Is interpolation benign for random forests?

Ludovic Arnould, Claire Boyer, Erwan Scornet

Is interpolation benign for regression random forests?

Ludovic Arnould[1], Claire Boyer[1,2], and Erwan Scornet[3]

[1]LPSM, Sorbonne Université, Paris, France
[2]MOKAPLAN, INRIA Paris
[3]CMAP, Ecole Polytechnique, Paris, France

Abstract

Statistical wisdom suggests that very complex models, interpolating training data, will be poor at predicting unseen examples. Yet, this aphorism has been recently challenged by the identification of benign overfitting regimes, specially studied in the case of parametric models: generalization capabilities may be preserved despite model high complexity. While it is widely known that fully-grown decision trees interpolate and, in turn, have bad predictive performances, the same behavior is yet to be analyzed for random forests. In this paper, we study the trade-off between interpolation and consistency for several types of random forest algorithms. Theoretically, we prove that interpolation regimes and consistency cannot be achieved for non-adaptive random forests. Since adaptivity seems to be the cornerstone to bring together interpolation and consistency, we study interpolating Median RF which are proved to be consistent in a noiseless scenario. Numerical experiments show that Breiman's random forests are consistent while exactly interpolating, when no bootstrap step is involved. We theoretically control the size of the interpolation area, which converges fast enough to zero, so that exact interpolation and consistency occur in conjunction.

1 Introduction

Random Forests (RF) [8] have proven to be very efficient algorithms, especially on tabular data sets. As any machine learning (ML) algorithm, Random Forests and Decision Trees have been analyzed and used according to the overfitting-underfitting trade-off. Regularization parameters have been introduced in order to control the variance while still reducing the bias. For instance, one can increase the variety of the constructed trees (by playing either with bootstrap samples or feature subsampling) or control the tree structure (by limiting either the number of points falling within each leaf or the maximum depth of all trees).

However, the paradigm stating that high model complexity leads to bad generalization capacity has been recently challenged: in particular, deeper and larger neural networks still empirically exhibit high predictive performances [15]. In such situations, overfitting can be qualified as "benign":

complex models, possibly leading to interpolation of the training examples, still generalize well on unseen data [4].

Regarding parametric methods, benign overfitting has been exhibited and well understood in linear regression [3, 25, 18]. Many researchers currently study the *implicit bias* or *implicit regularization* of stochastic gradient (SGD) strategies used during neural network training: the optimization of an over-parametrized one-hidden-layer neural network via SGD will converge to a minimum of minimal norm with good generalization properties in a regression setting [2], or with maximal margin in a classification setting [11].

Regarding non-parametric methods, practitioners have noticed the good performances of high-depth RFs for a long time (by default, several ML libraries such as the popular Scikit-Learn grow trees until pure leaves are reached). More recently, the use of interpolating (or very deep) trees for boosting and bagging methods has been advocated in [27]. Indeed, Wyner et al. [27] believe that the *self-averaging* process at hand in RF (or in boosting methods) also produces an implicit regularization that prevents the interpolating algorithm from overfitting. Note that the regularization properties of RF have also been studied in the light of their complexity [10] and tree depth [28]. [27] even argue that interpolation actually provides robustness against noise: (i) the interpolating estimator would grasp the main signal thanks to its averaging ability; (ii) its high complexity would allow it to locally interpolate a noisy point without damaging the estimated function globally. This argument is to be put in parallel with the results proved in [12, 6] where they show that an interpolating kernel method using a singular kernel (similar to $K(x) = ||x||^{-\alpha}\mathbb{1}_{||x||\leq 1}$) is consistent, reaching minimax convergence rate for β-Hölder regular functions.

Contributions In this paper, we study the trade-off between interpolation and consistency for several types of random forest algorithms. Theoretically, we prove that interpolation regimes and consistency cannot be achieved simultaneously for non-adaptive centered random forests (Section 3). The major problems in combining interpolation and consistency arise from empty cells in tree partitions. Therefore, we study Centered RF which do not take into account empty cells (in Section 3) as well as Kernel Random Forests (KeRF) that are built by averaging over all connected data points (Section 4). By neglecting empty cells, these methods are consistent for larger tree depths, which unfortunately does not meet the strict interpolation requirement. Since adaptivity seems to be the cornerstone to bring together interpolation and consistency, we study the interpolating Median RF, which is proved to be consistent in a noiseless scenario (Section 5). Numerical experiments show that Breiman random forests are consistent and interpolate exactly, when the whole data set is used to build each tree (Section 6). If bootstrap is used instead, we numerically show that Breiman random forests are consistent but do not interpolate anymore: however each weak learner in the forest is inconsistent while being an interpolator. Finally, we prove that the volume of the interpolation zone for an infinite Breiman RF (without bootstrap) tends to 0 at a polynomial rate in the number of samples n and an exponential rate in the dimension d (Section 6). This supports the idea that the decay of the interpolation volume is fast enough to retrieve consistency despite interpolation. All proofs are given in Appendix A and all details of the experiments are given in Appendix B.

2 Setting

Framework In a general non-parametric regression framework, we assume to be given a *training set* $\mathcal{D}_n := ((X_1, Y_1), ..., (X_n, Y_n))$, composed of i.i.d. copies of the generic random variable (X, Y), where the input X is assumed *throughout the paper* to be uniformly distributed over $[0,1]^d$, and $Y \in \mathbb{R}$ is the output. The underlying model is assumed to satisfy $Y = f^\star(X) + \varepsilon$, where $f^\star(x) = \mathbb{E}[Y|X = x]$ is the regression function and ε is a random centered noise of variance $\sigma^2 < \infty$. Given an input vector $x \in [0,1]^d$, the goal is then to predict the associated square integrable random response by estimating $f^\star(x)$. We measure the performance of any estimator f_n via its quadratic risk, also referred to the *generalization* error, defined as $\mathcal{R}(f_n) := \mathbb{E}\left[(f_n(X) - f^\star(X))^2\right]$. The asymptotic performance of an estimator f_n is assessed via its *consistency*, a property stating that $\lim_{n\to\infty} \mathcal{R}(f_n) = 0$.

Estimator A Random Forest (RF) is a predictor consisting of a collection of M randomized trees [see 9, for details about decision trees]. To build a forest, we generate $M \in \mathbb{N}^\star$ independent random variables $(\Theta_1, \ldots, \Theta_M)$, distributed as a generic random variable Θ, independent of $\mathcal{D}_n$. In our setting, Θ_j actually represents the successive random splitting directions and the resampling data mechanism in the j-th tree. The predicted value at the query point x given by the j-th tree is defined as

$$f_n(x, \Theta_j) = \sum_{i=1}^{n} \frac{\mathbb{1}_{X_i \in A_n(x,\Theta_j)} Y_i}{N_n(x, \Theta_j)} \mathbb{1}_{N_n(x,\Theta_j)>0}$$

where $A_n(x, \Theta_j)$ is the cell containing x and $N_n(x, \Theta_j)$ is the number of points falling into $A_n(x, \Theta_j)$. The (finite) forest estimate then results from the aggregation of M trees:

$$f_{M,n}(x, \mathbf{\Theta}_M) = \frac{1}{M} \sum_{m=1}^{M} f_n(x, \Theta_m),$$

where $\mathbf{\Theta}_M := (\Theta_1, ..., \Theta_M)$. By making the number M of trees grows towards infinity, we can consider instead the *infinite* forest estimate, which has also played an important role in the theoretical understanding of forests:

$$f_{\infty,n}(x) = \mathbb{E}_\Theta[f_n(x, \Theta)],$$

where $\mathbb{E}_\Theta$ denotes the expectation w.r.t. Θ, conditional on $\mathcal{D}_n$. This operation is justified by the law of large numbers [see 23, for more details].

Several random forests have been proposed depending on the type of randomness they contain (what Θ represents) and the type of decision trees they aggregate. Breiman forest is one of the most widely used random forests, which exhibits excellent predictive performances. Unfortunately, its behavior is difficult to theoretically analyze, because of the numerous complex mechanisms involved in the predictive process (data resampling, data-dependent splits, split randomization). Therefore, in this paper, we simultaneously study the consistency and interpolation properties of different simplified versions of RF, both adaptive (i.e. when trees are built in a data-dependent manner) and non-adaptive.

All forests include a *depth* parameter, denoted k_n, which limits the maximum length of each branch in a tree, thus limiting the number of leaves (up to 2^{k_n}). In this work, we analyze how the tuning of k_n allows us to adjust the *consistency* and *interpolation* characteristics of the forest. The classical notion of (exact) interpolation is defined below.

Definition 2.1 ((Exact) interpolation)**.** An estimator f_n is said to *interpolate* if for all training data (X_i, Y_i), we have $f_n(X_i) = Y_i$ almost surely.

Recall that the prediction of a single tree at a point x is given by the average of all Y_i such that X_i is contained in the leaf of x. Therefore, each tree within a forest can be parameterized in order to interpolate: it is sufficient to grow the tree until pure leaves (i.e. leaves containing labels of the same values) are reached. In any regression model with continuous random noise, we have $Y_i \neq Y_j$ for all $i \neq j$ almost surely. Therefore, an interpolating tree is a tree that contains at most one point per leaf.

As the final prediction of the random forest is made by averaging the predictions of all its trees, if all trees interpolate, the random forest interpolates as well. Consequently, throughout all the theoretical analysis, we consider RF built without sub-sampling: each tree is built using the whole dataset instead of bootstrap samples as in standard RF. We will discuss the empirical effect of bootstrap in Section 6.

Remark 2.2. In a classification setting, if we handle the problem of estimating the probabilities of being in a class, interpolation occurs as soon as there is no diversity within each leaf, i.e. each leaf is pure containing points from a single class. Indeed, consider a degenerated setting such as X uniformly distributed on $[0,1]^d$ and $Y \sim \mathcal{B}(p), p \in [1/2, 1)$ independent from X, then as soon as there are $t > 1$ points in a cell, the probability of not interpolating is greater than $1 - ((1-p)^t + p^t) > 0$. In such a setting, interpolation almost surely occurs when there is at most one data point per cell, similarly to the regression setting. For a more detailed analysis of the classification setting, see [20].

We start our analysis of interpolation and consistency of RF with the simple yet widely studied Centered Random Forest (CRF).

3 Centered RF

Centered Random Forests [7] are ensemble methods that are said to be non-adaptive since trees are built independently of the data: at each step of a centered tree construction, a feature is uniformly chosen among all possible d features and the split along the chosen feature is made at the center of the current cell. Then, the trees are aggregated to produce a CRF.

3.1 Interpolation in CRF

For CRF, forest interpolation is equivalent to tree interpolation, as shown below.

Lemma 3.1. *The CRF $f_{M,n}^{\mathrm{CRF}}$ interpolates if and only if all trees that form the CRF interpolate.*

Since CRF construction is non-adaptive, it is impossible to enforce exactly one observation per leaf. Hence trees do not interpolate and in turn, the interpolation regime (Definition 2.1) cannot be satisfied for CRF. This leads us to examine a weaker notion of interpolation in probability.

Proposition 3.2 (Probability of interpolation for a centered tree)**.** *Denote $\mathcal{I}_T$ the event "a centered tree of depth k_n interpolates the training data". Then, for all $n \geq 3$, fixing $k_n = \lfloor \log_2(\alpha_n n) \rfloor$, with $\alpha_n > 1$, one has*

$$e^{-\frac{n}{\alpha_n - 1}} \leq \mathbb{P}(\mathcal{I}_T) \leq e^{-\frac{n}{2(\alpha_n+1)}}.$$

According to Proposition 3.2, the probability that a tree interpolates tends to one if and only if $k_n = \lfloor \log_2(\alpha_n n) \rfloor$ with $\alpha_n = \omega(n)$[1]. Consequently, the regime $\alpha_n = \omega(n)$ completely characterizes the interpolation of a centered tree. Proposition 3.2 can be in turn used to control the interpolation probability of a centered RF.

Corollary 3.3 (Probability of interpolation for a CRF)**.** *We denote by $\mathcal{I}_F$ the event "a centered forest $f_{M,n}^{\mathrm{CRF}}(., \mathbf{\Theta}_M)$ interpolates". Then, for $k_n = \lfloor \log_2(\alpha_n n) \rfloor$ with $\alpha_n \geq 1$,*

$$\mathbb{P}(\mathcal{I}_F) \leq e^{-\frac{n}{2(\alpha_n+1)}}. \tag{1}$$

According to Corollary 3.3, the condition $\alpha_n = \omega(n)$ (corresponding to the interpolation of a single centered tree with an overwhelming probability) is necessary to ensure that w.h.p., the forest interpolates. Our analysis stresses that a tree depth of at least $k_n = 2\log_2(n)$ is required to obtain tree/forest interpolation.

In fact, choosing k_n of the order of $\log_2(n)$ characterizes another type of interpolation regime. To see this, consider a centered tree of depth k, whose leaves are denoted $L_1, \ldots, L_{2^k}$. The number of points falling into the leaf L_i is denoted $N_n(L_i)$. If X is uniformly distributed over $[0,1]^d$, then by construction, for a given leaf L_i,

$$\mathbb{P}(X \in L_i) = \frac{1}{2^k} \quad \text{and} \quad \mathbb{E}[N_n(L_i)] = \frac{n}{2^k}. \tag{2}$$

Definition 3.4 (Mean interpolation regime)**.** A CRF $f_{M,n}^{\mathrm{CRF}}$ satisfies the *mean interpolation regime* when each tree of $f_{M,n}$ has at least n leaves.

The mean interpolation regime is met for CRF if and only if $k_n \geq \log_2 n$. By Equation (2), this implies that for all leaves L_i, $\mathbb{E}[N_n(L_i)] \leq 1$, that is, each leaf contains at most one point in expectation. Therefore, one could say that trees interpolate in expectation in the mean interpolation regime.

[1] i.e. α_n asymptotically dominates n.

3.2 Inconsistency of the standard CRF

In both interpolation regimes (mean and in probability), trees need to be very deep, with a growing number of empty cells as n tends to infinity, eventually damaging the consistency of the overall CRF.

Proposition 3.5. *Suppose that* $\mathbb{E}\left[f^\star(X)^2\right] > 0$. *Then the infinite Centered Random Forest* $f_{\infty,n}^{\mathrm{CRF}}$ *of depth* $k_n \geq \lfloor \log_2 n \rfloor$ *is inconsistent.*

The non-consistency of the CRF stems from the fact that the probability for a random point X to fall in an empty cell does not converge to zero, introducing an irreducible bias in the excess risk.

Proposition 3.5 emphasizes the poor generalization capacities of the interpolating CRF (under any interpolating regime), which could be expected given its non-adaptive construction.

3.3 Consistency of void-free CRF under the mean interpolation regime

Since limiting the impact of empty cells seems crucial for consistency, we study a CRF that averages over non-empty cells only, which we call the *Void-Free CRF*. Note that if all cells are empty, the prediction remains arbitrary set to 0. Denoting $\Lambda_n(x, \boldsymbol{\Theta}_M)$ the number of non-empty leaves containing x in the forest with trees $\Theta_1, \ldots, \Theta_M$, the void-free CRF is written as

$$f_{M,n}^{\mathrm{VF}}(x, \boldsymbol{\Theta}_M) = \frac{1}{\Lambda_n(x, \boldsymbol{\Theta}_M)} \sum_{j=1}^{M} f_n(x, \Theta_j) \mathbb{1}_{N_n(x,\Theta_j)>0}.$$

We can also introduce an infinite version of the void-free CRF by letting M grow to infinity:

$$f_{\infty,n}^{\mathrm{VF}}(x, \Theta) = \mathbb{E}_\Theta \left[f_n(x, \Theta) | N_n(x, \Theta) > 0 \right]$$

Lemma 3.6. *Consider a finite void-free CRF* $f_{M,n}^{\mathrm{VF}}$ *of depth* $k \in \mathbb{N}$ *and* $x \in [0,1]^d$. *We denote* $\mathcal{E}_{M,n}(x)$ *the event "x falls into an empty cell in all trees of* $f_{M,n}^{\mathrm{VF}}$*". Then,*

$$\mathbb{P}\left(\mathcal{E}_{M,n}\right) \leq e^{-\frac{kn}{2^{k+1}}} + e^{-Md^{-k}}. \tag{3}$$

Consequently, if $k = \log_2(n)$ *and* $M = \omega(n^{\log_2 d})$, *then,*

$$\lim_{n\to\infty} \mathbb{P}\left(\mathcal{E}_{\infty,n}\right) = 0. \tag{4}$$

Theorem 3.7. *Grant that* $f^\star$ *is bounded and has bounded partial derivatives. Then, the infinite void-free-CRF of depth* $k = \log_2 n$ *is consistent in a noiseless setting* ($\sigma = 0$), *and*

$$\mathbb{E}\left[\left(f_{\infty,n}^{\mathrm{VF}}(X) - f^\star(X)\right)^2\right] \leq 2d \sum_{j=1}^{d} ||\partial f_j^\star||_\infty^2 n^{\log_2\left(1-\frac{1}{2d}\right)} + 2n^{-\frac{1}{\log(2)}}.$$

The overall rate is of order $O\left(n^{\log(1-1/2d)}\right)$ which is a typical approximation rate for CRF which is also found in Klusowski [17]. As a matter of fact, Theorem 3.7 highlights that empty cells do not limit the performance of the void-free-CRF.

Indeed, the problematic terms that arise in the theoretical derivations of classical CRF vs. void-free CRF are of different natures: the probability $\mathbb{P}\left(N_n(X,\Theta)=0\right)$ of falling into an empty leaf in a random tree of an (infinite) CRF compared to the probability $\mathbb{P}_{X,\mathcal{D}_n}\left[\forall\Theta, N_n(X,\Theta)=0\right]$ of falling into empty leaves in all trees in the (infinite) CRF.

Neglecting empty cells allows a CRF to become consistent in the mean interpolation regime. However, this construction introduces a conditioning over $N_n(x,\Theta)>0$ that prevented us from efficiently bounding the variance in the case of noisy samples. Therefore in the next section we analyse Centered Kernel RFs (KeRF) for which the aggregation rule is slightly different (the empty cells still being neglected).

4 Centered kernel RF

As formalized in [14] and developed in [1], slightly modifying the aggregation rule of tree estimates provides a kernel-type estimator. Instead of averaging the predictions of all centered trees, the construction of a Kernel RF (KeRF) is performed by growing all centered trees and then averaging along all points contained in the leaves in which x falls, i.e.

$$f_{M,n}^{\mathrm{KeRF}}(x,\mathbf{\Theta}_M) := \frac{\sum_{i=1}^n Y_i \sum_{m=1}^M \mathbb{1}_{X_i\in A_n(x,\Theta_m)}}{\sum_{i=1}^n \sum_{m=1}^M \mathbb{1}_{X_i\in A_n(x,\Theta_m)}}.$$

One of the benefits of this construction is to limit the influence of empty cells, which can be harmful both for consistency and interpolation (see Section 3). Letting $K_{M,n}$ be the connection function of the M finite forest defined by

$$K_{M,n}(x,\mathbf{z}) := \frac{1}{M}\sum_{j=1}^M \mathbb{1}_{\mathbf{z}\in A_n(x,\Theta_j)},$$

[24] shows that the KeRF can be rewritten as

$$f_{M,n}^{\mathrm{KeRF}}(x,\mathbf{\Theta}_M) = \frac{\sum_{i=1}^n Y_i K_{M,n}(x,\mathbf{X}_i)}{\sum_{i=1}^n K_{M,n}(x,\mathbf{X}_i)},$$

hence the name of kernel RF. In addition, it is shown that

$$\lim_{M\to\infty} K_{M,n}(x,z) := K_n(x,z),$$

where $K_n(x,z)=\mathbb{P}_\Theta\left[z\in A_n(x,\Theta)\right]$ which can be seen as the empirical probability that x and z are in the same cell w.r.t. a tree built according to Θ. Consequently, for all $x\in[0,1]^d$, the infinite

KeRF reads as

$$f_{\infty,n}^{\mathrm{KeRF}}(x) = \frac{\sum_{i=1}^{n} Y_i K_n(x, X_i)}{\sum_{i=1}^{n} K_n(x, X_i)}.$$

4.1 Interpolation Conditions

Since KeRF aggregates centered trees as CRF (but in a different way), the results of Section 3 can be extended to KeRF:

1. the mean interpolation regime is met for centered trees, and therefore for KeRF, as soon as $k_n \geq \log_2 n$;
2. a necessary condition to attain the KeRF interpolation in probability is $k_n > 2\log_2(n)$.

One can note that the depths required for both interpolation regimes are still large, leading to as many empty cells for KeRF as for classical CRF but the aggregation rule is such that they are not taken into account in KeRF predictions, which gives hope that consistency could be preserved.

4.2 Consistency

In this section, we study the convergence of the centered KeRF when k_n is of the order of $\log_2(n)$, i.e. under the *mean interpolation regime*. To this end, we consider extra hypotheses on the noise and on the regularity of $f^\star$.

Theorem 4.1. *Assume that $f^\star$ is Lipschitz continuous and that the additive noise ε is a centered Gaussian variable with finite variance σ^2. Then, the risk of the infinite centered KeRF of depth $k_n = \lfloor \log_2(n) \rfloor$ verifies, for all $n \geq 2$,*

$$\mathcal{R}(f_{\infty,n}^{\mathrm{KeRF}}) \leq C_d \log(n)^{-(d-11)/6},$$

with $C_d > 0$ a constant depending on $\sigma, d, \|f^\star\|_\infty$.

Theorem 4.1 states that the infinite centered KeRF estimator is consistent as soon as $d > 11$, with a slow convergence rate of $\log(n)^{-(d-11)/6}$. The proof is based on the general paradigm of bias-variance trade-off and is adapted from [24]. At first sight, one might think that the rate becomes better as the dimension d increases. This would be without thinking that the constant in front of it depends on the dimension, so that the established bound should be regarded for any fixed d.

Choosing $k_n = \lfloor \log_2(n) \rfloor$ in Theorem 4.1 allows us to have a mean interpolation regime concomitant with consistency for KeRF, therefore highlighting that consistency and mean interpolation are compatible. This is not the case for CRF for which the mean interpolation regime forbids convergence

(Proposition 3.5). If a "mean" overfitting regime is benign for the consistency of KeRF, it seems to be nonetheless malignant for the convergence rate. Indeed, Lin and Jeon [19] provides a lower bound on the optimal convergence rate of a non-adaptive RF (such as the CRF), scaling in $(\log n)^{-(d-1)}$. This leads us to believe that the convergence rate we obtain in Theorem 4.1 is marginally improvable.

Interpolation of kernel estimators has been recently studied with singular kernel by [6]. Since KeRF are kernel estimators, one can wonder how sharp is our bound (Theorem 4.1) compared to that of [6], which is minimax. Due to the spikiness of the singular kernel studied in [6], interpolation arises for any kernel bandwidth. The latter can be then tuned to reach minimax rates of consistency. The story is totally different for KeRF since interpolation occurs only for specific tree depths $k_n \geq \log(n)$ (where the depth parameter is closely related to the bandwidth of classical kernel estimates). Less latitude for choosing the depth then leads to sub-optimal rates of consistency (see Theorem 4.1). Of course, a better rate of consistency in $O(n^{1/(3+d\log 2)})$ could be obtained as in [24] when optimizing this depth parameter, but leaving the interpolation world.

4.3 Empirical results

We numerically assess the performance of KeRF in the mean interpolation regime.

Experimental framework We consider four different regression models, most of which have been already considered in [26]: Model 1 is additive without noise ($d = 2$), Model 2 is polynomial with interactions ($d = 8$), Model 3 is the sum of elementary terms that contain non-polynomial interactions ($d = 6$) and Model 4 ($d = 5$) corresponds to a generalized linear model. All models are specified in Appendix B. For each model, the simulated dataset is divided into a training set (80% of the data set) and a test set (the remaining 20%). We train a centered KeRF (with $M = 500$) of depth fixed to $\lfloor \log_2 n \rfloor + 1$ (mean interpolation regime) for different sample sizes n and evaluate the empirical quadratic risk on the test set.

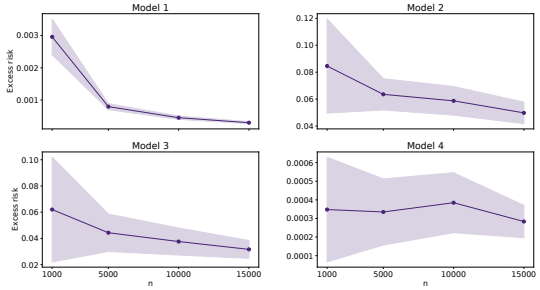

Figure 1: KeRF consistency results: excess risk w.r.t. sample sizes. For each sample size n, the experiment is repeated 30 times: we represent the mean over the 30 tries (bold line) and the mean $\pm$ std (filled zone).

Results On Figure 1, for all models, the risk decreases toward zero as the number of samples n increases (with slow convergence rates). These numerical results, even though obtained for a finite KeRF with a large number $M = 500$ of centered trees, support the theoretical consistency of the infinite KeRF in the mean interpolation regime (see Theorem 4.1).

5 Semi-Adaptive RF

So far, consistency has been analyzed in the mean interpolation regime. What about consistency with exact RF interpolation? To this end, we introduce semi-adaptive RF whose constructions depend on the training inputs X_i's (and not on the outputs Y_i's).

5.1 Semi-adaptive CRF

We first introduce a (semi-)*adaptive centered tree* which is a modified version of a centered tree, built by taking into account the positions of the X_i's, and thereby reduces the number of empty leaves. It is recursively grown: at each node, a feature is uniformly chosen among the set of all *separable* d features (a feature is separable if cutting this feature produces two non-empty cells) and the split is made in the middle of the current node along the chosen feature. If there are more than one point in the current node and none of the feature separates them, the splitting direction is uniformly chosen among all the separable features of the previous cut. The construction stops when all leaves contain 0 or 1 observation. The semi-*Adaptive Centered RF* (AdaCRF) results from a specific aggregation of such trees: for a given point x, the final prediction of the RF is given by averaging along all the trees for which x falls into a non-empty leaf.

5.1.1 Interpolation and depth

By construction, AdaCRF interpolates since all trees interpolate. AdaCRF adaptivity allows it to reach full growth while preserving a reasonable depth in probability. To show this, we grow an adaptive centered tree and measure, for a given point x, the depth $k_n(x)$ associated to the cell containing x.

Lemma 5.1. *For all $\alpha \in [0, 1)$, the depth k_n^{AdaCT} of a semi-adaptive centered tree verifies*

$$\lim_{n\to\infty} \mathbb{P}\left(k_n^{\mathrm{AdaCT}}(X) \in [\log(n) \pm \log^{1-\alpha}(n)]\right) = 1.$$

Lemma 5.1 states that the asymptotic behavior of $k_n^{\mathrm{AdaCT}}(X)$ is equivalent to $\log n$ up to a negligible factor. The $\log(n)$ equivalent matches the condition for the mean interpolation regime in the case of CRF exhibited in Section 3. Therefore, while AdaCRF has a depth of the same order as that of a classical CRF, its adaptivity nature ensures its interpolation. As seen hereafter, this adaptivity is not sufficient enough to preserve consistency while interpolating.

5.1.2 Inconsistency of AdaCRF

In this section we show that AdaCRF is not adaptive enough in order to preserve both consistency and interpolation.

Lemma 5.2. *Consider the infinite AdaCRF $f_{\infty,n}^{\mathrm{AdaCRF}}$. Denote $\mathcal{E}$ the event "X falls into an empty leaf of $f_{\infty,n}^{\mathrm{AdaCRF}}$". Let $c \in \mathbb{N}$ be such that $n \geq 2^c$, then for all d,*

$$\mathbb{P}\left(\mathcal{E}\right) \gtrsim (1-4^{-c})^d e^{-\frac{dn}{n-1}} \left(1 - \frac{n}{n-1} e^{-1} - e^{-1}\right)^d.$$

Lemma 5.2 states that too many points fall into an empty cell of the infinite AdaCRF: the above lower bound does not converge to 0 as n grows to infinity. This entails the inconsistency of AdaCRF, as the risk is lower bounded by $\mathbb{E}\left[f^\star(X)^2\right]\mathbb{P}\left(\mathcal{E}\right)$.

Proposition 5.3. *If $\mathbb{E}[f^\star(X)^2] > 0$, the infinite AdaCRF $f_{\infty,n}^{\mathrm{AdaCRF}}$ is inconsistent in an exact interpolation regime.*

AdaCRF is not consistent due to the number of empty cells that is not negligible enough and introduces a bias in the generalization error. To maintain both the interpolation and consistency properties, empty cells should be avoided by construction: one has to choose the split between two points and not anymore independently of the position of the X_i' s. Therefore, the considered RF should be more adaptive to the data, as in Median RF described below.

5.2 Median RF

The Median RF, studied e.g. in [13], is composed of median trees which first randomly choose the direction to cut over and then cut at the median of the data points contained in the current cells. In order to avoid points falling on a cell boundary, whenever the number of points n_c in the cell is odd, the cut is made at the quantile $(n_c+1)/2n_c$.

5.2.1 Consistency of Median RF in the interpolating regime

An interpolating infinite Median RF turns out to be consistent, at least in a noiseless scenario.

Theorem 5.4. *Under a noiseless setting ($\sigma = 0$), suppose that $f^\star$ is bounded and has bounded partial derivatives. Then, the infinite interpolating Median RF $f_{\infty,n}^{\mathrm{MedRF}}$ is consistent and verifies:*

$$\mathcal{R}\left(f_{\infty,n}^{\mathrm{MedRF}}\right) \leq 4d \left(1 - \frac{57}{64d}\right)^{\log_2(n+1)-2} \sum_{j=1}^{d} ||\partial_j f^\star||_\infty^2.$$

The proof is adapted from [13]. Theorem 5.4 is the first result to establish the consistency of an interpolating (adaptive) forest, theoretically supporting the self-regularization of RF allowing consistency and (exact) interpolation together. Note that the consistency achieved by Median RF cannot be obtained for CRF (adaptive or not) under the exact interpolation regime, even in the noiseless setting. Indeed, this is due to the non-negligible probability of falling into empty cells (see Propositions 3.2 and 5.3 resp. for CRF and AdaCRF), which cannot be coerced under any interpolation regimes.

We believe that Theorem 5.4 also stands in the more general setting of noisy data so that the limitation is mainly due to a technical obstruction within the proof. In details, when dealing with interpolating trees, the variance reduction does not come from averaging many points in the leaf of a given tree anymore, but rather from averaging single points from the leaves of many different trees. Hence, it requires to understand the geometry of the tree partitions and their intersections, which is an arduous task for generic RFs even those with data-independent cuts. This intuition is corroborated by a theoretical control of the interpolation volume and by numerical experiments (resp. Section 5.2.2 and 5.2.3).

5.2.2 Volume of the interpolation area

To go one step further in the analysis of Median RF, we introduce the notion of *interpolation area* defined below.

Definition 5.5. The ***interpolation area*** is the subspace of $[0,1]^d$ where the forest prediction depends only on one training point. For a given forest $f_{M,n}(.,\mathbf{\Theta}_M)$, the interpolation area is denoted by[2]

$$\mathcal{A}(f_{M,n}(.,\mathbf{\Theta}_M)) = \left\{ x \in [0,1]^d, \exists! X_i \in \mathcal{D}_n, X_i \in \bigcap_{m=1}^{M} A_n(x,\Theta_m) \right\}.$$

The interpolation zone is highly dependent on both the geometry of the training points X_i's and the construction of the trees. Analyzing the interpolation area for a finite Median RF turns out to be quite a challenging task. Therefore, we focus our study on the *core interpolation area* $\mathcal{A}_{min}$ written as

$$\mathcal{A}_{min} = \bigcap_{M \in \mathbb{N}, \mathbf{\Theta}_M} \mathcal{A}(f_{M,n}(.,\mathbf{\Theta}_M)).$$

The area $\mathcal{A}_{min}$ is nothing but the intersection of the interpolation zones of all possible forests, or equivalently of a forest containing all possible trees (and therefore all possible cuts). As an example note that in the case of centered trees, every cut may occur with a positive probability. Therefore, $\mathcal{A}_{min}$ matches the volume of the interpolation area of an infinite Median RF. In the following proposition, we control the volume $\mu(\mathcal{A}_{min})$ of the minimal interpolation zone of a Median RF, with μ the Lebesgue measure.

[2] the symbol $\exists!$ means "there exists a unique"

Proposition 5.6. *For all $n \geq 2$, for all $d \geq 2$, consider an infinite Median RF $f_{\infty,n}^{\mathrm{MedRF}}$. Then,*

$$\mathbb{E}_{\mathcal{D}_n}\left[\mu(\mathcal{A}_{min}^{\mathrm{MedRF}})\right] \leq 2\left(\frac{2}{n}\right)^{d-1}.$$

The volume of the interpolation area for an infinite Median RF tends to 0 polynomially in n and exponentially in d, and so does $\mathbb{E}_{\mathcal{D}_n}\left[\mathbb{P}\left(X \in \mathcal{A}_{min}\right)\right]$.

Remark 5.7. Apart from a very restricted zone, the prediction of a Median RF mostly relies on more than one training point. More specifically, this is a *necessary* condition for consistency: indeed, to reach consistency, the volume of the area where the prediction involves a finite number of points (*a fortiori* the interpolation zone) should tend to 0. In fact, the variance is of the order of σ^2 in such an area. Proposition 5.6 portends the predominant *self-averaging* property of adaptive RF, and hence underpins the idea of good capabilities of Median RF in interpolation regimes (with or without noise).

5.2.3 Empirical consistency results

We analyze the empirical performances of Median RF in noiseless and noisy settings on the same models as considered in Section 4.3. For each model, given a training set, we train Median RF (with $M = 500$ trees) until pure leaves are reached, and measure its excess risk on a test set.

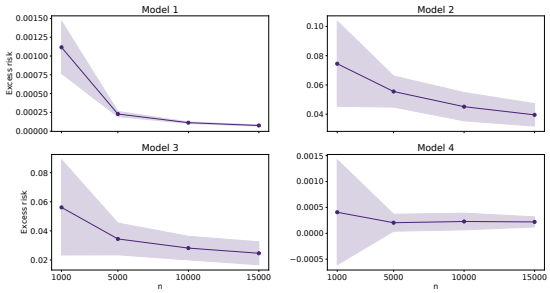

Figure 2: Consistency results for a Median RF with $M = 500$ trees: excess risk w.r.t. the sample size n. For each sample size, the experiment is repeated 30 times: we represent the mean over the 30 tries (bold line) and the mean $\pm$ std (filled zone).

Figure 2 shows that the excess risk of a Median RF decreases as n grows. These empirical performances lend support to the idea that Median RF are consistent even with a finite number of trees and beyond the noiseless setting.

6 Breiman RF

The widely-used Breiman RF is composed of several trees, built with CART methodology, each one trained on bootstrap samples, and for which the successive splitting directions and thresholds are chosen at each step (among a random subset of directions) in order to minimize the CART criterion (empirical variance for instance). Breiman forests are among the state-of-the-art ensemble methods in terms of predictive performance even if their adaptivity to the data remains a real hurdle to their theoretical analysis.

From the interpolation perspective, each CART being trained on a bootstrap sample, the RF interpolation is not ensured when considering fully-grown trees. Indeed, a tree cannot interpolate a point that is not chosen in the bootstrap step. For this reason, we focus our study on the volume of interpolation areas for Breiman RF without bootstrap and then analyze their empirical behavior in interpolating regimes through a battery of numerical experiments.

6.1 Interpolation

As a Breiman RF is built using both the X_i's and the Y_i's, it is difficult to determine the depth necessary to reach the interpolation state. Depending on the data, the latter can be of the order $k \approx \log_2(n)$ in the best case, if each cut creates approximately two groups of the same size), or $k \approx n$ in the worst case, if only one point is separated from the others at each step [low signal-to-noise ratios situations, see e.g., 16]. Note that by omitting the bootstrap step in the RF construction, the interpolation of Breiman forests directly results from aggregating fully-grown trees.

6.2 Volume of the interpolation zone

As shown in the next proposition, the volume of the minimal interpolation zone tends to 0 as n tends to infinity.

Proposition 6.1. *Consider an infinite Breiman forest constructed without bootstrap. Suppose that for a given configuration of the training data, all cuts have a probability strictly greater than 0 to appear. Then, the volume of the minimal interpolation zone verifies*

$$\mathbb{E}\left[\mu(\mathcal{A}_{min})\right] \leq \frac{1}{n^{d-1}} \left(1 - 2^{-n}\right)^d.$$

Similarly to the Median RF, the bound on the interpolation volume for a Breiman forest enjoys the same order of decay, improved by a constant exponential in the dimension. Since predictions cannot be accurate in the interpolation area in a noisy setting, it is necessary that the volume of this area decreases to zero in order to ensure the RF consistency (see Remark 5.7). Proposition 6.1 therefore suggests the good generalization properties of Breiman RF in interpolation regimes, as several training points are mostly used for prediction.

Setting the number of eligible features for splitting to 1 is sufficient to ensure the hypothesis on cuts in Proposition 6.1: one can obtain a tree in which all splits are performed along a single direction. This is a minor modification to the original algorithm and an easy one to implement since most ML libraries have a "max-feature" (as scikit-learn in Python) or "mtry" (in R) parameter that can be set to 1.

6.3 Empirical study

Interpolation volume We numerically evaluate the volume of the interpolation area of a Breiman RF (with 5000 trees, see Figure 11 in Appendix B.2 for details about this choice) when the sample size n increases.

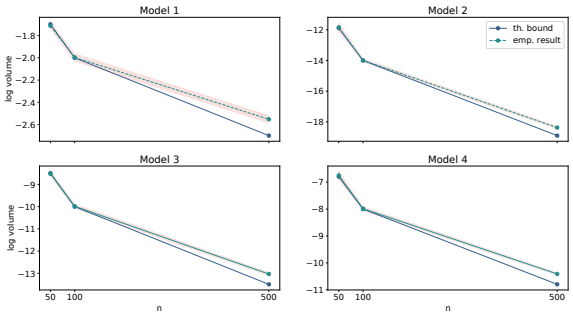

Figure 3: Log volume of the interpolation zone of a Breiman RF with 5000 Trees, max features set to 1, no bootstrap. Mean over 10 tries (red line) and mean ± std (filled zone). The theoretical bound (Proposition 6.1) is represented in green.

In Figure 3, the volume of the minimal interpolation zone is shown to tend polynomially fast to 0 (linear in the logarithmic scale) for all considered models as the dataset size increases, matching the behavior of the theoretical bound established in Proposition 6.1.

One could notice the slight gap between the theoretical and experimental curves, which actually reflects the gap between an infinite forest (for which Proposition 6.1 holds) and its approximation by a finite forest (5000 trees here). This gap naturally tends to increase with n (when the number of trees is fixed) as the approximation of the infinite RF by a finite one deteriorates with n.

Consistency We now present an empirical study of Breiman RF consistency in interpolation regimes. In the theoretical analysis, we have focused on a specific type of Breiman RF (without bootstrap and a *max-features* parameter equal to 1). We now examine the characteristics of Breiman forests with their default parameter and study the regularization processes that limit the noise sensitivity in the interpolation regime.

In order to reach a better estimation of the regression function, Breiman RF averages several CARTs while introducing randomness in the construction of each tree to diversify them. The first

randomization comes from the *bootstrap*: each tree is trained on a bootstrap sample (selecting n observations out of the n original ones, with replacement). The other randomization results from a random selection of splitting directions: at each node, a subset of $\{1, \dots, d\}$ of size *max-features* is randomly selected and the CART criterion is optimized along these directions only (setting *max-features* to 1 provides the maximum diversity whereas setting it to d results in the construction of a unique tree).

The benefit of these two aspects in the construction of the Breiman RF is numerically analyzed when using interpolating Breiman trees. In Figure 4, we measure the excess risk of two RFs with 500 trees and max-depth= `None`, where for the first one, bootstrap is used and the max-features parameter is set to 1, whereas the second one excludes bootstrap and sets the max-features parameter to $\lceil d/3 \rceil$ (default value in `randomForest` in `R`).

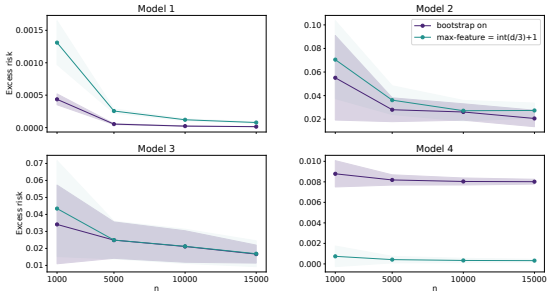

Figure 4: Consistency of two Breiman RF: excess risk w.r.t. sample size n. Parameters : 500 trees per forest, max-depth=None, max-features= d for the "bootstrap on" RF, bootstrap off for the "max-feature= $\lceil d/3 \rceil$" RF. Mean over 30 tries (bold lines) and mean $\pm$ std (filled zone).

In Figure 4, we observe that the excess risk decreases to 0 for all models and for both forests. Indeed, each randomizing process alone induces enough diversity across trees for the self-averaging property to be efficient, resulting in the consistency of the overall forests [see also 23, 20, 21, for insights about tree diversity in random forests].

However, when using bootstrap, consistency comes at the cost of leaving the interpolation regime, as only 2/3 of the data are used in average to build each tree (see Figures 12, 13 in Section B.2.3 for more details about the forest non-interpolation). In regards of this internal sampling selection, the aggregation of interpolating bagged trees results in smoothing the decision process of the entire forest, providing thereby a consistent but not interpolating estimate.

In turn, Breiman RF built with *max-features*= $\lceil d/3 \rceil$ seems consistent while preserving its interpolating behavior. Within this configuration, the final RF still interpolates the data but the volume of the interpolation zone is very small as shown in Figure 10. This is in line with the vision of a *locally spiky* estimator developed in [27] and [4]. Indeed, the influence of the averaging effect is locally null near the data training points, but increases with the distance from these points. Note that bootstrap and feature subsampling act differently. Bootstrap smoothens predictions by averaging different observations, even at points of the training set, which leads to an empty interpolation area. On the other hand, feature subsampling increases tree partition diversity, which reduces but does not annihilate the interpolation area of the overall forest.

In this regard, Breiman RF with *max-features*$= \lceil d/3 \rceil$ are similar to interpolating *spiky* non-singular kernel methods, as the ones introduced in [6], except for the leeway allowed for the hyperparameters tuning. Indeed, as underlined for non-adaptive centered forests, the depth k_n (i.e. the tuned parameter) is constrained to a strict range to ensure both consistency and interpolation. This is not the case for singular kernel methods, as they interpolate regardless of the window parameter value.

7 Conclusion

In this paper, we study both empirically and theoretically the tradeoff between interpolation and consistency of different types of random forests. In particular, we show that interpolation is harmless in the case of adaptive methods when the *self-averaging* process in the forest is sufficient to restrain the interpolation effect to a local influence.

Indeed, we prove that the Median RF reaches consistency and exact interpolation regimes in a noiseless scenario. This is the first result to prove that consistency and interpolation are not irreconcilable for such powerful learners. Breiman forests is also shown empirically to be consistent and interpolate when no bootstrap step is involved. This results from a fast decrease of the interpolation area, which limits the negative impact of interpolation on the overall consistency of the method.

We believe that the analysis of the interpolation zone of RF introduced in this article is a milestone for the understanding of RF prediction in interpolation regimes. Indeed the volume of the interpolation area is actually a roundabout way to measure the diversity in the constructed trees: if this volume is high, all trees end up building similar partitions. This diversity measure could also be used as a regularization tool to reduce the RF complexity by keeping only the most uncorrelated trees (in terms of partition) in a PCA fashion.

References

[1] Sylvain Arlot and Robin Genuer. Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*, 2014.

[2] Francis Bach and Lenaïc Chizat. Gradient descent on infinitely wide neural networks: Global convergence and generalization, 2021.

[3] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[4] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177*, 2021.

[5] Necdet Batir. Inequalities for the gamma function. *Archiv der Mathematik*, 91(6):554–563, 2008.

[6] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

[7] Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13:1063–1095, 2012.

[8] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[9] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[10] Sebastian Buschjäger and Katharina Morik. There is no double-descent in random forests. *arXiv preprint arXiv:2111.04409*, 2021.

[11] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

[12] Luc Devroye, Laszlo Györfi, and Adam Krzyżak. The hilbert kernel regression estimate. *Journal of Multivariate Analysis*, 65(2):209–227, 1998.

[13] Roxane Duroux and Erwan Scornet. Impact of subsampling and tree depth on random forests. *ESAIM: Probability and Statistics*, 22:96–128, 2018.

[14] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[16] Hemant Ishwaran. The effect of splitting on random forests. *Machine learning*, 99(1):75–118, 2015.

[17] Jason Klusowski. Sharp analysis of a simple model for random forests. In *International Conference on Artificial Intelligence and Statistics*, pages 757–765. PMLR, 2021.

[18] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR, 2020.

[19] Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

[20] Lucas Mentch and Siyu Zhou. Randomization as regularization: A degrees of freedom explanation for random forest success. *Journal of Machine Learning Research*, 21(171):1–36, 2020.

[21] Jaouad Mourtada, Stéphane Gaïffas, and Erwan Scornet. Minimax optimal rates for mondrian trees and forests. *The Annals of Statistics*, 48(4):2253–2276, 2020.

[22] Lawrence Bruce Richmond and Jeffrey Shallit. Counting abelian squares. *Electronic Journal of Combinatorics*, 2009.

[23] Erwan Scornet. On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146: 72–83, 2016.

[24] Erwan Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500, 2016.

[25] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

[26] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[27] Abraham J Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *The Journal of Machine Learning Research*, 18(1):1558–1590, 2017.

[28] Siyu Zhou and Lucas Mentch. Trees, forests, chickens, and eggs: when and why to prune trees in a random forest. *arXiv preprint arXiv:2103.16700*, 2021.

A Proofs

A.1 Proofs of Section 3

A.1.1 Proof of Proposition 3.2

As all the leaves have the same volume and the data points are independent and uniformly distributed, having at most one point per leaf is equivalent to distribute n balls into 2^k boxes containing at most one point with $2^k \geq n$ as can be seen on Figure 5.

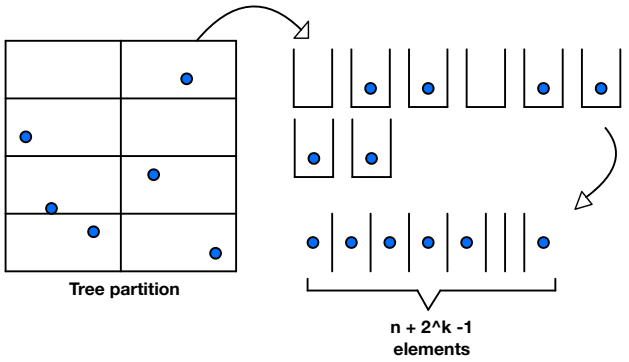

Figure 5: Computing the interpolation probability (depth $k = 3$, $n = 6$)

Thus

$$\begin{aligned}\mathbb{P}\left(\mathcal{I}_T\right) &= \frac{\binom{2^k}{n}}{\binom{n+2^k-1}{n}} \\ &= \frac{2^k!}{(2^k-n)!n!}\frac{n!(2^k-1)!}{(n+2^k-1)!}.\end{aligned}$$

With $k = \lfloor \log_2(\alpha_n n) \rfloor \in \mathbb{N}$, we have

$$\mathbb{P}\left(\mathcal{I}_T\right) = \frac{\alpha_n n}{(\alpha_n+1)n-1} \cdot \frac{\alpha_n n - 1}{(\alpha_n+1)n-2} \cdots \frac{(\alpha_n-1)n+1}{\alpha_n n}.$$

- We begin by computing the lower bound:

$$\begin{aligned}\mathbb{P}\left(\mathcal{I}_T\right) &\geq \frac{\alpha_n n}{(\alpha_n+1)n} \cdot \frac{\alpha_n n-1}{(\alpha_n+1)n-1} \cdots \frac{(\alpha_n-1)n+1}{\alpha_n n+1} \\ &\geq \left(\frac{\alpha_n-1}{\alpha_n}\right)^n \\ &\geq e^{n \log(1-\frac{1}{\alpha_n})} \\ &\geq e^{-\frac{n}{\alpha_n-1}} \xrightarrow[\alpha_n \to \infty]{} 1.\end{aligned}$$

- The computation of the upper bound is similar, note that for all $r \in \{0, ..., n\}$,

$$\frac{\alpha_n n-r}{(\alpha_n+1)n-r-1} \leq \frac{\alpha_n+1/2}{\alpha_n+1}.$$

  It follows that

$$\begin{aligned}\mathbb{P}\left(\mathcal{I}_T\right) &\leq \left(\frac{\alpha_n+1/2}{\alpha_n+1}\right) n \\ &\leq e^{n \log(\frac{\alpha_n+1/2}{\alpha_n+1})} \\ &\leq e^{-\frac{n}{2(\alpha_n+1)}}.\end{aligned}$$

A.1.2 Proof of Lemmma 3.1

Suppose that a tree f_r in the forest does not interpolate for a given point $X_s, s \in \{1, \ldots, n\}$, we write $f_r(X_s, \Theta_r) = Y_s + \xi$, $\xi \neq 0$. Then, by definition of $f_{M,n}^{\mathrm{CRF}}$,

$$\begin{aligned}f_{M,n}^{\mathrm{CRF}}(X_s, \boldsymbol{\Theta_M}) &= \frac{1}{M} \sum_{j=1}^{M} \sum_{i=1}^{n} Y_i W_{ij} \\ &= \frac{1}{M} \sum_{j=1, j \neq r}^{M} \sum_{i=1}^{n} (f^\star(X_i) + \varepsilon_i) W_{ij} + \frac{1}{M}(Y_s + \xi)\end{aligned}$$

where $W_{ij} := \frac{\mathbb{1}_{X_i \in A_n(X_s, \Theta_j)}}{N_n(X_s, \Theta_j)} \mathbb{1}_{N_n(X_s, \Theta_j) > 0}$. Therefore, $f_{M,n}^{\mathrm{CRF}}(X_s, \boldsymbol{\Theta_M}) = Y_s$ if and only if

$$\begin{aligned}&\frac{1}{M} \sum_{j=1, j \neq r}^{M} \sum_{i=1}^{n} (f^\star(X_i) + \varepsilon_i) W_{ij} + \frac{1}{M}(Y_s + \xi) = Y_s \\ \iff \quad &\frac{1}{M} \sum_{j=1, j \neq r}^{M} \sum_{i=1}^{n} (f^\star(X_i) + \varepsilon_i) W_{ij} = -\frac{\xi}{M} \\ \iff \quad &\frac{1}{M} \sum_{j=1, j \neq r}^{M} \sum_{i=1}^{n} \varepsilon_i W_{ij} = -\frac{\xi}{M} + C\end{aligned}$$

where C is random and independent from ε_i for all i. ξ was computed with the label noise of at least one point different from X_s (otherwise it would equal 0). We can write

$$\xi = C' + \frac{1}{M}\sum_{i=1}^{n} \varepsilon_i W_{ir}$$

where C' is independent from the ε_is . Finally, the forest interpolates at X_s if and only if

$$\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{n} \varepsilon_i W_{ij} = C + C'$$

However, as the noise is continuous and independent from W_{ij} for all i, j and from C, C' , this equality happens with a zero probability.

A.1.3 Proof of Corollary 3.3

As it is necessary for all trees to interpolation for the forest to interpolate, the probability that the forest interpolates is smaller than the probability that a single tree interpolates.

A.1.4 Proof of Proposition 3.5

Let $f_{\infty,n}^{\text{CRF}}$ be an infinite CRF with each tree containing at least $\alpha_n n$ leaves, with $\alpha_n > 1$. Let X be uniformly distributed on $[0,1]^d$. We write $\bar{f}_{n,\infty}^{\text{CRF}}(X) = \mathbb{E}\left[f_{\infty,n}^{\text{CRF}}(X)|X, X_1, ..., X_n\right]$. Then, denoting $\mathcal{E}$ the event "$N_{n,\infty}(X) = 0$" (or equivalently, "X falls into a non-empty leaf"),

$$\begin{aligned}
\mathcal{R}(f_{\infty,n}^{\text{CRF}}(X)) &= \mathbb{E}\left[\left(f_{\infty,n}^{\text{CRF}}(X) - f^\star(X)\right)^2)\right] && (5)\\
&\geq \mathbb{E}\left[\left(\bar{f}_{n,\infty}^{\text{CRF}}(X) - f^\star(X)\right)^2\right] && (6)\\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbb{E}_\Theta\left[W_i f^\star(X_i)\right] - \left(\mathbb{1}_{\mathcal{E}} + \mathbb{1}_{\mathcal{E}^c}\right) f^\star(X)\right)^2\right] && (7)\\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbb{E}_\Theta\left[W_i\left(f^\star(X_i) - f^\star(X)\right)\right] - \mathbb{1}_{\mathcal{E}^c} f^\star(X)\right)^2\right] && (8)\\
&\geq \mathbb{E}\left[f^\star(X)^2 \mathbb{1}_{\mathcal{E}^c}\right] && (9)\\
&\geq \mathbb{E}\left[f^\star(X)^2 \mathbb{P}\left(\mathcal{E}^c|X\right)\right]. && (10)
\end{aligned}$$

Besides,

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}^c|X\right) &= \mathbb{P}\left(N_{n,\infty}(X) = 0|X\right) && (11)\\
&= \left(1 - \frac{1}{\alpha_n n}\right)^n, && (12)
\end{aligned}$$

and as $\log(1-1/x) \geq -\frac{1}{x-1}$ for $x > 1$,

$$\left(1 - \frac{1}{\alpha_n n}\right)^n = e^{n \log\left(1-\frac{1}{\alpha_n n}\right)} \tag{13}$$

$$\geq e^{-\frac{n}{\alpha_n n - 1}}. \tag{14}$$

The above quantity does not tend to 0 when n tends to infinity. Therefore, if $\mathbb{E}\left[f^\star(X)^2\right] > 0$, the infinite CRF is inconsistent.

A.1.5 Proof of Lemma 3.6

To ease the notations, we write $\mathbb{P}\left(\mathcal{E}_M\right)$ instead of $\mathbb{P}\left(\mathcal{E}_{M,n}\right)$ throughout the proof. We have

$$\mathcal{E}_M(X) = \bigcap_{j=1}^{M} \left\{N_n(x, \Theta_j) = 0\right\}. \tag{15}$$

Given a dataset, we distinguish two situations: either x falls into an area where it cannot be connected to a point X_i for any tree, or the dataset is such that x could eventually be connected to a point X_i for a certain configuration of cuts within a tree. We write $\mathcal{E}_1(x)$ the ($\mathcal{D}_n$-measurable) event $\{\forall \Theta, N_n(x, \Theta) = 0\}$ and $\mathcal{E}_2(x) = \{\exists\, \Theta, N_n(x, \Theta) \neq 0\}$. Using these notations, we obtain

$$\mathbb{P}\left(\mathcal{E}_M(X)\right) = \mathbb{P}\left(\mathcal{E}_M(X) \cap \mathcal{E}_1(x)\right) + \mathbb{P}\left(\mathcal{E}_M(X) \cap \mathcal{E}_1(x)^c\right) \tag{16}$$

$$= \mathbb{P}\left(\mathcal{E}_1(x)\right) + \mathbb{P}\left(\mathcal{E}_M(X) \cap \mathcal{E}_2(x)\right) \tag{17}$$

where the first probability term of the second line is a probability taken over $\mathcal{D}_n$ only. We control it thanks to the following Lemma:

Lemma A.1. *We have*

$$\mathbb{P}\left(\mathcal{E}_1(x)\right) \leq e^{-\frac{n}{2^{k+1}}}.$$

Proof. The event $\mathcal{E}_1(x)$ happens if all points of the dataset fall into parts of the space that cannot connect to x for any tree. In order to compute its probability, we compute the size of the *connection area of* x for trees of depth k, denoted

$$Z_{c,k}(x) = \left\{z \in [0,1]^d \ : \exists \Theta, z \in A_n(x, \Theta)\right\}. \tag{18}$$

We recall that trees are built independently from the dataset and that all cuts are made in the middle of the current node for a uniformly chosen feature at each step. We denote $A(k_1, ..., k_d, x)$ the cell of x obtained by cutting k_j times along feature $X^{(j)}$ for all $j \in \{1, \ldots, d\}$. Then, the volume

of the connection area $Z_{k,c}$ of x is

$$\mu(Z_{c,k}(x)) = \mu\left(\bigcup_{\substack{0\leq k_1,\dots,k_d\leq k\\ \sum_j k_j = k}} A(k_1,...,k_d,x)\right) \tag{19}$$

$$\geq \mu\left(\bigcup_{\substack{0\leq k_1,k_2\leq k\\ k_1+k_2=k}} A(k_1,k_2,0,...,0,x)\right). \tag{20}$$

By σ-additivity of μ,

$$\mu\left(\bigcup_{\substack{0\leq k_1,k_2\leq k\\ k_1+k_2=k}} A(k_1,k_2,0,...,0,x)\right)$$

$$= \mu\Big(A(k,0,...,0,x)\Big) + \sum_{j=1}^{k} \mu\left(A(k-j,j,0,...,0,x) \setminus \bigcup_{\ell=0}^{j-1} A(k-\ell,\ell,0,...,0,x)\right). \tag{21}$$

Given the shape of the cells $A(k-j,j,0,...,0,x)$ for all $j \in \{1,\dots,d\}$, we have

$$A(k-j,j,0,...,0,x) \setminus \bigcup_{\ell=0}^{j-1} A(k-\ell,\ell,0,...,0,x)$$

$$= A(k-j,j,0,...,0,x) \setminus A(k-j+1,j-1,0,...,0,x). \tag{22}$$

Furthermore, note that, for all $j \in \{1,\dots,d\}$, the volume of each cell $A(k-j+1,j-1,0,...,0,x)$ is 2^{-k} (since k cuts have been performed). Therefore, for all $j \in \{1,\dots,k\}$,

1. $\mu(A(k-j,j,0,...,0,x)) = \mu(A(k-j+1,j-1,0,...,0,x)) = 2^{-k}$
2. $\mu\big((A(k-j,j,0,...,0,x) \cap A(k-j+1,j-1,0,...,0,x)\big) = \frac{\mu(A(k-j,j,0,...,0,x))}{2}$ as can be seen on Figure 6.

We deduce from these observations that, for all j,

$$\mu\big(A(k-j,j,0,...,0,x) \setminus A(k-j+1,j-1,0,...,0,x)\big) = \frac{\mu(A(k-j,j,0,...,0,x))}{2} \tag{23}$$

$$= 2^{-(k+1)} \tag{24}$$

Hence, combining equations (21), (22) and (23), we have

$$\mu\left(\bigcup_{\substack{0\leq k_1,k_2\leq k\\ k_1+k_2=k}} A(k_1,k_2,0,...,0,x)\right) = 2^{-k} + k2^{-(k+1)}. \tag{25}$$

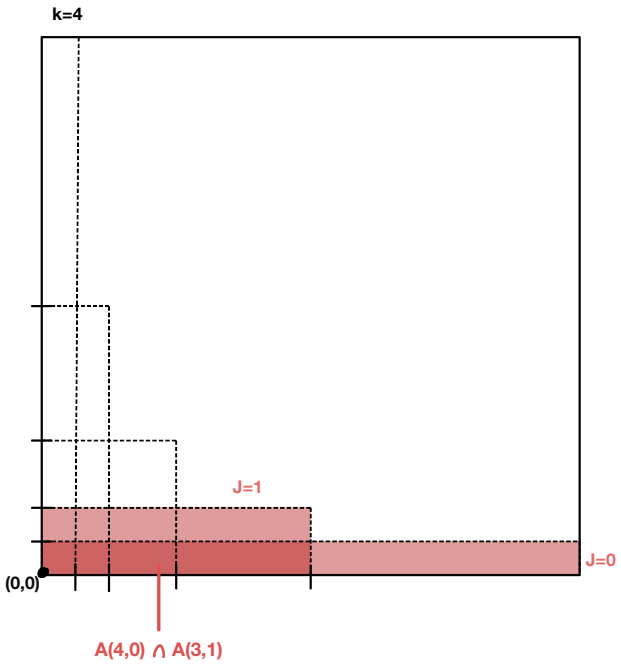

Figure 6: Volume of leaf intersection $\mu\big((A(k-j,j,x) \cap A(k-j+1,j-1,x)\big)$ in dimension 2 with $x = (0,0)$, $k = 4$ cuts and $j \in \{0,1\}$.

Consequently, using inequality (20),

$$\mu(Z_c(x)) \geq k2^{-(k+1)}. \tag{26}$$

Finally, as the X_i's are uniformly distributed on $[0,1]^d$ and $\mathcal{E}_1(x)$ is realized when none of the X_is fall into $Z_{c,k}(x)$,

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_1(x)\right) &= \mathbb{P}\left(\forall i \in \{1,\dots,n\}, X_i \notin Z_{c,k}(x)\right) && (27)\\
&= \left(1-\mu(Z_{c,k}(x))\right)^n && (28)\\
&\leq \left(1-k2^{-(k+1)}\right)^n && (29)\\
&= e^{n\log(1-k2^{-(k+1)})} && (30)\\
&\leq e^{-\frac{kn}{2^{k+1}}}. && (31)
\end{aligned}$$

□

Regarding the second term of (17), we have

$$\mathbb{P}\left(\mathcal{E}_M(x) \cap \mathcal{E}_2(x)\right) = \mathbb{P}\left(\left(\bigcap_{j=1}^{M} N_n(x, \Theta_j) = 0\right) \bigcap \left(\exists i \in \{1, \dots, n\}, X_i \in Z_{c,k}(x)\right)\right) \tag{32}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\exists i \in \{1,\dots,n\}, X_i \in Z_{c,k}(x)} \mathbb{1}_{\bigcap_{j=1}^{M} N_n(x,\Theta_j)=0} | \mathcal{D}_n\right]\right] \tag{33}$$

$$= \mathbb{E}\left[\mathbb{1}_{\exists i \in \{1,\dots,n\}, X_i \in Z_{c,k}(x)} \mathbb{P}\left(\bigcap_{j=1}^{M} N_n(x, \Theta_j) = 0 | \mathcal{D}_n\right)\right] \tag{34}$$

$$= \mathbb{E}\left[\mathbb{1}_{\exists i \in \{1,\dots,n\}, X_i \in Z_{c,k}(x)} (1 - p_n)^M\right] \tag{35}$$

where $p_n = \mathbb{P}_\Theta\left(N_n(x, \Theta) > 0 | \mathcal{D}_n\right)$ and where the last line is obtained by independence of the Θ_j's conditionally on $\mathcal{D}_n$. Note that

- either $\exists i \in \{1, \dots, n\}, X_i \in Z_{c,k}(x)$, then $p_n \geq d^{-k}$ since a tree connects x and a point in $Z_{c,k}(x)$ with probability at least d^{-k} (i.e. by choosing the right cut at each step);
- or $\nexists i \in \{1, \dots, n\}, X_i \in Z_{c,k}(x)$;

and consequently,

$$\mathbb{1}_{\exists i \in \{1,\dots,n\}, X_i \in Z_{c,k}(x)} (1 - p_n)^M \leq (1 - d^{-k})^M. \tag{36}$$

Overall,

$$\mathbb{P}\left(\mathcal{E}_M(x) \cap \mathcal{E}_2(x)\right) \leq \left(1 - d^{-k}\right)^M \tag{37}$$

$$\leq e^{-Md^{-k}}. \tag{38}$$

Finally, gathering Lemma A.1 and the last equation yields

$$\mathbb{P}\left(\mathcal{E}_M(x)\right) \leq e^{-\frac{kn}{2^{k+1}}} + e^{-Md^{-k}}. \tag{39}$$

A.1.6 Proof of Proposition 3.7

We follow the proof of [17]. As we are in a noiseless setting, the risk simply equals the bias.

$$\mathbb{E}\left[\left(f_{\infty,n}^{\mathrm{VF}}(X) - f^\star(X)\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{\mathbb{P}_\Theta\left(N_n(X, \Theta) > 0\right)} \sum_{i=1}^{n} f^\star(X_i) \mathbb{E}_\Theta[W_i \mathbb{1}_{N_n(X,\Theta)>0}] - f^\star(X)\right)^2\right].$$

We decompose $f^\star(X)$ as

$$f^\star(X) = \left(\mathbb{1}_{\mathbb{P}_\Theta(N_n(X,\Theta)>0)>0} + \mathbb{1}_{\mathbb{P}_\Theta(N_n(X,\Theta)>0)=0}\right) f^\star(X)$$

in order to write

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{\mathbb{P}_\Theta\left(N_n(X,\Theta>0)\right)}\sum_{i=1}^n f^\star(X_i)\mathbb{E}_\Theta[W_i\mathbb{1}_{N_n(X,\Theta)>0}] - f^\star(X)\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{\mathbb{P}_\Theta\left(N_n(X,\Theta>0)\right)}\sum_{i=1}^n \left(f^\star(X_i)-f^\star(X)\right)\mathbb{E}_\Theta[W_i\mathbb{1}_{N_n(X,\Theta)>0}] - f^\star(X)\mathbb{1}_{\mathbb{P}_\Theta(N_n(X,\Theta)>0)=0}\right)^2\right] \\
&\leq 2\mathbb{E}\left[\left(\frac{1}{\mathbb{P}_\Theta\left(N_n(X,\Theta>0)\right)}\sum_{i=1}^n \left(f^\star(X_i)-f^\star(X)\right)\mathbb{E}_\Theta[W_i\mathbb{1}_{N_n(X,\Theta)>0}]\right)^2\right] \\
&+ 2\mathbb{E}\left[\left(f^\star(X)\mathbb{1}_{\mathbb{P}_\Theta(N_n(X,\Theta)>0)=0}\right)^2\right] \qquad (40)
\end{aligned}$$

The second term of the last inequality verifies

$$\mathbb{E}\left[\left(f^\star(X)\mathbb{1}_{\mathbb{P}_\Theta(N_n(X,\Theta)>0)=0}\right)^2\right] \leq ||f^\star||_\infty^2 \mathbb{P}\left(\mathbb{P}_\Theta\left(N_n(X,\Theta)>0\right)=0\right). \qquad (41)$$

The above probability is taken over $\mathcal{D}_n$. The event $\{\mathbb{P}_\Theta\left(N_n(X,\Theta)>0\right)=0\}$ is $(X,\mathcal{D}_n)$-measurable, it corresponds to the situation where for any Θ, $N_n(X,\Theta)=0$, i.e. the dataset is such that it is impossible for a tree to connect X with one of the X_i's. This probability is controlled by Lemma A.1:

$$\mathbb{P}\left(\mathbb{P}_\Theta\left(N_n(X,\Theta)>0\right)=0\right) \leq e^{-\frac{kn}{2^{k+1}}}.$$

Following a computation from [17],

$$\begin{aligned}
|f^\star(X)-f^\star(X_i)| &\leq \sum_{j=1}^d ||\partial_j f^\star||_\infty |X_i^{(j)} - X^{(j)}| \\
&\leq \sum_{j=1}^d ||\partial_j f^\star||_\infty (b_j - a_j)
\end{aligned}$$

and

$$\begin{aligned}
\sum_{i=1}^n W_i|f^\star(X)-f^\star(X_i)|\mathbb{1}_{N_n(X,\Theta)>0} &\leq \sum_{i=1}^n W_i\mathbb{1}_{N_n(X,\Theta)>0}\sum_{j=1}^d ||\partial_j f^\star||_\infty(b_j-a_j) \qquad (42)\\
&\leq \mathbb{1}_{N_n(X,\Theta)>0}\sum_{j=1}^d ||\partial_j f^\star||_\infty(b_j-a_j). \qquad (43)
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\left(f^{\mathrm{VF}}_{\infty,n}(X)-f^\star(X)\right)^2\right] \leq 2\mathbb{E}\left[\left(\frac{1}{\mathbb{P}_\Theta(N_n(X,\Theta)>0)}\sum_{j=1}^d ||\partial_j f||_\infty \mathbb{E}_\Theta[\mathbb{1}_{N_n(X,\Theta)>0}(b_j-a_j)]\right)^2\right] + 2e^{-\frac{kn}{2^{k+1}}} \tag{44}$$

$$\leq 2d\sum_{j=1}^d ||\partial f_j^\star||_\infty^2 \mathbb{E}\left[\frac{1}{\mathbb{P}_\Theta(N_n(X,\Theta)>0)}\mathbb{E}_\Theta[\mathbb{1}_{N_n(X,\Theta)>0}(b_j-a_j)]^2\right] + 2e^{-\frac{kn}{2^{k+1}}} \tag{45}$$

$$\leq 2d\sum_{j=1}^d ||\partial f_j^\star||_\infty^2 \mathbb{E}\left[\mathbb{E}_\Theta[(b_j-a_j)]^2\right] + 2e^{-\frac{kn}{2^{k+1}}} \tag{46}$$

where the last inequality directly results from Cauchy-Schwarz inequality and the penultimate one from independence of $N_n(X,\Theta)$ from $b_j - a_j$ for all j.

Then, Jensen inequality yields

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_\Theta[(b_j-a_j)]^2\right] &\leq \mathbb{E}\left[b_j-a_j\right]^2 \\ &= \mathbb{E}\left[2^{-K_j(X)}\right]^2 \\ &= \mathbb{E}\left[\mathbb{E}\left[2^{-K_j(X)}|X\right]\right]^2\end{aligned}$$

where $K_j(X)$ is the number of splits made on feature j to produce the cell containing X. It is conditionally distributed as a binomial distribution of parameters $(k, 1/d)$. Hence,

$$\mathbb{E}\left[2^{-K_j(X)}|X\right] \leq \left(1-\frac{1}{2d}\right)^k.$$

To conclude,

$$\mathbb{E}\left[\left(f^{\mathrm{VF}}_{\infty,n}(X)-f^\star(X)\right)^2\right] \leq 2d\sum_{j=1}^d ||\partial f_j||_\infty^2 \left(1-\frac{1}{2d}\right)^k + 2e^{-\frac{kn}{2^{k+1}}}. \tag{47}$$

A.2 Proofs of Section 4 (Theorem 4.1)

In this section, we prove the consistency of the infinite KeRF estimator in the interpolating regime in expectancy (Theorem 4.1). We follow the proof given in [24] and first present two of its results.

Lemma A.2. *Let $k \in \mathbb{N}$ and consider an infinite centered random forest of depth k. Then, for all*

$x, \mathbf{z} \in [0,1]^d$,

$$K_n^{cc}(x,\mathbf{z}) = \sum_{\substack{k_1,\dots,k_d \\ \sum_{\ell=1}^d k_\ell = k}} \frac{k!}{k_1!\dots k_d!} \left(\frac{1}{d}\right)^k \prod_{j=1}^d \mathbb{1}_{\lceil 2^{k_j} x_j \rceil = \lceil 2^{k_j} z_j \rceil}.$$

Theorem A.3. *From [24]. Let f be a L-Lipschitz function. Then, for all k,*

$$\sup_{x\in[0,1]^d} \left| \frac{\int_{[0,1]^d} K_k^{cc}(x,\mathbf{z}) f^\star(\mathbf{z}) \mathrm{d}z_1 \dots \mathrm{d}z_d}{\int_{[0,1]^d} K_k^{cc}(x,\mathbf{z}) \mathrm{d}z_1 \dots \mathrm{d}z_d} - f^\star(x) \right| \leq Ld \left(1 - \frac{1}{2d}\right)^k.$$

Proof of Theorem 4.1. Let $x \in [0,1]^d$, $\|f^\star\|_\infty = \sup_{x\in[0,1]^d} |f^\star(x)|$ and recall that

$$f_{\infty,n}^{\mathrm{KeRF}}(x) = \frac{\sum_{i=1}^n Y_i K_k^{cc}(x,X_i)}{\sum_{i=1}^n K_k^{cc}(x,X_i)}.$$

Thus, letting

$$\begin{aligned}
A_n(x) &= \frac{1}{n}\sum_{i=1}^n \left(\frac{Y_i K_k^{cc}(x,X_i)}{\mathbb{E}\left[K_k^{cc}(x,X)\right]} - \frac{\mathbb{E}\left[Y K_k^{cc}(x,X)\right]}{\mathbb{E}\left[K_k^{cc}(x,X)\right]} \right), \\
B_n(x) &= \frac{1}{n}\sum_{i=1}^n \left(\frac{K_k^{cc}(x,X_i)}{\mathbb{E}\left[K_k^{cc}(x,X)\right]} - 1 \right), \\
\text{and } M_n(x) &= \frac{\mathbb{E}\left[Y K_k^{cc}(x,X)\right]}{\mathbb{E}\left[K_k^{cc}(x,X)\right]},
\end{aligned}$$

the estimate $f_{\infty,n}^{\mathrm{KeRF}}(x)$ can be rewritten as

$$f_{\infty,n}^{\mathrm{KeRF}}(x) = \frac{M_n(x) + A_n(x)}{1 + B_n(x)},$$

which leads to

$$f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x) = \frac{M_n(x) - f^\star(x) + A_n(x) - B_n(x) f^\star(x)}{1 + B_n(x)}.$$

According to Theorem A.3, we have

$$\begin{aligned}
|M_n(x) - f^\star(x)| &= \left| \frac{\mathbb{E}\left[f^\star(X) K_k^{cc}(x,X)\right]}{\mathbb{E}\left[K_k^{cc}(x,X)\right]} + \frac{\mathbb{E}\left[\varepsilon K_k^{cc}(x,X)\right]}{\mathbb{E}\left[K_k^{cc}(x,X)\right]} - f^\star(x) \right| \\
&\leq \left| \frac{\mathbb{E}\left[f^\star(X) K_k^{cc}(x,X)\right]}{\mathbb{E}\left[K_k^{cc}(x,X)\right]} - f^\star(x) \right| \\
&\leq C \left(1 - \frac{1}{2d}\right)^k,
\end{aligned}$$

where $C = Ld$. Take $\alpha \in]0, 1/2]$. Let $\mathcal{C}_\alpha(x)$ be the event on which $\big\{|A_n(x)|, |B_n(x)| \leq \alpha\big\}$. On the event $\mathcal{C}_\alpha(x)$, we have

$$\begin{aligned}|f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x)|^2 &\leq 8|M_n(x) - f^\star(x)|^2 + 8|A_n(x) - B_n(x)f^\star(x)|^2 \\ &\leq 8C^2\left(1 - \frac{1}{2d}\right)^{2k} + 8\alpha^2(1 + \|f^\star\|_\infty)^2.\end{aligned}$$

Thus,

$$\mathbb{E}[|f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x)|^2 \mathbb{1}_{\mathcal{C}_\alpha(x)}] \leq 8C^2\left(1 - \frac{1}{2d}\right)^{2k} + 8\alpha^2(1 + \|f^\star\|_\infty)^2. \tag{48}$$

Consequently, to find an upper bound on the rate of consistency of $f_{\infty,n}^{\mathrm{KeRF}}$, we just need to upper bound

$$\begin{aligned}\mathbb{E}\Big[|f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x)|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(x)}\Big] &\leq \mathbb{E}\Big[\Big|\max_{1\leq i\leq n} Y_i + f^\star(x)\Big|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(x)}\Big] \\ &\quad \text{(since } f_{\infty,n}^{\mathrm{KeRF}} \text{ is a local averaging estimate)} \\ &\leq \mathbb{E}\Big[\Big|2\|m\|_\infty + \max_{1\leq i\leq n} \varepsilon_i\Big|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(x)}\Big] \\ &\leq \left(\mathbb{E}\left[2\|m\|_\infty + \max_{1\leq i\leq n} \varepsilon_i\right]^4 \mathbb{P}\left[\mathcal{C}_\alpha^c(x)\right]\right)^{1/2} \\ &\quad \text{(by Cauchy-Schwarz inequality)} \\ &\leq \left(\left(16\|m\|_\infty^4 + 8\mathbb{E}\Big[\max_{1\leq i\leq n} \varepsilon_i\Big]^4\right) \mathbb{P}\left[\mathcal{C}_\alpha^c(x)\right]\right)^{1/2}.\end{aligned}$$

Simple calculations on Gaussian tails show that one can find a constant $C' > 0$ such that for all n,

$$\mathbb{E}\Big[\max_{1\leq i\leq n} \varepsilon_i\Big]^4 \leq C'(\log n)^2.$$

Thus, there exists C'' such that, for all $n > 1$,

$$\mathbb{E}\Big[|f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x)|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(x)}\Big] \leq C''(\log n)(\mathbb{P}\left[\mathcal{C}_\alpha^c(x)\right])^{1/2}. \tag{49}$$

The last probability $\mathbb{P}\left[\mathcal{C}_\alpha^c(x)\right]$ can be upper bounded by using Chebyshev's inequality. Indeed, with

respect to $A_n(x)$,

$$\begin{aligned}
\mathbb{P}\big[|A_n(x)| > \alpha\big] &\leq \frac{1}{n\alpha^2}\mathbb{E}\left[\frac{YK_k^{cc}(x,X)}{\mathbb{E}\left[K_k^{cc}(x,X)\right]} - \frac{\mathbb{E}\left[YK_k^{cc}(x,X)\right]}{\mathbb{E}\left[K_k^{cc}(x,X)\right]}\right]^2 \\
&\leq \frac{1}{n\alpha^2}\frac{1}{(\mathbb{E}\left[K_k^{cc}(x,X)\right])^2}\mathbb{E}\left[Y^2 K_k^{cc}(x,X)^2\right] \\
&\leq \frac{2}{n\alpha^2}\frac{1}{(\mathbb{E}\left[K_k^{cc}(x,X)\right])^2}\Bigg(\mathbb{E}\left[f^\star(X)^2 K_k^{cc}(x,X)^2\right] \\
&\qquad + \mathbb{E}\left[\varepsilon^2 K_k^{cc}(x,X)^2\right]\Bigg) \\
&\leq \frac{2(\|f^\star\|_\infty^2 + \sigma^2)}{n\alpha^2}\frac{\mathbb{E}\left[K_k^{cc}(x,X)^2\right]}{(\mathbb{E}\left[K_k^{cc}(x,X)\right])^2}. \qquad (50)
\end{aligned}$$

Meanwhile with respect to $B_n(x)$, we obtain, still by Chebyshev's inequality,

$$\mathbb{P}\big[|B_n(x)| > \alpha\big] \leq \frac{1}{n\alpha^2}\mathbb{E}\left[\frac{K_k^{cc}(x,X_i)^2}{\mathbb{E}\left[K_k^{cc}(x,X)\right]^2}\right] \qquad (51)$$

which matches the control made by [24]. Note that from here, the control on $\mathbb{P}\big[|A_n(x)| > \alpha\big]$ (50) and on $\mathbb{P}\big[|B_n(x)| > \alpha\big]$ (51) will depart from the work of [24].

First, we need a Lemma to upper bound $\mathbb{E}\left[K_k^{cc}(x,X)^2\right]$.

Lemma A.4. *For all k,*

$$\mathbb{E}\left[K_k^{cc}(x,X)^2\right] \leq v_k + 2^{-\frac{k}{d}-1} + C_2\left(\frac{2}{d}\right)^k k^{d+1/2}$$

where C_2 is a constant depending only on d and $v_k \approx C_1/(2^k k^{(d-1)/2})$ with C_1 also a constant depending only on d.

Proof of Lemma A.4. We know that

$$\mathbb{E}\left[K_k^{cc}(x,X)\right] = \frac{1}{2^k} \;\geq\; \mathbb{E}\left[K_k^{cc}(x,X)^2\right] \;\geq\; \mathbb{E}\left[K_k^{cc}(x,X)\right]^2 = \frac{1}{2^{2k}}, \qquad (52)$$

but we need a tighter upper bound on $\mathbb{E}\left[K_k^{cc}(x,X)^2\right]$. From Lemma A.2, we know that

$$\mathbb{E}\left[K_k^{cc}(x,X)^2\right] = \mathbb{E}\left[\left(\sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1!\ldots k_d!}\left(\frac{1}{d}\right)^k \prod_{j=1}^d \mathbb{1}_{\lceil 2^{k_j} x_j\rceil = \lceil 2^{k_j} z_j\rceil}\right)^2\right]. \qquad (53)$$

Developing the square within the expectation, we obtain two terms, the first one being the sum of the squares:

$$A := \mathbb{E}\left[\sum_{\substack{k_1,\dots,k_d \\ \sum_{j=1}^d k_j = k}} \left(\frac{k!}{k_1!\dots k_d!}\right)^2 \left(\frac{1}{d}\right)^{2k} \prod_{j=1}^d \mathbb{1}_{\lceil 2^{k_j} x_j \rceil = \lceil 2^{k_j} X_j \rceil}\right] \tag{54}$$

$$= \sum_{\substack{k_1,\dots,k_d \\ \sum_{j=1}^d k_j = k}} \left(\frac{k!}{k_1!\dots k_d!}\right)^2 \left(\frac{1}{d}\right)^{2k} \prod_{j=1}^d \mathbb{P}\left(\lceil 2^{k_j} x_j \rceil = \lceil 2^{k_j} X_j \rceil\right). \tag{55}$$

Note that, for all j,

$$\mathbb{P}\left(\lceil 2^{k_j} x_j \rceil = \lceil 2^{k_j} X_j \rceil\right) = \left(\frac{1}{2}\right)^{k_j},$$

and

$$\prod_{j=1}^d \left(\frac{1}{2}\right)^{k_j} = \left(\frac{1}{2}\right)^k.$$

Therefore,

$$A = \sum_{\substack{k_1,\dots,k_d \\ \sum_{j=1}^d k_j = k}} \left(\frac{k!}{k_1!\dots k_d!}\right)^2 \left(\frac{1}{d}\right)^{2k} \left(\frac{1}{2}\right)^k. \tag{56}$$

Thanks to [22], we know that

$$\sum_{\substack{k_1,\dots,k_d \\ \sum_{j=1}^d k_j = k}} \left(\frac{k!}{k_1!\dots k_d!}\right)^2 \approx \frac{d^{2k+d/2}}{k^{(d-1)/2}}. \tag{57}$$

The second term corresponds to the sum of cross-products:

$$B := \mathbb{E}\left[\sum_{\substack{(k_1,\dots,k_d) \\ \neq (l_1,\dots,l_d), \\ \sum_{j=1}^d k_j = \sum_{j=1}^d l_j = k}} \frac{k!}{k_1!\dots k_d!} \frac{k!}{l_1!\dots l_d!} \left(\frac{1}{d}\right)^{2k} \prod_{j=1}^d \mathbb{1}_{\lceil 2^{k_j} x_j \rceil = \lceil 2^{k_j} X_j \rceil} \mathbb{1}_{\lceil 2^{l_j} x_j \rceil = \lceil 2^{l_j} X_j \rceil}\right] \tag{58}$$

$$= \sum_{\substack{(k_1,\dots,k_d) \\ \neq (l_1,\dots,l_d), \\ \sum_{j=1}^d k_j = \sum_{j=1}^d l_j = k}} \frac{k!}{k_1!\dots k_d!} \frac{k!}{l_1!\dots l_d!} \left(\frac{1}{d}\right)^{2k} \mathbb{P}\left(\bigcap_{j=1}^d \left(\left(\lceil 2^{k_j} x_j \rceil = \lceil 2^{k_j} X_j \rceil\right) \cap \left(\lceil 2^{l_j} x_j \rceil = \lceil 2^{l_j} X_j \rceil\right)\right)\right).$$

A small computation yields

$$\mathbb{P}\left(\bigcap_{j=1}^{d}\left((\lceil 2^{k_j}x_j\rceil=\lceil 2^{k_j}X_j\rceil)\cap(\lceil 2^{l_j}x_j\rceil=\lceil 2^{l_j}X_j\rceil)\right)\right)=\left(\frac{1}{2}\right)^{k+\sum_{j=1}^{d}l_j\mathbb{1}_{l_j\neq k_j}}. \tag{59}$$

Therefore,

$$B=\left(\frac{1}{d}\right)^{2k}\left(\frac{1}{2}\right)^{k}\sum_{\substack{(k_1,\dots,k_d)\\ \neq(l_1,\dots,l_d),\\ \sum_{j=1}^{d}k_j=\sum_{j=1}^{d}l_j=k}}\frac{k!}{k_1!\dots k_d!}\frac{k!}{l_1!\dots l_d!}\left(\frac{1}{2}\right)^{\sum_{j=1}^{d}l_j\mathbb{1}_{l_j\neq k_j}}. \tag{60}$$

As $d>2$, we can write $k=qd+r$ with $(q,r)\in\mathbb{N}\times\{0,...,d-1\}$. Denoting $\mathcal{K}_q=\{l=(l_1,\dots,l_d)|\sum_{(l_1,\dots,l_d)\neq(k_1,\dots,k_d)}l_j\geq q\}$, we can write:

$$B=\left(\frac{1}{2d^2}\right)^{k}\sum_{\substack{(k_1,\dots,k_d)\\ \sum_{j=1}^{d}k_j=k}}\frac{k!}{k_1!\dots k_d!}\sum_{l\in\mathcal{K}_q}\frac{k!}{l_1!\dots l_d!}\left(\frac{1}{2}\right)^{\sum_{j=1}^{d}l_j\mathbb{1}_{l_j\neq k_j}} \tag{61}$$

$$+\left(\frac{1}{2d^2}\right)^{k}\sum_{\substack{(k_1,\dots,k_d)\\ \sum_{j=1}^{d}k_j=k}}\frac{k!}{k_1!\dots k_d!}\sum_{l\notin\mathcal{K}_q}\frac{k!}{l_1!\dots l_d!}\left(\frac{1}{2}\right)^{\sum_{j=1}^{d}l_j\mathbb{1}_{l_j\neq k_j}} \tag{62}$$

$$=\left(\frac{1}{2d^2}\right)^{k}(B_1+B_2). \tag{63}$$

Then, regarding B_1, as we sum over $l\in\mathcal{K}_q$,

$$\frac{k!}{k_1!\dots k_d!}\frac{k!}{l_1!\dots l_d!}\left(\frac{1}{2}\right)^{\sum_{j=1}^{d}l_j\mathbb{1}_{l_j\neq k_j}}\leq\frac{k!}{k_1!\dots k_d!}\frac{k!}{l_1!\dots l_d!}\left(\frac{1}{2}\right)^{q} \tag{64}$$

$$\leq\frac{k!}{k_1!\dots k_d!}\frac{k!}{l_1!\dots l_d!}\left(\frac{1}{2}\right)^{\frac{k}{d}-1}. \tag{65}$$

Regarding B_2, there are at most $d-1$ integers summing at most at $q-1$. Therefore for a fixed $k_1,\dots,k_d$, we have

$$\sum_{l\notin\mathcal{K}}\frac{k!}{l_1!\dots l_d!}\left(\frac{1}{2}\right)^{\sum_{j=1}^{d}l_j\mathbb{1}_{l_j\neq k_j}}\leq\sum_{l\notin\mathcal{K}}\frac{k!}{l_1!\dots l_d!}. \tag{66}$$

As a first remark, for $s \in \{2, \dots, d-2\}$, $k!/(l_{i_1}! \dots l_{i_s}!)$ is maximal when all l_is are equal. We will use Γ function verifying $\Gamma(n+1) = n!$ for all $n \in \mathbb{N}$. Using an inequality from [5], we know that

$$\frac{k!}{l_1! \dots l_d!} \leq \frac{k!}{l_{i_1}! \dots l_{i_s}!} \tag{67}$$

$$\leq \frac{k!}{\Gamma(k/s+1)^s} \tag{68}$$

$$\leq \sqrt{2\pi} s^{s+s/2} \frac{\sqrt{k} k^k e^{-k} e^{1/12k}}{k^{\frac{s}{2}} k^k s^{-k} e^{-k} e^{-\frac{s}{6k/s+3/8}}} \tag{69}$$

$$\leq 2C_s k^{-\frac{s-1}{2}} s^k, \tag{70}$$

with C_s a constant depending only on s. Note that when we are not in $\mathcal{K}$, we can choose s k_js such that their sum is greater than $k-q+1$. For $l \notin \mathcal{K}$, we denote $\mathcal{K}_l = \{l = (l'_{i_1}, ..., l_{i'_{d-s}}) | l_j \neq k_j\}$. We obtain

$$B_2 = \sum_{p=2}^{q-1} \sum_{s=2}^{d-2} \sum_{\substack{(k_1,\dots,k_d) \\ \sum_{i=1}^s k_{j_i} = k-p \\ \sum_{i=1}^d k_i = k}} \frac{k!}{k_1! \dots k_d!} \sum_{\substack{l' \in \mathcal{K}_l \\ l_{j_1} = k_{j_1}, \dots, l_{j_s} = k_{j_s} \\ \text{fixed} \\ \sum_{i=1}^d l_i = k}} \frac{k!}{l_1! \dots l_d!} \tag{71}$$

$$\leq \sum_{p=2}^{q-1} \sum_{s=2}^{d-2} \sum_{\substack{(k_1,\dots,k_d) \\ \sum_{i=1}^s k_{j_i} = k-p \\ \sum_{i=1}^d k_i = k}} \frac{k!}{k_1! \dots k_d!} \sum_{\substack{l' \in \mathcal{K}_l \\ l_{j_1} = k_{j_1}, \dots, l_{j_s} = k_{j_s} \\ \text{fixed} \\ \sum_{i=1}^d l_i = k}} 2C_s k^{-\frac{s-1}{2}} s^k. \tag{72}$$

The number of terms in the third sum over the d-uplet with s elements being fixed equals

$$\binom{d}{s} \binom{(p-1)+d-s-1}{p-1} \tag{73}$$

and is maximum for $p = q-1$. Therefore,

$$B_2 \leq \sum_{p=2}^{q-1} \sum_{s=2}^{d-2} \sum_{\substack{(k_1,\dots,k_d) \\ \sum_{i=1}^s k_{j_i} = k-p \\ \sum_{i=1}^d k_i = k}} \frac{k!}{k_1! \dots k_d!} C'_s \frac{(q+d-s-3)!}{(q-2)!} k^{-\frac{s-1}{2}} s^k, \tag{74}$$

with $C'_s > 0$ a constant depending only on d. Recall that $q = \lfloor k/d \rfloor$. Note that $\frac{(q+d-s-2)!}{(q-1)!}$ is in $O(q^{d-s-2})$. The maximal term of the previous sum is therefore reached for $s = 2$. Overall, we find

$$B_2 \leq \sum_{p=2}^{q-1} \sum_{s=2}^{d-2} \sum_{\substack{(k_1,\dots,k_d) \\ \sum_{i=1}^s k_{j_i} = k-p \\ \sum_{i=1}^d k_i = k}} \frac{k!}{k_1! \dots k_d!} C_2 \cdot k^d \cdot k^{-1/2} \cdot 2^k \tag{75}$$

$$= \sum_{\substack{(k_1,\dots,k_d) \\ \sum_{i=1}^d k_i = k}} \frac{k!}{k_1! \dots k_d!} C_2 \cdot k^d \cdot k^{-1/2} \cdot 2^k, \tag{76}$$

with $C_2 > 0$ a constant depending only on d.

Finally, as

$$\sum_{\substack{k_1,\dots,k_d \\ \sum_{j=1}^d k_j = k}} \left(\frac{k!}{k_1! \dots k_d!} \right) = d^k,$$

we obtain

$$B_2 \leq C_2 \cdot (2d)^k \cdot \sqrt{k} \cdot k^d. \tag{77}$$

□

Using the previous Lemma, we have

$$\mathbb{P}\left(|A_n(x)| > \alpha\right) \leq 2M_1^2 \frac{2^k}{n\alpha^2} \left(v_k + 2^{-\frac{k}{d}-1} + C_2 \left(\frac{2}{d}\right)^k k^{d+1/2} \right), \tag{78}$$

where $v_k \approx C_d/(k^{(d-1)/2})$, and

$$\mathbb{P}\left(|B_n(x)| > \alpha\right) \leq \frac{2^k}{n\alpha^2} \left(v_k + 2^{-\frac{k}{d}-1} + C_2 \left(\frac{2}{d}\right)^k k^{d+1/2} \right).$$

Thus, the probability of $\mathcal{C}_\alpha(x)$ is given by

$$\begin{aligned}
\mathbb{P}\big[\mathcal{C}_\alpha(x)\big] &\geq 1 - \mathbb{P}\big(|A_n(x)| \geq \alpha\big) - \mathbb{P}\big(|B_n(x)| \geq \alpha\big) \\
&\geq 1 - \left(\frac{2^k}{n} \frac{2M_1^2}{\alpha^2} + \frac{2^k}{n\alpha^2} \right) \left(v_k + 2^{-\frac{k}{d}-1} + C_2 \left(\frac{2}{d}\right)^k k^{d+1/2} \right).
\end{aligned}$$

Consequently, according to inequality (49), we obtain

$$\mathbb{E}\Big[|f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x)|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(x)} \Big] \leq C_2(\log n) \left(\frac{2^k}{n} \frac{2M_1^2}{\alpha^2} + \frac{2^k}{n\alpha^2} \right)^{1/2} \left(v_k + 2^{-\frac{k}{d}-1} + C_2 \left(\frac{2}{d}\right)^k k^{d+1/2} \right)^{1/2}.$$

Then using inequality (48),

$$\begin{aligned}
&\mathbb{E}\Big[f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x) \Big]^2 \\
&\leq \mathbb{E}\Big[|f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x)|^2 \mathbb{1}_{\mathcal{C}_\alpha(x)} \Big] + \mathbb{E}\Big[|f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x)|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(x)} \Big] \\
&\leq 8C_1^2 \left(1 - \frac{1}{2d}\right)^{2k} + 8\alpha^2(1 + \|m\|_\infty)^2 \\
&+ C_2(\log n) \left(\frac{2^k}{n} \frac{2M_1^2}{\alpha^2} + \frac{2^k}{n\alpha^2} \right)^{1/2} \left(v_k + 2^{-\frac{k}{d}-1} + C_2 \left(\frac{2}{d}\right)^k k^{d+1/2} \right)^{1/2}.
\end{aligned}$$

Optimizing the right hand side in α
($\alpha^3 = C_2(\log n)\left(\frac{2^k}{n}2M_1^2 + \frac{2^k}{n}\right)^{1/2}\left(v_k + 2^{-\frac{k}{d}-1} + C_2\left(\frac{2}{d}\right)^k k^{d+1/2}\right)^{1/2}(1+\|m\|_\infty)^{-2}/16$), we get

$$\mathbb{E}\Big[f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x)\Big]^2$$
$$\leq 8C_1^2\left(1-\frac{1}{2d}\right)^{2k} + C_3(\log n)^{2/3}\left(\frac{2^k}{n}2M_1^2+\frac{2^k}{n}\right)^{1/3}\left(v_k + 2^{-\frac{k}{d}-1} + C_2\left(\frac{2}{d}\right)^k k^{d+1/2}\right)^{1/3}.$$

for some constant $C_3 > 0$. Choosing $k_n = \lfloor \log_2(n) \rfloor$, we obtain:

$$\mathbb{E}\Big[f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x)\Big]^2 \tag{79}$$
$$\leq C_d n^{2\log(1-\frac{1}{d})} + C_d(\log n)^{2/3}\left(w_n + \frac{n^{-1/2d}}{2} + n^{-\log_2(d-2)}(\log n)^{d+1/2}\right)^{1/3}, \tag{80}$$

with $C_d > 0$ and $w_n \approx \log(n)^{-(d-1)/2}$. Finally,

$$\mathbb{E}\Big[f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x)\Big]^2 \leq C_d\left(\log(n)^2 w_n + \frac{\log(n)^2 n^{-1/d}}{2} + n^{-\log_2(d-2)}(\log n)^{d+5/2}\right)^{1/3} + C_d n^{2\log(1-\frac{1}{2d})}.$$

□

A.3 Proofs of section 5

A.3.1 Proof of Lemma 5.1

To begin with, we have

$$\mathbb{P}\left(k_n(X) \geq k\right) = \mathbb{P}\left(\bigcap_{i=1}^{k-1} N_{i,\Theta}(X) \geq 2\right) \tag{81}$$
$$= \mathbb{P}\left(N_{1,\Theta} \geq 2\right)\prod_{i=2}^{k-1}\mathbb{P}\left(N_{i,\Theta}(X) \geq 2 | N_{i-1,\Theta}(X) \geq 2\right) \tag{82}$$
$$= \mathbb{P}\left(N_{k-1,\Theta}(X) \geq 2\right) \tag{83}$$
$$= \mathbb{E}\left[\mathbb{P}\left(N_{k-1,\Theta}(X) \geq 2 | X\right)\right] \tag{84}$$
$$= 1 - \left(1-\frac{1}{2^{k-1}}\right)^n - \frac{n}{2^{k-1}}\left(1-\frac{1}{2^{k-1}}\right)^{n-1}. \tag{85}$$

Studying the right-hand side of equality (85) with $k = (1-\log^{-\alpha}(n))\log_2(n)$ and using the inequalities

$$\exp(-\frac{n}{2^k-1}) \leq \exp(n\log(1-\frac{1}{2^k})) \leq \exp(-\frac{n}{2^k})$$

yields the result.

A.3.2 Proof of Lemma 5.2

Given a dataset, a point X falls into an empty leaf of $m_{n,\infty}$ if and only if the probability w.r.t Θ to connect 0 and a point is 0. In particular, if along each direction $X^{(j)}$, the tree built by cutting along $X^{(j)}$ only (which appears a.s. in the infinite forest) does not connect X and a point of the dataset, then X falls into an empty leaf of $m_{n,\infty}$ (the reciprocal is true as well). As drawing i.i.d. uniform X_i's in $[0,1]^d$ is the same as drawing them i.i.d. coordinate per coordinate, we can start by focusing on the direction $X^{(1)}$. Mathematically, we have $\mathcal{E}_0 = \bigcap_{j=1}^{d} \mathcal{E}_0^{(j)}$ where $\mathcal{E}_0^{(j)}$ is the event $\{0$ falls into an empty leaf for all trees that cut only along direction $X^{(j)}\}$. To lighten the notation, we write X instead of $X^{(1)}$ during the proof.

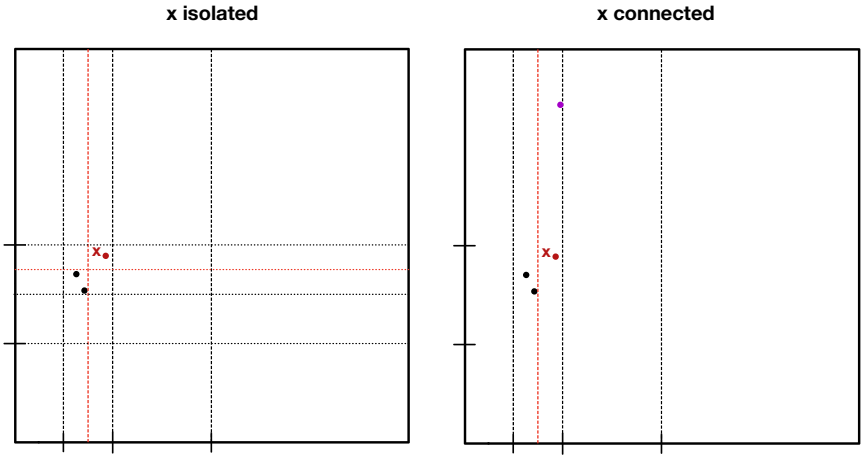

Figure 7: Condition to isolate x for all trees. In the first case, x cannot connect to any point whatever the chosen cuts whereas in the second case, x can connect to the purple point if a tree cuts along the first direction only.

As seen on Figure 7, X cannot be connected to any point along direction X if and only if, denoting $N_n(A)$ the number of points X_i falling into a set $A \subset [0,1]^d$, either

1. there exist γ_1 a real and $[\alpha_1, \beta_1]$ an interval, with $1 \geq \beta_1 \geq \alpha_1 \geq X \geq \gamma_1$ and $N_n([\alpha_1, \beta_1]) \geq 2$ and $N_n([\alpha_1, \gamma_1]) = 0$ where γ_1 is the closest split to X under X.
2. the symmetrized situation of the previous one.

In order to compute the probability of the first situation described above, we introduce base-2

notations: we define U_n the unique base-2 sequence of X defined as

$$X = \sum_{s=1}^{\infty} 2^{-s} \mathbb{1}_{U_s=1}.$$

To choose α_1, we consider the splits defined from the base-2 sequence of X: if ι^1 (resp. ι^0) is the set of indexes of the successive 1 in the base-2 representation of X, then, for all $q \in \iota^0$, we define $\alpha_{1,q} = \sum_{s=1}^{q-1} 2^{-s} \mathbb{1}_{U_s=1} + 2^{-q}$. The base-2 representation of $\alpha_{1,q}$ corresponds to the one of X truncated at index $q-1$ and where the 0 at position q is replaced by a 1. As a consequence, $X \leq \alpha_{1,q}$ for all q. Then, following the previous notations, $\beta_{1,q}$ simply equals $\alpha_{1,q} + 2^{-q}$ and $\gamma_{1,q} = \alpha_{1,q} - 2^{-q}$.

We begin by conducting the computations conditional on X:

$$\mathbb{P}\left(\mathcal{E}_0^{(1)}\right) = \mathbb{E}\left[\mathbb{P}_{\mathcal{D}_n}\left(\mathcal{E}_0^{(1)}\right)\right]. \tag{86}$$

We have

$$\mathbb{P}_{\mathcal{D}_n}\left(\mathcal{E}_0^{(1)}\right) \geq \mathbb{P}_{\mathcal{D}_n}\left(\bigcup_{q\in\iota^0} N_n([\alpha_{1,q}, \beta_{1,q}]) \geq 2 \cap N_n([\gamma_{1,q}, \alpha_{1,q}]) = 0\right) \tag{87}$$

$$= \sum_{q\in\iota^0} \mathbb{P}_{\mathcal{D}_n}\left(N_n([\alpha_{1,q}, \beta_{1,q}]) \geq 2 \cap N_n([\gamma_{1,q}, \alpha_{1,q}]) = 0\right). \tag{88}$$

The first inequality comes from the fact that we do not take into account the symmetric situation, the probability of which scales similarly. Then, for all $q \in \iota^0$,

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}_n}\left(N_n([\alpha_{1,q}, \beta_{1,q}]) \geq 2 \cap N_n([\gamma_{1,q}, \alpha_{1,q}]) = 0\right) \\
&= \mathbb{P}_{\mathcal{D}_n}\left(N_n([\alpha_{1,q}, \beta_{1,q}]) \geq 2 \middle| N_n([\gamma_{1,q}, \alpha_{1,q}]) = 0\right) \mathbb{P}_{\mathcal{D}_n}\left(N_n([\gamma_{1,q}, \alpha_{1,q}]) = 0\right) \\
&= \left(1 - \mathbb{P}_{\mathcal{D}_n}\left(N_n([\alpha_{1,q}, \beta_{1,q}]) \leq 1 \middle| N_n([\gamma_{1,q}, \alpha_{1,q}]) = 0\right)\right)\left(1 - 2^{-q}\right)^n \\
&= \left(1 - \binom{n}{1} \frac{2^{-q}}{1-2^{-q}} \left(1 - \frac{2^{-q}}{1-2^{-q}}\right)^{n-1} - \left(1 - \frac{2^{-q}}{1-2^{-q}}\right)^n\right)\left(1 - 2^{-q}\right)^n \\
&= \left(1 - \frac{n}{2^q - 1} \left(1 - \frac{1}{2^q - 1}\right)^{n-1} - \left(1 - \frac{1}{2^q - 1}\right)^n\right)\left(1 - 2^{-q}\right)^n
\end{aligned}$$

as the X_i's are uniform and i.i.d.

The above quantity does not tend to 0 when $q \approx \log_2 n$. Given $c \leq \log_2 n$ a constant, we write A_n the event $\{\exists s \in (\lfloor \log_2 n \rfloor \pm c), U_s = 0\}$. As X follows a uniform distribution over $[0, 1]$, $\mathbb{P}(A_n) = 1 - 2^{-2c}$.

$$\mathbb{P}\left(\mathcal{E}_0^{(1)}\right) \geq \mathbb{E}\left[\left(1-\frac{n}{2^q-1}\left(1-\frac{1}{2^q-1}\right)^{n-1}-\left(1-\frac{1}{2^q-1}\right)^n\right)\left(1-2^{-q}\right)^n \mathbb{1}_{A_n}\right]$$
$$\gtrsim (1-4^{-c})e^{-\frac{n}{n-1}}\left(1-\frac{n}{n-1}e^{-1}-e^{-1}\right).$$

Finally,

$$\mathbb{P}\left(\mathcal{E}\right) \gtrsim (1-4^{-c})^d e^{-\frac{dn}{n-1}}\left(1-\frac{n}{n-1}e^{-1}-e^{-1}\right)^d.$$

A.3.3 Proof of Proposition 5.3

Note that the infinite AdaCRF is written as

$$f_{\infty,n}^{\text{AdaCRF}}(x) = \frac{1}{\mathbb{P}_\Theta\left(N_n(x,\Theta)>0\right)}\mathbb{E}_\Theta[f_n(x,\Theta)\mathbb{1}_{N_n(x,\Theta)>0}]$$
$$= \mathbb{E}_\Theta[f_n(x,\Theta)|N_n(x,\Theta)>0]$$

where $f_n(.,\Theta)$ denotes a single adaptive centered tree.

$f_{\infty,n}^{\text{AdaCRF}}$ is obtained by applying the law of large numbers to $f_{M,n}^{\text{AdaCRF}}$ (see [23]) for details). Therefore,

$$\mathbb{E}\left[\left(f_{\infty,n}^{\text{AdaCRF}}(X)-f^\star(X)\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{\mathbb{P}_\Theta\left(N_n(X,\Theta)>0\right)}\mathbb{E}_\Theta[f_n(X,\Theta)\mathbb{1}_{N_n(x,\Theta)>0}]-f^\star(X)\right)^2\right]$$
$$= \mathbb{E}\left[\left(\frac{1}{\mathbb{P}_\Theta\left(N_n(x,\Theta)>0\right)}\mathbb{E}_\Theta[f_n(X,\Theta)\mathbb{1}_{N_n(x,\Theta)>0}]-f^\star(X)\left(\mathbb{1}_{\mathcal{E}}+\mathbb{1}_{\mathcal{E}^\perp}\right)\right)^2\right]$$
$$\geq \mathbb{E}\left[f^\star(X)^2\mathbb{1}_{\mathcal{E}^c}\right]$$

with $\mathcal{E}$ the event $\{\forall\Theta, N_n(X,\Theta)=0\}$ which is $(X,X_1,\ldots,X_n)$ measurable. Note that when $\mathcal{E}$ is satisfied, $\mathbb{P}_\Theta\left(N_n(x,\Theta)>0\right)=0$.

From Lemma 5.2, we know that $\mathbb{E}\left[f^\star(X)^2\mathbb{1}_{\mathcal{E}^c}\right]$ does not tend to 0 as n tends to infinity. Therefore $f_{\infty,n}^{\text{AdaCRF}}$ is inconsistent.

A.3.4 Proof of Theorem 5.4

As we are in a noiseless setting, the risk simply equals the bias:

$$\begin{aligned}\mathbb{E}\left[\left(f_{\infty,n}^{\mathrm{MedRF}}(X)-f^\star(X)\right)^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^n f^\star(X_i)\mathbb{E}_\Theta[W_i]-f^\star(X)\right)^2\right]\\ &= \mathbb{E}\left[\left(\sum_{i=1}^n \left(f^\star(X_i)-f^\star(X)\right)\mathbb{E}_\Theta[W_i]\right)^2\right].\end{aligned}$$

The second term of the last inequality verifies

$$\mathbb{E}\left[\left(f^\star(X)\mathbb{1}_{\mathbb{P}_\Theta(N_n(X,\Theta)>0)=0}\right)^2\right] \leq ||f^\star||_\infty^2 \mathbb{P}\left(\mathbb{P}_\Theta\left(N_n(X,\Theta)>0\right)=0\right). \tag{89}$$

Following a computation from [17],

$$\begin{aligned}|f^\star(X)-f^\star(X_i)| &\leq \sum_{\ell=1}^d ||\partial_\ell f^\star||_\infty |X_i^{(j)}-X^{(j)}|\\ &\leq \sum_{\ell=1}^d ||\partial_\ell f^\star||_\infty (b_\ell-a_\ell)\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^n W_i|f^\star(X)-f^\star(X_i)| &\leq \sum_{i=1}^n W_i \sum_{\ell=1}^d ||\partial_\ell f^\star||_\infty (b_\ell-a_\ell) &(90)\\ &\leq \sum_{\ell=1}^d ||\partial_\ell f^\star||_\infty (b_\ell-a_\ell). &(91)\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\left(f_{\infty,n}^{\mathrm{MedRF}}(X)-f^\star(X)\right)^2\right] &\leq 2\mathbb{E}\left[\left(\sum_{\ell=1}^d ||\partial_\ell f^\star||_\infty \mathbb{E}_\Theta[(b_\ell-a_\ell)]\right)^2\right] &(92)\\ &\leq 2d\sum_{\ell=1}^d ||\partial f_\ell^\star||_\infty^2 \mathbb{E}\left[\mathbb{E}_\Theta[(b_\ell-a_\ell)]^2\right] &(93)\end{aligned}$$

where the last inequality directly results from Cauchy-Schwarz inequality.

Then, Jensen inequality yields

$$\mathbb{E}\left[\mathbb{E}_\Theta[(b_\ell-a_\ell)]^2\right] \leq \mathbb{E}\left[b_\ell-a_\ell\right]^2.$$

We now adapt a Lemma from [13] to control the volume of the leaves.

Lemma A.5. *For all $\ell \in \{1, \dots, d\}$ and depth $k \in \mathbb{N}^*$, we have*

$$\mathbb{E}\Big[(b_\ell - a_\ell)^2\Big] \leq 2\left(1 - \frac{57}{64d}\right)^{k-1}.$$

Proof of Lemma A.5. Let us fix $x \in [0,1]^d$ and denote by $n_0, n_1, \dots, n_k$ the number of points in the successive cells containing x (for example, n_0 is the number of points in the root of the tree, that is $n_0 = a_n$). Note that $n_0, n_1, \dots, n_k$ depends on $\mathcal{D}_n$ and Θ, but to lighten notations, we omit these dependencies.

Recalling that $b_l - a_l$ is the length of the ℓ-th side of the cell containing x, this quantity can be compared to a product of independent data distributions: the n_{j+1}-th statistical order of a uniform sample $Z_1, \dots, Z_j$ follows a distribution of parameters $(n_{j+1}, n_j - n_{j+1} + 1)$ and corresponds to the length of the cell at depth j if the cut is made on the n_j-th points. If the number of points n_j is even and the cut is made between two points, then the length of the current node j will be less than the $n_j + 1$-th statistical order. Therefore we have the inequality

$$\mathbb{E}\big[(b_l - a_l)^2\big] \leq \prod_{j=1}^{k} \mathbb{E}\Big[\big[B(n_j+1, n_{j-1}-n_j)\big]^{2\delta_{\ell,j}(x,\Theta)\eta_j}\big[B(n_j+2, n_{j-1}-n_j-1)\big]^{2\delta_{\ell,j}(x,\Theta)(1-\eta_j)}\Big]$$

where $B(\alpha, \beta)$ denotes the beta distribution of parameters α and β, and the indicator $\delta_{\ell,j}(x,\Theta)$ equals to 1 if the j-th split of the cell containing x is performed along the ℓ-th dimension (and 0 otherwise) and $\eta_j = \mathbb{1}_{\{n_j \% 2 = 1\}}$. Note that as $\mathbb{E}\left[B(\alpha,\beta)^q\right] \leq \mathbb{E}\left[B(\alpha+1,\beta-1)^q\right]$ for $q \in \{1,2\}$, we have

$$\begin{aligned}
\mathbb{E}\big[(b_l - a_l)^2\big] &\leq \prod_{j=1}^{k} \mathbb{E}\Big[\mathbb{E}\Big[\big[B(n_j+2, n_{j-1}-n_j-1)\big]^{2\delta_{\ell,j}(x,\Theta)}\big|\delta_{\ell,j}(x,\Theta)\Big]\Big] \\
&= \prod_{j=1}^{k} \mathbb{E}\Big[\mathbb{1}_{\delta_{\ell,j}(x,\Theta)=0} + \mathbb{E}\big[B(n_j+2, n_{j-1}-n_j-1)\big]^2 \mathbb{1}_{\delta_{\ell,j}(x,\Theta)=1}\Big] \\
&= \prod_{j=1}^{k} \left(\frac{d-1}{d} + \frac{1}{d}\mathbb{E}\big[B(n_j+2, n_{j-1}-n_j-1)\big]^2\right) \\
&= \prod_{j=1}^{k} \left(\frac{d-1}{d} + \frac{1}{d}\frac{(n_j+2)(n_j+3)}{(n_{j-1}+1)(n_{j-1}+2)}\right) \\
&\leq \prod_{j=1}^{k} \left(\frac{d-1}{d} + \frac{1}{4d}\frac{(n_{j-1}+4+1/2)(n_{j-1}+6+1/2)}{(n_{j-1}+1)(n_{j-1}+2)}\right)
\end{aligned}$$

where the first inequality stems from the relation $n_j \leq (n_{j-1}+1)/2$ for all $j \in \{1, \dots, k\}$. Further-

more,

$$g(n_{j-1}) := \frac{(n_{j-1}+4+1/2)(n_{j-1}+6+1/2)}{(n_{j-1}+1)(n_{j-1}+2)}$$

is a decreasing function of n_{j-1} and $g(2) = 221/48$. As $n_{j-1} \geq 3$ for all $j \leq k-1$, we have

$$\frac{(n_{j-1}+4)(n_{j-1}+6)}{(n_{j-1}+1)(n_{j-1}+2)} \leq \frac{57}{16}. \tag{94}$$

Finally,

$$\begin{aligned}\mathbb{E}\big[V_\ell(x,\Theta)^2\big] &\leq \left(1+\frac{29}{192d}\right)\prod_{j=1}^{k-1}\left(1-\frac{1}{d}+\frac{57}{64d}\right)\\ &\leq 2\left(1-\frac{57}{64d}\right)^{k-1}.\end{aligned}$$

□

As the cut are performed at the median of the samples of the current node, it is clear that the depth of $f_{\infty,n}^{\mathrm{MedRF}}$ at point X is greater than $\log_2(\frac{n+1}{2})$. Therefore,

$$\mathbb{E}\left[\left(f_{\infty,n}^{\mathrm{MedRF}}(X)-f^\star(X)\right)^2\right] \leq 4d\left(1-\frac{57}{64d}\right)^{\log_2(n+1)-2}\sum_{\ell=1}^{d}||\partial f_\ell^\star||_\infty^2. \tag{95}$$

A.3.5 Proof of Proposition 5.6

It is possible to conduct a one-dimensional analysis and then to extend the result to the multi-dimensional case by a simple multiplication. Indeed all the leaves are determined coordinate per coordinate, therefore the interpolation area is the product of all interpolation areas along each direction.

Let $Z_1, \ldots, Z_n$ be n i.i.d. random variables uniformly distributed over $[0,1]$. As in the infinite Median RF the trees built by cutting along one direction only appear a.s. , the length of the cell along the chosen direction is less than $Z_{(k+1)} - Z_{(k-1)}$ where $Z_{(i)}$ indicates the i-th statistical order. Moreover, it is known that $Z_{(k)}$ follows a Beta distributions of parameters $(k, n-k+1)$. Therefore,

$$\mathbb{E}\left[Z_{(k+1)} - Z_{(k-1)}\right] = \frac{k+1}{n+1} - \frac{k-1}{n+1} \tag{96}$$

$$\leq \frac{2}{n}. \tag{97}$$

Now, as $X_1, ..., X_n$ are i.i.d. and uniformly distributed over $[0,1]^d$, for any data point $x \in [0,1]^d$ we simply have that

$$\mathbb{E}\left[\mu(\mathcal{A}_{min,x})\right] \leq \frac{2^d}{n^d}.$$

Finally, since by definition all interpolation zones are disjoint and the interpolation area is the union of n interpolation areas, we have

$$\mathbb{E}\left[\mu(\mathcal{A}_{min})\right] \leq \frac{2^d}{n^{d-1}}$$

which ends the proof.

A.4 Proofs of section 6

Proof of 6.1. Before diving into the computations, let us recall two facts about Breiman RF construction. First, when a CART cuts between two points, the cut is made at the middle of these two points. Second, assume that all the cuts are possible, i.e. that the probability of cutting between all pairs of successive points along all dimensions is strictly positive. Therefore, for a given point X_i, one can define the minimal interpolation zone $\mathcal{A}_{min,X_i} := \bigcap_{M\in\mathbb{N},\boldsymbol{\Theta}_M} \mathcal{A}_{X_i,\boldsymbol{\Theta}_M}$ around X_i. The boundaries of this area are given for each direction by the cuts between X_i and its *neighbor points* respectively to the considered direction, as illustrated on Figure 8.

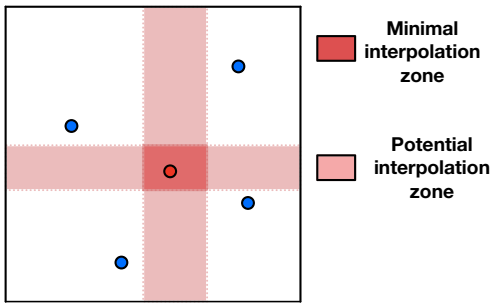

Figure 8: Different interpolation zones of a data point (in red).

1. The interpolation zone is the union of n interpolation zones, each one containing a single X_i. We denote $\mathcal{A}(m_{M,n}(.,\boldsymbol{\Theta}_M)) = \mathcal{A}_{X_1,\boldsymbol{\Theta}_M} \cup ... \cup \mathcal{A}_{X_n,\boldsymbol{\Theta}_M}$ with $\mathcal{A}_{X_i,\Theta_M} = \{x \in [0,1]^d, m_{M,n}(x,\boldsymbol{\Theta}_M) = Y_i\}$. We begin with a one-dimensional analysis. We denote $X_i^{(j)}$ the j-th feature of X_i, for all $j \in \{1,\dots,d\}$ and $i \in \{1,\dots,n\}$ and we focus on the first variable $X^{(1)}$. As $X_1,\dots,X_n$ are i.i.d. and follow a uniform distribution over $[0,1]^d$, $X_1^{(1)},...,X_n^{(1)}$ are i.i.d. and uniformly distributed on $[0,1]$. For the ease of notation, we define $Z_1 := X_1^{(1)},...,Z_n := X_n^{(1)}$. Let $x = Z_n$. The length (volume) of $\mathcal{A}_{min,Z_n}$ restricted to the first dimension is simply given by the sum of the distance from x to its closest point on the left side and to its closest point on the right side (divided by 2 as the cut are made in the middle of two points). Therefore,

$$\mu(A_{min,x}) = \frac{1}{2}\left(x - \max_{\{Z_i, Z_i \leq x\}\cup\{0\}} Z_i + \min_{\{Z_i, Z_i \geq x\}\cup\{1\}} Z_i - x\right). \tag{98}$$

All computations are made conditionally on x. Denoting N_x the cardinal of the set $\{Z_i : Z_i \leq x \text{ with } 1 \leq i < n\}$, we have for any $t \in [0, x/2)$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{2}\left(x - \max_{\{Z_i, Z_i \leq x\} \cup \{0\}} Z_i\right) \leq t \,|x\right) \\
&= 1 - \mathbb{P}\left(\max_{\{Z_i, Z_i \leq x\} \cup \{0\}} Z_i < x - 2t \,|x\right) \\
&= 1 - \mathbb{E}\left[\mathbb{E}\left[\mathbb{P}\left((Z_{i_1} < x - 2t) \cap ... \cap (Z_{i_{N_x}} < x - 2t) | N_x, Z_{i_1} \leq x, ..., Z_{i_{N_x}|x} \leq x\right)\right] |x\right] \\
&= 1 - \mathbb{E}\left[(Z_1 < x - 2t | Z_1 \leq x)^{N_x} | x\right] \\
&= 1 - \sum_{k=0}^{n-1} \mathbb{P}\left(N_x = k|x\right) \mathbb{P}\left(Z_1 < x - 2t | Z_1 \leq x | x\right)^k \\
&= 1 - \sum_{k=0}^{n-1} \mathbb{P}\left(N_x = k|x\right) \left(\frac{x - 2t}{x}\right)^k \\
&= 1 - \left((1 - x) + x\left(\frac{x - 2t}{x}\right)\right)^{n-1} \\
&= 1 - (1 - 2t)^{n-1}
\end{aligned}$$

where the penultimate equality is obtained by noticing that N_x is a binomial of parameters $(n-1, x)$ and computing its probability-generating function.
So for all $t \geq 0$,

$$\mathbb{P}\left(\frac{1}{2}\left(x - \max_{\{Z_i, Z_i \leq x\} \cup \{0\}} Z_i\right) \leq t | x\right) = 1 - (1 - 2t)^{n-1} \mathbb{1}_{t < x/2}.$$

By symmetry,

$$\mathbb{P}\left(\frac{1}{2}\left(\min_{\{Z_i, Z_i \geq x\} \cup \{1\}} Z_i - x\right) \leq t | x\right) = 1 - (1 - 2t)^{n-1} \mathbb{1}_{t > (1-x)/2}.$$

Overall, using the fact that for any variable Z with cumulative function F_Z, $\mathbb{E}[Z] = \int (1 - F_Z)$, we have

$$\begin{aligned}
\mathbb{E}\left[\mu(\mathcal{A}_{min,x}) | x\right] &= \int_0^{x/2} (1 - 2u)^{n-1} du + \int_0^{(1-x)/2} (1 - 2u)^{n-1} du \\
&= \frac{1}{2n}\left(2 - (1 - x)^n - x^n\right) \\
&\leq \frac{1}{n}\left(1 - \frac{1}{2^n}\right).
\end{aligned}$$

Now, as $X_1, ..., X_n$ are i.i.d. and uniformly distributed over $[0, 1]^d$, for any data point $x \in [0, 1]^d$ we simply have that

$$\mathcal{A}_{min,x} = \bigtimes_{j=1}^{d} \mathcal{A}_{min,x^{(j)}}.$$

Therefore,

$$\mathbb{E}\left[\mu(\mathcal{A}_{min,x})\right] \leq \frac{1}{n^d}\left(1-2^{-n}\right)^d.$$

Finally, since by definition all interpolation zones are disjoint, we have

$$\mathbb{E}\left[\mu(\mathcal{A}_{min})\right] \leq \frac{1}{n^{d-1}}\left(1-2^{-n}\right)^d.$$

2. It is enough to notice that the minimal interpolation zone is the intersection of all the potential interpolation zones. It is reached when the forest contains all the possible cuts. Then, as the probability of any given cut appearing is strictly greater than 0 by hypothesis, the probability of its appearance in the infinite forest is one. Therefore almost surely, when M grows to infinity, the interpolation zone of the forest reaches the minimal interpolation zone.

□

B Experiment supplementary

For all experiments, we introduce the following regression models.

- **Model 1:** $d = 2$, $Y = 2X_1^2 + \exp(-X_2^2)$
- **Model 2:** $d = 8$, $Y = X_1X_2 + X_3^2 - X_4X_5 + X_6X_7 - X_8^2 + \mathcal{N}(0, 0.5)$
- **Model 3:** $d = 6$, $Y = X_1^2 + X_2^2X_3e^{-|X_4|} + X_5 - X_6 + \mathcal{N}(0, 0.5)$
- **Model 4:** $d = 5$, $Y = 1/(1 + \exp(-10 * (\sum_{i=1}^{d} X_i - 1/2))) + \mathcal{N}(0, 0.05)$

All the experiments are conducted using Python3. We use Scikit-learn RandomForestRegressor class to implement the Breiman RF model. We coded CRF, KeRF and AdaCRF models ourselves, mainly relying on *numpy* and *joblib* libraries for computation optimisation.

B.1 Consistency experiments

For all consistency experiments, the dataset was divided into a train dataset (80% of the data) and a test dataset (20%) of the data.

The parameters of the estimators were set as follows:

- all RF estimators have 500 *trees* to mimic the behavior of the infinite RF.
- the *max depth* parameter is set to *None* for all RF estimators, which corresponds to growing each tree until pure leaves.
- parameter *bootstrap* is set to *False* for all estimators in order preserve the interpolation property, or set to *True* when specified.
- all other parameters are set to default value for Breiman RF.

B.1.1 Consistency of Breiman RF with max-feature= 1

On Figure 9, we see that the excess risk of a Breiman RF with the max-features parameter set to 1 is decreasing towards 0 as n increases. This RF seems consistent for all models.

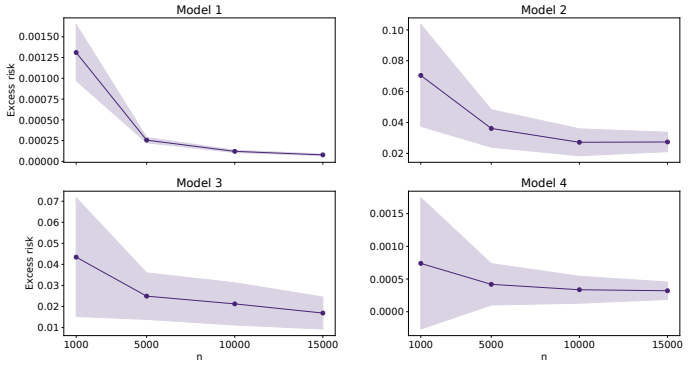

Figure 9: Consistency of Breiman RF: excess risk w.r.t sample size. RF parameters: 500 trees, max-depth set to None, max-features= 1, no bootstrap. Mean over 30 tries(doted line) and std (filled zone).

B.2 Interpolation experiments

B.2.1 Volume of the interpolation zone w.r.t sample size n

We plot on Figure 10 the log-volume of the interpolation zone of a Breiman RF with the max-features parameter set to $\lceil d/3 \rceil$ (the default value proposed in `R randomForest` package). The volume decreases polynomially in n but slower than when max-features= 1 (Figure 3) which is to be expected: choosing max-features= 1 should increase the diversity of the splits and therefore reduce the volume of the interpolation zone.

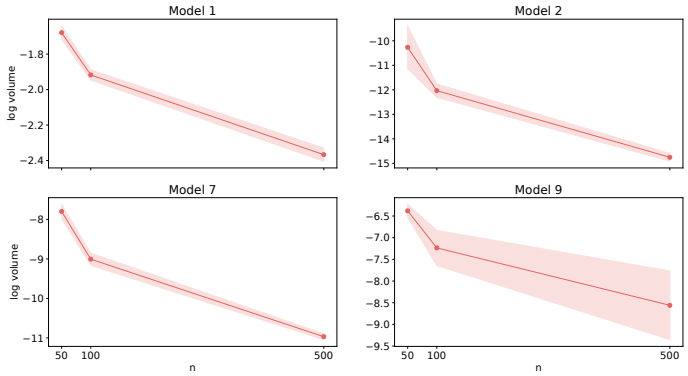

Figure 10: Log volume of Breiman RF interpolation zone w.r.t. sample size n. RF parameters: 500 trees, no bootstrap, max features = $\lceil d/3 \rceil$. Mean over 10 tries (bold line) and std (filled zone).

B.2.2 Volume of the interpolation zone w.r.t number of trees M

In this section, we empirically measure how fast decreases the volume of the interpolation zone of a Breiman RF when its number of trees M increases, and how close the interpolation zone gets from the minimal interpolation zone.

To this end, for a fixed sample size $n = 500$, we numerically evaluate the volume of the interpolation area when the number M of trees in the forest grows. This volume is anticipated to be a non-increasing function of M (for $M = 1$, note that the interpolation volume is 1, the volume of $[0, 1]^d$), but its decrease rate highly depends on the data geometry, making its theoretical evaluation difficult. The numerical results in Figure 11 show a fast decay towards zero of the interpolation volume for all models, already tiny from $M = 500$ trees. Furthermore, it seems to converge to the theoretical bound (dotted line) derived in Proposition 6.1 for an infinite RF with a max-feature parameter equal to 1.

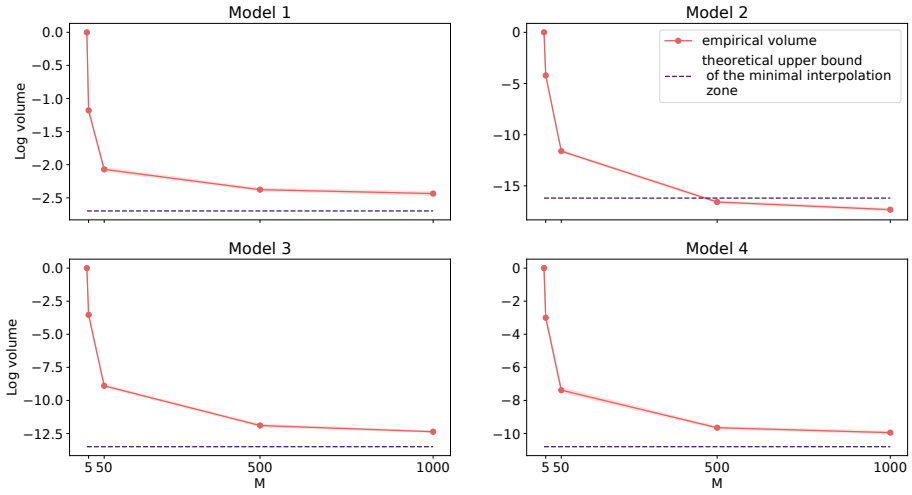

Figure 11: Log volume of Breiman RF interpolation zone w.r.t. the number M of trees. RF parameters: no bootstrap, max features = 1. Mean over 10 tries (bold line) and std (filled zone). Sample size $n = 500$.

B.2.3 Analysis of the interpolation property of Breiman RF with bootstrap

In this experiment, we try to measure how close a Breiman RF with bootstrap on is from exactly interpolating (with other parameters being 500 trees, max-depth set to None, max-features= d). To this end, we measure the difference between the true train labels (the Y_is) and the predicted ones (the $\hat{Y}_i$s) by computing

$$I_{\text{loss}} := \frac{1}{n} \sum_{i=1}^{n} \frac{|Y_i - \hat{Y}_i|}{Y_i}.$$

The closer is this quantity to 0, the closer is the forest from interpolating. On Figure 12, we plot different quantiles of the above quantity as n varies.

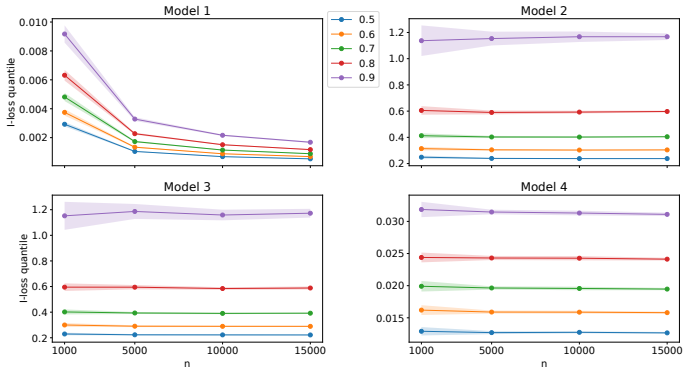

Figure 12: I_{loss} of a Breiman RF w.r.t sample size n. RF parameters: 500 trees, bootstrap on, max-features= d, max-depth set to None. Mean over 30 tries (doted lines) and std (filled zones).

For instance, if we take the 0.8-quantile in red on Figure 12 and look at the upper-right plot (model 2), we read that the I_{loss} roughly equals 0.6 for 80% of the points. This quantity seems globally constant in n. Finally, the quantiles are smaller in the case of a strong signal-to-noise ratio (models 1 and 4) than in the case of a bigger one (models 2 and 3).

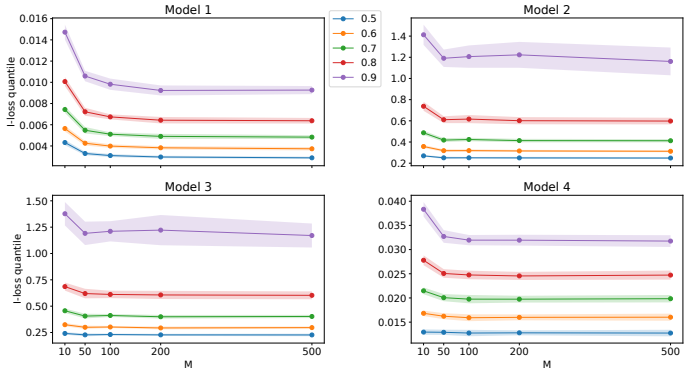

Figure 13: I_{loss} of a Breiman RF w.r.t number of trees. Parameters: bootstrap on, max-features= d, max-depth set to None. Sample size $n = 1000$. Mean over 30 tries (doted lines) and std (filled zones).

On Figure 13, we also plot the quantiles of the I_{loss} for the four different models while the number of trees varies. Adding trees does not significantly change the value of the different quantiles.