$$

where $|E|$ *denotes the volume of* $E \subset \mathbb{R}^n$. *Denote formally* $f = \sum_{\Omega \in \mathcal{F}} w_{j(\Omega)} \psi_\Omega$, *and assume that* $|f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{F})} < \infty$, *where*

$$\frac{1}{\tau} = \alpha + \frac{1}{p}.$$

Then, for the M*-term approximation (10) we have*

$$\sigma_M(f) := \|f - f_M\|_p \leq C\left(p, \alpha, \rho\right) JM^{-\alpha} |f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{F})} \quad . \tag{18}$$

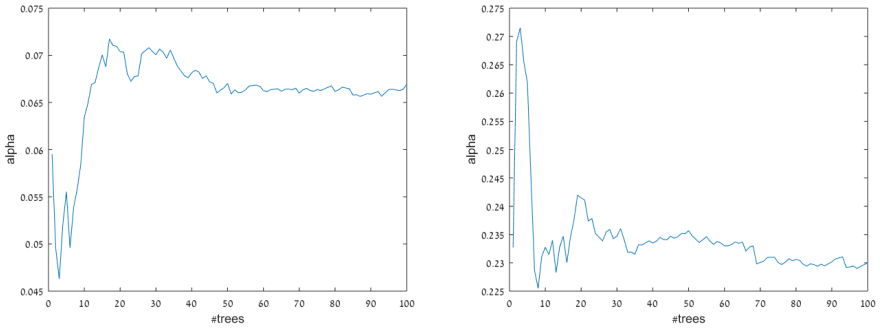
One important contribution of this work is the attempt to generalize to the setting of machine learning, the function space theoretical perspective. There are several candidate numeric methods to estimate the critical 'weak-type' Besov smoothness index α from the given data. That is, the maximal α for which the Besov norm is finite. Our goal is to estimate the true smoothness of the underlying function, removing influences of noise and outliers if exist within the given dataset. One potential method is to use the equivalence (16) and then search for a transient value of τ for which $N_\tau\left(f, \mathcal{F}\right)$ becomes 'infinite'. However, we choose to generalize the numeric algorithm of (DeVore et. al. 1992) and estimate the critical index α using a numeric exponential fit of the error σ_M in (18). We found that it is somewhat more robust to fit each decision tree in the forest with an estimated smoothness index α_j and then average to obtain the estimated forest smoothness α. Thus, based on (18), we model the error function by $\sigma_{j,m} \sim c_j m^{-\alpha_j}$ for unknown c_j, α_j, where $\sigma_{j,m}$ is the approximation error when using the m most significant wavelets of the jth tree. First, notice that we can estimate $c_j \sim \sigma_{j,1}$. Then, using $\int_1^M m^{-u} dm = \left(M^{1-u} - 1\right)/(1-u)$, we estimate α_j by

$$\min_{\alpha_j} \left| \frac{M^{1-\alpha_j} - 1}{1 - \alpha_j} \sigma_{j,1} - \sum_{m=1}^{M-1} \sigma_{j,m} \right|. \tag{19}$$

Similarly to (DeVore et. al. 1992), we select only M significant terms, to avoid fitting the tail of the exponential expression. This is done by discarding wavelets that are overfitting the error on the Out Of Bag (OOB) samples (see Figure 3). Let us see some examples of how this works in practice. As can be seen in Figure 4, the estimate of the Besov index of two target functions using (19) stabilizes after a relatively small number of trees are added.

Next, we show that when an underlying function is not 'well clustered' and has a sharp transition of values across the boundary of two domains, then the Besov index is limited in the general case and suffers from the curse of dimensionality. Again, it should make sense to the practitioners, that such a function can be learnt but with more effort, e.g. trees with higher depth.

Lemma 5 *Let* $f\left(x\right) = \mathbf{1}_{\tilde{\Omega}}\left(x\right)$, *where* $\tilde{\Omega} \subset [0,1]^n$ *is a compact domain with a smooth boundary. Then,* $f \in B_\tau^{\alpha,r}\left(\mathcal{T}_I\right)$, *for* $\alpha < 1/p\left(n-1\right)$, $\tau^{-1} = \alpha + 1/p$, *and any* $r \geq 1$, *where* $\mathcal{T}_I$



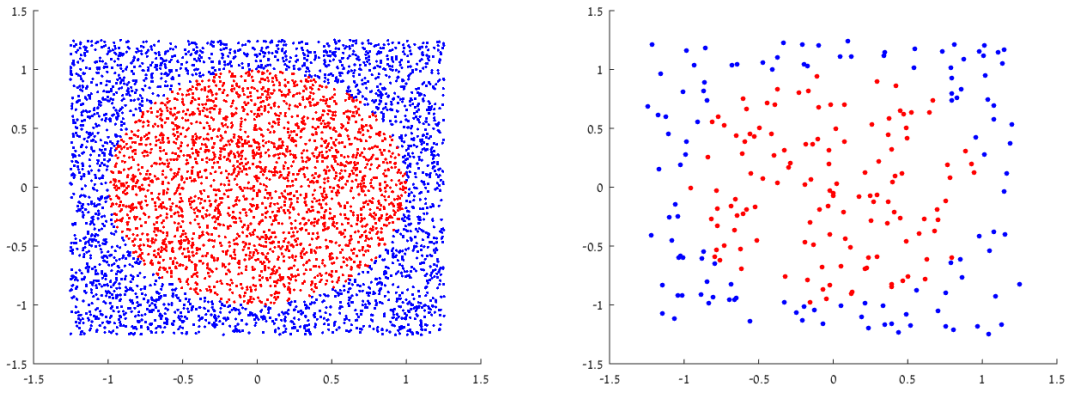
(a) "Red Wine" (b) "Airfoil"

Figure 4: Estimation of the Besov critical smoothness index

is the tree with isotropic dyadic partitions, creating dyadic cubes of side lengths 2^{-k} *on the level* nk*.*

Proof See Appendix.

We note that in the general case, when subdivisions along main axes are used, the non-adaptive tree of the above lemma is almost best possible. That is, one cannot hope for significantly higher smoothness index using an adaptive tree with subdivisions along main axes. In Figure 5(a) we see 5000 random points and in (b) 250 random points, sampled from a uniform distribution taking a response value of $f(x) = \mathbf{1}_{\tilde{\Omega}}(x)$, where $\tilde{\Omega} \subset \mathbb{R}^2$, is the unit sphere.



(a) 5,000 sampled points (b) 250 sampled points

Figure 5: Dataset created by random sampling points of the indicator function of a unit sphere

By Lemma 4.3, the lower bound for the critical Besov exponent of f is $\alpha = 0.5$, for $p = 2$. This should correlate with the intuition of machine learning practitioners: the dataset does have two well defined clusters, but the boundary between the clusters (boundary of the sphere) is a non-trivial curve and any classification algorithm will need to learn the geometry of the curve.

In Figure 6 we see a plot of the numeric calculation of the α Besov index for given number of sampling points of f. We see relatively fast convergence to $\alpha = 0.51$. As discussed, our method attempts to capture the geometric properties of the 'true' underlying function that is potentially buried in the noisy input data. To show this, we constructed from a dataset of 10k samples of f, a ten dimensional dataset, by adding additional eight noisy features, with uniform distribution in [0,1] and no bearing on the response variable. The numeric computation in this example was again,$\alpha = 0.51$, which demonstrates that the method is stable under this noisy embedding in $\mathbb{R}^n$ as well.

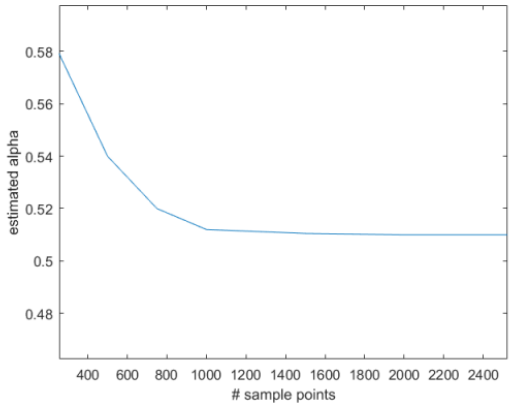


Figure 6: Numeric calculation of the α Besov index for given number of sampling points of the indicator function of a unit sphere.

5. Wavelet-based variable importance

In many cases, there is a need to understand in greater detail in what way the different variables influence the response variable (Guyon and Elisseff 2003). Which of the possibly hundreds of parameters is more critical? What are the interactions between the significant variables? Also, the property of obtaining fewer features that provide equivalent prediction could be used for feature engineering and for 'feature budget algorithms' such as in (Feng et. al. 2015), (Vens and Costa 2011). As described in (Genuer et. al. 2010), the use of RF for variable importance detection has several advantages.

There are several existing Variable Importance (VI) quantification methods that use RF. A popular approach for measuring the importance of a variable is summing the total decrease in node impurities when splitting on the variable, averaged over all trees (RF in R), (Hastie et. al. 2009). As suggested in the RF package documentation of the R language (RF in R): "For classification, the node impurity is measured by the Gini Index. For regression; it is measured by the residual sum of squares". Although not stated specifically in (RF in R), it is common practice to multiply the information gain of each node by its size (Raileanu and Stoffel 2004), (Du and Zhan 2002), (Rokach and Maimon 2005). Additional methods for variable importance measure are the 'Permutation Importance' measure (Genuer et. al. 2010), or similarly 'OOB randomization' (Hastie et. al. 2009). With these latter two methods, sequential predictions of RF are done, when each time one feature is being permuted as the rest of the features remain. Then, the measure for variable importance is the difference in prediction accuracy before and after a feature is permuted in MSE terms.

However, both 'Impurity gain' and 'Permutation' have some pitfalls that should be considered, when used for variable importance. As shown by (Strobl et. al. 2006), the 'Impurity gain' tends to be in favor of variables with more varying values. As shown in (Strobl et. al. 2008), 'Permutation' tends to overestimate the variable importance of highly correlated variables.

The wavelet-based VI is derived by imposing a restriction on the adaptive re-ordering of the wavelet components (11), such that they must appear in 'feature related blocks'. To make this precise, let $\{x \in \mathbb{R}^n, f(x)\}$ be a dataset and let $\tilde{f}$ represent the RF decomposition, as in (8). We evaluate the importance of the i-th feature by

$$S_i^\tau := \frac{1}{J}\sum_{j=1}^{J}\sum_{\Omega\in\mathcal{T}_j\cap V_i}\|\psi_\Omega\|_2^\tau, \quad i=1,\ldots,n, \tag{20}$$

where, $\tau > 0$ and V_i is the set of child domains formed by partitioning their parent domain along the ith variable. This allows us to score the variables, using the ordering $S_{i_1}^\tau \geq S_{i_2}^\tau \geq \cdots$. Recall that our wavelet-based approach transforms classification problems into the functional setting (see section 2) by mapping each label l_k to a vertex $\vec{l_k} \in \mathbb{R}^{L-1}$ of a regular simplex. Therefore, in classification problems, the wavelet norms in (20) are given by (7) which implies that we provide a unified approach to VI.

It is crucial to observe that from an approximation theoretical perspective, the more suitable choice in (20) is $\tau = 1$, since with this choice, the ordering is related to ordering

the variables by the approximation error of their corresponding wavelet subset

$$\begin{aligned}\min_{1\le i\le n}\left\|\tilde{f}-\frac{1}{J}\sum_{j=1}^{J}\sum_{\Omega\in\mathcal{T}_j\cap V_i}\psi_\Omega\right\|_2 &= \min_{1\le i\le n}\left\|\frac{1}{J}\sum_{k\neq i}\sum_{j=1}^{J}\sum_{\Omega\in\mathcal{T}_j\cap V_k}\psi_\Omega\right\|_2\\ &\le \min_{1\le i\le n}\frac{1}{J}\sum_{k\neq i}\sum_{j=1}^{J}\sum_{\Omega\in\mathcal{T}_j\cap V_k}\|\psi_\Omega\|_2\\ &= \min_{1\le i\le n}\sum_{k\neq i}S_k^1\\ &= \sum_{1\le k\le n}S_k^1-\max_{1\le i\le n}S_i^1.\end{aligned}$$

What is interesting is that, in regression problems, when using piecewise constant approximation in (1),(4), the VI score (20) with $\tau=2$, is in fact exactly as in (Louppe et. al. 2013) when variance is used as the impurity measure. To see this, for any dataset $\{x\in\mathbb{R}^n, f(x)\}$ and domain $\tilde{\Omega}$ of an RF, denote briefly

$$K_{\tilde{\Omega}}:=\#\left\{x_i\in\tilde{\Omega}\right\}\quad,\quad \mathrm{Var}\left(\tilde{\Omega}\right)=\frac{1}{\#\left\{x_i\in\tilde{\Omega}\right\}}\sum_{x_i\in\tilde{\Omega}}\left(f(x_i)-C_{\tilde{\Omega}}\right)^2.$$

For any domain Ω of a RF, with children Ω',Ω'', the variance impurity measure is

$$\Delta(\Omega):=\mathrm{Var}(\Omega)-\frac{K_{\Omega'}}{K_\Omega}\mathrm{Var}\left(\Omega'\right)-\frac{K_{\Omega''}}{K_\Omega}\mathrm{Var}\left(\Omega''\right).$$

The importance of the variable i (up to normalization by the size of the dataset) is defined in (Louppe et. al. 2013) by

$$\frac{1}{J}\sum_{j=1}^{J}\sum_{\text{children of}\,\Omega\,\text{in}\,\mathcal{T}_j\cap V_i}K_\Omega\Delta(\Omega). \tag{21}$$

Theorem 6 *The variable importance methods of (20) and (21) are identical for $\tau=2$.*

Proof For any domain Ω and its two children Ω',Ω'',

$$\begin{aligned}K_\Omega\Delta(\Omega) &= K_\Omega\left(\mathrm{Var}(\Omega)-\frac{K_{\Omega'}}{K_\Omega}\mathrm{Var}\left(\Omega'\right)-\frac{K_{\Omega''}}{K_\Omega}\mathrm{Var}\left(\Omega''\right)\right)\\ &= \sum_{x_i\in\Omega}\left(f(x_i)-C_\Omega\right)^2-\sum_{x_i\in\Omega'}\left(f(x_i)-C_{\Omega'}\right)^2-\sum_{x_i\in\Omega''}\left(f(x_i)-C_{\Omega''}\right)^2\\ &= \|\psi_{\Omega'}\|_2^2+\|\psi_{\Omega''}\|_2^2.\end{aligned}$$

Therefore,

$$\frac{1}{J}\sum_{j=1}^{J}\sum_{\text{children of }\Omega\text{ in }\mathcal{T}_j\cap V_i} K_\Omega \Delta\left(\Omega\right) = S_i^2.$$

$$\diamondsuit$$

Further to the choice of $\tau = 1$ over $\tau = 2$ in (20), the novelty of the wavelet-based VI approach is targeted at difficult noisy datasets. In these cases, one should compute VI at various degrees of approximation, using only subsets of 'significant' nodes, by thresholding out wavelet components with norm below some $\varepsilon > 0$

$$S_i^1\left(\varepsilon\right) := \sum_{j=1}^{J}\sum_{\Omega\in\mathcal{T}_j\cap V_i,\ \|\psi_\Omega\|\geq\varepsilon} \|\psi_\Omega\|_2. \tag{22}$$

As pointed out, a popular RF approach for identifying important variables is summing the total decrease in node impurities when splitting on the variable, averaged over all trees (RF in R), (Hastie et. al. 2009). However, this method may not be reliable in situations where potential predictor variables vary in their scale of measurement or their number of categories (Strobl et. al. 2006). This restriction is very limiting in practice, as in many cases binary variables such as 'Gender' are very descriptive where less descriptive variables (or noise) may vary with many values.

To demonstrate this problem, we follow the experiment suggested in (Strobl et. al. 2006). We set a number of samples to $m = 120$, where each sample has two explanatory independent variables: $x_1 \sim N(0,1)$ and $x_2 \sim Ber(0.5)$. A correlation between $y = f(x_1, x_2)$ and x_2 is established by:

$$y \sim \left\{ \begin{array}{ll} Ber(0.7), & x_2 = 0, \\ Ber(0.3), & x_2 = 1. \end{array} \right\} \tag{23}$$

In accordance with the point made in (Strobl et. al. 2006), when applying the VI of (RF in R), (Hastie et. al. 2009) we observe that the important variable is the 'noisy' uncorrelated feature x_1. As shown in Example 1, while we may obtain many false partitions along the noise, with high probability their wavelet norm is controlled, and relatively small. In Figure 7 we see a histogram of the wavelet norms (taken from one of the RF trees) for the example (23). We see that the wavelet norms of the important variable x_2 are larger, but that there exists a long tail of wavelet norms relating to x_1. Therefore, applying the thresholding strategy (22) as part of the feature importance estimation could be advantageous in such a case.

We now address the choice of ϵ in (22). So as to remove the noisy wavelet components from the VI scoring process, we choose the threshold as norm of the M-th wavelet, where

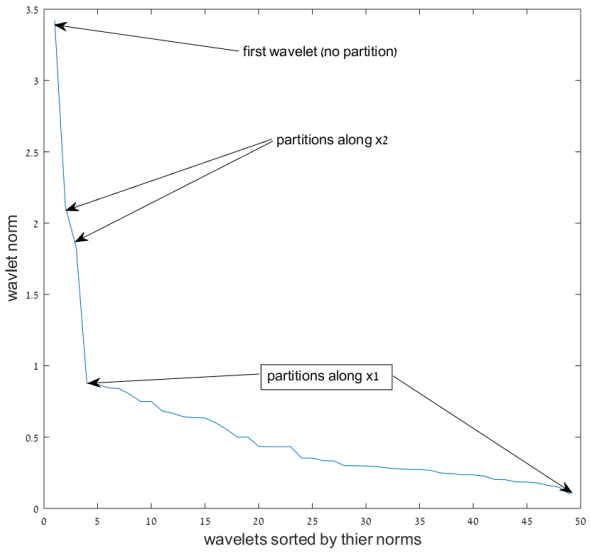


Figure 7: wavelets norms taken from one of the RF trees constructed for the example (23)

M is the selected using the M-term wavelet that minimizes the approximation error on the validation set $\{x_i, f(x_i)\}_{i=1,..k}$ by

$$\varepsilon = \|\psi_M\|_2 , \qquad s.t \ \min_M \left\{ \sum_{i=1}^{k} \left(f(x_i) - \frac{1}{J} \sum_{m=1}^{M} \psi_{\Omega_{k_m}}(x_i) \right)^2 \right\}. \tag{24}$$

The calculation of ϵ for the "Pima diabetes" dataset using a validation set is depicted in Figure 8. In Section 6.2 we demonstrate the advantage of the wavelet-based thresholding technique in VI on several datasets.

6. Applications and Experimental Results

For our experimental results, we implemented C# code that supports RF construction, Besov index analysis, wavelet decompositions of RF and applications such as wavelet-based VI, etc. (source code is available, see link in (Wavelet RF code)). The algorithms are executed on the Amazon Web Services cloud, using up to 120 CPUs. Most datasets are taken from the UCI repository (UCI repository), which allows us to compare our results to previous work.

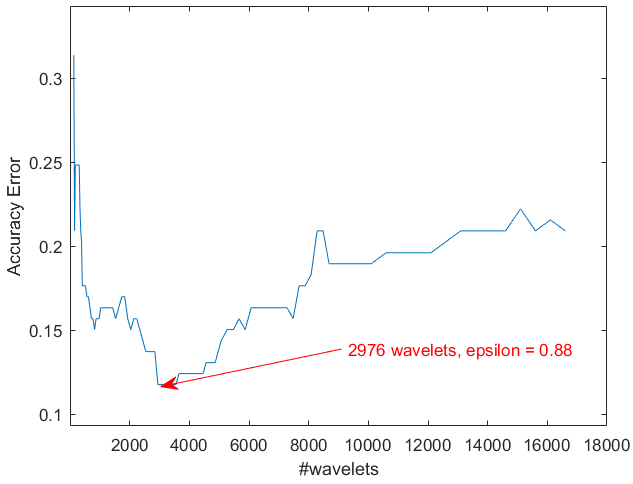


Figure 8: "Pima diabetes" - Choice of ϵ in (22) using the validation set

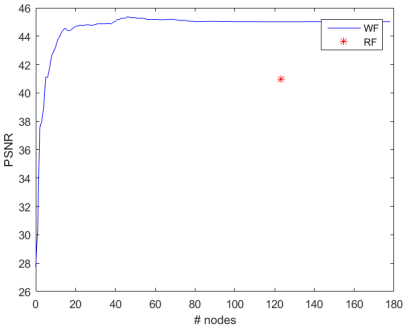
6.1 Ensemble Compression

In applications, constructed predictive models, such as RF, need to be stored, transmitted and applied to new data. In such cases the size of the model becomes a consideration, especially when using many trees to predict large amounts of incoming new data over distributed architectures. Furthermore, as presented in (Geurts and Gilles 2011), the number of total nodes of the RF and the average tree depth impact the memory requirements and evaluation performance of the ensemble.

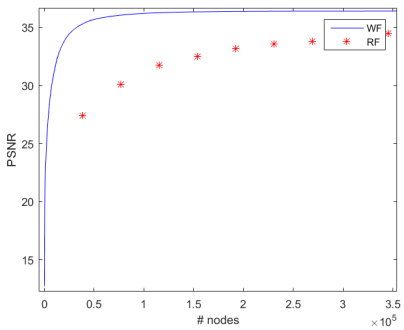
In order to demonstrate the correlation between the Besov index of the underlying function and the 'complexity' of these datasets we need to compare on the same scale different datasets of different sizes and dimensions. Therefore, we replaced the commonly used metrics in machine learning such as MSE (Mean Square Error) by the normalized PSNR (Peak Signal To Noise Ratio) metric which is commonly used in the context of signal processing. For a given dataset $\{x_i, f(x_i)\}$ and an approximation $\{x_i, f_A(x_i)\}$ PSNR is defined by

$$\text{PSNR}:=10\cdot\log_{10}\frac{\max_{i,j}\left\{\left|f(x_i)-f(x_j)\right|^2\right\}}{\frac{1}{\#\{x_i\}}\sum_i\left(f(x_i)-f_A(x_i)\right)^2}.$$

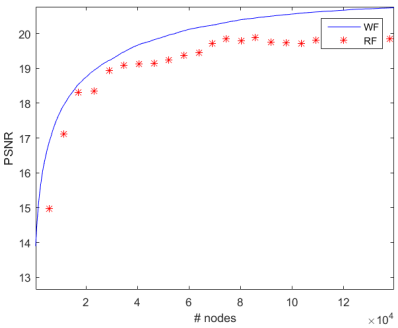
Observe that higher PSNR implies smaller error. In Figure 9 we observe the rate-distortion performance measured on validation points in a fivefold cross validation of $M-term$ wavelet approximation and standard RF, as trees are added. It can be seen that for functions that are smoother in 'weak-type' sense (e.g. higher α), wavelet approximation outperforms the standard RF. Table 1 below shows an extensive list of more datasets.



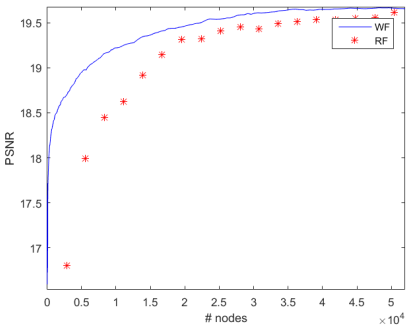
(a) "Record linkage" dataset, $\alpha = 0.99$



(b) "CT Slice" dataset, $\alpha = 0.51$



(c) "Parkinson" dataset , $\alpha = 0.11$



(d) "Red Wine quality" dataset , $\alpha = 0.07$

Figure 9: PSNR of four UCI data sets.

We now compare wavelet-based compression with existing RF pruning strategies. As stated by (Kulkarni and Sinha 2012), most of the current efforts in pruning RF are based on 'Over-produce-and-Choose' strategy, where the forest is grown to a fixed number of trees, and then only a subset of these trees are chosen by a 'leave one out strategy' as in (Martinez-Muoz et. al. 2009), (Yang et. al. 2012). For each dataset we first computed a point at which the graph of wavelet approximation error begins to 'flatten out' on the validation set. We then used this target error pre-saturation point for both wavelet shrinkage and the pruning methods that aim for a minimal number of nodes to achieve it on a validation set of fivefold cross validation. To this end, we have generated RF with 100 decision trees with 80% bagging and $\sqrt{n}$ hyper-parameter. The two pruning strategies are based on a 'leave one out' strategy as presented in (Yang et. al. 2012). In this approach trees are recursively omitted according to their correspondence with the rest of the ensemble (based on the correspondence of the margins in classification and MSE in regression). We have collected the results of the experiment described above applied to 12 UCI datasets in Table 1. The datasets for classification are marked (C) and regression (R). One may observe from Table 1 that the wavelet-based method performs better than conventional pruning. Also, as expected, there is significant correlation between the performance of compression and the function smoothness. That is, the compression is more effective for smoother functions.

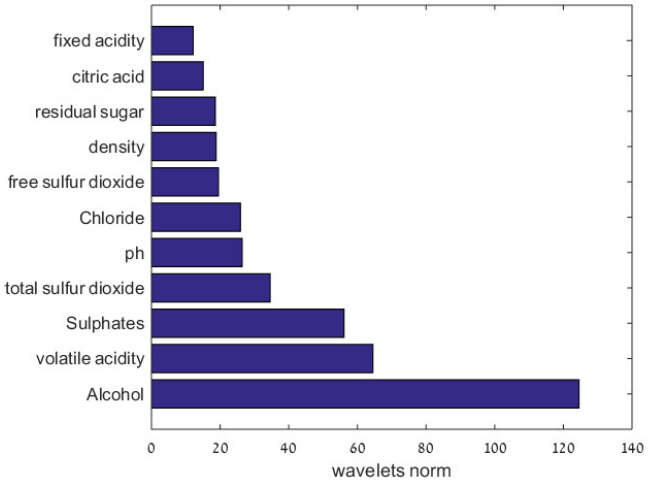
Note that when computing an M-term wavelet approximation, some components may be unconnected as depicted in Figure 2. Obviously, any compression of the wavelet approximation would need to encode the nodal data associated with these unconnected components. Therefore, we enforce connectivity on any wavelet approximation we compute, by adding all wavelet components along the tree paths leading to the selected significant wavelet components. Thus, the wavelet compression appearing in Table 1 is in the form of a collection of J connected subtrees.

6.2 Variable Importance

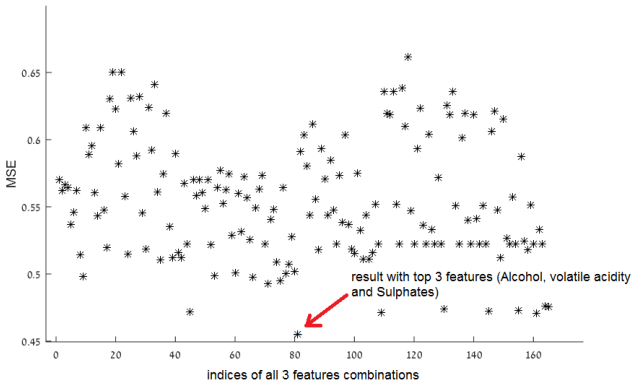
We first demonstrate how the wavelet-based method (20) with $\tau = 1$, succeeds in identifying the features with the highest impact on the prediction. In Figure 10 (a), we show an histogram of VI scores for the "Red Wine" dataset using the wavelet-based approach (20). As can be seen, the top three features in the histogram are 'Alcohol', 'Volatile acidity' and 'Sulphates'. We then constructed RFs for all the possible triple combinations of features of the UCI repository 'Wine Quality' dataset (165 simulations) using 100 trees and 80% bagging. In Figure 10(b), we can see the MSE of each of these RFs. One can verify that the triple 'Alcohol', 'Volatile acidity' and 'Sulphates' (index 81) has the smallest error, as identified by our wavelet-based method.

Table 1: Compression - number of nodes required to reach the error pre-saturation point

Dataset	**error**	**Pruning Min-D** (Yang et. al. 2012)		**Pruning Mean-D** (Yang et. al. 2012)		**Wavelet subtrees**		α
		#trees	#nodes	#trees	#nodes	#trees	#nodes	
Record linkage (C)	2%	1	123	1	123	1	6	0.99
CT Slice (R)	2.9 MSE	2	77042	2	76396	2	5141	0.51
Titanic (C)	17%	3	711	10	2248	1	34	0.42
Balanced scale (C)	22%	1	185	1	185	1	55	0.34
Concrete (R)	15 MSE	19	2297	8	966	3	64	0.32
Magic Gamma (C)	13%	9	26793	5	14961	3	1657	0.25
Airfoil (R)	3.2 MSE	5	4533	3	7487	3	1929	0.23
California Housing (R)	0.5 MSE	4	65436	9	149863	4	7292	0.2
EEG (C)	8%	7	17845	11	28355	6	12808	0.15
Parkinson (R)	3.2 MSE	18	103822	19	110187	12	20947	0.11
Wine quality (R)	0.4 MSE	14	39350	13	36439	12	29089	0.07
Year Prediction (R)	88 MSE	21	10657799	24	12201588	19	9300284	0.02



(a) Wavelet-based feature importance histogram



(b) Error of RFs constructed over all possible 3 feature subsets

Figure 10: Wavelet-based variable importance of the UCI Red wine data set

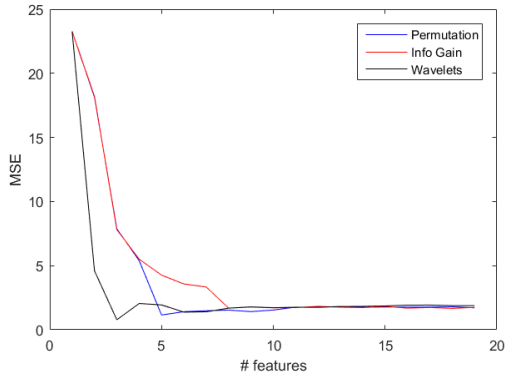
Next, we show that the wavelet-based VI approach, in particular, the noise-removal variant using (22) with $\epsilon > 0$, can provide a better estimation of VI than the existing methods in R (RF in R). Note that we apply the wavelet-based VI method in classification problems as well, competing with, for example, the standard Gini-based algorithm of R. To this end we employ a test methodology used in (Feng et. al. 2015). Using each VI method we first calculate a corresponding VI score of the features. Each method uses an RF with 100 trees and 80% bagging. However, the wavelet-based method was computed using our implementation, based on (22) while the Permutation and Information Gain based methods were applied using R. After each method 'decides' on the order of the features by importance, we iterate by adding features one-by-one, where at the k-th iteration, only the selected first k features are used for prediction. Here also, we used wavelet-based choice of k most important features to construct a wavelet-based best prediction, while for the other methods, we used their choice of k most important features as the input for an R based RF. The results of fivefold cross validation are presented in Figure 11. For example, in Figure 11(a), we see that on the "Parkinson" dataset, the wavelet-based method reaches better prediction using the first three features it selected. This is due to fact that the wavelet-based method selected different features ('Age', 'Time' and 'Gender') than the other methods.

6.3 Classification and regression tasks

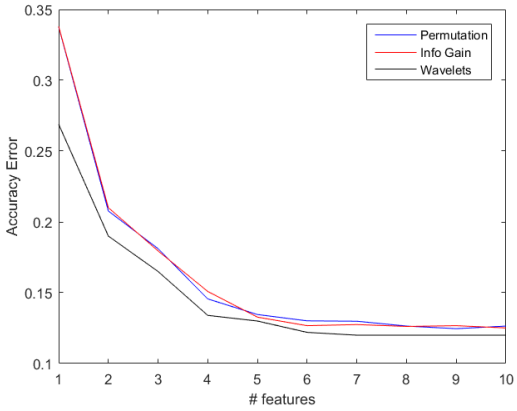
In this Section we focus on difficult datasets, such as small sets or with high bias, bad features, mis-labeling and outliers (see for example "Pima Diabetes" dataset with only 768 samples with 8 attributes in Figure 11(c)) and show that in such cases the wavelet-based approach provides smaller predictive errors.

We begin with a demonstration of a case of 'false labeling' using the R machine learning benchmark "Spirals" (Spiral dataset). From the given dataset we create a dataset with mis-labeling by randomly replacing 20% of the values of the response variable. The original and noisy datasets are rendered in Figure 12. We then compare the predictive performance of the standard RF and the M-term wavelet approximation (11), where optimal M values are computed automatically as depicted in Figure 3. We also compare the M-term performance to a minimal node size restriction as in (Biau and Scornet 2016), setting this value to 5, as in (Denil et. al. 2014). We perform RF construction with 1000 trees and 5 fold cross validation.

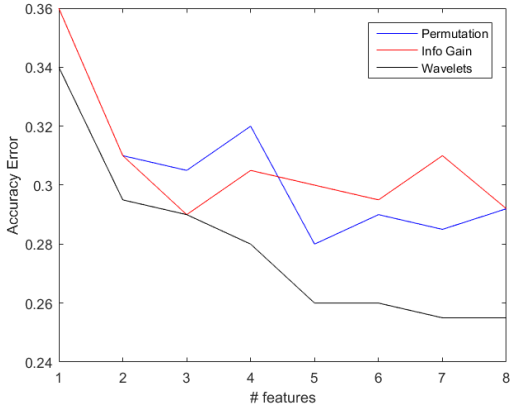
When the training dataset contains 'false labeling', the correspondence with the testing set is reduced. Trying to restrict the tree depth, can potentially miss the geometry of the underlying function, while too many levels can lead to overfitting. As seen in Table 2, the wavelet approach selects the right significant components from any tree and any level and thus outperforms the standard RF method. Observe that the added value of the wavelet



(a) "Parkinson" dataset, $\epsilon = 1.74$.

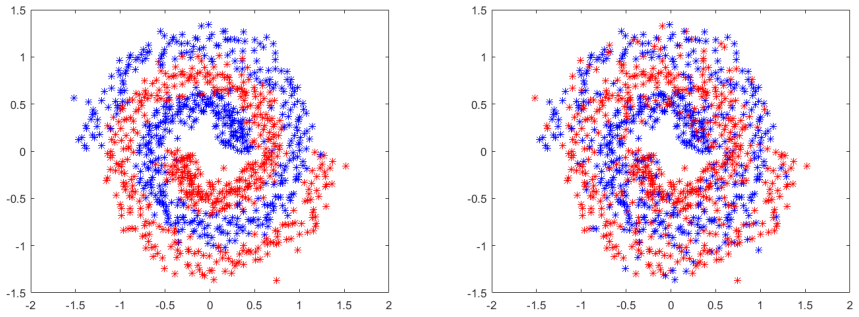


(b) "Magic gamma" dataset, $\epsilon = 0.57$.



(c) "Pima Diabetes" dataset,$\epsilon = 0.93$.

Figure 11: Comparisons of performance of standard VI methods used in R, with the wavelet-based method (22)



(a) Original set (b) Set with amplified mis-labeling

Figure 12: 'Spirals' dataset (Spiral dataset)

approach is more significant in the second case with more 'false labeling' in the training set.

Table 2: 'Spirals' dataset - Classification results.

	Wavelet error	RF error	Pruned RF error
Original spiral set	$12.2 \pm 0.9\%$	$14.4 \pm 1.1\%$	$15.9 \pm 0.8\%$
Set with amplified mis-labeling	$13.9 \pm 1.2\%$	$17.8 \pm 1.3\%$	$22.7 \pm 1.6\%$

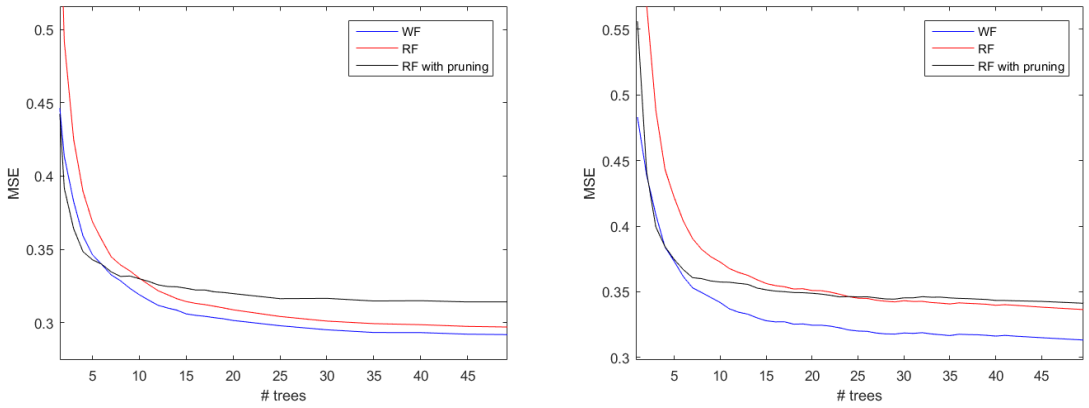
Next, we compare the performance of wavelet-based regression with state-of-the-art method on a challenging problem. The authors of (Denil et. al. 2014) provide comparative results of different pruning strategies for the difficult "Wine Quality" dataset. Learning this dataset is challenging since the data is very biased and depends on the personal taste of the wine experts. In Table 3 below, we collect the results of (Biau 2012), (Biau et. al. 2008), (Breiman 2001) and (Denil et. al. 2014) (as listed in (Denil et. al. 2014)). The RFs are all constructed of 1000 trees and fivefold cross validation is applied. We follow the notation presented in (Denil et. al. 2014) and use the abbreviation that was provided for each method variation ('+','F','S','NB', 'T'). In our RF implementation, we used bootstrapping with 80% and randomized$\sqrt{n}$ features at each node. M was selected automatically using 10 percent of the training set.

Another form of a challenging dataset is when some of the features are extremely noisy or uncorrelated with the response variable. As shown in (Strobl et. al. 2006) (see the discussion in Section 5), in such cases, RF partitions sometimes are influenced by these variables and the constructed ensemble is of lower quality. To explore the impact of our approach on such datasets, we used the "Poker Hand" dataset from the UCI repository (UCI repository) in two modes: with and without a very non-descriptive feature "instance

Table 3: Performance comparison on the "Wine Quality"

Algorithm	MSE
Biau08	0.53
Biau12	0.59
Biau12+T	0.57
Biau12+S	0.57
Denil	0.48
Denil+F	0.48
Denil+S	0.41
Breiman	0.4
Breiman+NB	0.39
Wavelets	**0.36**

id". As can be seen from Figure 13, the wavelet method significantly outperforms the standard RF regression, especially in the second scenario with the 'bad' feature included.



(a) 'Bad' feature excluded (b) 'Bad' feature included

Figure 13: The impact of a bad feature on the regression of the "Poker Hand" dataset

Acknowledgments

The authors would like to thank the reviewers for their careful reading of several versions of this work and valuable comments which resulted in a substantially revised manuscript. This work was supported by an 'AWS in Education Grant award'.

References

Alani D., Averbuch A. and Dekel S., Image coding using geometric wavelets, *IEEE transactions on image processing* 16:69-77, 2007.

Alpaydin E., *Introduction to machine learning*, MIT Press, 2004.

Avery M., Literature Review for Local Polynomial Regression, http://www4.ncsu.edu/ mravery/AveryReview2.pdf.

Bernard S., Adam S. and Heutte L., Dynamic random forests, *Pattern Recognition Letters* 33:1580-1586.

Biau G., Analysis of a random forests model, *Journal of Machine Learning Research* 13: 1063-1095, 2012.

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

Biau G. and Scornet E., A random forest guided tour, *TEST* 25(2):197-227, 2016.

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

Breiman L., Bagging predictors, *Machine Learning* 24(2):123-140, 1996.

Breiman L, Friedman J., Stone C. and Olshen R., *Classification and Regression Trees*, Chapman and Hall/CRC, 1984.

Boulesteix A., Janitza S., Kruppa J. and König I., Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2(6):493-507, 2012.

Chen H., Tino P. and Yao X., Predictive Ensemble Pruning by Expectation Propagation, *IEEE journal of knowledge and data engineering* 21:999-1013, 2009.

Christensen O., *An introduction to Frames and Riesz Bases*, Birkäuser, 2002.

Criminisi A., Shotton J. and Konukoglu E., Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning, *Microsoft Research technical report*, report TR-2011-114, 2011.

Dahmen W., Dekel S. and Petrushev P., Two-level-split decomposition of anisotropic Besov spaces, *Constructive approximation* 31:149-194, 2001.

Daubechies I., *Ten lectures on wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics,1992.

Dekel S., Gershtansky I., Active Geometric Wavelets, In *Proceedings of Approximation Theory XIII 2010*, 95-109, 2012.

Dekel S. and Leviatan D., Adaptive multivariate approximation using binary space partitions and geometric wavelets, *SIAM Journal on Numerical Analysis* 43:707-732, 2005.

Denil M., Matheson D. and De Freitas N., Narrowing the gap Random forests in theory and in practice, In *Proceedings of the 31st International Conference on Machine Learning* 32, 2014.

DeVore R., Nonlinear approximation, *Acta Numerica* 7:51-150, 1998.

DeVore R. and Lorentz G., *Constructive approximation*, Springer Science and Business, 1993.

DeVore R., Jawerth B. and Lucier B., Image compression through wavelet transform coding, *IEEE transactions on information theory* 38(2):719-746, 1992.

Du W. and Zhan Z., Building decision tree classifier on private data, In *Proceedings of the IEEE international conference on Privacy, security and data mining* 14:1-8, 2002.

Elad M., *Sparse and redundant representations: from theory to applications in signal and image processing*, Springer Science and Business Media, 2010.

Feng N., Wang J. and Saligrama V., Feature-Budgeted Random Forest, In Proceedings of The 32nd International Conference on Machine Learning, 1983-1991, 2015.

Kelley P. and Barry R., Sparse spatial autoregressions, *Statistics and Probability Letters* 33(3):291-297, 1997.

Gavish M., Nadler B., Coifman R., Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning, In *Proceedings of the 27th International Conference on Machine Learning*, 367-374, 2010.

Genuer R., Poggi J. and Christine T., Variable selection using Random Forests, *Pattern Recognition Letters* 31(14): 2225-2236, 2010.

Geurts P. and Gilles L., Learning to rank with extremely randomized trees, In *JMLR: Workshop and Conference Proceedings* 14:49-61, 2011.

Guyon I. and Elisseff A., An introduction to variable and feature selection, *Journal of Machine Learning Research* 3:1157-1182, 2003.

Hastie T., Tibshirani R. and Friedman J., *The elements of statistical learning*, Springer, 2009.

Joly A., Schnitzler F.,Geurts P. and Wehenkel L., L1-based compression of random forest models, In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 375-380, 2012.

Karaivanov B. and Petrushev P., Nonlinear piecewise polynomial approximation beyond Besov spaces, *Applied and computational harmonic analysis* 15:177-223, 2003.

Kulkarni V. and Sinha P., Pruning of Random Forest classifiers: A survey and future directions, In *International Conference on data science and engineering*, 64-68, 2012.

Lee A., Nadler B. and Wasserman L., Treelets: an adaptive multi-scale basis for sparse unordered data, *Annals of Applied Statistics* 2(2):435-471, 2008.

Loh W., Classification and regression trees, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1(1):14-23, 2011.

Louppe G., Wehenkel L., Sutera A. and Geurts P., Understanding variable importances in forests of randomized trees, *Advances in Neural Information Processing Systems* 26:431-439, 2013.

Mallat S., *A Wavelet tour of signal processing, 3rd edition (the sparse way)*, Acadmic Press, 2009.

Martinez-Muoz G., Hernández-Lobato D. and Suarez A., An analysis of ensemble pruning techniques based on ordered aggregation, *IEEE Transactions on pattern analysis and machine intelligence* 31:245-259, 2009.

Radha H., Vetterli M. and Leonardi R., Image compression using binary space partitioning trees, *IEEE transactions on image processing* 5:1610-1624, 1996.

Raileanu L. and Stoffel K., Theoretical comparison between the Gini index and information gain criteria, *Annals of Mathematics and Artificial Intelligence* 41(1):77-93, 2004.

'Random Forest' package in R,
http://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Rokach L. and Maimon O., Top-down induction of decision trees classifiers-a survey, *IEEE transactions on systems, man, and cybernetics, part C: applications and reviews*35(4):476-487, 2005.

Salembier P. and Garrido L., Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval, *IEEE transactions on image processing*9:561-576, 2000.

Spiral dataset, http://www.inside-r.org/packages/cran/mlbench/docs/mlbench.spirals.

Strobl C., Boulesteix A., Zeileis A. and Hothorn T., Bias in random forest variable importance measures, In *Workshop on Statistical Modelling of Complex Systems*, 2006.

Strobl C., Boulesteix A., Kneib T., Augustin T. and Zeileis A., Conditional variable importance for random forests, *BMC bioinformatics* 9(1):1-11, 2008.

UCI machine learning repository, http://archive.ics.uci.edu/ml/.

Vens C. and Costa F., Random forest based feature induction, In *IEEE international conference on data mining*, 744-753, 2011.

Yang F.,Lu. W., Luo L. and Li T., Margin optimization based pruning for random forest, *Neurocomputing* 94:54-63, 2012.

Wavelet-based Random Forest source code, https://github.com/orenelis/WaveletsForest.git.

Appendix

Proof of Lemma 1

Denoting briefly for any domain $\tilde{\Omega}$, $K_{\tilde{\Omega}} := \#\left\{x_i \in \tilde{\Omega}\right\}$ we have

$$\begin{aligned}
\sum_{x_i\in\Omega}\left(f\left(x_i\right)-C_\Omega\right)^2 - V_\Omega &= \sum_{x_i\in\Omega}\left(f\left(x_i\right)-C_\Omega\right)^2 - \sum_{x_i\in\Omega'}\left(f\left(x_i\right)-C_{\Omega'}\right)^2 - \sum_{x_i\in\Omega''}\left(f\left(x_i\right)-C_{\Omega''}\right)^2 \\
&= \sum_{x_i\in\Omega'}\left[\left(f\left(x_i\right)-C_\Omega\right)^2 - \left(f\left(x_i\right)-C_{\Omega'}\right)^2\right] + \\
&\quad \sum_{x_i\in\Omega''}\left[\left(f\left(x_i\right)-C_\Omega\right)^2 - \left(f\left(x_i\right)-C_{\Omega''}\right)^2\right] \\
&= 2\left(C_{\Omega'}-C_\Omega\right)\sum_{x_i\in\Omega'} f\left(x_i\right) + K_{\Omega'}\left(C_\Omega^2 - C_{\Omega'}^2\right) \\
&\quad + 2\left(C_{\Omega''}-C_\Omega\right)\sum_{x_i\in\Omega''} f\left(x_i\right) + K_{\Omega''}\left(C_\Omega^2 - C_{\Omega''}^2\right) \\
&= 2\left(C_{\Omega'}-C_\Omega\right)K_{\Omega'}C_{\Omega'} + K_{\Omega'}\left(C_\Omega^2 - C_{\Omega'}^2\right) \\
&\quad + 2\left(C_{\Omega''}-C_\Omega\right)K_{\Omega''}C_{\Omega''} + K_{\Omega''}\left(C_\Omega^2 - C_{\Omega''}^2\right) \\
&= K_{\Omega'}\left(C_{\Omega'}-C_\Omega\right)^2 + K_{\Omega''}\left(C_{\Omega''}-C_\Omega\right)^2 \\
&= \|\psi_{\Omega'}\|_2^2 + \|\psi_{\Omega''}\|_2^2 = W_\Omega.
\end{aligned}$$

Now, since $\sum_{x_i\in\Omega}\left(f\left(x_i\right)-C_\Omega\right)^2$ is independent of the selection of the partition of Ω and since W_Ω is always positive, the search for minimizing V_Ω is equivalent to maximizing W_Ω.

$\diamondsuit$

Proof of Example 1

1. For any attribute $j \neq k$ we denote $m_1 := \#\left\{x_i \in \Omega'\right\}$ and $m_2 := m - m_1$. Hence, for any $\delta \in (0,1)$, applying the Hoeffding bound gives w.p. $\geq 1-\delta$

$$\left|C_{\Omega'} - \frac{1}{2}\right| \leq \sqrt{\frac{\log(2/\delta)}{2m_1}}, \qquad \left|C_{\Omega''} - \frac{1}{2}\right| \leq \sqrt{\frac{\log(2/\delta)}{2m_2}}. \tag{25}$$

Note, that we can write $C_\Omega = \frac{m_1}{m}C_{\Omega'} + \frac{m_2}{m}C_{\Omega''}$. Thus, using (25) we get w.p. $\geq 1-\delta$

$$\begin{aligned}
\left(C_\Omega - C_{\Omega'}\right)^2 &= \frac{m_2^2}{m^2}\left(C_{\Omega''} - C_{\Omega'}\right)^2 \\
&\leq \frac{m_2^2}{m^2}\left(\frac{1}{m_1}+\frac{1}{m_2}\right)\log\left(2/\delta\right).
\end{aligned}$$

Therefore, w.p. $\geq 1-\delta$,

$$\begin{aligned}\|\psi_{\Omega'}\|_2^2 &= m_1 \left(C_\Omega - C_{\Omega'}\right)^2 \\ &\leq \frac{m_1 m_2^2}{m^2}\left(\frac{1}{m_1}+\frac{1}{m_2}\right)\log\left(2/\delta\right) \\ &= \left(\frac{m_2^2}{m^2}+\frac{m_1 m_2}{m^2}\right)\log\left(2/\delta\right) \\ &\leq 2\log\left(2/\delta\right).\end{aligned}$$

2. Observe that for the case $j=k$, a subdivision that minimizes (1) is $x_k = 1/2$. Denote $m_1 = \#\left\{x_i \in \Omega : \quad y_i = 1\right\}$. Applying the Hoeffding bound with $\delta \in (0,1)$ yields w.p. $\geq 1-\delta$

$$\left|m_1 - \frac{m}{2}\right| \leq \sqrt{\frac{\log(2/\delta)}{2m}}.$$

If Ω' is the subset of $[0,1]^n$ where $x_k > 1/2$, then $C_{\Omega'} = 1$ and $\|\psi_{\Omega'}\|_2^2 = m_1\left(1-\frac{m_1}{m}\right)^2$. Plugging into the bound above we conclude that w.p. $\geq 1-\delta$,

$$\begin{aligned}\|\psi_{\Omega'}\|_2^2 &\geq \left(\frac{m}{2}-\sqrt{\frac{\log(2/\delta)}{2m}}\right)\left(1-\frac{\frac{m}{2}+\sqrt{\frac{\log(2/\delta)}{2m}}}{m}\right)^2 \\ &= \left(\frac{m}{2}-\sqrt{\frac{\log(2/\delta)}{2m}}\right)^3 \Big/ m^2.\end{aligned}$$

$\diamondsuit$

Proof of Theorem 4 We prove the case $1 < p < \infty$ (the case $0 < p \leq 1$ is easier). We need to show two essential properties. First, for any $\Omega' \in \mathcal{F}$ and any $x \in \Omega'$, denoting $\Lambda := \{\Omega \in \mathcal{F} : x \in \Omega, |\Omega| \geq |\Omega'|\}$, we have

$$\sum_{\Omega\in\Lambda}\left(\frac{|\Omega'|}{|\Omega|}\right)^{1/p} \leq C\left(\rho,p\right)J. \tag{26}$$

Indeed, using (17), recursively for all domains on lower levels intersecting with Ω' we have

$$\begin{aligned}\sum_{\Omega\in\Lambda}\left(\frac{|\Omega'|}{|\Omega|}\right)^{1/p} &\leq \sum_{k=0}^{\infty} J\rho^{k/p} \\ &\leq \frac{J}{1-\rho^{1/p}}.\end{aligned}$$

Secondly, we need the property that

$$\|\psi_\Omega\|_\infty \leq c\,|\Omega|^{-1/p}\,\|\psi_\Omega\|_p \quad , \quad \forall \Omega \in \mathcal{F}. \tag{27}$$

It is easy to see property (27) for the case $r = 1$, where $\psi_\Omega = \mathbf{1}_\Omega C_\Omega$, but it is also known for the general case of $r \geq 1$ and convex domains (see e.g (Dekel and Leviatan 2005)). This allows us to prove the following Lemma

Lemma 7 *For $1 < p < \infty$, let $F(x) = \sum_{i=1}^{I} w_{j(\Omega_i)}\psi_{\Omega_i}(x)$, $\Omega_i \in \mathcal{F}$, where $\left\|w_{j(\Omega_i)}\psi_{\Omega_i}\right\|_p \leq L$. Then*

$$\|F\|_p \leq cJLI^{1/p}. \tag{28}$$

Proof Applying property (27) gives

$$\begin{aligned}\|F\|_p &\leq \left\|\sum_{i=1}^{I}\left\|w_{j(\Omega_i)}\psi_{\Omega_i}\right\|_\infty \mathbf{1}_{\Omega_i}(\cdot)\right\|_p \\ &\leq L\left\|\sum_{i=1}^{I}|\Omega_i|^{-1/p}\,\mathbf{1}_{\Omega_i}(\cdot)\right\|_p.\end{aligned}$$

We define

$$\Gamma(x) := \begin{cases} \min_{1\leq i\leq I}\{|\Omega_i| : x \in \Omega_i\}, & x \in \bigcup_{i=1}^{I}\Omega_i, \\ 0, & \text{else.}\end{cases}$$

Then, (26) yields

$$\sum_{i=1}^{I}|\Omega_i|^{-1/p}\,\mathbf{1}_{\Omega_i}(x) \leq cJ\Gamma(x)^{-1/p}, \quad \forall x \in \Omega_0.$$

Thus,

$$\begin{aligned}\|F\|_p &\leq cL\left\|\Gamma(\cdot)^{-1/p}\right\|_p \\ &= cJL\left(\int_{\bigcup\Omega_i}\Gamma(x)^{-1}\,dx\right)^{1/p} \\ &\leq cJL\left(\sum_{i=1}^{I}|\Omega_i|^{-1}\int_{\Omega_i}dx\right)^{1/p} = cJLI^{1/p}.\end{aligned}$$

$\diamondsuit$

We now proceed with the proof of the Theorem. Observe that we may use (16), that is, $|f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{F})} \sim N_\tau(f,\mathcal{F})$. For $\nu = 1, 2, \cdots$, denote

$$\Xi_\nu := \left\{\Omega \in \mathcal{F} : 2^{-\nu}\mathcal{N}_\tau(f,\mathcal{F}) \le w_{j(\Omega)} \|\psi_\Omega\|_p < 2^{-\nu+1}\mathcal{N}_\tau(f,\mathcal{F})\right\}.$$

Recall that for any non-negative discrete sequence $\beta = \{\beta_k\}_{k=1}^\infty$, the weak-$l_\tau$ norm $\|\beta\|_{wl_\tau}$, is defined as the infimum (if exists) over all $A > 0$, for which

$$\#\{\beta_k : \beta_k > \varepsilon\}\,\varepsilon^\tau \le A^\tau, \quad \forall \varepsilon > 0.$$

Since $\|\beta\|_{wl_\tau} \le \|\beta\|_{l_\tau}$, this implies that

$$\#\Xi_m \le \sum_{\nu \le m} \#\Xi_\nu = \#\bigcup_{\nu \le m} \Xi_\nu \le 2^{m\tau}.$$

Let $F_\nu(x) := \sum_{\Omega \in \Xi_\nu} w_{j(\Omega)}\psi_\Omega(x)$. For the special case $M := \sum_{\nu \le m} \#\Xi_\nu$, we have by (28)

$$\begin{aligned}
\|f - f_M\|_p &\le \left\| \sum_{\nu=m+1}^\infty F_\nu \right\|_p \\
&\le \sum_{\nu=m+1}^\infty \|F_\nu\|_p \\
&\le cJ \sum_{\nu=m+1}^\infty 2^{-\nu}\mathcal{N}_\tau(f,\mathcal{F})\,(\#\Xi_\nu)^{1/p} \\
&\le cJ\mathcal{N}_\tau(f,\mathcal{F}) \sum_{\nu=m+1}^\infty 2^{-\nu(1-\tau/p)} \\
&\le cJ\mathcal{N}_\tau(f,\mathcal{F})\, M^{-(1/\tau-1/p)} = cJ\mathcal{N}_\tau(f,\mathcal{F})\, M^{-\alpha}.
\end{aligned}$$

Extending this result for any $M \ge 1$ is standard (using a larger leading constant). This completes the proof.

$\diamondsuit$

Proof of Lemma 3 Since there are a finite number of boxes, there exists a, possibly unbalanced, binary tree that after at most $K2^n$ partitions, has also the boxes $\{B_k\}$ as nodes of the tree. Since the modulus of smoothness of order r of polynomials of degree $r-1$ is zero (DeVore 1998), (Devore and Lorentz 1993), for any of these box nodes we have that

$$\omega_r(f, B_k)_\tau = \omega_r(P_k, B_k)_\tau = 0.$$

Similarly, for any descendant node $\Omega' \subset B_k$, for some $1 \leq k \leq K$,

$$\omega_r \left(f, \Omega'\right)_\tau = \omega_r \left(P_k, \Omega'\right)_\tau = 0.$$

For any node Ω such that $\Omega \cap B_k = \emptyset$, $1 \leq k \leq K$, we have

$$\omega_r \left(f, \Omega\right)_\tau = \omega_r \left(0, \Omega\right)_\tau = 0.$$

We may then conclude that $\omega_r \left(f, \Omega\right)_\tau \neq 0$, for only a finite low-level subset Λ of the tree nodes, each strictly containing at least one B_k. Therefore, for any $\alpha > 0$,

$$\begin{aligned}
|f|_{B_\tau^{\alpha,r}} &= \left(\sum_{\Omega \in \mathcal{T}} \left(|\Omega|^{-\alpha} \omega_r \left(f, \Omega\right) \right)^\tau \right)^{1/\tau} \\
&= \left(\sum_{\Omega \in \Lambda} \left(|\Omega|^{-\alpha} \omega_r \left(f, \Omega\right)_\tau \right)^\tau \right)^{1/\tau} \\
&\leq 2^r \|f\|_\tau \left(\min_k |B_k| \right)^{-\alpha} \left(K 2^n\right)^{1/\tau},
\end{aligned}$$

where we have used the inequality

$$\omega_r \left(f, \Omega\right)_\tau \leq 2^r \|f\|_{L_\tau(\Omega)} \leq 2^r \|f\|_{L_\tau(\Omega_0)}.$$

$\diamondsuit$

Proof of Lemma 5 As stated, the tree $\mathcal{T}_I$ with isotropic dyadic partitions, creates dyadic cubes of side lengths 2^{-k} at the level nk. Let us denote by $D := \{D_k\}_{k=0}^\infty$, the collection of dyadic cubes of $[0,1]^n$, where D_k is the collection of cubes with side lengths 2^{-k}. Observe that any domain $\Omega' \in \mathcal{T}_I$, at a level $nk < l < n\left(k+1\right)$, is contained in some dyadic cube $\Omega \in \mathcal{T} \cap D_k$ at the level nk. Also, from the properties of the modulus of smoothness, $\Omega' \subset \Omega \Rightarrow \omega_r \left(f, \Omega'\right)_\tau \leq \omega_r \left(f, \Omega\right)_\tau$. Combining these two observations gives

$$\left|\Omega'\right|^{-\alpha} \omega_r \left(f, \Omega'\right)_\tau \leq 2^{n\alpha} |\Omega|^{-\alpha} \omega_r \left(f, \Omega\right)_\tau.$$

Next, observe that for any $\Omega \in D_k$

$$\omega_r \left(f, \Omega\right)_\tau \begin{cases} = 0, & \Omega \cap \partial\tilde{\Omega} = \emptyset, \\ \leq 2^{-kn/\tau}, & \Omega \cap \partial\tilde{\Omega} \neq \emptyset, \end{cases}$$

where $\partial\tilde{\Omega}$ is the boundary of $\tilde{\Omega}$. Therefore,

$$|f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{T})} \leq c\left(n,\alpha,\tau\right)\left(\sum_{\Omega\in D}\left(|\Omega|^{-\alpha}\,\omega_r\left(f,\Omega\right)_\tau\right)^\tau\right)^{1/\tau}$$
$$\leq c\left(n,\alpha,\tau,r\right)\left(\sum_{k=0}^{\infty}2^{kn(\alpha\tau-1)}\#\left\{\Omega\in D_k:\Omega\cap\partial\tilde{\Omega}\neq\emptyset\right\}\right)^{1/\tau}.$$

Thus, it remains to estimate the maximal number of dyadic cubes of side length 2^{-k} that can intersect a smooth boundary of a domain $\tilde{\Omega}\subset[0,1]^n$. For sufficiently large k, only one connected component of the boundary $\partial\tilde{\Omega}$ intersects a dyadic cube $\Omega\in D_k$, in similar manner to an hyperplane of dimension $n-1$ with surface area $\leq c2^{-k(n-1)}$. Therefore, for sufficiently large k

$$\#\left\{\Omega\in D_k:\Omega\cap\partial\tilde{\Omega}\neq\emptyset\right\}\leq c2^{k(n-1)}.$$

This gives

$$|f|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{T})}\leq c\left(n,\alpha,\tau,r,\partial\tilde{\Omega}\right)\left(\sum_{k=0}^{\infty}2^{kn(\alpha\tau-1)}2^{k(n-1)}\right)^{1/\tau}.$$

Therefore, if $\tau^{-1}=\alpha+1/p$, then

$$\alpha<\frac{1}{p\left(n-1\right)}\Rightarrow|f|_{\mathcal{B}_\tau^{\alpha,1}(\mathcal{T})}<\infty.$$

$$\diamondsuit$$