Is interpolation benign for random forest regression?

Ludovic Arnould, Claire Boyer, Erwan Scornet

Is interpolation benign for random forest regression?

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 (RF). In this paper, we study the trade-off between interpolation and consistency for several types of RF algorithms. Theoretically, we prove that interpolation regimes and consistency cannot be achieved simultaneously for several non-adaptive RF. 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 the interpolating regime. This is the first result conciliating interpolation and consistency for RF, highlighting that the averaging effect introduced by feature randomization is a key mechanism, sufficient to ensure the consistency in the interpolation regime and beyond. Numerical experiments show that Breiman's RF 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, giving a necessary condition for exact interpolation and consistency to occur in conjunction.

1 Introduction

Random Forests [RF, 9] 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 [16]. 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, 28, 20] and investigated in the context of neural networks [6]. 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 [12].

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 discussed by [27] and [31]. While [27] criticize the relevancy of interpolating random forests, Wyner et al. [31] 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 [11] and tree depth [32]. This phenomenon can be put in parallel with the results proved in [13] and [7] 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. More recently, Wang and Scott [30] showed the consistency of interpolating kernel methods, defined on Riemannian manifolds, whose kernels can be written as weighted random partition kernels on the sphere (similarly to the kernel random forest methods defined in Section 4).

Contributions and outline In this paper, we study the trade-off between interpolation and consistency in the context of regression, for different types of RF:

- Centered RF (Section 3). We prove theoretically that interpolation regimes and consistency cannot be achieved simultaneously for non-adaptive centered RF . The major problem arises from empty cells in tree partitions. Therefore, we also study a slightly modified Centered RF that does not take into account empty cells;
- Kernel RF (Section 4). We then study a more refined version of the CRF, the Kernel Random Forest (KeRF), built by averaging over all connected data points. By neglecting empty cells, this method is consistent for larger tree depths, but does not meet the exact interpolation requirement yet;
- Median RF (Section 5). Since adaptivity seems to be the cornerstone to conciliate interpolation and consistency, we study the interpolating Median RF, which is proved to be consistent in the exact interpolation regime . For the first time, it is shown that the averaging effect of the feature randomization inside RF (without boostrap) is sufficient to "average the noise out" (interpolating trees being sensitive to the noise), i.e. to decrease the variance towards 0. The bias of interpolating trees can be still classically controlled;

- Breiman RF (Section 6). Numerical experiments show that Breiman RF are consistent when exactly interpolating, i.e. when the whole data set is used to build each fully-grown tree (no bootstrap). It seems that the key randomization mechanism at work in RF is sufficient to reach consistency in spite of interpolation. Finally, we prove that the volume of the interpolation zone (where noise sensitivity is maximum) for an infinite Breiman RF tends to 0 at an exponential rate in the dimension d. This supports the idea that the decay of the interpolation volume could be fast enough to retrieve consistency despite interpolation.

		Conditions for consistency			
		Regardless of the noise scenario		In a noisy scenario	
		Managing the empty cells issue	Controlling the bias	Controlling the variance	Decreasing volume of the interpolation zone
Mean interpolation regime (non-adaptive RF)	Centered RF	✗	✓	✓	
	Void-free CRF	✓	✓	?	
	Centered KeRF	✓	✓	✓	
Exact interpolation (semi-adaptive and adaptive RF)	Median RF	✓	✓	✓	✓
	Breiman RF	✓	?	?	✓

Figure 1: Summary of theoretical contributions.

Please refer to Figure 1 for an overview of theoretical contributions. All proofs and details on numerical experiments are given in Appendix B and C.

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 a random noise satisfying, almost surely, $\mathbb{E}[\varepsilon|X] = 0$ and $\mathbb{V}[\varepsilon|X] \leq \sigma^2 < \infty$, for some $\sigma^2 \geq 0$. Given an input $x \in [0,1]^d$, the goal is to estimate the associated response $f^\star(x)$. We measure the performance of an estimator f_n via its *excess risk*, defined as $\mathcal{R}(f_n) := \mathbb{E}\left[(f_n(X) - f^\star(X))^2\right]$, and its consistency property.

Definition 2.1 (Consistency)**.** An estimator f_n is **consistent** when $\lim_{n\to\infty} \mathcal{R}(f_n) = 0$.

Estimator A Random Forest (RF) is a predictor consisting of a collection of M randomized trees [see 10, 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 letting M tending to infinity, we can consider the *infinite* forest estimate, $f_{\infty,n}(x) = \mathbb{E}_\Theta[f_n(x, \Theta)]$, which has also played an important role in the theoretical understanding of random forests [see 25, for more details]. Here, $\mathbb{E}_\Theta$ denotes the expectation w.r.t. Θ, conditional on $\mathcal{D}_n$.

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 RF, 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 2.2 ((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 2.3. In a classification setting, it is possible to obtain pure leaves with more than one point per cell (see [22] for more details).

3 Centered RF

We start our analysis of interpolation and consistency of RF with the simple yet widely studied Centered Random Forest [CRF, see 8]. CRF are ensemble methods 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. Although simpler, the study of the mechanisms at hand in non-adaptive RF already provides good insights about the inner behaviour of more general RF.

3.1 Interpolation in CRF

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 2.2) 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 $\alpha_n \in \mathbb{N} \setminus \{0, 1\}$, one has*

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

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 $\alpha_n \geq 1$,*

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

Therefore, the condition $\alpha_n = \omega(n)$ (corresponding to the interpolation of a single centered tree with high probability) is necessary to ensure that w.h.p., the RF 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

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

points falling into the leaf L_i is denoted $N_n(L_i)$. Since X is uniformly distributed over $[0,1]^d$, then, for all $i = 1, \dots, 2^k$,

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

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, i.e. if and only if $k_n \geq \log_2 n$.

By Equation (2), the mean interpolation regime implies that for all leaves L_i, $\mathbb{E}\left[N_n(L_i)\right] \leq 1$: one could say that trees interpolate in expectation, in the mean interpolation regime.

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

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. Indeed, 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.

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 predictions in empty leaves are arbitrary set to 0. Denoting $\Lambda_n(x, \boldsymbol{\Theta}_M)$ the number of non-empty leaves containing x in the forest with trees $\Theta_1, \dots, \Theta_M$, the void-free CRF is written as

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

The problematic terms that arise in the theoretical derivations of classical CRF vs. void-free CRF are of different natures: the probability $\mathbb{P}\left(N_n(X, \boldsymbol{\Theta}_M) = 0\right)$ of falling into an empty leaf in a random tree of an (infinite) CRF compared to the probability $\mathbb{P}\left[\forall m \in \{1, \dots, M\}, N_n(X, \Theta_m) = 0\right]$ of falling into empty leaves in all trees in the (infinite) CRF. Lemma 3.6 below controls this last term.

Lemma 3.6. *Consider a finite void-free CRF $f^{\mathrm{VF}}_{M,n}(\cdot, \mathbf{\Theta}_M)$ of depth $k \in \mathbb{N}$. Let $x \in [0,1]^d$ and denote $\mathcal{E}_{M,n}(x)$ the event "for all $m \in \{1, \dots, M\}, N_n(x, \Theta_m) = 0$". Then,*

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

Consequently, if $k = \lfloor \log_2(n) \rfloor$ and $M_n = \omega(n^{\log_2 d})$, then $\lim_{n\to\infty} \mathbb{P}\left(\mathcal{E}_{M_n,n}(x)\right) = 0$.

As previously, the infinite void-free CRF is defined as $f^{\mathrm{VF}}_{\infty,n}(x) = \mathbb{E}_\Theta\left[f_n(x,\Theta) | N_n(x,\Theta) > 0\right]$.

Theorem 3.7. *Assume that $f^\star$ has bounded partial derivatives. Then, the infinite void-free-CRF of depth $k = \lfloor \log_2 n \rfloor$ is consistent in a noiseless setting ($\sigma = 0$), and, for all $n > 1$,*

$$\mathcal{R}\left(f^{\mathrm{VF}}_{\infty,n}(X)\right) \leq \qquad C_d \left(\frac{n}{\log_2 n}\right)^{2\log_2\left(1-\frac{1}{2d}\right)} + (C_d + 2)\, n^{-1/(2\ln 2)},$$

where $C_d = 4d\left(\sum_{j=1}^d ||\partial f_j^\star||^2_\infty\right)$.

The overall rate is of order $O\left(n^{2\log(1-1/2d)}\right)$ which is a typical approximation rate for CRF, see Klusowski [19]. As a matter of fact, Theorem 3.7 highlights that empty cells do not limit the performance of the void-free-CRF in the mean interpolation regime.

However, this construction introduces a conditioning over $N_n(x,\Theta) > 0$ that prevents us from bounding the variance in the case of noisy samples. Therefore, in the next section, we analyze Centered Kernel RF (KeRF) with a different aggregation rule (empty cells still being neglected).

4 Centered kernel RF

As formalized in [15] 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.

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

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). As earlier, the infinite KeRF is defined as,

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

where $K_n(x,z) = \mathbb{P}_\Theta\left[z \in A_n(x,\Theta)\right]$ is the probability that x and z are in the same cell w.r.t. a tree built according to Θ [see 26, for details].

Interpolation conditions Since KeRF aggregates centered trees as CRF (but in a different way), the results of Section 3 can be extended to KeRF: (*i*) the mean interpolation regime is met for centered trees (hence for KeRF) when $k_n \geq \log_2 n$; (*ii*) 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.

Consistency We study the convergence of the centered KeRF 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 independent from X 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 $d > 5$, for all n large enough,*

$$\mathcal{R}(f_{\infty,n}^{\mathrm{KeRF}}) \leq 8L^2 d^2 n^{2\log_2(1-\frac{1}{d})} + C_d (\log_2 n)^{-\frac{d-5}{6}} (\log_2(\log_2 n))^{d/3},$$

where $C_d > 0$ is a constant dependent on σ^2 and made explicit in the proof.

Theorem 4.1 states that the infinite centered KeRF estimator is consistent as soon as $d > 5$, with a slow convergence rate of $\log(n)^{-(d-5)/6}$. The proof is based on the general paradigm of bias-variance trade-off and is adapted from [26]. At first sight, one might think that the rate becomes better as the dimension d increases. However, the constant term highly depends on the dimension, so that the established bound should be regarded for a 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 [21] provides a lower bound on the convergence rate of a deep 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 [7]. Since KeRF are kernel estimators, one can wonder how sharp is our bound (Theorem 4.1) compared to that of [7], which is minimax. Due to the spikiness of the singular kernel studied in [7], 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 [26] when optimizing this depth parameter, but leaving the interpolation world.

We numerically assess the performance of KeRF in the mean interpolation regime (see Appendix C).

5 Semi-Adaptive RF: Median RF

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

The Median RF, studied e.g. in [14, 19], is composed of median trees that first randomly choose the direction to cut over and then cut at the median of the data points contained in the current cells. In our analysis, for any cell containing n_c observations, the median is set as the middle of the segment of two consecutive order statistics: $X_{(n_c/2)}$ and $X_{(n_c/2+1)}$ for an even number of observations, $X_{(\frac{n_c-1}{2})}$ and $X_{(\frac{n_c+1}{2})}$ otherwise.

5.1 Consistency

In order to obtain consistency for an adaptive RF, one needs to control two terms: the bias and the variance terms. On the one hand, the bias is roughly controlled by the diameter of the leaf times the supremum of the derivatives of $f^\star$ in the leaf. In the interpolating regime, the depth is maximum so the diameter of the leaf is minimum and therefore the bias is smoothly upper bounded.

On the other hand, a "low" depth regime is usually required to control the variance term, so that each leaf of each tree contains an infinite number of points when n tends to $+\infty$. This directly "averages the noise out" and decreases the variance towards 0 within each tree. However, in the interpolating case, each leaf contains only one point and we can only rely on the averaging effect of the RF, induced by the random splitting mechanism, to upper bound the variance. Studying the effect of the random splitting mechanism in full generality remains challenging. However, as increasing the dimension also increases the diversity of the trees within the RF, it should naturally be easier to control the variance of an interpolating Median RF in an asymptotic high-dimensional setting, as we prove and discuss in Appendix B.5.1.

The following theorem establishes the consistency of the interpolating Median RF in the general setting of noisy data and fixed input dimension.

Theorem 5.1. *Suppose that $f^\star$ has bounded partial derivatives and that n is a power of two. Then, the infinite interpolating Median RF $f_{\infty,n}^{\mathrm{MedRF}}$ is consistent and verifies:*

$$\mathcal{R}\left(f_{\infty,n}^{\mathrm{MedRF}}\right) \leq C_1 d \left(\sum_{\ell=1}^{d} ||\partial_\ell f^\star||_\infty^2\right) \left(1 - \frac{3}{4d}\right)^{\log_2 n} + \sigma^2 C_{2,d} (\log_2 n)^{-(d-1)/2},$$

where C_1 and $C_{2,d}$ are explicit constants, the former being independent of the dimension d (see the proof for the exact computations).

The control of the bias term follows the general approach used in [14] with substantial technical refinements. On the other hand, we propose a more general approach for the control of the variance

inspired by [8, 19], where we derive explicit bounds specifically designed for Median RF. Note that the consistency achieved by Median RF cannot be obtained for CRF under the interpolation regime due to the non-negligible probability of falling into empty cells (see Proposition 3.2).

Theorem 5.1 is the first result ensuring consistency of RF despite exact interpolation. It is even more impressive considering that bootstrap is off so that the averaging process in the RF is only due to feature subsampling. More specifically, when dealing with interpolating trees, the variance reduction does not come from averaging many points in the leaf of a given tree anymore (since the tree depth is no longer limited), but results from averaging single points from the leaves of different trees.

If interpolation remains compatible with consistency in the case of Median RF, it nevertheless damages the convergence rate. Indeed, it has been proved that, for all α small enough, the convergence rate of Median RF with trees of depth $k = (1-\alpha)\log_2(n)$ is $n^{-\alpha}$ [see Theorem 3 in 19]. In the case of interpolating Median RF, Theorem 5.1 highlights a phase transition when $k = \log_2(n)$, as the convergence rate is driven by the variance term, which is of order $(\log_2 n)^{-(d-1)/2}$. While being very slow, this rate is close to the lower bound $(\log_2 n)^{-(d-1)}$ established for non-adaptive interpolating RF [21]. Actually, by assuming $\log_2(n) \geq d$, our proof can be directly modified so that our upper bound matches the lower bound of Lin and Jeon [21] [using the second statement of Lemma S.1 in 19, instead of the first one].

Note that Theorem 5.1 does not contradict Proposition 1 in [27], as the condition therein is not proved to be satisfied for interpolating median RF (nor for interpolating CRF).

We also provide numerical experiments (resp. Section 5.2 and C.1.2) that illustrate the consistency of the interpolating Median RF.

5.2 Volume of the interpolation area

In this section, we aim at quantifying the volume of the interpolation area of a Median RF, which is a prerequisite for the RF consistency. To pursue our analysis, we first give a rigorous definition of the interpolation area.

Definition 5.2. The ***interpolation area*** is the subspace of $[0,1]^d$ where the forest prediction depends only on one training point. For a given forest $f_{M,n}(.,\boldsymbol{\Theta}_M)$, the interpolation area is denoted by[2]

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

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

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

to be quite a challenging task. Therefore, we focus our study on the *core interpolation area* $\mathcal{A}_{min}$ written as

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

The area $\mathcal{A}_{min}$ is 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 median 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 Lebesgue measure (denoted by μ) of the core interpolation area $\mathcal{A}_{min}^{\text{MedRF}}$ of an infinite Median RF.

Proposition 5.3. *For all $n \geq 2$, for all $d \geq 2$, consider an infinite Median RF. Then,*

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

The volume of the core interpolation area of an infinite Median RF tends to 0 polynomially in n and exponentially in d.

Remark 5.4. 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: the volume of the area where the prediction involves only a finite number of points (*a fortiori* the interpolation zone) should tend to 0. Indeed, by decomposing the risk as $R(f_n(X)\mathbb{1}_{X \in \mathcal{A}_{min}^{\text{MedRF}}}) + R(f_n(X)\mathbb{1}_{X \notin \mathcal{A}_{min}^{\text{MedRF}}})$, the first term is at least of the order $\sigma^2\mu(\mathcal{A}_{min}^{\text{MedRF}})$. Therefore, it is not possible to cancel out the noise of the training dataset when only a finite number of points is used for the prediction. The noise in such an area remains of order σ^2. Proposition 5.3 portends the predominant *self-averaging* property of adaptive RF, and hence underpins the idea of good capabilities of Median RF in interpolation regimes.

6 Breiman RF

The widely-used Breiman RF is composed of several CART [10], each one trained on a bootstrap sample, 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. Breiman RF exhibit excellent 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.

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., 18]. Note that by omitting the bootstrap in the RF construction, the interpolation of Breiman RF directly results from aggregating fully-grown trees.

Volume of the interpolation zone As shown in the next proposition, the volume of the core interpolation area of Breiman RF 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 5.4). 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, easy 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.

In Appendix C, we numerically evaluate the volume of the interpolation zone and compare it to the theoretical bounds in Proposition 6.1.

Empirical study of 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 parameters 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 average 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 $\{1, \ldots, d\}$ 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 2, we measure the excess risk of two RFs with 2000 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`).

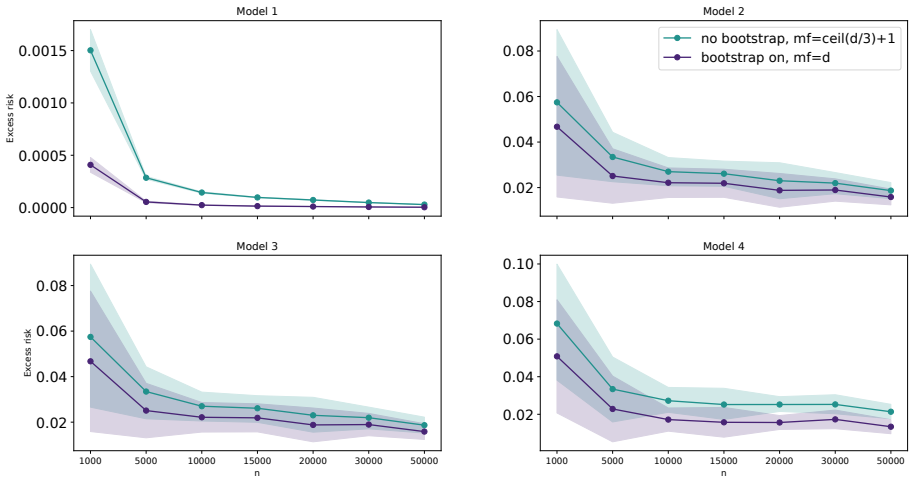

Figure 2: Consistency of two Breiman RF: excess risk w.r.t. sample size n. Mean over 10 tries (bold lines) and mean $\pm$ std (filled zone), when using 2000 trees per forest, and max-depth=None. See Appendix C for the model definitions.

In Figure 2, 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 25, 22, 23, 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 18, 19 in Section C.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 16. This is in line with the vision of a *locally spiky* estimator developed in [31] 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.

Remark 6.2. One of the advantage of using deep (interpolating) trees is that it allows the RF to build more diversified trees. Indeed, the number of possible trees roughly grows exponentially with regard to the depth (also depending on n, d and the `max-features` parameter). Especially when `max-features` is low, this should improve the averaging effect of the RF which is of particular interest when dealing with noisy data.

In this regard, Breiman RF with *max-features*$= \lceil d/3 \rceil$ are similar to interpolating *spiky* non-singular kernel methods, as studied in [7], 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: when dealing with non-adaptive RF (CRF), empty cells prevent consistency; so that aggregating only non-empty leaves (void-free CRF) leads to convergence rates, only in a noiseless scenario. In a noisy setting, the kernel RF aggregates leaves differently (also avoiding empty ones). For kernel RF, we establish a (slow) consistency rate in the mean interpolation regime. We then study semi-adaptive RF that are closer to those used in practice and that present the advantage of being able to exactly interpolate the training data. The convergence of the median RF in the exact interpolation regime is established, showing the power of such architecture (even when used without bootstrap). Our study also shows that a prerequisite for consistency is that the minimal interpolation zone tends to zero as n tends to infinity. We theoretically analyze this quantity for median and Breiman forests, emphasizing that interpolation might occur in conjunction with consistency if the volume of such areas vanishes fast enough. An experimental study supports the concomitance of consistency and interpolation in Breiman RF, when no bootstrap step is involved.

Contrary to Nadaraya-Watson methods involving singular kernels that interpolate regardless of the bandwidth parameter, RF interpolate only for a specific choice of the depth, thus restricting the regime in which interpolation and consistency occur in concordance. Overall, most simple RF versions were relevant to study RF consistency when the tree depth was limited but are not actually sufficient to handle deeper trees corresponding to interpolation regimes. For adaptive forests, increasing the tree depth towards the interpolation regime results in a reduced bias, and the variance reduction phenomenon only results from the split randomization effect. The higher the dimension, the more diversified the trees, the stronger the averaging effect and the variance reduction. Analyzing the strength of this phenomenon, which highly depends on the very shape of tree partitions, is the cornerstone to prove the consistency of adaptive RF in a general regression setting. We believe that interpolation remains benign for the consistency of adaptive RF, but can damage their convergence rate (this was the case for KeRF in the mean interpolation regime and for Median RF in the exact interpolation regime), at least when bootstrap is not used.

The analysis of the interpolation zone of RF introduced in this article is an important tool 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, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[7] 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.

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

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

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

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

[12] 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.

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

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

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

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

[17] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986, 2022. doi: 10.1214/21-AOS2133. URL `https://doi.org/10.1214/21-AOS2133`.

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

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

[20] 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.

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

[22] 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.

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

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

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

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

[27] Cheng Tang, Damien Garreau, and Ulrike von Luxburg. When do random forests fail? *Advances in neural information processing systems*, 31, 2018.

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

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

[30] Yutong Wang and Clayton D Scott. Consistent interpolating ensembles via the manifold-hilbert kernel. *arXiv preprint arXiv:2205.09342*, 2022.

[31] 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.

[32] 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 Summary of contributions

		Conditions for consistency			
		Regardless of the noise scenario		In a noisy scenario	
		Managing the empty cells issue	Controlling the bias	Controlling the variance	Decreasing volume of the interpolation zone
Mean interpolation regime (non-adaptive RF)	Centered RF	✗	✓	✓	
	Void-free CRF	✓	✓	?	
	Centered KeRF	✓	✓	✓	
Exact interpolation (semi-adaptive and adaptive RF)	Median RF	✓	✓	✓	✓
	Breiman RF	✓	?	?	✓

Figure 3: Summary of theoretical contributions

B Proofs

B.1 Reminders and notations

Tree and RF estimator: We recall the prediction of the given by the j-th tree of the RF at point x:

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

where $A_n(x, \Theta_j)$ is the cell containing x and $N_n(x, \Theta_j)$ is the number of points falling into $A_n(x, \Theta_j)$. It is also written as follows:

$$f_n(x, \Theta_j) = \sum_{i=1}^{n} W_{ni}(x, \Theta_j) Y_i,$$

where $W_{ni}(x, \Theta_j) = \frac{\mathbb{1}_{X_i \in A_n(x,\Theta_j)}}{N_n(x,\Theta_j)} \mathbb{1}_{N_n(x,\Theta_j)>0}$. 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 $\mathbf{\Theta}_M := (\Theta_1, ..., \Theta_M)$.

B.2 Proofs of Section 3 (Centered RF)

B.2.1 Proof of Lemma 3.1 (Link between tree and forest interpolation)

First, it is clear that if all trees of a forest interpolate, the forest interpolates. Now, suppose that the forest $f_{M,n}^{\mathrm{CRF}}$ interpolates a training point $X_s, s \in \{1, \dots, n\}$. Then, by definition of $f_{M,n}^{\mathrm{CRF}}$,

$$\begin{aligned} f_{M,n}^{\mathrm{CRF}}(X_s, \boldsymbol{\Theta_M}) &= \frac{1}{M} \sum_{j=1}^{M} \sum_{i=1}^{n} Y_i W_{ni}(X_s, \Theta_j) \\ &= \sum_{i=1}^{n} Y_i \left(\frac{1}{M} \sum_{j=1}^{M} W_{ni}(X_s, \Theta_j) \right) \\ &= Y_s, \end{aligned}$$

where $W_{ni}(X_s, \Theta_j) := \frac{\mathbb{1}_{X_i \in A_n(X_s, \Theta_j)}}{N_n(X_s, \Theta_j)} \mathbb{1}_{N_n(X_s, \Theta_j) > 0}$. Consequently,

$$f_{M,n}^{\mathrm{CRF}}(X_s, \boldsymbol{\Theta_M}) = Y_s \tag{4}$$

$$\iff \quad Y_s \left(\frac{1}{M} \sum_{j=1}^{M} W_{ns}(X_s, \Theta_j) - 1 \right) + \sum_{i \neq s} Y_i \left(\frac{1}{M} \sum_{j=1}^{M} W_{ni}(X_s, \Theta_j) \right) = 0. \tag{5}$$

For (5) to hold almost surely, it is necessary that it holds conditional on $X_1, \dots, X_n, \Theta_1, \dots, \Theta_M$. Since, for all $j \in \{1, \dots, M\}$, the terms $W_{ni}(X_s, \Theta_j)$ are measurable with respect to $X_1, \dots, X_n, \Theta_1, \dots, \Theta_M$ and Y_s is independent of $(Y_i, i \neq s)$ given $X_1, \dots, X_n, \Theta_1, \dots, \Theta_M$, equality (5) leads to, for all $i \neq s$,

$$\frac{1}{M} \sum_{j=1}^{M} W_{ns}(X_s, \Theta_j) = 1, \quad \text{and} \quad \frac{1}{M} \sum_{j=1}^{M} W_{ni}(X_s, \Theta_j) = 0. \tag{6}$$

Since all weights $W_{ni}(X, \Theta)$ take values in $[0, 1]$, we have, for all $j \in \{1, \dots, M\}$ and for all $i \neq s$

$$W_{ns}(X_s, \Theta_j) = 1 \quad \text{and} \quad W_{ni}(X_s, \Theta_j) = 0. \tag{7}$$

Finally, for all $j \in \{1, \dots, M\}$, the prediction of the jth tree at X_s is given by

$$f_n^{\mathrm{CRF}}(X_s, \Theta_j) = \sum_{i=1}^{n} W_{ni}(X_s, \Theta_j) Y_i \tag{8}$$

$$= Y_s, \tag{9}$$

and therefore all trees of the forest interpolate the point X_s.

B.2.2 Proof of Proposition 3.2 (Probability of interpolation for a centered tree)

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

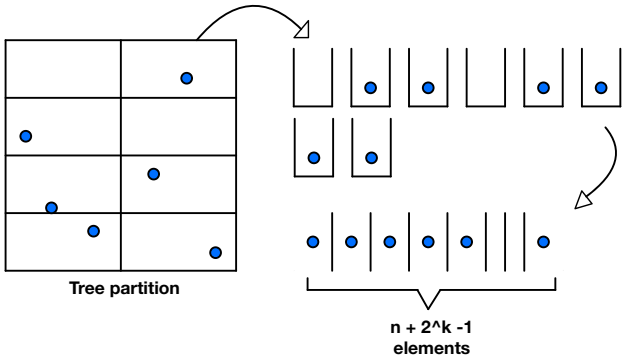

Figure 4: Computing the interpolation probability (depth $k = 3$, $n = 6$)

one point with $2^k \geq n$ as can be seen on Figure 4. Recalling that $\mathcal{I}_T$ is the event "a centered tree of depth k_n interpolates the training data", we have

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

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

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

In the general case where $k = \lfloor \log_2(\alpha_n n) \rfloor$, that is $\alpha_n n/2 \leq 2^k \leq \alpha_n n$, we can lower bound the probability of the event $\mathcal{I}_T$ as

$$\begin{aligned}\mathbb{P}\left(\mathcal{I}_T\right) &= \frac{2^k \times (2^k-1) \times \ldots \times (2^k-n+1)}{(2^k+n-1) \times (2^k+n-2) \times \ldots 2^k} \geq \left(\frac{2^k-n+1}{2^k+n-1}\right)^n \geq \left(\frac{2^k-n}{2^k+n}\right)^n \\ &\geq \exp\left(n \log\left(\frac{2^k-n}{2^k+n}\right)\right) \geq \exp\left(n \log\left(1-\frac{2n}{2^k+n}\right)\right) \geq \exp\left(-n\left(\frac{2}{\frac{2^k}{n}-1}\right)\right) \\ &\geq \exp\left(-\left(\frac{4n}{\alpha_n-2}\right)\right)\end{aligned}$$

since $\log(1-x) \geq -x/(1-x)$ and provided that $\alpha_n > 2$ for the last inequality. To upper bound the

probability, note that, for all $r \in \{1, \dots, \lfloor n/2 \rfloor\}$

$$\frac{2^k - n + r}{2^k + n - r} \leq \frac{2^k - n + \frac{n}{2}}{2^k + n - \frac{n}{2} - 1} \leq \frac{2^k - \frac{n}{2}}{2^k + \frac{n}{2} - 1},$$

and, for all $r \in \{1, ..., n\}$,

$$\frac{2^k - n + r}{2^k + n - r} \leq 1.$$

Therefore, one can also upper bound the probability as

$$\begin{aligned}
\mathbb{P}\left(\mathcal{I}_T\right) &= \frac{2^k \times (2^k - 1) \times \ldots \times (2^k - n + 1)}{(2^k + n - 1) \times (2^k + n - 2) \times \ldots 2^k} \\
&\leq \left(\frac{2^k - \frac{n}{2}}{2^k + \frac{n}{2} - 1}\right)^{\lfloor n/2 \rfloor} \\
&\leq \exp\left(\left\lfloor \frac{n}{2} \right\rfloor \log\left(1 - \frac{n-1}{2^k + \frac{n}{2} - 1}\right)\right) \\
&\leq \exp\left(-\left\lfloor \frac{n}{2} \right\rfloor \left(\frac{\frac{n}{2}}{2^k + \frac{n}{2} - 1}\right)\right) \\
&\leq \exp\left(-\left\lfloor \frac{n}{2} \right\rfloor \left(\frac{\frac{1}{2}}{\frac{2^k}{n} + \frac{1}{2}}\right)\right) \\
&\leq \exp\left(-\left\lfloor \frac{n}{2} \right\rfloor \left(\frac{1}{2\alpha_n + 1}\right)\right),
\end{aligned}$$

for all $n \geq 2$. Finally, for all $n \geq 2$, and for all $\alpha_n > 2$,

$$\exp\left(-\frac{4n}{\alpha_n - 2}\right) \leq \mathbb{P}\left(\mathcal{I}_T\right) \leq \exp\left(-\left\lfloor \frac{n}{2} \right\rfloor \left(\frac{1}{2\alpha_n + 1}\right)\right).$$

B.2.3 Proof of Corollary 3.3 (Probability of interpolation for a CRF)

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.

B.2.4 Proof of Proposition 3.5 (CRF inconsistency)

Let $f_{\infty,n}^{\mathrm{CRF}}$ be an infinite CRF with each tree of depth $k_n \geq \log_2(\alpha_n n)$, that is each tree has at least $\alpha_n n$ leaves, with $\alpha_n n > 1$. Let X be uniformly distributed on $[0,1]^d$. We write $\bar{f}_{n,\infty}^{\mathrm{CRF}}(X) = \mathbb{E}\left[f_{\infty,n}^{\mathrm{CRF}}(X) | X, X_1, ..., X_n\right]$. Then, denoting $\mathcal{E}$ the event "$N_{n,\infty}(X) = 0$" (or equivalently, "X falls

into a non-empty leaf"),

$$\mathcal{R}(f^{\mathrm{CRF}}_{\infty,n}(X)) = \mathbb{E}\left[\left(f^{\mathrm{CRF}}_{\infty,n}(X) - f^\star(X)\right)^2\right] \tag{10}$$

$$\geq \mathbb{E}\left[\left(\bar{f}^{\mathrm{CRF}}_{n,\infty}(X) - f^\star(X)\right)^2\right] \tag{11}$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbb{E}_\Theta\left[W_{ni}(X,\Theta) f^\star(X_i)\right] - \left(\mathbb{1}_{\mathcal{E}} + \mathbb{1}_{\mathcal{E}^c}\right) f^\star(X)\right)^2\right] \tag{12}$$

$$= \mathbb{E}\left[\left(\mathbb{1}_{\mathcal{E}^c} \sum_{i=1}^{n} \mathbb{E}_\Theta\left[W_{ni}(X,\Theta)\left(f^\star(X_i) - f^\star(X)\right)\right] - \mathbb{1}_{\mathcal{E}} f^\star(X)\right)^2\right] \tag{13}$$

$$\geq \mathbb{E}\left[f^\star(X)^2 \mathbb{1}_{\mathcal{E}}\right] \tag{14}$$

$$\geq \mathbb{E}\left[f^\star(X)^2 \mathbb{P}\left(\mathcal{E}|X\right)\right]. \tag{15}$$

Besides,

$$\mathbb{P}\left(\mathcal{E}|X\right) = \mathbb{P}\left(N_{n,\infty}(X) = 0|X\right) \tag{16}$$

$$\geq \left(1 - \frac{1}{\alpha_n n}\right)^n, \tag{17}$$

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

$$\left(1 - \frac{1}{\alpha_n n}\right)^n = e^{n \log\left(1 - \frac{1}{\alpha_n n}\right)} \tag{18}$$

$$\geq e^{-\frac{n}{\alpha_n n - 1}}. \tag{19}$$

Thus,

$$\mathcal{R}(f^{\mathrm{CRF}}_{\infty,n}(X)) \geq e^{-\frac{n}{\alpha_n n - 1}} \mathbb{E}\left[f^\star(X)^2\right], \tag{20}$$

which tends to 0 if and only if α_n tends to zero as n tends to infinity. Since, by assumptions, α_n does not tend to zero and $\mathbb{E}\left[f^\star(X)^2\right] > 0$, the infinite CRF is inconsistent.

B.2.5 Proof of Lemma 3.6 (Probability of falling into an empty cell of the void-free CRF)

Recall that $\mathcal{E}_{M,n}(x)$ is the event "for all $m \in \{1, \ldots, M\}, N_n(x, \Theta_m) = 0$". We have

$$\mathcal{E}_{M,n}(x) = \bigcap_{j=1}^{M} \left\{N_n(x, \Theta_j) = 0\right\}. \tag{21}$$

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 be connected to a point

X_i for a certain configuration of cuts within a tree. We write $\mathcal{E}_{1,n}(x)$ the ($\mathcal{D}_n$-measurable) event $\{\forall\theta, N_n(x,\theta) = 0\}$. Consequently, we have $\mathcal{E}_{1,n}(x)^c = \{\exists\, \theta, N_n(x,\theta) \neq 0\}$. Using these notations, we obtain

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}_{M,n}(x)\right) &= \mathbb{P}\left(\mathcal{E}_{M,n}(x) \cap \mathcal{E}_{1,n}(x)\right) + \mathbb{P}\left(\mathcal{E}_{M,n}(X) \cap \mathcal{E}_{1,n}(x)^c\right) &(22)\\ &= \mathbb{P}\left(\mathcal{E}_{1,n}(x)\right) + \mathbb{P}\left(\mathcal{E}_{M,n}(x) \cap \mathcal{E}_{1,n}(x)^c\right) &(23)\end{aligned}$$

where the first probability term of the second line is a probability taken over $\mathcal{D}_n$ only, since $\mathcal{E}_{1,n}(x)$ does not depend on Θ. We control this probability thanks to the following Lemma.

Lemma B.1. *For all $x \in [0,1]^d$, we let $\mathcal{E}_{1,n}(x)$ be the event $\{\forall\theta, N_n(x,\theta) = 0\}$. Then, we have*

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

Proof. Let $x \in [0,1]^d$. The event $\mathcal{E}_{1,n}(x)$ happens if all the 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

$$Z_{c,k}(x) = \left\{z \in [0,1]^d \ : \exists\theta, z \in A_n(x,\theta)\right\}. \tag{24}$$

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 $Z_{c,k}$ of x is

$$\begin{aligned}\mu(Z_{c,k}(x)) &= \mu\left(\bigcup_{\substack{0\leq k_1,\ldots,k_d\leq k\\ \sum_j k_j = k}} A(k_1, ..., k_d, x)\right) &(25)\\ &\geq \mu\left(\bigcup_{\substack{0\leq k_1,k_2\leq k\\ k_1+k_2=k}} A(k_1, k_2, 0, ..., 0, x)\right). &(26)\end{aligned}$$

By σ-additivity of μ,

$$\begin{aligned}&\mu\left(\bigcup_{\substack{0\leq k_1,k_2\leq k\\ k_1+k_2=k}} A(k_1, k_2, 0, ..., 0, x)\right)\\ &= \mu\Big(A(k, 0, ..., 0, x)\Big) + \sum_{j=1}^{k} \mu\left(A(k-j, j, 0, ..., 0, x) \setminus \bigcup_{\ell=0}^{j-1} A(k-\ell, \ell, 0, ..., 0, x)\right). &(27)\end{aligned}$$

Given the shape of the cells $A(k-j, j, 0, ..., 0, x)$, for all $j \in \{1, \ldots, d\}$, we have (see Figure 5)

$$\begin{aligned}&A(k-j, j, 0, ..., 0, x) \setminus \bigcup_{\ell=0}^{j-1} A(k-\ell, \ell, 0, ..., 0, x)\\ =&A(k-j, j, 0, ..., 0, x) \setminus A(k-j+1, j-1, 0, ..., 0, x). &(28)\end{aligned}$$

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. $\mu\big((A(k-j, j, 0, ..., 0, x) \cap A(k-j+1, j-1, 0, ..., 0, x)\big) = \frac{\mu(A(k-j,j,0,\dots,0,x))}{2}$ as can be seen on Figure 5.

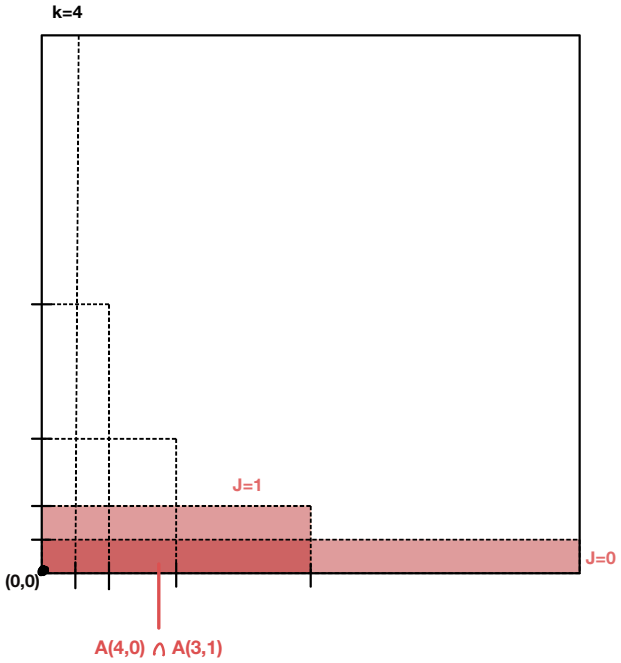

Figure 5: Volume of leaf intersection $\mu\big((A(k-j, j, x) \cap A(k-j+1, j-1, x)\big)$ in dimension 2 with $x = (0, 0)$, $k = 4$ cuts and $j \in \{0, 1\}$.

We deduce from these facts that, for all j,

$$\mu\big(A(k-j, j, 0, ..., 0, x) \setminus A(k-j+1, j-1, 0, ..., 0, x)\big) = \frac{\mu(A(k-j, j, 0, ..., 0, x))}{2} \tag{29}$$

$$= 2^{-(k+1)} \tag{30}$$

Hence, combining equations (27), (28) and (29), we have

$$\mu\left(\bigcup_{\substack{0 \le k_1, k_2 \le k \\ k_1 + k_2 = k}} A(k_1, k_2, 0, ..., 0, x)\right) = 2^{-k} + k2^{-(k+1)}. \tag{31}$$

Consequently, using inequality (26),

$$\mu(Z_{c,k}(x)) \geq k2^{-(k+1)}. \tag{32}$$

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

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_{1,n}(x)\right) &= \mathbb{P}\left(\forall i \in \{1,\dots,n\}, X_i \notin Z_{c,k}(x)\right) && (33)\\
&= \left(1-\mu(Z_{c,k}(x))\right)^n && (34)\\
&\leq \left(1-k2^{-(k+1)}\right)^n && (35)\\
&= e^{n\log(1-k2^{-(k+1)})} && (36)\\
&\leq e^{-\frac{kn}{2^{k+1}}}. && (37)
\end{aligned}$$

□

Regarding the second term of (23), we have

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_{M,n}(x)\cap\mathcal{E}_{2,n}(x)\right) &= \mathbb{P}\left(\left(\bigcap_{j=1}^{M} N_n(x,\Theta_j)=0\right)\bigcap\left(\exists i\in\{1,\dots,n\}, X_i\in Z_{c,k}(x)\right)\right) && (38)\\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\exists i\in\{1,\dots,n\}, X_i\in Z_{c,k}(x)}\mathbb{1}_{\bigcap_{j=1}^{M} N_n(x,\Theta_j)=0}|\mathcal{D}_n\right]\right] && (39)\\
&= \mathbb{E}\left[\mathbb{1}_{\exists i\in\{1,\dots,n\}, X_i\in Z_{c,k}(x)}\mathbb{P}\left(\bigcap_{j=1}^{M} N_n(x,\Theta_j)=0|\mathcal{D}_n\right)\right] && (40)\\
&= \mathbb{E}\left[\mathbb{1}_{\exists i\in\{1,\dots,n\}, X_i\in Z_{c,k}(x)}(1-p_n)^M\right] && (41)
\end{aligned}$$

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, if $\exists i\in\{1,\dots,n\}, X_i\in Z_{c,k}(x)$, 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). Hence,

$$\mathbb{1}_{\exists i\in\{1,\dots,n\}, X_i\in Z_{c,k}(x)}(1-p_n)^M \leq (1-d^{-k})^M, \tag{42}$$

which leads to

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_{M,n}(x)\cap\mathcal{E}_{1,n}(x)^c\right) &\leq \left(1-d^{-k}\right)^M && (43)\\
&\leq e^{-Md^{-k}}. && (44)
\end{aligned}$$

Finally, gathering Lemma B.1 and inequality (44) yields

$$\mathbb{P}\left(\mathcal{E}_{M,n}(x)\right) \leq e^{-\frac{kn}{2^{k+1}}} + e^{-Md^{-k}}. \tag{45}$$

B.2.6 Proof of Proposition 3.7 (Consistency of void-free-CRF in a noiseless setting)

Recall that, in a noiseless setting (that is, for all i, $Y_i = f^\star(X_i)$), the risk of the Void-free CRF can be written as

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

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

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\mathbb{1}_{\mathbb{P}_\Theta(N_n(X,\Theta)>0)>0}}{\mathbb{P}_\Theta\left(N_n(X,\Theta>0)\right)}\sum_{i=1}^n f^\star(X_i)\mathbb{E}_\Theta[W_{ni}(X,\Theta)\mathbb{1}_{N_n(X,\Theta)>0}] - f^\star(X)\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{\mathbb{1}_{\mathbb{P}_\Theta(N_n(X,\Theta)>0)>0}}{\mathbb{P}_\Theta\left(N_n(X,\Theta>0)\right)}\sum_{i=1}^n \left(f^\star(X_i) - f^\star(X)\right)\mathbb{E}_\Theta[W_{ni}(X,\Theta)\mathbb{1}_{N_n(X,\Theta)>0}] - f^\star(X)\mathbb{1}_{\mathbb{P}_\Theta(N_n(X,\Theta)>0)=0}\right)^2\right] \\
&\leq 2\mathbb{E}\left[\left(\frac{\mathbb{1}_{\mathbb{P}_\Theta(N_n(X,\Theta)>0)>0}}{\mathbb{P}_\Theta\left(N_n(X,\Theta>0)\right)}\sum_{i=1}^n \left(f^\star(X_i) - f^\star(X)\right)\mathbb{E}_\Theta[W_{ni}(X,\Theta)\mathbb{1}_{N_n(X,\Theta)>0}]\right)^2\right] \\
&\quad + 2\mathbb{E}\left[\left(f^\star(X)\mathbb{1}_{\mathbb{P}_\Theta(N_n(X,\Theta)>0)=0}\right)^2\right] \qquad (46)
\end{aligned}$$

The second term of the last inequality verifies

$$\mathbb{E}\left[\left(f^\star(X)\mathbb{1}_{\mathbb{P}_\Theta(N_n(X,\Theta)>0)=0}\right)^2\right] \leq ||f^\star||^2_\infty \mathbb{P}\left(\mathbb{P}_\Theta\left(N_n(X,\Theta)>0\right)=0\right). \qquad (47)$$

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 B.1:

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

Denoting by $\mu\left(A_n^{(j)}(x,\Theta)\right)$ the length of the jth side of the cell containing x and following a

computation from [19],

$$\sum_{i=1}^{n} W_{ni}(X,\Theta)|f^\star(X)-f^\star(X_i)|\mathbb{1}_{N_n(X,\Theta)>0} \leq \sum_{i=1}^{n} W_{ni}(X,\Theta)\left(\sum_{j=1}^{d}||\partial_j f^\star||_\infty |X_i^{(j)}-X^{(j)}|\right)\mathbb{1}_{N_n(X,\Theta)>0} \tag{48}$$

$$\leq \sum_{i=1}^{n} W_{ni}(X,\Theta)\mathbb{1}_{N_n(X,\Theta)>0}\sum_{j=1}^{d}||\partial_j f^\star||_\infty (b_j-a_j) \tag{49}$$

$$\leq \mathbb{1}_{N_n(X,\Theta)>0}\sum_{j=1}^{d}||\partial_j f^\star||_\infty \mu\left(A_n^{(j)}(X,\Theta)\right). \tag{50}$$

Therefore,

$$\mathbb{E}\left[\left(f_{\infty,n}^{\mathrm{VF}}(X)-f^\star(X)\right)^2\right] \leq 2\mathbb{E}\left[\left(\frac{1}{\mathbb{P}_\Theta(N_n(X,\Theta)>0)}\sum_{j=1}^{d}||\partial_j f||_\infty \mathbb{E}_\Theta\left[\mathbb{1}_{N_n(X,\Theta)>0}\mu\left(A_n^{(j)}(X,\Theta)\right)\right]\right)^2\right] + 2e^{-\frac{kn}{2^{k+1}}} \tag{51}$$

$$\leq 2d\sum_{j=1}^{d}||\partial f_j^\star||_\infty^2 \mathbb{E}\left[\frac{1}{\mathbb{P}_\Theta(N_n(X,\Theta)>0)^2}\mathbb{E}_\Theta\left[\mathbb{1}_{N_n(X,\Theta)>0}\mu\left(A_n^{(j)}(X,\Theta)\right)\right]^2\right] + 2e^{-\frac{kn}{2^{k+1}}}. \tag{52}$$

Note that the length $\mu\left(A_n^{(j)}(X,\Theta)\right)$ of the j-th side of the cell $A_n(X,\Theta)$ and the event $\{N_n(X,\Theta)>0\}$ are not independent conditional on $X_1,...,X_n,X$. Indeed, given the geometry of the dataset, it is possible that cutting along the jth direction isolates X from the dataset. Therefore its length should be computed conditional on the event $\{N_n(X,\Theta)>0\}$.

To this aim, we denote for all $\kappa\in\mathbb{N}$, $A_{n,\kappa}(X,\Theta)$ the cell containing X at depth κ in a centered tree built with the extra randomness Θ. Conditional on $N_n(X,\Theta)>0$, the jth direction can be chosen to split along if and only if it does not isolate X from the points of the dataset. Thus, we denote by $E_{n,\kappa}(j,X,\Theta)$ the event "In a centered tree built with the randomized cuts Θ, at depth κ, splitting the cell containing X along the jth direction does not isolate X". Then,

$$\mathbb{E}_\Theta\left[\mathbb{1}_{N_n(X,\Theta)>0}\mu\left(A_n^{(j)}(X,\Theta)\right)\right] = \mathbb{E}_\Theta\left[\mathbb{1}_{N_n(X,\Theta)>0}\mu\left(A_n^{(j)}(X,\Theta)\right)\left(\mathbb{1}_{E_{n,\kappa}(j,X,\Theta)^c}+\mathbb{1}_{E_{n,\kappa}(j,X,\Theta)}\right)\right] \tag{53}$$

$$\leq \mathbb{E}_\Theta\left[\mathbb{1}_{N_n(X,\Theta)>0}\mathbb{1}_{E_{n,\kappa}(j,X,\Theta)^c}\right] \tag{54}$$

$$+\mathbb{E}_\Theta\left[\mathbb{1}_{N_n(X,\Theta)>0}\mathbb{1}_{E_{n,\kappa}(j,X,\Theta)}\mu\left(A_n^{(j)}(X,\Theta)\right)\right], \tag{55}$$

since $\mu\left(A_n^{(j)}(X,\Theta)\right)\leq 1$. We denote $A_{n,\kappa}^{(j),\mathrm{left}}(X,\Theta)$ (resp. $A_{n,\kappa}^{(j),\mathrm{right}}(X,\Theta)$) the left (resp. right) daughter of the cell $A_{n,\kappa}(X,\Theta)$ that has been split along the jth direction (note that the whole cell

is considered here, not only the projection on the j-th side). Then,

$$\begin{aligned}
&\mathbb{E}_\Theta\left[\mathbb{1}_{N_n(X,\Theta)>0}\mathbb{1}_{E_{n,\kappa}(j,X,\Theta)^c}\right] \\
&= \mathbb{P}_\Theta\left(E_{n,\kappa}(j,X,\Theta)^c \middle| N_n(X,\Theta)>0\right)\mathbb{P}_\Theta\left(N_n(X,\Theta)>0\right) &(56)\\
&= \mathbb{P}_\Theta\left(\left(N_n(A_{n,\kappa}^{(j),\text{left}}(X,\Theta))=0\right)\cap\left(X\in A_{n,\kappa}^{(j),\text{right}}(X,\Theta)\right)\middle| N_n(X,\Theta)>0\right)\mathbb{P}_\Theta\left(N_n(X,\Theta)>0\right) \\
&\quad + \mathbb{P}_\Theta\left(\left(N_n(A_{n,\kappa}^{(j),\text{right}}(X,\Theta))=0\right)\cap\left(X\in A_{n,\kappa}^{(j),\text{left}}(X,\Theta)\right)\middle| N_n(X,\Theta)>0\right)\mathbb{P}_\Theta\left(N_n(X,\Theta)>0\right) &(57)\\
&\leq 2\mathbb{P}_\Theta\left(N_n(A_{n,\kappa}^{(j),\text{left}}(X,\Theta))=0\middle| N_n(X,\Theta)>0\right)\mathbb{P}_\Theta\left(N_n(X,\Theta)>0\right). &(58)
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\mathbb{E}_\Theta\left[\mathbb{1}_{N_n(X,\Theta)>0}\mathbb{1}_{E_{n,\kappa}(j,X,\Theta)}\mu\left(A_n^{(j)}(X,\Theta)\right)\right] \\
&\leq \mathbb{E}_\Theta\left[\mu\left(A_n^{(j)}(X,\Theta)\right)\middle| E_{n,\kappa}(j,X,\Theta), N_n(X,\Theta)>0\right]\mathbb{P}_\Theta\left(N_n(X,\Theta)>0\right). &(59)
\end{aligned}$$

Denoting $K_{j,\kappa}(X,\Theta)$ the number of splits made on feature j up to depth κ to produce the cell containing X, we obtain

$$\mathbb{E}_\Theta\left[\mu\left(A_n^{(j)}(X,\Theta)\right)\middle| E_{n,\kappa}(j,X,\Theta), N_n(X,\Theta)>0\right] \leq \mathbb{E}_\Theta\left[2^{-K_{j,\kappa}(X,\Theta)}\middle| E_{n,\kappa}(j,X,\Theta), N_n(X,\Theta)>0\right]. \tag{60}$$

We denote $\delta_j(X,\Theta)\in\{0,1\}^k$ the vector indicating at which depth the jth direction is chosen for splitting, that is $\delta_{j,\ell}(X,\Theta)=1$ if and only if the jth feature is used for splitting at depth ℓ. We have

$$K_{j,\kappa}(X,\Theta)=\sum_{\ell=1}^{\kappa}\delta_{j,\ell}(X,\Theta).$$

For $\ell=1,\ldots,\kappa$, the random variables $\delta_{j,\ell}(X,\Theta)$ are distributed as Bernoulli random variables. Conditional on $E_{n,\kappa}(j,X,\Theta)$ and $N_n(X,\Theta)>0$, we know that for all $\ell=1,\ldots,\kappa$, the jth direction was eligible for splitting at level ℓ. Therefore, the probability of selecting the jth direction at any level $1\leq\ell\leq\kappa$, is $p_\ell\geq 1/d$ (at worst, all variables are eligible for splitting, leading to $p_\ell=1/d$). Besides, conditional on $E_{n,\kappa}(j,X,\Theta)$ and $N_n(X,\Theta)>0$, the random variables $\delta_{j,\ell}(X,\Theta)$ are independent by construction of the centered forest. Indeed, conditional on $E_{n,\kappa}(j,X,\Theta)$ and $N_n(X,\Theta)>0$, the jth direction can be chosen up to depth κ (independence is broken only when the direction cannot be chosen at a given depth as the following one will not be chosen either). Then,

$$\begin{aligned}
\mathbb{E}_\Theta\left[2^{-K_{j,\kappa}(X,\Theta)}\middle| E_{n,\kappa}(j,X,\Theta), N_n(X,\Theta)>0\right] &= \prod_{\ell=1}^{\kappa}\mathbb{E}_\Theta\left[2^{-\delta_{j,\ell}(X,\Theta)}\middle| E_{n,\kappa}(j,X,\Theta), N_n(X,\Theta)>0\right] &(61)\\
&= \prod_{\ell=1}^{\kappa}\left(\frac{p_\ell}{2}+(1-p_\ell)\right) &(62)\\
&\leq \left(1-\frac{1}{2d}\right)^{\kappa}. &(63)
\end{aligned}$$

Therefore, injecting Equations (58) and (63) into (55), we get

$$\frac{\mathbb{E}_\Theta\left[\mathbb{1}_{N_n(X,\Theta)>0}\mu\left(A_n^{(j)}(X,\Theta)\right)\right]}{\mathbb{P}_\Theta\left(N_n(X,\Theta)>0\right)} \leq 2\mathbb{P}_\Theta\left(N_n(A_{n,\kappa}^{(j),\text{left}}(X,\Theta))=0\middle|N_n(X,\Theta)>0\right)+\left(1-\frac{1}{2d}\right)^\kappa, \tag{64}$$

which implies

$$\left(\frac{\mathbb{E}_\Theta\left[\mathbb{1}_{N_n(X,\Theta)>0}\mu\left(A_n^{(j)}(X,\Theta)\right)\right]}{\mathbb{P}_\Theta\left(N_n(X,\Theta)>0\right)}\right)^2 \leq 4\mathbb{P}_\Theta\left(N_n(A_{n,\kappa}^{(j),\text{left}}(X,\Theta))=0\middle|N_n(X,\Theta)>0\right)+2\left(1-\frac{1}{2d}\right)^{2\kappa}, \tag{65}$$

using $(a+b)^2 \leq 2a^2+2b^2 \leq 2a^2+2b$ if $b \leq 1$. Plugging-in this expression into (52) leads to

$$\begin{aligned}\mathbb{E}\left[\left(f_{\infty,n}^{\text{VF}}(X)-f^\star(X)\right)^2\right] \leq 4d\sum_{j=1}^d ||\partial f_j^\star||_\infty^2\left(1-\frac{1}{2d}\right)^{2\kappa}+2e^{-\frac{kn}{2^{k+1}}} \\ +8d\sum_{j=1}^d ||\partial f_j^\star||_\infty^2\mathbb{P}\left(N_n(A_{n,\kappa}^{(j),\text{left}}(X,\Theta))=0\middle|N_n(X,\Theta)>0\right).\end{aligned} \tag{66}$$

Then,

$$\begin{aligned}&\mathbb{P}\left(N_n(A_{n,\kappa}^{(j),\text{left}}(X,\Theta))=0\middle|N_n(X,\Theta)>0\right) && (67)\\ &=\mathbb{E}\left[\mathbb{P}\left(N_n(A_{n,\kappa}^{(j),\text{left}}(X,\Theta))=0\middle|N_n(X,\Theta)>0,N_n(A_{n,\kappa}(X,\Theta)),X,\Theta\right)|N_n(X,\Theta)>0\right] \\ &=\mathbb{E}\left[2^{-N_n(A_{n,\kappa}(X,\Theta))}|N_n(X,\Theta)>0\right] && (68)\\ &\leq 2\mathbb{E}\left[2^{-N_n(A_{n,\kappa}(X,\Theta))}\right]. && (69)\end{aligned}$$

The last line is obtained by making the expectation explicit and noting that $\mathbb{P}\left(N_n(X,\Theta)>0\right)^{-1} \leq 1/(1-e^{-1}) \leq 2$. Furthermore, conditional on X,Θ, $N_n(A_{n,\kappa}(X,\Theta))$ is distributed as a binomial of parameters n and $\mu\left(A_{n,\kappa}(X,\Theta)\right)=2^{-\kappa}$. Thus,

$$\begin{aligned}\mathbb{P}\left(N_n(A_{n,\kappa}^{(j),\text{left}}(X,\Theta))=0\middle|N_n(X,\Theta)>0\right) &\leq 2\mathbb{E}\left[2^{-N_n(A_{n,\kappa}(X,\Theta))}\right] && (70)\\ &\leq 2\mathbb{E}\left[\mathbb{E}\left[2^{-N_n(A_{n,\kappa}(X,\Theta))}|X,\Theta\right]\right] && (71)\\ &\leq 2\left(1-\frac{\mu\left(A_{n,\kappa}(X,\Theta)\right)}{2}\right)^n && (72)\\ &= 2\left(1-2^{-\kappa-1}\right)^n && (73)\\ &\leq 2\exp\left(-\frac{n}{2^{\kappa+1}}\right). && (74)\end{aligned}$$

Overall,

$$\mathbb{E}\left[\left(f^{\mathrm{VF}}_{\infty,n}(X)-f^\star(X)\right)^2\right] \leq 4d\left(\sum_{j=1}^d ||\partial f_j^\star||_\infty^2\right)\left(\left(1-\frac{1}{2d}\right)^{2\kappa}+4\exp\left(-\frac{n}{2^{\kappa+1}}\right)\right)+2\exp\left(-\frac{kn}{2^{k+1}}\right). \tag{75}$$

Choosing $\kappa = \log_2(n) - \log_2(\log_2(n))$, that is $2^\kappa = n/(\log_2(n))$, we obtain

$$\exp\left(2\kappa\log\left(1-\frac{1}{2d}\right)\right)+4\exp\left(-\frac{n}{2^{\kappa+1}}\right) \leq \left(\frac{n}{\log_2 n}\right)^{2\log_2\left(1-\frac{1}{2d}\right)}+4n^{-1/(2\ln 2)}. \tag{76}$$

Consequently, recalling that $k = \lfloor \log_2(n) \rfloor$,

$$\mathbb{E}\left[\left(f^{\mathrm{VF}}_{\infty,n}(X)-f^\star(X)\right)^2\right] \leq 4d\left(\sum_{j=1}^d ||\partial f_j^\star||_\infty^2\right)\left(\left(\frac{n}{\log_2 n}\right)^{2\log_2\left(1-\frac{1}{2d}\right)}+4n^{-1/(2\ln 2)}\right)+2n^{-1/(2\ln 2)} \tag{77}$$

$$\leq C_d\left(\frac{n}{\log_2 n}\right)^{2\log_2\left(1-\frac{1}{2d}\right)}+(C_d+2)\,n^{-1/(2\ln 2)}, \tag{78}$$

where $C_d = 4d\left(\sum_{j=1}^d ||\partial f_j^\star||_\infty^2\right)$.

B.3 Proofs of Section 4 (Theorem 4.1)

In this section, we prove the consistency of the infinite KeRF estimator in the mean interpolating regime (Theorem 4.1). We follow the proof given in [26] and first present two of its results.

Lemma B.2 ([26])**.** *Let $k \in \mathbb{N}$ and consider an infinite centered random forest of depth k. Then, for all $x, z \in [0,1]^d$,*

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

Theorem B.3 ([26])**.** *Let $f^\star$ be a L-Lipschitz function. Then, for all k,*

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

Proof of Theorem 4.1. Let $x \in [0,1]^d$ and recall that

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

Thus, letting

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

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 B.3, we have

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

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

$$\mathbb{E}[|f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x)|^2 \mathbb{1}_{\mathcal{C}_\alpha(x)}] \leq 8C^2 \left(1 - \frac{1}{2d}\right)^{2k} + 8\alpha^2 (1 + \|f^\star\|_\infty)^2. \tag{79}$$

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

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

According to Lemma B.5, there exists a constant $C' > 0$ such that, for all n,

$$\mathbb{E}\Big[\max_{1\leq i\leq n}\varepsilon_i^4\Big] \leq C'\sigma^4(\log n)^2. \tag{80}$$

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

$$\mathbb{E}\Big[|f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x)|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(x)}\Big] \leq C''\sigma^2(\log n)(\mathbb{P}\left[\mathcal{C}_\alpha^c(x)\right])^{1/2}. \tag{81}$$

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 $A_n(x)$,

$$\begin{aligned}
\mathbb{P}\big[|A_n(x)| > \alpha\big] &\leq \frac{1}{n\alpha^2}\mathbb{E}\left[\frac{YK_k(x,X)}{\mathbb{E}\left[K_k(x,X)\right]} - \frac{\mathbb{E}\left[YK_k(x,X)\right]}{\mathbb{E}\left[K_k(x,X)\right]}\right]^2 \\
&\leq \frac{1}{n\alpha^2}\frac{1}{(\mathbb{E}\left[K_k(x,X)\right])^2}\mathbb{E}\Big[Y^2K_k(x,X)^2\Big] \\
&\leq \frac{2}{n\alpha^2}\frac{1}{(\mathbb{E}\left[K_k(x,X)\right])^2}\Big(\mathbb{E}\Big[f^\star(X)^2K_k(x,X)^2\Big] \\
&\qquad + \mathbb{E}\Big[\varepsilon^2K_k(x,X)^2\Big]\Big) \\
&\leq \frac{2(\|f^\star\|_\infty^2 + \sigma^2)}{n\alpha^2}\frac{\mathbb{E}\left[K_k(x,X)^2\right]}{(\mathbb{E}\left[K_k(x,X)\right])^2} \qquad (82) \\
&= \frac{C_0}{n\alpha^2}\frac{\mathbb{E}\left[K_k(x,X)^2\right]}{(\mathbb{E}\left[K_k(x,X)\right])^2} \qquad (83)
\end{aligned}$$

with $C_0 = 2(\|f^\star\|_\infty^2 + \sigma^2)$ a constant. Meanwhile with respect to $B_n(x)$, we obtain, still by Chebyshev's inequality,

$$\mathbb{P}\big[|B_n(x)| > \alpha\big] \leq \frac{1}{n\alpha^2}\frac{\mathbb{E}\left[K_k(x,X)^2\right]}{(\mathbb{E}\left[K_k(x,X)\right])^2} \tag{84}$$

which matches the control made by [26]. Consequently,

$$\mathbb{P}\left[\mathcal{C}_\alpha^c(x)\right] \leq \mathbb{P}\big[|A_n(x)| > \alpha\big] + \mathbb{P}\big[|B_n(x)| > \alpha\big] \tag{85}$$

$$\leq \frac{C_0+1}{n\alpha^2} \frac{\mathbb{E}\left[K_k(x,X)^2\right]}{(\mathbb{E}\left[K_k(x,X)\right])^2}. \tag{86}$$

Besides, for all $x \in [0,1]^d$, for all k, $\mathbb{E}\left[K_k^{cc}(x,X)\right] = \frac{1}{2^k}$ (see in [26] the proof of theorem VI.1 p.11). Since $K_k(x,X) \leq 1$, we know that

$$\mathbb{E}\left[K_k^{cc}(x,X)\right] = \frac{1}{2^k} \;\geq\; \mathbb{E}\left[K_k^{cc}(x,X)^2\right] \;\geq\; \left(\mathbb{E}\left[K_k^{cc}(x,X)\right]\right)^2 = \frac{1}{2^{2k}}, \tag{87}$$

which leads to

$$\mathbb{P}\left[\mathcal{C}_\alpha^c(x)\right] \leq 2^{2k}\left(\frac{C_0+1}{n\alpha^2}\right)\mathbb{E}\left[K_k(x,X)^2\right], \tag{88}$$

but to pursue, we need a tighter upper bound on $\mathbb{E}\left[K_k^{cc}(x,X)^2\right]$ than that obtained from (87). Such a control is provided in Lemma B.4 below, which is original, and departs from the work of [26].

Lemma B.4. *For all $d \geq 2$, for all k large enough, for all $x \in [0,1]^d$,*

$$\mathbb{E}\left[K_k^{cc}(x,X)^2\right] \leq 2^{-k} k^{-\frac{d-1}{2}} \left(C_1 + C_2\left(\log_2(k)\right)^d\right), \tag{89}$$

where

$$C_1 = 1 + \frac{2d^{d/2}}{(4\pi)^{(d-1)/2}} \quad \textit{and} \quad C_2 = 5^d\left(\frac{d-1}{2}\right)^d. \tag{90}$$

Proof of Lemma B.4. From Lemma B.2, we know that

$$\mathbb{E}\left[K_k^{cc}(x,X)^2\right] = \mathbb{E}\left[\left(\sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1!\ldots k_d!}\left(\frac{1}{d}\right)^k \prod_{j=1}^d \mathbb{1}_{\lceil 2^{k_j} x^{(j)}\rceil = \lceil 2^{k_j} X^{(j)}\rceil}\right)^2\right]. \tag{91}$$

Developing the square within the expectation, we obtain two terms, the first one A being the sum of squares and the second one, B, being the cross-product terms. The first term A takes the form

$$A := \mathbb{E}\left[\sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \left(\frac{k!}{k_1!\ldots k_d!}\right)^2 \left(\frac{1}{d}\right)^{2k} \prod_{j=1}^d \mathbb{1}_{\lceil 2^{k_j} x^{(j)}\rceil = \lceil 2^{k_j} x^{(j)}\rceil}\right] \tag{92}$$

$$= \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \left(\frac{k!}{k_1!\ldots k_d!}\right)^2 \left(\frac{1}{d}\right)^{2k} \prod_{j=1}^d \mathbb{P}\left(\lceil 2^{k_j} x^{(j)}\rceil = \lceil 2^{k_j} X^{(j)}\rceil\right). \tag{93}$$

Note that, for all j, $\mathbb{P}\left(\lceil 2^{k_j} x^{(j)} \rceil = \lceil 2^{k_j} X^{(j)} \rceil\right) = 2^{-k_j}$, and $\prod_{j=1}^{d} 2^{-k_j} = 2^{-k}$. Therefore,

$$A = \sum_{\substack{k_1,\dots,k_d \\ \sum_{j=1}^d k_j = k}} \left(\frac{k!}{k_1!\dots k_d!}\right)^2 \left(\frac{1}{d}\right)^{2k} 2^{-k}. \tag{94}$$

Thanks to [24], we know that, for all $d \geq 2$,

$$\sum_{\substack{k_1,\dots,k_d \\ \sum_{j=1}^d k_j = k}} \left(\frac{k!}{k_1!\dots k_d!}\right)^2 \underset{k\to+\infty}{\sim} \frac{d^{2k+d/2}}{(4\pi k)^{(d-1)/2}}. \tag{95}$$

Therefore, for all k large enough, we have

$$\sum_{\substack{k_1,\dots,k_d \\ \sum_{j=1}^d k_j = k}} \left(\frac{k!}{k_1!\dots k_d!}\right)^2 \leq \frac{2d^{2k+d/2}}{(4\pi k)^{(d-1)/2}}. \tag{96}$$

Thus, letting $C_1 = 2d^{d/2}/(4\pi)^{(d-1)/2}$, for all k large enough,

$$A \leq C_1 2^{-k} k^{-(d-1)/2}. \tag{97}$$

Regarding the second term B,

$$B := \mathbb{E}\left[\sum_{\substack{(k_1,\dots,k_d) \\ \neq(\ell_1,\dots,\ell_d), \\ \sum_{j=1}^d k_j = \sum_{j=1}^d \ell_j = k}} \frac{k!}{k_1!\dots k_d!}\frac{k!}{\ell_1!\dots \ell_d!}\left(\frac{1}{d}\right)^{2k} \prod_{j=1}^{d} \mathbb{1}_{\lceil 2^{k_j} x^{(j)} \rceil = \lceil 2^{k_j} X^{(j)} \rceil} \mathbb{1}_{\lceil 2^{\ell_j} x^{(j)} \rceil = \lceil 2^{\ell_j} X^{(j)} \rceil}\right] \tag{98}$$

$$= \sum_{\substack{(k_1,\dots,k_d) \\ \neq(\ell_1,\dots,\ell_d), \\ \sum_{j=1}^d k_j = \sum_{j=1}^d \ell_j = k}} \frac{k!}{k_1!\dots k_d!}\frac{k!}{\ell_1!\dots \ell_d!}\left(\frac{1}{d}\right)^{2k} \mathbb{P}\left(\bigcap_{j=1}^{d} \left(\left(\lceil 2^{k_j} x^{(j)} \rceil = \lceil 2^{k_j} X^{(j)} \rceil\right) \cap \left(\lceil 2^{\ell_j} x^{(j)} \rceil = \lceil 2^{\ell_j} X^{(j)} \rceil\right)\right)\right).$$

A small computation yields

$$\mathbb{P}\left(\bigcap_{j=1}^{d}\left((\lceil 2^{k_j}x^{(j)}\rceil = \lceil 2^{k_j}X^{(j)}\rceil)\cap(\lceil 2^{\ell_j}x^{(j)}\rceil = \lceil 2^{\ell_j}X^{(j)}\rceil)\right)\right)$$

$$= \mathbb{P}\left(\bigcap_{j=1}^{d}\lceil 2^{\ell_j}x^{(j)}\rceil = \lceil 2^{\ell_j}X^{(j)}\rceil \middle| \forall j, \lceil 2^{k_j}x^{(j)}\rceil = \lceil 2^{k_j}X^{(j)}\rceil\right)2^{-k} \tag{99}$$

$$= 2^{-k}\prod_{j=1}^{d}\mathbb{P}\left(\lceil 2^{\ell_j}x^{(j)}\rceil = \lceil 2^{\ell_j}X^{(j)}\rceil \middle| \lceil 2^{k_j}x^{(j)}\rceil = \lceil 2^{k_j}X^{(j)}\rceil\right) \tag{100}$$

$$= 2^{-k}2^{-\sum_{j=1}^{d}(\ell_j-k_j)\mathbb{1}_{\ell_j\geq k_j}} \tag{101}$$

$$= 2^{-\sum_{j=1}^{d}k_j(\mathbb{1}_{\ell_j\geq k_j}+\mathbb{1}_{\ell_j<k_j})-\sum_{j=1}^{d}(\ell_j-k_j)\mathbb{1}_{\ell_j\geq k_j}} \tag{102}$$

$$= 2^{-\sum_{j=1}^{d}k_j\mathbb{1}_{\ell_j<k_j}-\sum_{j=1}^{d}\ell_j\mathbb{1}_{\ell_j\geq k_j}} \tag{103}$$

$$= 2^{-\sum_{j=1}^{d}\max(k_j,\ell_j)}. \tag{104}$$

Therefore,

$$B = \left(\frac{1}{d}\right)^{2k}\sum_{\substack{(k_1,\ldots,k_d)\\ \neq(\ell_1,\ldots,\ell_d),\\ \sum_{j=1}^{d}k_j=\sum_{j=1}^{d}\ell_j=k}}\frac{k!}{k_1!\ldots k_d!}\frac{k!}{\ell_1!\ldots\ell_d!}\left(\frac{1}{2}\right)^{\sum_{j=1}^{d}\max(k_j,\ell_j)}. \tag{105}$$

$$= \left(\frac{1}{d}\right)^{2k}\sum_{\substack{(k_1,\ldots,k_d)\\ \neq(\ell_1,\ldots,\ell_d),\\ \sum_{j=1}^{d}k_j=\sum_{j=1}^{d}\ell_j=k}}\frac{k!}{k_1!\ldots k_d!}\frac{k!}{\ell_1!\ldots\ell_d!}\left(\frac{1}{2}\right)^{k+\frac{1}{2}\sum_{j=1}^{d}|k_j-\ell_j|} \tag{106}$$

$$= \left(\frac{1}{2d^2}\right)^{k}\sum_{\substack{(k_1,\ldots,k_d)\\ \neq(\ell_1,\ldots,\ell_d),\\ \sum_{j=1}^{d}k_j=\sum_{j=1}^{d}\ell_j=k}}\frac{k!}{k_1!\ldots k_d!}\frac{k!}{\ell_1!\ldots\ell_d!}\left(\frac{1}{2}\right)^{\frac{1}{2}\sum_{j=1}^{d}|k_j-\ell_j|}. \tag{107}$$

For all $q > 0$, define the set $\mathcal{K}_q = \{\boldsymbol{\ell} = (\ell_1, \dots, \ell_d), \boldsymbol{k} = (k_1, \dots, k_d) | \sum_{j=1}^{d} |k_j - \ell_j| \geq 2q\}$, so that

$$
\begin{aligned}
B = & \left(\frac{1}{2d^2}\right)^k \sum_{\substack{(\boldsymbol{k},\boldsymbol{\ell}) \in \mathcal{K}_q \\ \boldsymbol{\ell} \neq \boldsymbol{k} \\ \sum_{j=1}^d k_j = \sum_{j=1}^d \ell_j = k}} \frac{k!}{k_1! \dots k_d!} \frac{k!}{\ell_1! \dots \ell_d!} \left(\frac{1}{2}\right)^{\frac{1}{2}\sum_{j=1}^d |k_j - \ell_j|} \\
& + \left(\frac{1}{2d^2}\right)^k \sum_{\substack{(\boldsymbol{k},\boldsymbol{\ell}) \notin \mathcal{K}_q \\ \boldsymbol{\ell} \neq \boldsymbol{k} \\ \sum_{j=1}^d k_j = \sum_{j=1}^d \ell_j = k}} \frac{k!}{k_1! \dots k_d!} \frac{k!}{\ell_1! \dots \ell_d!} \left(\frac{1}{2}\right)^{\frac{1}{2}\sum_{j=1}^d |k_j - \ell_j|} \\
= & B_1 + B_2. \qquad (108)
\end{aligned}
$$

Regarding B_1, we have

$$
\begin{aligned}
B_1 & \leq \left(\frac{1}{2d^2}\right)^k \sum_{\substack{(\boldsymbol{k},\boldsymbol{\ell}) \in \mathcal{K}_q \\ \boldsymbol{\ell} \neq \boldsymbol{k} \\ \sum_{j=1}^d k_j = \sum_{j=1}^d \ell_j = k}} \frac{k!}{k_1! \dots k_d!} \frac{k!}{\ell_1! \dots \ell_d!} 2^{-q} \qquad (109) \\
& \leq \left(\frac{1}{2d^2}\right)^k 2^{-q} \left(\sum_{\boldsymbol{k}, \sum_{j=1}^d k_j = k} \frac{k!}{k_1! \dots k_d!} \right) \left(\sum_{\boldsymbol{\ell}, \sum_{j=1}^d \ell_j = k} \frac{k!}{\ell_1! \dots \ell_d!} \right) \qquad (110) \\
& \leq 2^{-k-q}, \qquad (111)
\end{aligned}
$$

as

$$
\sum_{\boldsymbol{k}, \sum_{j=1}^d k_j = k} \frac{k!}{k_1! \dots k_d!} = d^k. \qquad (112)
$$

We now define, for all $\boldsymbol{k}, \mathcal{K}_q(\boldsymbol{k}) := \{\boldsymbol{\ell} = (\ell_1, \dots, \ell_d), \sum_{j=1}^{d} \ell_j = k, \sum_{j=1}^{d} |k_j - \ell_j| \geq 2q\}$. Regarding B_2, we have

$$
\begin{aligned}
B_2 & \leq \left(\frac{1}{2d^2}\right)^k \sum_{\substack{\boldsymbol{k}, \boldsymbol{\ell} \notin \mathcal{K}_q \\ \boldsymbol{\ell} \neq \boldsymbol{k} \\ \sum_{j=1}^d k_j = \sum_{j=1}^d \ell_j = k}} \frac{k!}{k_1! \dots k_d!} \frac{k!}{\ell_1! \dots \ell_d!} \qquad (113) \\
& = \left(\frac{1}{2d^2}\right)^k \sum_{\boldsymbol{k}, \sum_{j=1}^d k_j = k} \frac{k!}{k_1! \dots k_d!} \sum_{\substack{\boldsymbol{\ell} \notin \mathcal{K}_q(\boldsymbol{k}) \\ \boldsymbol{\ell} \neq \boldsymbol{k} \\ \sum_{j=1}^d k_j = \sum_{j=1}^d \ell_j = k}} \frac{k!}{\ell_1! \dots \ell_d!}. \qquad (114) \\
& \qquad (115)
\end{aligned}
$$

Note that for all $\boldsymbol{\ell}$, $\frac{k!}{\ell_1!\dots\ell_d!}$ is maximal when $\max_i \ell_i$ is minimal. Therefore, for all $k \geq 2d$,

$$\frac{k!}{\ell_1!\dots\ell_d!} = \frac{k!}{\Gamma(\ell_1+1)\dots\Gamma(\ell_d+1)} \tag{116}$$

$$\leq \frac{k!}{\Gamma(\lfloor k/d \rfloor+1)\dots\Gamma(\lfloor k/d \rfloor+1)} \tag{117}$$

$$\leq \frac{k!}{\Gamma(k/d)^d}. \tag{118}$$

Using an inequality from [5], we obtain

$$\begin{aligned}\frac{k!}{\Gamma(k/d)^d} &\leq \frac{k^{k+1/2}e^{-k}}{k^k d^{-k} e^{-k} k^{d/2}} \\ &\leq d^k k^{-(d-1)/2}.\end{aligned}$$

Overall, for all $k \geq 2d$,

$$B_2 \leq \left(\frac{1}{2d}\right)^k \sum_{\boldsymbol{k}, \sum_{j=1}^d k_j = k} \frac{k!}{k_1!\dots k_d!} \sum_{\substack{\boldsymbol{\ell} \notin \mathcal{K}_q(\boldsymbol{k}) \\ \boldsymbol{\ell} \neq \boldsymbol{k} \\ \sum_{j=1}^d k_j = \sum_{j=1}^d \ell_j = k}} k^{-(d-1)/2} \tag{119}$$

$$\leq k^{-(d-1)/2} \left(\frac{1}{2d}\right)^k \sum_{\boldsymbol{k}, \sum_{j=1}^d k_j = k} \frac{k!}{k_1!\dots k_d!} \mathrm{Card}(\mathcal{K}_q(\boldsymbol{k})). \tag{120}$$

We now want to upper bound the cardinal of $\mathcal{K}_q(\boldsymbol{k})$. Denoting by $B_{L_1}(0, 2q)$ the ball of radius $2q$ with respect to the L_1 norm, note that

$$\mathrm{Card}(\mathcal{K}_q(\boldsymbol{k})) \leq \mathrm{Card}(\{x \in \mathbb{N}^d \cap B_{L_1}(\boldsymbol{k}, 2q)\}) \tag{121}$$

$$\leq \mathrm{Card}(\{x \in \mathbb{N}^d \cap B_{L_1}(0, 2q)\}). \tag{122}$$

Since,

$$B_{L_1}(0, c) \subset B_{L_\infty}(0, c) \subset B_{L_\infty}(0, \lceil c \rceil),$$

we have,

$$\begin{aligned}\mathrm{Card}(\mathcal{K}_q(\boldsymbol{k})) &\leq \mathrm{Card}(\{x \in \mathbb{N}^d \cap B_{L_\infty}(0, \lceil 2q \rceil)\}) \\ &\leq (2\lceil 2q \rceil + 1)^d \\ &\leq (4q+3)^d.\end{aligned}$$

Thus, we have, for all $k \geq 2d$,

$$B_2 \leq k^{-(d-1)/2} \left(\frac{1}{2d}\right)^k (4q+3)^d \sum_{\boldsymbol{k}, \sum_{j=1}^d k_j = k} \frac{k!}{k_1!\dots k_d!} \tag{123}$$

$$\leq k^{-(d-1)/2} (4q+3)^d 2^{-k}, \tag{124}$$

as

$$\sum_{\boldsymbol{k},\sum_{j=1}^{d} k_j = k} \frac{k!}{k_1! \dots k_d!} = d^k. \tag{125}$$

Finally, for all q, we have

$$B = B_1 + B_2 \tag{126}$$
$$\leq 2^{-k-q} + k^{-(d-1)/2} (4q+3)^d 2^{-k}. \tag{127}$$

Let $q = \left(\frac{d-1}{2}\right) \log_2(k)$. For all $q \geq 3$, that is for all $k \geq 2^{6/(d-1)}$, and for all $k \geq 2d$,

$$B \leq 2^{-k} \left(k^{-\frac{d-1}{2}} + k^{-(d-1)/2} C_2 \left(\log_2(k)\right)^d \right), \tag{128}$$

where

$$C_2 = 5^d \left(\frac{d-1}{2} \right)^d. \tag{129}$$

Finally, for all k large enough

$$\mathbb{E}\left[K_k^{cc}(x,X)^2\right] \leq A + B_1 + B_2 \tag{130}$$
$$\leq 2^{-k} k^{-\frac{d-1}{2}} \left(C_1 + 1 + C_2 \left(\log_2(k)\right)^d \right). \tag{131}$$

□

According to inequality (88) and Lemma B.4, we have, for all k large enough

$$\mathbb{P}\left[\mathcal{C}_\alpha^c(x)\right] \leq \frac{C_0+1}{n\alpha^2} 2^k k^{-\frac{d-1}{2}} \left(C_1 + C_2 \left(\log_2(k)\right)^d \right). \tag{132}$$

Consequently, according to inequality (81), we obtain, for all k large enough

$$\begin{aligned}
\mathbb{E}\Big[|f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x)|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(x)}\Big] &\leq C'' \sigma^2 \log n \left(\frac{C_0+1}{n\alpha^2} 2^k k^{-\frac{d-1}{2}} \left(C_1 + C_2 \left(\log_2(k)\right)^d \right) \right)^{1/2} \\
&\leq C'' \sigma^2 (C_0+1)^{1/2} (\max(C_1, C_2))^{1/2} \frac{\log n}{n^{1/2}\alpha} 2^{k/2} k^{-\frac{d-1}{4}} \left(\left(1 + \left(\log_2(k)\right)^d \right) \right)^{1/2} \\
&\leq C_3 \frac{\log n}{n^{1/2}\alpha} 2^{k/2} k^{-\frac{d-1}{4}} \left(\log_2(k)\right)^{d/2},
\end{aligned}$$

where $C_3 = C'' \sigma^2 (C_0+1)^{1/2} (2\max(C_1, C_2))^{1/2}$. Then using inequality (79), for all k large enough

$$\begin{aligned}
&\mathbb{E}\Big[f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x) \Big]^2 \\
&\leq \mathbb{E}\Big[|f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x)|^2 \mathbb{1}_{\mathcal{C}_\alpha(x)}\Big] + \mathbb{E}\Big[|f_{\infty,n}^{\mathrm{KeRF}}(x) - f^\star(x)|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(x)}\Big] \\
&\leq 8L^2 d^2 \left(1 - \frac{1}{2d}\right)^{2k} + 8\alpha^2 (1 + \|f^\star\|_\infty)^2 \\
&\quad + C_3 \sigma^2 (\log n) \frac{2^{k/2}}{\alpha n^{1/2}} k^{-\frac{d-1}{4}} (\log_2 k)^{d/2}.
\end{aligned}$$

Optimizing the right hand side in α, that is choosing

$$\alpha^3 = (\log n)\frac{2^{k/2}}{n^{1/2}}k^{-\frac{d-1}{4}}(\log_2 k)^{d/2}\frac{C_3}{8(1+\|f^\star\|_\infty)^2}, \tag{133}$$

we get

$$\mathbb{E}\Big[f^{\mathrm{KeRF}}_{\infty,n}(x) - f^\star(x)\Big]^2 \leq 8L^2d^2\left(1-\frac{1}{2d}\right)^{2k} + 4C_3^{2/3}(1+\|f^\star\|_\infty)^{2/3}(\log n)^{2/3}\frac{2^{k/3}}{n^{1/3}}k^{-\frac{d-1}{6}}(\log_2 k)^{d/3}.$$

Choosing $k_n = \log_2(n)$, we obtain, for all n large enough,

$$\mathbb{E}\Big[f^{\mathrm{KeRF}}_{\infty,n}(x) - f^\star(x)\Big]^2 \leq 8L^2d^2n^{2\log_2(1-\frac{1}{d})} + 4C_3^{2/3}(1+\|f^\star\|_\infty)^{2/3}(\log n)^{2/3}(\log_2 n)^{-\frac{d-1}{6}}(\log_2(\log_2 n))^{d/3}. \tag{134}$$

Finally,

$$\mathbb{E}\Big[f^{\mathrm{KeRF}}_{\infty,n}(x) - f^\star(x)\Big]^2 \leq 8L^2d^2n^{2\log_2(1-\frac{1}{d})} + C_4(\log_2 n)^{-\frac{d-5}{6}}(\log_2(\log_2 n))^{d/3}. \tag{135}$$

with

$$C_4 = 18\times 2^{2/3}\times(\log 2)^{2/3}C''^{2/3}(\|f^\star\|_\infty^2+\sigma^2+1)(\max(C_1,C_2))^{1/3}. \tag{136}$$

□

Lemma B.5. *Consider n i.i.d. random variables $\varepsilon_1,\dots,\varepsilon_n$, distributed as $\mathcal{N}(0,1)$. Then, for all $n \geq 21$,*

$$\mathbb{E}\Big[\max_{1\leq i\leq n}\varepsilon_i^4\Big] \leq 32e(\log n)^2.$$

Proof. We have, for all $p \geq 1$,

$$\mathbb{E}\left[\max_{1\leq i\leq n}|\varepsilon_i|^4\right] \leq \left(\mathbb{E}\left[\max_{1\leq i\leq n}|\varepsilon_i|^{4p}\right]\right)^{1/p} \leq \left(\mathbb{E}\left[\sum_{i=1}^n|\varepsilon_i|^{4p}\right]\right)^{1/p}, \tag{137}$$

using Jensen's inequality (by concavity of $x\mapsto x^{1/p}$ for $p\geq 1$). The p-th moment of a Gaussian variable $\mathcal{N}(0,1)$ can be computed as follows

$$\begin{aligned}\mathbb{E}\left[|\varepsilon_1|^p\right] &= \int_0^\infty \mathbb{P}\left[|\varepsilon|^p \geq u\right]\mathrm{d}u &(138)\\ &= \int_0^\infty \mathbb{P}\left[|\varepsilon| \geq t\right]pt^{p-1}\mathrm{d}t &(139)\\ &\leq \int_0^\infty 2\exp(-t^2/2)pt^{p-1}\mathrm{d}t, &(140)\end{aligned}$$

using classical tail inequalities for Gaussian variables. Now, setting $s = t^2/2$ and recalling that $\Gamma(z) = \int_0^\infty \exp(-t)t^{z-1}\mathrm{d}t$, we have

$$\int_0^\infty 2\exp(-t^2/2)pt^{p-1}\mathrm{d}t = 2p\int_0^\infty \exp(-s)(2s)^{\frac{p-2}{2}}\mathrm{d}s \tag{141}$$

$$= 2p2^{\frac{p-2}{2}}\Gamma(p/2). \tag{142}$$

According to Theorem 2.2 in Batir [5], we have, for all $x > 0$

$$\Gamma(x+1) < \sqrt{2\pi}x^x \exp(-x)\left(x^2 + \frac{x}{3} + \frac{1}{18}\right)^{1/4}. \tag{143}$$

Let

$$f : x \mapsto \exp(-x)\left(x^2 + \frac{x}{3} + \frac{1}{18}\right), \tag{144}$$

one can show that f is non-increasing on $[1/2, \infty)$. Thus, for all $x \geq 1/2$,

$$\Gamma(x+1) < \sqrt{2\pi}x^x f(1/2)^{1/4} \tag{145}$$

$$< \sqrt{2\pi}x^x \exp(-1/2)\left(\frac{1}{2}\right)^{1/4} \tag{146}$$

$$< 2x^x. \tag{147}$$

Hence, for all $p \geq 3$,

$$\mathbb{E}\left[|\varepsilon_1|^p\right] \leq 4p2^{\frac{p-2}{2}}(p/2)^{p/2}, \tag{148}$$

which leads to

$$\mathbb{E}\left[\max_{1\leq i\leq n}|\varepsilon_i|^4\right] \leq \left(\mathbb{E}\left[\sum_{i=1}^n |\varepsilon_i|^{4p}\right]\right)^{1/p} \tag{149}$$

$$\leq n^{1/p}\left(16p2^{\frac{4p-2}{2}}(2p)^{2p}\right)^{1/p} \tag{150}$$

$$\leq 16n^{1/p}p^2\left(\frac{p}{2}\right)^{1/p} \tag{151}$$

$$\leq 32n^{1/p}p^2. \tag{152}$$

Choosing $p = \log n$ yields, for all $n \geq e^3$,

$$\mathbb{E}\left[\max_{1\leq i\leq n}|\varepsilon_i|^4\right] \leq 32e(\log n)^2. \tag{153}$$

□

B.4 Proofs of Section 5 (Semi-adaptive forests)

Lemma B.6. *For all $\alpha \in [0,1)$, the depth k_n^{AdaCT} of a semi-adaptive centered tree verifies*

$$\lim_{n\to\infty} \mathbb{P}\left(k_n^{\mathrm{AdaCT}}(X,\Theta) \in [\log_2(n) \pm \log_2^{1-\alpha}(n)]\right) = 1.$$

Lemma B.6 states that the asymptotic behavior of $k_n^{\mathrm{AdaCT}}(X,\Theta)$ is equivalent to $\log_2 n$ 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.

B.4.1 Proof of Lemma B.6

For all $0 \le j \le k$, we let $A_{j,n}(X,\Theta)$ be the cell containing X in the tree truncated at level j. Similarly, we let $N_{j,n}(X,\Theta)$ the number of observations in this cell. Then,

$$\begin{aligned}
\mathbb{P}\left(k_n(X,\Theta) \ge k\right) &= \mathbb{P}\left(N_{k-1,n}(X,\Theta) \ge 2\right) && (154)\\
&= \mathbb{E}\left[\mathbb{P}\left(N_{k-1,n}(X,\Theta) \ge 2 | X,\Theta\right)\right] && (155)\\
&= 1 - \left(1 - \frac{1}{2^{k-1}}\right)^n - \frac{n}{2^{k-1}}\left(1 - \frac{1}{2^{k-1}}\right)^{n-1}. && (156)
\end{aligned}$$

Using the inequality $\log(1-x) \le -x$ for all $x \in [0,1)$ yields,

$$\begin{aligned}
\mathbb{P}\left(k_n(X,\Theta) \ge k\right) &\ge 1 - \exp\left(-\frac{n}{2^{k-1}}\right) - \frac{n}{2^{k-1}}\exp\left(-\frac{n-1}{2^{k-1}}\right) && (157)\\
&\ge 1 - \left(1 + \frac{n}{2^{k-1}}\right)\exp\left(-\frac{n}{2^{k-1}}\right). && (158)
\end{aligned}$$

Letting $k = (1-\varepsilon_n)\log_2(n)$ in (158) yields

$$\mathbb{P}\left(k_n(X,\Theta) \ge k\right) \ge 1 - (1 + 2n^{\varepsilon_n})\exp\left(-2n^{\varepsilon_n}\right). \quad (159)$$

Note that, setting $\varepsilon_n = c_1(\log_2 n)^{-\alpha}$ for any $\alpha \in [0,1)$ implies that

$$n^{\varepsilon_n} = \exp\left(\varepsilon_n \log n\right) \quad (160)$$

tends to infinity. Therefore, for all $c_1 > 0$ and all $\alpha_i n[0,1)$,

$$\lim_{n\to\infty} \mathbb{P}\left(k_n(X,\Theta) \ge \log_2(n) - c_1(\log_2 n)^{-\alpha}\right) = 1. \quad (161)$$

Besides,

$$\begin{aligned}
\mathbb{P}\left(k_n(X,\Theta) \le k\right) &= 1 - \mathbb{P}\left(k_n(X,\Theta) > k\right) && (162)\\
&= \left(1 - \frac{1}{2^k}\right)^n - \frac{n}{2^k}\left(1 - + \frac{1}{2^k}\right)^{n-1}. && (163)
\end{aligned}$$

Using the inequality $\log(1-x) \geq -x/(1-x)$ for all $x \in [0,1)$, we have

$$\mathbb{P}\left(k_n(X,\Theta) \leq k\right) \geq \exp\left(-\frac{n}{2^k-1}\right) + \frac{n}{2^k}\exp\left(-\frac{n-1}{2^k-1}\right) \tag{164}$$

$$\geq \left(1+\frac{n}{2^k}\right)\exp\left(-\frac{n}{2^k-1}\right). \tag{165}$$

Letting $k = (1+\varepsilon_n)\log_2(n)$ in (165) yields

$$\mathbb{P}\left(k_n(X,\Theta) \geq k\right) \geq \left(1+2n^{-\varepsilon_n}\right)\exp\left(-\frac{n}{n^{1+\varepsilon_n}-1}\right) \tag{166}$$

$$\geq \left(1+2n^{-\varepsilon_n}\right)\exp\left(-\frac{n^{-\varepsilon_n}}{1-\frac{1}{n^{1+\varepsilon_n}}}\right), \tag{167}$$

which tends to 1 for the choice $\varepsilon_n = c_2(\log_2 n)^{-\alpha}$, for any $\alpha \in [0,1)$ and any $c_2 > 0$.

B.4.2 Proof of Theorem 5.1 (Consistency of Median RF)

Preliminary results In all the preliminary results, we use the fact that the spacing between two consecutive order statistics, that originate from an i.i.d. sample uniformly distributed on $[0,1]$ of size n_j is distributed as a beta distribution $\mathcal{B}(1,n_j)$. We also recall that, for all α,β,

$$\mathbb{V}\left[\mathcal{B}(\alpha,\beta)\right] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \quad \text{and} \quad \mathbb{E}\left[\mathcal{B}(\alpha,\beta)\right] = \frac{\alpha}{\alpha+\beta}. \tag{168}$$

Lemma B.7 (Control of a cell side of a fully-developed median RF)**.** *Assume that $n \geq 16$ is a power of two. For all $x \in [0,1]^d$, for all $\ell \in \{1,\dots,d\}$ and depth $k \in \mathbb{N}^*$, with $k \leq \lfloor \log_2 n \rfloor$, we have*

$$\mathbb{E}\left[\mu\left(A_{k,n}^{(\ell)}(x,\Theta)\right)^2\right] \leq C_1\left(1-\frac{3}{4d}\right)^k, \tag{169}$$

with $C_1 \leq 256\exp\left(\frac{42+\sqrt{5}}{2-\sqrt{2}}\right)$.

Proof of Lemma B.7. Fix $x \in [0,1]^d$. For all ℓ, let $\boldsymbol{\delta}_\ell(x,\Theta)$ be the vector whose components are defined as $\delta_{j,\ell}(x,\Theta) = 1$ if the j-th cut is made along direction ℓ and 0 otherwise. Without loss of generality, we let $\ell = 1$ and fix $x \in [0,1]^d$. For all $j \in \{0,\dots,k\}$, we denote $A_{j,n}^{(1)}(x,\Theta)$ the cell containing x at level j, projected onto the first direction, and $n_j = n2^{-j}$ the number of observations falling into this cell.

Recall that we consider the median forest in which splits are performed at the middle of two consecutive order statistics in a cell, so that each resulting cell contains exactly the same number of observations. With these notations in mind, we want to upper bound, for all j,

$$\mathbb{E}\Big[\mu\left(A_{j,n}^{(1)}(x,\Theta)\right)^2 | \boldsymbol{\delta}_1(x,\Theta)\Big],$$

where, for now, the split randomization $\boldsymbol{\delta}_1(x,\Theta)$ is considered fixed and may be omitted in the notations. Let us fix $j \leq k-1$, define

$$A^{(1)}_{j,n}(x,\Theta) = [M_{1,j}, M_{2,j}],$$

and assume that the next cut is made along the first axis at position M_j. Then,

$$\begin{aligned}
&\mu\left(A^{(1)}_{j+1,n}(x,\Theta)\right)^2 \\
&= (M_j - M_{1,j})^2 \mathbb{1}_{x\in[M_{1,j},M_j]} + (M_{2,j} - M_j)^2 \mathbb{1}_{x\in[M_j,M_{2,j}]} && (170)\\
&= (M_j - M_{1,j})^2 + \left((M_{2,j} - M_j)^2 - (M_j - M_{1,j})^2\right) \mathbb{1}_{x\in[M_j,M_{2,j}]} && (171)\\
&= (M_j - M_{1,j})^2 + (M_{1,j} + M_{2,j} - 2M_j)\,(M_{2,j} - M_{1,j})\mathbb{1}_{x\in[M_j,M_{2,j}]}. && (172)
\end{aligned}$$

We denote $X'_1, ..., X'_{n_j}$ the points contained in the cell $A^{(1)}_{j,n}(x,\Theta)$. Note that the second term in (172) can be decomposed as

$$\begin{aligned}
&(M_{1,j} + M_{2,j} - 2M_j)\,(M_{2,j} - M_{1,j})\mathbb{1}_{x\in[M_j,M_{2,j}]} \\
&= \left(X'_{(1)} + X'_{(n_j)} - X'_{(n_j/2)} - X'_{(n_j/2+1)} + M_{1,j} - X'_{(1)} + M_{2,j} - X'_{(n_j)}\right)(M_{2,j} - M_{1,j})\mathbb{1}_{x\in[M_j,M_{2,j}]} && (173)\\
&= \left(\frac{X'_{(1)} + X'_{(n_j)}}{2} - X'_{(n_j/2)} + \frac{X'_{(1)} + X'_{(n_j)}}{2} - X'_{(n_j/2+1)} + M_{1,j} - X'_{(1)} + M_{2,j} - X'_{(n_j)}\right) \\
&\quad \times (M_{2,j} - M_{1,j})\mathbb{1}_{x\in[M_j,M_{2,j}]} && (174)\\
&\leq \left(\frac{X'_{(1)} + X'_{(n_j)}}{2} - X'_{(n_j/2)} + \frac{X'_{(1)} + X'_{(n_j)}}{2} - X'_{(n_j/2+1)} + M_{2,j} - X'_{(n_j)}\right)(M_{2,j} - M_{1,j}). && (175)
\end{aligned}$$

Injecting (175) into (172), taking the expectation and using Cauchy-Schwarz inequality leads to

$$\begin{aligned}
\mathbb{E}\left[\mu\left(A^{(1)}_{j+1,n}(x,\Theta)\right)^2\right] \leq\ & \mathbb{E}\left[(M_j - M_{1,j})^2\right] \\
&+ \left(\mathbb{E}\left[\left(\frac{X'_{(1)} + X'_{(n_j)}}{2} - X'_{(n_j/2)}\right)^2\right]\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right]\right)^{1/2} \\
&+ \left(\mathbb{E}\left[\left(\frac{X'_{(1)} + X'_{(n_j)}}{2} - X'_{(n_j/2+1)}\right)^2\right]\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right]\right)^{1/2} \\
&+ \left(\mathbb{E}\left[\left(M_{2,j} - X'_{(n_j)}\right)^2\right]\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right]\right)^{1/2}. && (176)
\end{aligned}$$

Considering the second term, we have

$$\mathbb{E}\left[\left(\frac{X'_{(1)}+X'_{(n_j)}}{2}-X'_{(n_j/2)}\right)^2\right]=\mathbb{E}\left[\left(\frac{X'_{(n_j)}-X'_{(1)}}{2}-(X'_{(n_j/2)}-X'_{(1)})\right)^2\right] \tag{177}$$

$$=\mathbb{E}\left[(X'_{(n_j)}-X'_{(1)})^2\mathbb{E}\left[\left(\frac{1}{2}-\frac{(X'_{(n_j/2)}-X'_{(1)})}{(X'_{(n_j)}-X'_{(1)})}\right)^2|X'_{(1)},X'_{(n_j)}\right]\right]. \tag{178}$$

where

$$\frac{(X'_{(n_j/2)}-X'_{(1)})}{(X'_{(n_j)}-X'_{(1)})}|X'_{(1)},X'_{(n_j)}\sim\mathcal{B}\left(\frac{n_j}{2}-1,\frac{n_j}{2}\right),$$

with $\mathbb{E}[\mathcal{B}(\frac{n_j}{2}-1,\frac{n_j}{2})]=\frac{n_j-2}{2(n_j-1)}$.

Thus,

$$\mathbb{E}\left[\left(\frac{1}{2}-\frac{(X'_{(n_j/2)}-X'_{(1)})}{(X'_{(n_j)}-X'_{(1)})}\right)^2|X'_{(1)},X'_{(n_j)}\right]=\left(\frac{1}{2}-\frac{n_j-2}{2(n_j-1)}\right)^2+\mathbb{V}\left[\mathcal{B}\left(\frac{n_j}{2}-1,\frac{n_j}{2}\right)\right] \tag{179}$$

$$=\frac{1}{4(n_j-1)^2}+\frac{1}{4}\frac{n_j-2}{(n_j-1)^2} \tag{180}$$

$$=\frac{1}{4(n_j-1)}. \tag{181}$$

Consequently,

$$\mathbb{E}\left[\left(\frac{X'_{(1)}+X'_{(n_j)}}{2}-X'_{(n_j/2)}\right)^2\right]=\frac{1}{4(n_j-1)}\mathbb{E}\left[(X'_{(n_j)}-X'_{(1)})^2\right] \tag{182}$$

$$\leq\frac{1}{4(n_j-1)}\mathbb{E}\left[(M_{2,j}-M_{1,j})^2\right]. \tag{183}$$

Similarly,

$$\mathbb{E}\left[\left(\frac{X'_{(1)}+X'_{(n_j)}}{2}-X'_{(n_j/2+1)}\right)^2\right]=\frac{1}{4(n_j-1)}\mathbb{E}\left[(X'_{(n_j)}-X'_{(1)})^2\right] \tag{184}$$

$$\leq\frac{1}{4(n_j-1)}\mathbb{E}\left[(M_{2,j}-M_{1,j})^2\right]. \tag{185}$$

By Lemma B.8,

$$\mathbb{E}\left[\left(M_{2,j}-X'_{(n_j)}\right)^2\right]\leq\frac{5}{(n_j-1)^2}\mathbb{E}\left[(M_{2,j}-M_{1,j})^2\right].$$

Gathering all previous inequalities into (172) yields

$$\mathbb{E}\left[\mu\left(A^{(1)}_{j+1,n}(x,\Theta)\right)^2\right] \leq \mathbb{E}\left[(M_j - M_{1,j})^2\right] + \frac{1}{\sqrt{n_j - 1}}\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right] + \frac{\sqrt{5}}{(n_j-1)}\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right]. \tag{186}$$

Considering the first term in (186), we have

$$(M_j - M_{1,j})^2 = \left(\frac{X'_{(n_j/2)} + X'_{(n_j/2+1)}}{2} - X'_{(1)} + X'_{(1)} - M_{1,j}\right)^2 \tag{187}$$

$$\leq \left(X'_{(n_j/2+1)} - X'_{(1)} + X'_{(1)} - M_{1,j}\right)^2 \tag{188}$$

$$\leq \left(X'_{(n_j/2+1)} - X'_{(1)}\right)^2 + \left(X'_{(1)} - M_{1,j}\right)^2 + 2\left(X'_{(n_j/2+1)} - X'_{(1)}\right)\left(X'_{(1)} - M_{1,j}\right).$$

Taking the expectation and using Cauchy-Schwarz inequality, we obtain

$$\mathbb{E}\left[(M_j - M_{1,j})^2\right] \leq \mathbb{E}\left[\left(X'_{(n_j/2+1)} - X'_{(1)}\right)^2\right] + \mathbb{E}\left[\left(X'_{(1)} - M_{1,j}\right)^2\right] + 2\left(\mathbb{E}\left[\left(X'_{(n_j/2+1)} - X'_{(1)}\right)^2\right]\mathbb{E}\left[\left(X'_{(1)} - M_{1,j}\right)^2\right]\right)^{1/2}. \tag{189}$$

Now,

$$\mathbb{E}\left[\left(X'_{(n_j/2+1)} - X'_{(1)}\right)^2\right] = \mathbb{E}\left[\left(X'_{(n_j)} - X'_{(1)}\right)^2 \mathbb{E}\left[\left(\frac{X'_{(n_j/2+1)} - X'_{(1)}}{X'_{(n_j)} - X'_{(1)}}\right)^2 | X'_{(1)}, X'_{(n_j)}\right]\right], \tag{190}$$

where

$$\mathbb{E}\left[\left(\frac{X'_{(n_j/2+1)} - X'_{(1)}}{X'_{(n_j)} - X'_{(1)}}\right)^2 | X'_{(1)}, X'_{(n_j)}\right] = \mathbb{E}\left[\mathcal{B}\left(\frac{n_j}{2}, \frac{n_j}{2} - 1\right)^2\right] \tag{191}$$

$$= \mathbb{V}\left[\mathcal{B}\left(\frac{n_j}{2}, \frac{n_j}{2} - 1\right)\right] + \left(\mathbb{E}\left[\mathcal{B}\left(\frac{n_j}{2}, \frac{n_j}{2} - 1\right)\right]\right)^2 \tag{192}$$

$$= \frac{\frac{n_j}{2}\left(\frac{n_j}{2} - 1\right)}{(n_j - 1)^2 n_j} + \left(\frac{n_j/2}{n_j - 1}\right)^2 \tag{193}$$

$$= \frac{1}{4}\frac{n_j - 2}{(n_j - 1)^2} + \left(\frac{1}{2}\frac{n_j}{n_j - 1}\right)^2 \tag{194}$$

$$= \frac{1}{4}\frac{n_j^2 + n_j - 2}{(n_j - 1)^2} \tag{195}$$

$$\leq \frac{1}{4}\frac{(n_j + 1/2)^2}{(n_j - 1)^2}. \tag{196}$$

Therefore,

$$\mathbb{E}\left[\left(X'_{(n_j/2+1)} - X'_{(1)}\right)^2\right] \leq \frac{1}{4}\frac{(n_j+1/2)^2}{(n_j-1)^2}\mathbb{E}\left[\left(X'_{(n_j)} - X'_{(1)}\right)^2\right]. \tag{197}$$

Injecting this expression into (189), we have

$$\begin{aligned}\mathbb{E}\left[(M_j - M_{1,j})^2\right] \leq \frac{1}{4}\frac{(n_j+1/2)^2}{(n_j-1)^2}\mathbb{E}\left[\left(X'_{(n_j)} - X'_{(1)}\right)^2\right] + \mathbb{E}\left[\left(X'_{(1)} - M_{1,j}\right)^2\right] \\ + \frac{(n_j+1/2)}{(n_j-1)}\left(\mathbb{E}\left[\left(X'_{(n_j)} - X'_{(1)}\right)^2\right]\mathbb{E}\left[\left(X'_{(1)} - M_{1,j}\right)^2\right]\right)^{1/2}.\end{aligned} \tag{198}$$

According to Technical Lemma B.8, we have

$$\mathbb{E}\left[\left(X'_{(1)} - M_{1,j}\right)^2\right] \leq \frac{5}{(n_j-1)^2}\mathbb{E}\left[M_{2,j} - M_{1,j}\right].$$

Hence,

$$\begin{aligned}&\mathbb{E}\left[(M_j - M_{1,j})^2\right] \\ &\leq \frac{1}{4}\frac{(n_j+1/2)^2}{(n_j-1)^2}\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right] + \frac{5}{(n_j-1)^2}\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right] \\ &\quad + \frac{(n_j+1/2)}{(n_j-1)}\left(\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right]\frac{5}{(n_j-1)^2}\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right]\right)^{1/2} &(199)\\ &\leq \left(\frac{1}{4}\frac{(n_j+1/2)^2}{(n_j-1)^2} + \frac{5}{(n_j-1)^2} + \frac{(n_j+1/2)\sqrt{5}}{(n_j-1)^2}\right)\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right] &(200)\\ &\leq \frac{1}{4}\frac{(n_j+1/2)^2}{(n_j-1)^2}\left(1 + \frac{20}{(n_j+1/2)^2} + \frac{4\sqrt{5}}{(n_j+1/2)}\right)\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right] &(201)\\ &\leq \frac{1}{4}\left(1 + \frac{3}{2(n_j-1)}\right)^2\left(1 + \frac{20}{(n_j+1/2)^2} + \frac{4\sqrt{5}}{(n_j+1/2)}\right)\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right] &(202)\\ &\leq \frac{1}{4}\left(1 + \frac{9}{2(n_j-1)}\right)\left(1 + \frac{20}{(n_j+1/2)^2} + \frac{4\sqrt{5}}{(n_j+1/2)}\right)\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right], &(203)\end{aligned}$$

for all $n_j \geq 4$, since $(1+x)^2 \leq 1+3x$ if $x \leq 1$.

Consequently,

$$\begin{aligned}
&\mathbb{E}\left[(M_j - M_{1,j})^2\right] \\
&\leq \frac{1}{4}\left(1 + \frac{9}{2(n_j-1)}\right)\left(1 + \frac{30}{n_j - 1}\right)\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right] && (204)\\
&\leq \frac{1}{4}\left(1 + \frac{69}{2(n_j-1)} + \frac{90}{(n_j-1)^2}\right)\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right] && (205)\\
&\leq \frac{1}{4}\left(1 + \frac{35+6}{n_j - 1}\right)\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right] && (206)\\
&\leq \frac{1}{4}\left(1 + \frac{41}{n_j - 1}\right)\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right], && (207)
\end{aligned}$$

for all $n_j \geq 16$. Recall that, until now, we have fixed $\boldsymbol{\delta}_1(x,\Theta)$ and omitted the explicit conditioning in the proof to lighten notations. Thus, plugging-in the previous inequality into (186) yields, for all $n_j \geq 16$,

$$\begin{aligned}
\mathbb{E}\left[\mu\left(A^{(1)}_{j+1,n}(x,\Theta)\right)^2 |\boldsymbol{\delta}_1(x,\Theta)\right] &\leq \frac{1}{4}\left(1 + \frac{41}{n_j - 1}\right)\mathbb{E}\left[(M_{2,j} - M_{1,j})^2 |\boldsymbol{\delta}_1(x,\Theta)\right] \\
&\quad + \frac{1}{\sqrt{n_j - 1}}\mathbb{E}\left[(M_{2,j} - M_{1,j})^2 \big|\boldsymbol{\delta}_1(x,\Theta)\right] + \frac{\sqrt{5}}{(n_j-1)}\mathbb{E}\left[(M_{2,j} - M_{1,j})^2 \big|\boldsymbol{\delta}_1(x,\Theta)\right] && (208)\\
&\leq \frac{1}{4}\left(1 + \frac{42+\sqrt{5}}{\sqrt{n_j - 1}}\right)\mathbb{E}\left[\mu\left(A^{(1)}_{j,n}(x,\Theta)\right)^2 \big|\boldsymbol{\delta}_1(x,\Theta)\right]. && (209)
\end{aligned}$$

Recall that $\boldsymbol{\delta}_1(x,\Theta)$ is the vector whose components are defined as $\delta_{j,1}(x,\Theta) = 1$ if the j-th cut is made along the first direction and 0 otherwise. We let $K_1 = \|\boldsymbol{\delta}_1(x,\Theta)\|_1$ be the number of times the first direction is split. By induction, we have

$$\begin{aligned}
\mathbb{E}\left[\mu\left(A^{(1)}_{k,n}(x,\Theta)\right)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\mu\left(A^{(1)}_{k,n}(x,\Theta)\right)^2 |\boldsymbol{\delta}_1(x,\Theta)\right]\right] && (210)\\
&\leq \mathbb{E}\left[\prod_{\substack{j:\delta_{j,1}=1,\\ j\leq k-4}} \frac{1}{4}\left(1 + \frac{42+\sqrt{5}}{\sqrt{n_j - 1}}\right)\right] && (211)\\
&\leq 4^4\mathbb{E}\left[4^{-K_1}\prod_{\substack{j:\delta_{j,1}=1,\\ j\leq k-4}} \left(1 + \frac{42+\sqrt{5}}{\sqrt{n_j - 1}}\right)\right]. && (212)
\end{aligned}$$

The product can be upper bounded as follows, with $C = 42 + \sqrt{5}$,

$$\log\left(\prod_{j,\delta_{j,l}=1, j\leq k-4}\left(1+\frac{C}{\sqrt{n_j-1}}\right)\right) \leq \log\left(\prod_{\substack{j:\delta_{j,1}=1,\\ j\leq k-4}}\left(1+\frac{C}{\sqrt{n_{j+1}}}\right)\right) \tag{213}$$

$$= \sum_{\substack{j:\delta_{j,1}=1,\\ j\leq k-4}} \log\left(1+\frac{C\sqrt{2}\cdot 2^{j/2}}{n^{1/2}}\right) \tag{214}$$

$$\leq \frac{C\sqrt{2}}{n^{1/2}}\sum_{j=0}^{k-4} 2^{j/2} \tag{215}$$

$$\leq \frac{C\sqrt{2}}{\sqrt{2}-1}\frac{2^{(k-3)/2}}{n^{1/2}} \tag{216}$$

$$\leq \frac{C}{2\sqrt{2}-2}. \tag{217}$$

Thus,

$$\mathbb{E}\left[\mu\left(A_{k,n}^{(1)}(x,\Theta)\right)^2\right] \leq 4^4 \exp\left(\frac{C}{2\sqrt{2}-2}\right)\mathbb{E}\left[4^{-K_1}\right]. \tag{218}$$

Since $K_1 \sim \text{Bin}(k, 1/d)$, we have

$$\mathbb{E}\left[4^{-K_1}\right] = \left(1-\frac{1}{d}+\frac{1}{4d}\right)^k \tag{219}$$

$$= \left(1-\frac{3}{4d}\right)^k. \tag{220}$$

Finally,

$$\mathbb{E}\left[\mu\left(A_{k,n}^{(1)}(x,\Theta)\right)^2\right] \leq 4^4 \exp\left(\frac{C}{2\sqrt{2}-2}\right)\left(1-\frac{3}{4d}\right)^k, \tag{221}$$

with $C = 42 + \sqrt{5}$.

□

Lemma B.8 (Technical Lemma)**.** *1. Let $x \in [0,1]^d$ and consider the cell $A_{n,j}(x,\Theta)$ containing x at depth $j \leq k-1$. W.l.o.g. restrict the study to the one-dimensional cell $A_{n,j}^{(1)}(x,\Theta)$ corresponding to the cell $A_{n,j}(x,\Theta)$ along the first dimension only, and set $A_{n,j}^{(1)}(x,\Theta) = [M_{1,j}; M_{2,j}]$. The one-dimensional cell $A_{n,j}^{(1)}(x,\Theta)$ contains n_j points denoted $X'_1, \dots X'_{n_j}$*

(random subsample of the initial training sample). Call $X'_{(1)}, \dots, X'_{(n_j)}$ the ordered version of $X'_1, \dots X'_{n_j}$. Then,

$$\mathbb{E}\left[\left(X'_{(1)} - M_{1,j}\right)^2\right] \leq \frac{5}{(n_j-1)^2}\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right],$$

and

$$\mathbb{E}\left[\left(M_{2,j} - X'_{(n_j)}\right)^2\right] \leq \frac{5}{(n_j-1)^2}\mathbb{E}\left[(M_{2,j} - M_{1,j})^2\right].$$

2. *Consider now the cell $A_{n,j}(X_1, \Theta)$ containing X_1 at depth $j \leq k-1$. W.l.o.g. restrict the study to the one-dimensional cell $A^{(1)}_{n,j}(X_1, \Theta)$ corresponding to the cell $A_{n,j}(X_1, \Theta)$ along the first dimension only, and set $A^{(1)}_{n,j}(X_1, \Theta) = [M_{1,j}; M_{2,j}] = [M_{1,j}(X_1, \Theta); M_{2,j}(X_1, \Theta)]$. The one-dimensional cell $A^{(1)}_{n,j}(X_1, \Theta)$ contains n_j points denoted $\{X'_1, \dots X'_{n_j-1}\} \cup \{X_1\}$ (random subsample of the initial training sample containing X_1 and projected on the first axis). Call $X'_{(1)}, \dots, X'_{(n_j-1)}$ the ordered version of $X'_1, \dots X'_{n_j-1}$. Then,*

$$\mathbb{E}\left[X'_{(1)} - M_{1,j} | X_1\right] \leq \frac{1}{n_j}\mathbb{E}\left[(M_{2,j} - M_{1,j}) | X_1\right],$$

and

$$\mathbb{E}\left[M_{2,j} - X'_{(n_j-1)} | X_1\right] \leq \frac{1}{n_j}\mathbb{E}\left[(M_{2,j} - M_{1,j}) | X_1\right].$$

Proof of Lemma B.8.

Notations W.l.o.g. consider the following development according to the first direction only. Let $x \in [0,1]^d$. Recall that we consider the cell $A^{(1)}_{j,n}(x, \Theta) = [M_{1,j}; M_{2,j}]$ containing x at depth $j \leq k-1$. The cut at $M_{1,j}$ (resp. $M_{2,j}$) has been obtained at an anterior depth $j_1 \leq j$ (resp. $j_2 \leq j$), as the middle of two order statistics of a previous subsample:

$$M_{1,j} = \frac{M_{1,j,-} + M_{1,j,+}}{2} \quad \text{and} \quad M_{2,j} = \frac{M_{2,j,-} + M_{2,j,+}}{2},$$

with $M_{1,j,-} < M_{1,j,+}$ and $M_{2,j,-} < M_{2,j,+}$. The following computations can be also conducted in a similar way when $M_{1,j} = 0$ or $M_{2,j} = 1$. The current cell $A^{(1)}_{j,n}(x, \Theta)$, includes now n_j points of the original training sample, which are denoted by $X'_1, \dots, X'_{n_j}$. Remark that as $M_{1,j,-}$ and $M_{2,j,+}$ refer to anterior order statistics of a previous subsample (including the points $X'_1, \dots, X'_{n_j}$), then $X'_1, \dots, X'_{n_j}$ are i.i.d. uniformly distributed in $[M_{j,1,-}; M_{2,j,+}]$. Denote by $X'_{(1)}, \dots, X'_{(n_j)}$, the ordered statistics of the current subsample $X'_1, \dots, X'_{n_j}$ in $A^{(1)}_{j,n}(x, \Theta)$ for some fixed x.

First statement - Control of $\mathbb{E}[(X'_{(1)} - M_{1,j})^2]$. We have

$$\mathbb{E}\left[\left(X'_{(1)} - M_{1,j}\right)^2\right] \leq 2\mathbb{E}\left[(M_{1,j,+} - M_{1,j})^2\right] + 2\mathbb{E}\left[\left(X'_{(1)} - M_{1,j,+}\right)^2\right]. \tag{222}$$

Note that, by definition of $M_{1,j}$, the quantity $M_{1,j,+} - M_{1,j}$ corresponds to a half spacing between two points in the cell previously built by cutting on the first direction at depth j_1, denoted $A^{(1)}_{j_1,n}(x,\Theta)$. By construction, the spacings between two consecutive points in $A^{(1)}_{j_1,n}(x,\Theta)$ were the same in distribution. Since points have been removed between $A^{(1)}_{j,n}(x,\Theta)$ and $A^{(1)}_{j_1,n}(x,\Theta)$, the spacings are larger between consecutive points in $A^{(1)}_{j,n}(x,\Theta)$ than between consecutive points in $A^{(1)}_{j_1,n}(x,\Theta)$. This leads to

$$M_{1,j,+} - M_{1,j} = \frac{M_{1,j,+} - M_{1,j,-}}{2} \leq \frac{X'_{(2)} - X'_{(1)}}{2}.$$

Therefore, since all variables are bounded,

$$\begin{aligned}
\mathbb{E}\left[\left(M_{1,j,+} - M_{1,j}\right)^2\right] &\leq \frac{1}{4}\mathbb{E}\left[\left(X'_{(2)} - X'_{(1)}\right)^2\right] \\
&\leq \frac{1}{4}\mathbb{E}\left[(X'_{(n_j)} - X'_{(1)})^2 \mathbb{E}\left[\frac{\left(X'_{(2)} - X'_{(1)}\right)^2}{\left(X'_{(n_j)} - X'_{(1)}\right)^2} | X'_{(1)}, X'_{(n_j)}\right]\right].
\end{aligned}$$

Regarding the inner expectation,

$$\begin{aligned}
\mathbb{E}\left[\frac{\left(X'_{(2)} - X'_{(1)}\right)^2}{\left(X'_{(n_j)} - X'_{(1)}\right)^2} | X'_{(1)}, X'_{(n_j)}\right] &= \mathbb{E}\left[\mathcal{B}\left(1, n_j - 2\right)^2\right] && (223)\\
&= \mathbb{V}\left[\mathcal{B}\left(1, n_j - 2\right)\right] + \left(\mathbb{E}\left[\mathcal{B}\left(1, n_j - 2\right)\right]\right)^2 && (224)\\
&= \frac{n_j - 2}{(n_j - 1)^2 n_j} + \left(\frac{1}{n_j - 1}\right)^2 && (225)\\
&\leq \frac{2}{(n_j - 1)^2}. && (226)
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{E}\left[\left(M_{1,j,+} - M_{1,j}\right)^2\right] &\leq \frac{1}{2(n_j - 1)^2}\mathbb{E}\left[\left(X'_{(n_j)} - X'_{(1)}\right)^2\right] && (227)\\
&\leq \frac{1}{2(n_j - 1)^2}\mathbb{E}\left[\left(M_{2,j} - M_{1,j}\right)^2\right]. && (228)
\end{aligned}$$

Regarding the second term in (222), we have

$$\begin{aligned}
\mathbb{E}\left[\left(X'_{(1)} - M_{1,j,+}\right)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(X'_{(1)} - M_{1,j,+}\right)^2 | M_{1,j,+}, X_{M_{2,j,-}}\right]\right] \\
&= \mathbb{E}\left[\left(M_{2,j,-} - M_{1,j,+}\right)^2 \mathbb{E}\left[\left(\frac{X'_{(1)} - M_{1,j,+}}{M_{2,j,-} - M_{1,j,+}}\right)^2 | M_{1,j,+}, M_{2,j,-}\right]\right] \\
&\leq \mathbb{E}\left[\left(M_{2,j,-} - M_{1,j,+}\right)^2 \mathbb{E}\left[\mathcal{B}(1, n_j - 1)^2\right]\right] \\
&\leq \frac{2}{n_j^2}\mathbb{E}\left[\left(M_{2,j,-} - M_{1,j,+}\right)^2\right] \\
&\leq \frac{2}{n_j^2}\mathbb{E}\left[\left(M_{2,j} - M_{1,j}\right)^2\right].
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{E}\left[\left(X'_{(1)} - M_{1,j}\right)^2\right] &\leq \left(\frac{1}{(n_j-1)^2} + \frac{4}{n_j^2}\right)\mathbb{E}\left[\left(M_{2,j} - M_{1,j}\right)^2\right] \\
&\leq \frac{5}{(n_j-1)^2}\mathbb{E}\left[\left(M_{2,j} - M_{1,j}\right)^2\right].
\end{aligned}$$

The second point of the first statement can be proved in the exact same manner.

Second statement - Control of $\mathbb{E}[X'_{(1)} - M_{1,j}|X_1]$. In this part, we study the cell $A_{n,j}^{(1)}(X_1, \Theta)$. The cell $A_{n,j}^{(1)}(X_1, \Theta)$ contains n_j data points (including X_1). We denote by $X'_1, \ldots, X'_{n_j-1}$ the observations falling into $A_{n,j}^{(1)}(X_1, \Theta)$, different from X_1. Note that, these $n_j - 1$ observations are still i.i.d. uniformly distributed in $[M_{1,j,-}; M_{2,j,+}]$. We denote by $X'_{(1)}, \ldots, X'_{(n_j-1)}$, the subsample $X'_1, \ldots, X'_{n_j-1}$. We have

$$M_{2,j} - M_{1,j} = M_{2,j} - X'_{(n_j-1)} + \sum_{q=1}^{n_j-2}\left(X'_{(n_j-q)} - X'_{(n_j-q-1)}\right) + X'_{(1)} - M_{1,j}. \tag{229}$$

Thus,

$$\mathbb{E}\left[X'_{(1)} - M_{1,j}|X_1\right] + \mathbb{E}\left[M_{2,j} - X'_{(n_j-1)}|X_1\right] = \mathbb{E}\left[M_{2,j} - M_{1,j}|X_1\right] - (n_j - 2)\mathbb{E}\left[X'_{(2)} - X'_{(1)}|X_1\right]. \tag{230}$$

The variables $X'_{(1)}$ and $X'_{(2)}$ being order statistics of a subsample independent of X_1, one gets

$$\mathbb{E}\left[X'_{(2)} - X'_{(1)}|X_1\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{X'_{(2)} - X'_{(1)}}{M_{2,j,+} - M_{1,j,-}}|X_1, M_{1,j,-}, M_{2,j,+}\right](M_{2,j,+} - M_{1,j,-})|X_1\right] \tag{231}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\mathcal{B}(1, n_j - 1)\right](M_{2,j,+} - M_{1,j,-})|X_1\right] \tag{232}$$

$$= \frac{1}{n_j}\mathbb{E}\left[(M_{2,j,+} - M_{1,j,-})|X_1\right] \tag{233}$$

$$\geq \frac{1}{n_j}\mathbb{E}\left[(M_{2,j} - M_{1,j})|X_1\right]. \tag{234}$$

Finally,

$$\mathbb{E}\left[X'_{(1)} - M_{1,j}|X_1\right] + \mathbb{E}\left[M_{2,j} - X'_{(n_j-1)}|X_1\right] \leq \left(1 - \frac{n_j - 2}{n_j}\right)\mathbb{E}\left[(M_{2,j} - M_{1,j})|X_1\right], \tag{235}$$

$$= \frac{2}{n_j}\mathbb{E}\left[(M_{2,j} - M_{1,j})|X_1\right], \tag{236}$$

and, by symmetry,

$$\mathbb{E}\left[X'_{(1)} - M_{1,j}|X_1\right] \leq \frac{1}{n_j}\mathbb{E}\left[(M_{2,j} - M_{1,j})|X_1\right], \tag{237}$$

and

$$\mathbb{E}\left[M_{2,j} - X'_{(n_j-1)}|X_1\right] \leq \frac{1}{n_j}\mathbb{E}\left[(M_{2,j} - M_{1,j})|X_1\right]. \tag{238}$$

□

Lemma B.9 (Control of the leaf side and volume of a fully developed median RF)**.** *Assume that $n \geq 4$ is a power of two. Consider a median tree of depth k and denote $A_{n,k}(X_1, \Theta)$ the leaf containing X_1. For all $\ell \in \{1, \ldots, d\}$, we denote K_ℓ the number of splits along the ℓ-th direction. Let also $\boldsymbol{\delta}_\ell(X_1, \Theta)$ be the vector whose components are defined as $\delta_{j,\ell}(X_1, \Theta) = 1$ if the j-th cut of the cell $A_n^{(\ell)}(X_1, \Theta)$ is made along direction ℓ and 0 otherwise. Then,*

$$\mathbb{E}\left[\mu(A_{n,k}^{(\ell)}(X_1, \Theta))|X_1, \boldsymbol{\delta}_\ell(X_1, \Theta)\right] \leq 2^{-K_\ell+2} \prod_{\substack{j:\delta_{j,\ell}=1,\\ j\leq k-2}} \left(1 + \frac{2}{\sqrt{n_j - 1}}\right). \tag{239}$$

In particular, letting $C_2 = 4\exp(5/(\sqrt{2}-1))$, we have

$$\mathbb{E}\left[\mu(A_{n,k}^{(\ell)}(X_1, \Theta))|X_1, \boldsymbol{\delta}_\ell(X_1, \Theta)\right] \leq C_2\ 2^{-K_\ell}, \tag{240}$$

and

$$\mathbb{E}\left[\mu(A_{n,k}(X_1, \Theta))|X_1, \boldsymbol{\delta}_1(X_1, \Theta), \ldots, \boldsymbol{\delta}_d(X_1, \Theta)\right] \leq C_2 2^{-k}. \tag{241}$$

Proof. We write $A_{n,j}^{(\ell)}(X_1,\Theta)=[M_{1,j},M_{2,j}]$ the cell of the RF containing X_1 along the direction ℓ, at depth j. To lighten the notations, we omit the dependencies in X_1, in Θ and in ℓ. We also write $X'_1,...,X'_{n_j}$ the data points contained in the cell $A_{n,j}^{(\ell)}(X_1,\Theta)$, and we denote by $X'_{(1)},...,X'_{(n_j-1)}$ the ordered version of $\{X'_1,...,X'_{n_j}\}\setminus\{X_1\}$. We suppose that the next cut is occurring on the ℓ-th direction and compute the size of the new cell containing X_1, $A_{n,j+1}^{(\ell)}(X_1,\Theta)$, so that 4 different events are possible:

1. X_1 is in the first "part" of the cell, i.e. $X_1\in\left[M_{1,j},X'_{(n_j/2-1)}\right]$;
2. X_1 is in the second "part" of the cell, i.e. , $X_1\in\left[X'_{(n_j/2+1)},M_{2,j}\right]$;
3. X_1 is in the "middle (left)" of the cell, i.e. $X_1\in\left[X'_{(n_j/2-1)},X'_{(n_j/2)}\right]$;
4. X_1 is in the "middle (right)" of the cell, $X_1\in\left[X'_{(n_j/2)},X'_{(n_j/2+1)}\right]$.

The length of the following cell can be therefore decomposed with respect to the previous events:

$$
\begin{aligned}
&\mu\left(A_{n,j+1}^{(\ell)}(X_1,\Theta)\right)\\
&=\left(\frac{X'_{(n_j/2-1)}+X'_{(n_j/2)}}{2}-M_{1,j}\right)\mathbb{1}_{X_1\in\left[M_{1,j},X'_{(n_j/2-1)}\right]}\\
&\quad+\left(M_{2,j}-\frac{X'_{(n_j/2)}+X'_{(n_j/2+1)}}{2}\right)\mathbb{1}_{X_1\in\left[X'_{(n_j/2+1)},M_{2,j}\right]}\\
&\quad+\left(\frac{X_1+X'_{(n_j/2)}}{2}-M_{1,j}\right)\mathbb{1}_{X_1\in\left[X'_{(n_j/2-1)},X'_{(n_j/2)}\right]}\\
&\quad+\left(M_{2,j}-\frac{X'_{(n_j/2)}+X_1}{2}\right)\mathbb{1}_{X_1\in\left[X'_{(n_j/2)},X'_{(n_j/2+1)}\right]} &(242)\\
&\leq\left(X'_{(n_j/2)}-M_{1,j}\right)\mathbb{1}_{X_1\in\left[M_{1,j},X'_{(n_j/2-1)}\right]}+\left(M_{2,j}-X'_{(n_j/2)}\right)\mathbb{1}_{X_1\in\left[X'_{(n_j/2+1)},M_{2,j}\right]} &(243)\\
&\quad+\left(X'_{(n_j/2)}-M_{1,j}\right)\mathbb{1}_{X_1\in\left[X'_{(n_j/2-1)},X'_{(n_j/2)}\right]}+\left(M_{2,j}-X'_{(n_j/2)}\right)\mathbb{1}_{X_1\in\left[X'_{(n_j/2)},X'_{(n_j/2+1)}\right]} &(244)\\
& &(245)
\end{aligned}
$$

$$
\begin{aligned}
&\mu\left(A^{(\ell)}_{n,j+1}(X_1,\Theta)\right)\\
&\leq \left(X'_{(n_j/2)}-M_{1,j}\right)\mathbb{1}_{X_1\in[M_{1,j},X'_{(n_j/2)}]}+\left(M_{2,j}-X'_{(n_j/2)}\right)\mathbb{1}_{X_1\in[X'_{(n_j/2)},M_{2,j}]} &(246)\\
&=\left(X'_{(n_j/2)}-M_{1,j}\right)+2\left(\frac{M_{1,j}+M_{2,j}}{2}-X'_{(n_j/2)}\right)\mathbb{1}_{X_1\in[X'_{(n_j/2)},M_{2,j}]} &(247)\\
&=\left(X'_{(n_j/2)}-X'_{(1)}\right)+\left(X'_{(1)}-M_{1,j}\right)+2\left(\frac{X'_{(1)}+X'_{(n_j-1)}}{2}-X'_{(n_j/2)}\right)\mathbb{1}_{X_1\in[X'_{(n_j/2)},M_{2,j}]}\\
&\quad-\left(\left(X'_{(1)}+X'_{(n_j-1)}\right))-(M_{1,j}+M_{2,j})\right)\mathbb{1}_{X_1\in[X'_{(n_j/2)},M_{2,j}]} &(248)\\
&\leq\left(X'_{(n_j/2)}-X'_{(1)}\right)+\left(X'_{(1)}-M_{1,j}\right)+2\left(\frac{X'_{(1)}+X'_{(n_j-1)}}{2}-X'_{(n_j/2)}\right)\mathbb{1}_{X_1\in[X'_{(n_j/2)},M_{2,j}]}\\
&\quad+(M_{2,j}-X'_{(n_j-1)}). &(249)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}\left[\mu\left(A^{(\ell)}_{n,j+1}(X_1,\Theta)\right)\middle|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right]\\
&\leq\mathbb{E}\left[(X'_{(n_j-1)}-X'_{(1)})\mathbb{E}\left[\frac{X'_{(n_j/2)}-X'_{(1)}}{X'_{(n_j-1)}-X'_{(1)}}|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),X'_{(1)},X'_{(n_j-1)}\right]\middle|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right]\\
&\quad+2\mathbb{E}\left[\mathbb{E}\left[\left|\frac{X'_{(n_j-1)}+X'_{(1)}}{2}-X'_{(n_j/2)}\right|\middle|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),X'_{(1)},X'_{(n_j-1)}\right]\middle|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right]\\
&\quad+\mathbb{E}\left[X'_{(1)}-M_{1,j}\middle|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right]+\mathbb{E}\left[M_{2,j}-X'_{(n_j-1)}\middle|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right]. &(250)
\end{aligned}
$$

Regarding the first term of (250), remark that by property of the uniform distribution, conditional on $X'_{(1)}$ and $X'_{(n_j-1)}$, the order statistics between 1 and n_j-1 follow Beta distributions independently from X_1 and $\boldsymbol{\delta}_\ell(X_1,\Theta)$. Therefore,

$$
\frac{X'_{(n_j/2)}-X'_{(1)}}{X'_{(n_j-1)}-X'_{(1)}}|X'_{(1)},X'_{(n_j-1)}\sim\mathcal{B}(n_j/2-1,n_j/2-1),
$$

so that

$$
\mathbb{E}\left[\frac{X'_{(n_j/2)}-X'_{(1)}}{X'_{(n_j-1)}-X'_{(1)}}\middle|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),X'_{(1)},X'_{(n_j-1)}\right]=\frac{n_j/2-1}{2(n_j/2-1)}=\frac{1}{2}.
$$

Overall the first term of (250) verifies

$$\begin{aligned}
&\mathbb{E}\left[(X'_{(n_j-1)}-X'_{(1)})\mathbb{E}\left[\frac{X'_{(n_j/2)}-X'_{(1)}}{X'_{(n_j-1)}-X'_{(1)}}\big|X_1,\boldsymbol{\delta}_\ell,X'_{(1)},X'_{(n_j-1)}\right]\Big|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right]\\
&=\mathbb{E}\left[\frac{X'_{(n_j-1)}-X'_{(1)}}{2}\big|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right] &(251)\\
&\leq\mathbb{E}\left[\frac{M_{2,j}-M_{1,j}}{2}\big|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right]. &(252)
\end{aligned}$$

Regarding the second term of (250), we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{X'_{(n_j-1)}+X'_{(1)}}{2}-X'_{(n_j/2)}\right|\Big|X_1,\boldsymbol{\delta}_\ell,X'_{(1)},X'_{(n_j-1)}\right]\\
&=\mathbb{E}\left[\left|\frac{X'_{(n_j-1)}-X'_{(1)}}{2}-\left(X'_{(n_j/2)}-X'_{(1)}\right)\right|\Big|X_1,\boldsymbol{\delta}_\ell,X'_{(1)},X'_{(n_j-1)}\right] &(253)\\
&=\left(X'_{(n_j-1)}-X'_{(1)}\right)\mathbb{E}\left[\left|\frac{1}{2}-\frac{X'_{(n_j/2)}-X'_{(1)}}{X'_{(n_j-1)}-X'_{(1)}}\right|\Big|X_1,\boldsymbol{\delta}_\ell,X'_{(1)},X'_{(n_j-1)}\right] &(254)\\
&=(X'_{(n_j-1)}-X'_{(1)})\mathbb{E}\left[\left|\frac{1}{2}-\mathcal{B}(n_j/2-1,n_j/2-1)\right|\right] &(255)\\
&\leq(X'_{(n_j-1)}-X'_{(1)})\sqrt{\mathbb{E}\left[\left|\mathcal{B}(n_j/2-1,n_j/2-1)-\frac{1}{2}\right|^2\right]} &(256)\\
&\leq\frac{M_{2,j}-M_{1,j}}{2\sqrt{n_j-1}}, &(257)
\end{aligned}$$

where the last inequality is simply obtained by computing the variance of a Beta distribution.

Therefore,

$$\begin{aligned}
&2\mathbb{E}\left[\mathbb{E}\left[\left|\frac{X'_{(n_j-1)}+X'_{(1)}}{2}-X'_{(n_j/2)}\right|\Big|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),X'_{(1)},X'_{(n_j-1)}\right]\Big|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right]\\
&\leq\frac{1}{2\sqrt{n_j-1}}\mathbb{E}\left[M_{2,j}-M_{1,j}\Big|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right]. &(258)
\end{aligned}$$

The third and fourth terms of (250) have the same expression, controlled by Lemma B.8:

$$\begin{aligned}
\mathbb{E}\left[X'_{(1)}-M_{1,j}\Big|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right]&=\mathbb{E}\left[M_{2,j}-X'_{(n_j-1)}\Big|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right] &(259)\\
&\leq\frac{1}{n_j}\mathbb{E}\left[M_{2,j}-M_{1,j}\big|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right]. &(260)
\end{aligned}$$

Finally, gathering (252), (258) and (260) yields

$$\mathbb{E}\left[\mu\left(A_{n,j+1}^{(\ell)}(X_1,\Theta)\right)\middle|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right] \leq \mathbb{E}\left[M_{2,j}-M_{1,j}\middle|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right]\left(\frac{1}{2}+\frac{1}{2\sqrt{n_j-1}}+\frac{2}{n_j}\right) \tag{261}$$

$$\leq \frac{1}{2}\left(1+\frac{5}{\sqrt{n_j-1}}\right)\mathbb{E}\left[M_{2,j}-M_{1,j}\middle|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right] \tag{262}$$

$$= \frac{1}{2}\left(1+\frac{5}{\sqrt{n_j-1}}\right)\mathbb{E}\left[\mu\left(A_{n,j}^{(\ell)}(X_1,\Theta)\right)\middle|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right] \tag{263}$$

for all $n_j \geq 4$. An iterative product yields

$$\mathbb{E}\left[\mu\left(A_{n,k}^{(\ell)}(X_1,\Theta)\right)\middle|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right] \leq \mathbb{E}\left[\prod_{\substack{j:\delta_{j,\ell}=1,\\ j\leq k-2}}\frac{1}{2}\left(1+\frac{5}{\sqrt{n_j-1}}\right)\middle|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right] \tag{264}$$

$$= \prod_{\substack{j:\delta_{j,\ell}=1,\\ j\leq k-2}}\frac{1}{2}\left(1+\frac{5}{\sqrt{n_j-1}}\right) \tag{265}$$

$$= 2^{-K_\ell+2}\prod_{\substack{j:\delta_{j,\ell}=1,\\ j\leq k-2}}\left(1+\frac{5}{\sqrt{n_j-1}}\right), \tag{266}$$

which proves the first statement. Recalling that $n_j = n2^{-j}$,

$$\sum_{j=0}^{k}\log\left(1+\frac{5}{\sqrt{n}}2^{j/2}\right) \leq \frac{5}{\sqrt{n}}\frac{2^{(k+1)/2}-1}{\sqrt{2}-1} \tag{267}$$

$$\leq \frac{5}{\sqrt{2}-1}\frac{2^{(\log_2 n)/2}}{\sqrt{n}} \tag{268}$$

$$= \frac{5}{\sqrt{2}-1}, \tag{269}$$

we have

$$\mathbb{E}\left[\mu\left(A_{n,k}^{(\ell)}(X_1,\Theta)\right)\middle|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right] \leq 2^{-K_\ell+2}\exp\left(\frac{5}{\sqrt{2}-1}\right), \tag{270}$$

which proves the second statement. Note that

$$\mathbb{E}\left[\mu\left(A_{n,k}(X_1,\Theta)\right)\left|X_1,\boldsymbol{\delta}_1(X_1,\Theta),\ldots,\boldsymbol{\delta}_d(X_1,\Theta)\right.\right] \tag{271}$$

$$=\mathbb{E}\left[\prod_{\ell=1}^{d}\mu\left(A_{n,k}^{(\ell)}(X_1,\Theta)\right)\left|X_1,\boldsymbol{\delta}_1(X_1,\Theta),\ldots,\boldsymbol{\delta}_d(X_1,\Theta)\right.\right] \tag{272}$$

$$=\prod_{\ell=1}^{d}\mathbb{E}\left[\mu\left(A_{n,k}^{(\ell)}(X_1,\Theta)\right)\left|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right.\right] \tag{273}$$

$$\leq\prod_{\ell=1}^{d}\prod_{\substack{j:\delta_{j,\ell}=1,\\ j\leq k-2}}\frac{1}{2}\left(1+\frac{5}{\sqrt{n_j-1}}\right) \tag{274}$$

$$\leq\prod_{j\leq k-2}\frac{1}{2}\left(1+\frac{5}{\sqrt{n_j-1}}\right) \tag{275}$$

$$\leq 4\times 2^{-k}\prod_{j\leq k-2}\left(1+\frac{5}{\sqrt{n_j-1}}\right) \tag{276}$$

$$\leq 4\times 2^{-k}\exp\left(\frac{5}{\sqrt{2}-1}\right). \tag{277}$$

□

B.5 Proof of the main result (median RF consistency)

Theorem B.10 (Upper bound on the risk of the median forest)**.** *Consider a generic pair (X,Y) of random variables such that $Y=f^\star(X)+\varepsilon$, where $||\partial_\ell f^\star||_\infty^2$ exists for all $\ell\in\{1,\ldots,d\}$, X is uniformly distributed on $[0,1]^d$ and the noise ε satisfies, almost surely, $\mathbb{E}[\varepsilon|X]=0$ and $\mathbb{V}[\varepsilon|X]\leq\sigma^2$. Consider $n\geq 16$ i.i.d. observations, where n is a power of two, distributed as the generic pair (X,Y). Then, the risk of the infinite median forest trained on this data set satisfies*

$$\mathbb{E}\left[\left(f_{\infty,n}^{\mathrm{MedRF}}(X)-f^\star(X)\right)^2\right]\leq C_1 d\left(\sum_{\ell=1}^{d}||\partial_\ell f^\star||_\infty^2\right)\left(1-\frac{3}{4d}\right)^{\log_2 n}+\sigma^2 C_{2,d}(\log_2 n)^{-(d-1)/2}, \tag{278}$$

with

$$C_1=1024\exp\left(\frac{42+\sqrt{5}}{2-\sqrt{2}}\right)\quad\text{and}\quad C_{2,d}=2\left(32\ \exp\left(\frac{5}{\sqrt{2}-1}\right)\ \right)^d d^{d/2}. \tag{279}$$

In particular, the infinite median forest is consistent, that is

$$\lim_{n\to\infty}\mathbb{E}\left[\left(f_{\infty,n}^{\mathrm{MedRF}}(X)-f^\star(X)\right)^2\right]=0. \tag{280}$$

Proof. We begin with a simple bias/variance decomposition:

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

where the penultimate line comes from the fact that

$$\sum_{i=1}^n \mathbb{E}_\Theta\left[W_{ni}(X,\Theta)\right] = \mathbb{E}_\Theta\left[\sum_{i=1}^n W_{ni}(X,\Theta)\right] = 1, \tag{281}$$

(since all leaves contain exactly one observation), and the last line results from a null cross product.

Controlling the bias We have,

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i=1}^n \mathbb{E}_\Theta\left[W_{ni}(X,\Theta)\right]\left(f^\star(X_i) - f^\star(X)\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}_\Theta\left[\sum_{i=1}^n W_{ni}(X,\Theta)(f^\star(X_i) - f^\star(X))\right]\right)^2\right] &(282)\\
&\leq \mathbb{E}\left[\left(\sum_{i=1}^n W_{ni}(X,\Theta)(f^\star(X_i) - f^\star(X))\right)^2\right] &(283)\\
&\leq \mathbb{E}\left[\left(\sum_{i=1}^n \mathbb{1}_{X\in A_n(X_i,\Theta)}(f^\star(X_i) - f^\star(X))\right)^2\right] &(284)\\
&\leq \mathbb{E}\left[\sum_{i=1}^n \mathbb{1}_{X\in A_n(X_i,\Theta)}\left(f^\star(X_i) - f^\star(X)\right)^2\right], &(285)
\end{aligned}$$

because $W_{ni}(X,\Theta) = \mathbb{1}_{X \in A_n(X_i,\Theta)}$ and by applying twice Jensen inequality (third and fifth lines). Noticing that,

$$\begin{aligned}\mathbb{1}_{X \in A_n(X_i,\Theta)} |f^\star(X) - f^\star(X_i)| &\leq \sum_{\ell=1}^{d} ||\partial_\ell f^\star||_\infty |X_i^{(\ell)} - X^{(\ell)}| \mathbb{1}_{X \in A_n(X_i,\Theta)} \\ &\leq \sum_{\ell=1}^{d} ||\partial_\ell f^\star||_\infty \mu(A_n^{(\ell)}(X,\Theta)) \mathbb{1}_{X \in A_n(X_i,\Theta)},\end{aligned}$$

we get,

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{X \in A_n(X_i,\Theta)} \left(f^\star(X_i) - f^\star(X)\right)^2\right] &\leq \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{X \in A_n(X_i,\Theta)} \left(\sum_{\ell=1}^{d} ||\partial_\ell f^\star||_\infty \mu(A_n^{(\ell)}(X,\Theta))\right)^2\right] \\ &\leq \mathbb{E}\left[\left(\sum_{\ell=1}^{d} ||\partial_\ell f^\star||_\infty \mu(A_n^{(\ell)}(X,\Theta))\right)^2\right] \qquad (286) \\ &\leq \left(\sum_{\ell=1}^{d} ||\partial_\ell f^\star||_\infty^2\right) \sum_{\ell=1}^{d} \mathbb{E}\left[\mu(A_n^{(\ell)}(X,\Theta))^2\right]. \qquad (287)\end{aligned}$$

where the last inequality directly results from Cauchy-Schwarz inequality. By Lemma B.7, since $k = \lfloor \log_2 n \rfloor$,

$$\left(\sum_{\ell=1}^{d} ||\partial_\ell f^\star||_\infty^2\right) \sum_{\ell=1}^{d} \mathbb{E}\left[\mu(A_n^{(\ell)}(X,\Theta))^2\right] \leq Cd \left(\sum_{\ell=1}^{d} ||\partial_\ell f^\star||_\infty^2\right) \left(1 - \frac{3}{4d}\right)^{\log_2 n}, \qquad (288)$$

with

$$C = 1024 \exp\left(\frac{42 + \sqrt{5}}{2 - \sqrt{2}}\right). \qquad (289)$$

Controlling the variance Following Biau [8], the variance term of the median forest writes

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbb{E}_\Theta\left[W_{ni}(X,\Theta)\right] \varepsilon_i\right)^2\right] &= \mathbb{E}\left[\sum_{i=1}^{n} \left(\mathbb{E}_\Theta\left[W_{ni}(X,\Theta)\right]\right)^2 \varepsilon_i^2\right] \qquad (290) \\ &= \mathbb{E}\left[\sum_{i=1}^{n} \left(\mathbb{E}_\Theta\left[W_{ni}(X,\Theta)\right]\right)^2 \mathbb{E}\left[\varepsilon_i^2 | X, X_1, \ldots, X_n\right]\right] \qquad (291) \\ &\leq \mathbb{E}\left[\sum_{i=1}^{n} \left(\mathbb{E}_\Theta\left[W_{ni}(X,\Theta)\right]\right)^2 \sigma^2\right] \qquad (292) \\ &\leq \sigma^2 n \mathbb{E}\left[\left(\mathbb{E}_\Theta\left[W_{n1}(X,\Theta)\right]\right)^2\right], \qquad (293)\end{aligned}$$

where we have used the fact that the cross products are null (since $\mathbb{E}[\varepsilon_i|X_i] = 0$). Since each leaf of the median tree contains exactly one observation, denoting Θ' an i.i.d. copy of Θ, we have

$$\begin{aligned}\left(\mathbb{E}_\Theta\left[W_{n1}(X,\Theta)\right]\right)^2 &= \mathbb{E}_\Theta\left[W_{n1}(X,\Theta)\right]\mathbb{E}_{\Theta'}\left[W_{n1}(X,\Theta')\right]\\ &= \mathbb{E}_{\Theta,\Theta'}\left[W_{n1}(X,\Theta)W_{n1}(X,\Theta')\right]\\ &= \mathbb{E}_{\Theta,\Theta'}\left[\mathbb{1}_{X\in A_n(X_1,\Theta)}\mathbb{1}_{X\in A_n(X_1,\Theta')}\right].\end{aligned}$$

Consequently,

$$\mathbb{E}\left[\left(\sum_{i=1}^n \mathbb{E}_\Theta\left[W_{ni}(X,\Theta)\right]\varepsilon_i\right)^2\right] \le \sigma^2 n\mathbb{E}\left[\mathbb{1}_{X\in A_n(X_1,\Theta)}\mathbb{1}_{X\in A_n(X_1,\Theta')}\right].$$

For all ℓ, we let $A_n^{(\ell)}(X_1,\Theta)$ be the cell $A_n(X_1,\Theta)$ projected onto the ℓ-th dimension. Let also $\boldsymbol{\delta}_\ell(X_1,\Theta)$ be the vector whose components are defined as $\delta_{j,l} = 1$ if the j-th cut of the cell $A_n^{(\ell)}(X_1,\Theta)$ is made along direction ℓ and 0 otherwise. We define similarly $\boldsymbol{\delta}_\ell(X_1,\Theta')$ for the cell $A_n^{(\ell)}(X_1,\Theta')$. We also let $K_\ell = \|\boldsymbol{\delta}_\ell(X_1,\Theta)\|_1$ (resp. K'_ℓ) be the number of times the ℓ-th direction is split in the tree built with Θ (resp. Θ'). Then,

$$\begin{aligned}&\mathbb{E}\left[\left(\sum_{i=1}^n \mathbb{E}_\Theta\left[W_{ni}(X,\Theta)\right]\varepsilon_i\right)^2\right]\\ &\le \sigma^2 n\mathbb{E}\left[\mathbb{1}_{X\in A_n(X_1,\Theta)\cap A_n(X_1,\Theta')}\right]\\ &= \sigma^2 n\mathbb{E}\left[\prod_{\ell=1}^d \mu\left(A_n^{(\ell)}(X_1,\Theta)\cap A_n^{(\ell)}(X_1,\Theta')\right)\right]\\ &= \sigma^2 n\mathbb{E}\left[\mathbb{E}\left[\prod_{\ell=1}^d \mu\left(A_n^{(\ell)}(X_1,\Theta)\cap A_n^{(\ell)}(X_1,\Theta')\right)\middle| X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),\boldsymbol{\delta}_\ell(X_1,\Theta')\right]\right]\\ &= \sigma^2 n\mathbb{E}\left[\prod_{\ell=1}^d \mathbb{E}\left[\mu\left(A_n^{(\ell)}(X_1,\Theta)\cap A_n^{(\ell)}(X_1,\Theta')\right)\middle| X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),\boldsymbol{\delta}_\ell(X_1,\Theta')\right]\right].\end{aligned}$$

The last equality is obtained by conditional independence: indeed, as the X_is are uniformly distributed, the positions of the coordinates do not influence each others. Therefore only the number of cuts along the other directions will influence the length the cell along a given direction, hence the conditional independence. Now,

$$\begin{aligned}&\mathbb{E}\left[\mu\left(A_n^{(\ell)}(X_1,\Theta)\cap A_n^{(\ell)}(X_1,\Theta')\right)\middle| X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),\boldsymbol{\delta}_\ell(X_1,\Theta')\right]\\ &\le \mathbb{E}\left[\min(\mu(A_n^{(\ell)}(X_1,\Theta)),\mu(A_n^{(\ell)}(X_1,\Theta'))\middle| X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),\boldsymbol{\delta}_\ell(X_1,\Theta')\right]\\ &= \frac{1}{2}\left(\mathbb{E}\left[\mu(A_n^{(\ell)}(X_1,\Theta))\middle| X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right] + \mathbb{E}\left[\mu(A_n^{(\ell)}(X_1,\Theta'))\middle| X_1,\boldsymbol{\delta}_\ell(X_1,\Theta')\right]\right)\\ &\quad - \frac{1}{2}\mathbb{E}\left[|\mu(A_n^{(\ell)}(X_1,\Theta)) - \mu(A_n^{(\ell)}(X_1,\Theta'))|\middle| X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),\boldsymbol{\delta}_\ell(X_1,\Theta')\right].\end{aligned}$$

Moreover,

$$
\begin{aligned}
&\mathbb{E}\left[\left|\mu(A_n^{(\ell)}(X_1,\Theta))-\mu(A_n'^{(\ell)}(X_1,\Theta))\right|\middle| X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),\boldsymbol{\delta}_\ell(X_1,\Theta')\right]\\
&=\mathbb{E}\left[\left|\mu(A_n^{(\ell)}(X_1,\Theta))-\mu(A_n^{(\ell)}(X_1,\Theta'))\right|\left(\mathbb{1}_{K_\ell<K_\ell'}+\mathbb{1}_{K_\ell\geq K_\ell'}\right)\middle| X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),\boldsymbol{\delta}_\ell(X_1,\Theta')\right]\\
&\geq\mathbb{E}\left[\left(\mu(A_n^{(\ell)}(X_1,\Theta))-\mu(A_n^{(\ell)}(X_1,\Theta'))\right)\middle| X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),\boldsymbol{\delta}_\ell(X_1,\Theta')\right]\mathbb{1}_{K_\ell<K_\ell'}\\
&\quad+\mathbb{E}\left[\left(\mu(A_n^{(\ell)}(X_1,\Theta'))-\mu(A_n^{(\ell)}(X_1,\Theta))\right)\middle| X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),\boldsymbol{\delta}_\ell(X_1,\Theta')\right]\mathbb{1}_{K_\ell\geq K_\ell'}\\
&\geq\left(\mathbb{E}\left[\mu(A_n^{(\ell)}(X_1,\Theta))|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right]-\mathbb{E}\left[\mu(A_n^{(\ell)}(X_1,\Theta'))|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta')\right]\right)\mathbb{1}_{K_\ell<K_\ell'}\\
&\quad+\left(\mathbb{E}\left[\mu(A_n^{(\ell)}(X_1,\Theta'))|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta')\right]-\mathbb{E}\left[\mu(A_n^{(\ell)}(X_1,\Theta))|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right]\right)\mathbb{1}_{K_\ell\geq K_\ell'}.
\end{aligned}
$$

Letting $B_\ell=\mathbb{E}\left[\mu(A_n^{(\ell)}(X_1,\Theta))|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta)\right]$ and $B_\ell'=\mathbb{E}\left[\mu(A_n^{(\ell)}(X_1,\Theta'))|X_1,\boldsymbol{\delta}_\ell(X_1,\Theta')\right]$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\mu\left(A_n^{(\ell)}(X_1,\Theta)\cap A_n^{(\ell)}(X_1,\Theta')\right)\middle| X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),\boldsymbol{\delta}_\ell(X_1,\Theta')\right]\\
&\leq\frac{1}{2}\left(B_\ell+B_\ell'\right)-\frac{1}{2}(B_\ell-B_\ell')\mathbb{1}_{K_\ell<K_\ell'}-\frac{1}{2}(B_\ell'-B_\ell)\mathbb{1}_{K_\ell\geq K_\ell'}\\
&\leq B_\ell\mathbb{1}_{K_\ell\geq K_\ell'}+B_\ell'\mathbb{1}_{K_\ell<K_\ell'}.
\end{aligned}
$$

Now, according to Lemma B.9, letting $C_2=4\exp(5/(\sqrt{2}-1))$, we have $B_\ell\leq C_2 2^{-K_\ell}$ and $B_\ell'\leq C_2 2^{-K_\ell'}$. Therefore,

$$
\begin{aligned}
\mathbb{E}\left[\mu\left(A_n^{(\ell)}(X_1,\Theta)\cap A_n^{(\ell)}(X_1,\Theta')\right)\middle| X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),\boldsymbol{\delta}_\ell(X_1,\Theta')\right]&\leq C_2 2^{-K_\ell}\mathbb{1}_{K_\ell\geq K_\ell'}+C_2 2^{-K_\ell'}\mathbb{1}_{K_\ell<K_\ell'}\\
&\leq C_2 2^{-\max(K_\ell,K_\ell')}.
\end{aligned}
$$

Overall,

$$
\begin{aligned}
\mathbb{E}\left[\left(\sum_{i=1}^n\mathbb{E}_\Theta\left[W_{ni}(X,\Theta)\right]\varepsilon_i\right)^2\right]&\leq\sigma^2 n\mathbb{E}\left[\prod_{\ell=1}^d\mathbb{E}\left[\mu\left(A_n^{(\ell)}(X_1,\Theta)\cap A_n^{(\ell)}(X_1,\Theta')\right)\middle| X_1,\boldsymbol{\delta}_\ell(X_1,\Theta),\boldsymbol{\delta}_\ell(X_1,\Theta')\right]\right] &(294)\\
&\leq\sigma^2 n\mathbb{E}\left[\prod_{\ell=1}^d C_2 2^{-\max(K_\ell,K_\ell')}\right] &(295)\\
&\leq\sigma^2 C_2^d n\ \mathbb{E}\left[2^{-\sum_{\ell=1}^d\max(K_\ell,K_\ell')}\right] &(296)\\
&\leq\sigma^2 C_2^d n\ \ 2^{-k_n}\mathbb{E}\left[2^{-\sum_{\ell=1}^d|K_\ell-K_\ell'|}\right], &(297)
\end{aligned}
$$

since

$$\sum_{\ell=1}^{d} \max(K_\ell, K'_\ell) = \frac{1}{2}\sum_{\ell=1}^{d} K_\ell + \frac{1}{2}\sum_{\ell=1}^{d} K'_\ell + \frac{1}{2}\sum_{\ell=1}^{d} |K_\ell - K'_\ell|$$
$$= k_n + \frac{1}{2}\sum_{\ell=1}^{d} |K_\ell - K'_\ell|.$$

According to Lemma S.1 from Klusowski [19] (see Supplementary Materials), one has

$$\mathbb{E}\left[2^{-\sum_{\ell=1}^{d}|K_\ell - K'_\ell|}\right] \leq \frac{8^d\ d^{d/2}}{k_n^{(d-1)/2}}. \tag{298}$$

Finally, combining (297) and (298), the variance of the median forest is upper bounded by

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbb{E}_\Theta\left[W_{ni}(X,\Theta)\right]\varepsilon_i\right)^2\right] \leq \sigma^2 C_2^d n\ 2^{-k_n}\frac{8^d\ d^{d/2}}{k_n^{(d-1)/2}} \tag{299}$$
$$\leq 2\sigma^2\left(8\ C_2\ d^{1/2}\right)^d (\log_2 n)^{-(d-1)/2}, \tag{300}$$

since $k_n = \lfloor \log_2(n) \rfloor$. All in all,

$$\mathbb{E}\left[\left(f_{\infty,n}^{\mathrm{MedRF}}(X) - f^\star(X)\right)^2\right]$$
$$\leq Cd\left(\sum_{\ell=1}^{d} ||\partial_\ell f^\star||_\infty^2\right)\left(1 - \frac{3}{4d}\right)^{\log_2 n} + 2\sigma^2\left(8\ C_2\ d^{1/2}\right)^d (\log_2 n)^{-(d-1)/2}$$

with $C_2 = 4\exp(5/(\sqrt{2}-1))$ and

$$C = 1024\exp\left(\frac{42+\sqrt{5}}{2-\sqrt{2}}\right). \tag{301}$$

□

B.5.1 Controlling the variance of an interpolating Median RF in an asymptotic high-dimensional setting

The following result shows the decrease of the variance of the Median RF under an asymptotic high-dimensional framework. It is also numerically illustrated in Section C.1.5.

Proposition B.11. *For all $d > \log_2 n$, the variance of the infinite interpolating Median RF $f_{\infty,n}^{\mathrm{MedRF}}$ verifies*

$$V(f_{\infty,n}^{\mathrm{MedRF}}) = \mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbb{E}_\Theta\left[W_{ni}(X,\Theta)\right]\varepsilon_i\right)^2\right] \leq \frac{4C_2^2\sigma^2}{n} + 2C_2\sigma^2\left(1 - \exp\left(-\frac{\log_2^2 n}{d - \log_2 n}\right)\right),$$

where $C_2 = 4\exp\left(5/(\sqrt{2}-1)\right)$. *Suppose that the input dimension* d *dominates* $\log_2^2 n$ *asymptotically* ($d \gg \log_2^2 n$), *then the variance tends to 0 (as* n, d *tends to infinity), with a rate of the order of* $\max(\frac{\log^2 n}{d}, \frac{1}{n})$.

The proof is given below. This results shows that the Median RF benefits from an increase of the dimension as it will improve its averaging effect and help to reduce the variance. Of course, in such a setting, the variance is only one part of the story, and a control on the bias becomes a real hindrance (as the approximation error may explode), unless extra model assumptions are formulated. For instance, consider for any input dimension d the case of a linear model, i.e. $Y = X^\top \theta + \varepsilon$ for $\theta \in \mathbb{R}^d$ and such that $\|\theta\|_2 \leq C/\sqrt{d}$, with $C > 0$ a constant. One can actually show that in such a setting, the bias term remains bounded as n (and d) grows towards infinity (using for example the analysis conducted in the next theorem). This echoes in particular the behavior of ridgeless least squares estimator in modern interpolation regimes [see, 17].

Proof of Proposition B.11. A typical bias-variance decomposition yields (see e.g. [8])

$$V(f_{\infty,n}^{\mathrm{MedRF}}) \leq \sigma^2 n \mathbb{P}\left(X \in A_n(X_1, \Theta) \cap A_n(X_1, \Theta')\right) \tag{302}$$

with Θ' an independent copy of Θ. Recalling that the depth is chosen as $k = \lfloor \log_2 n \rfloor$. Consider the event

$$E = E(\Theta, \Theta', X_1, k) := \{\Theta \text{ and } \Theta' \text{ do not cut on common directions on the path to } X_1\}.$$

Denote $M(\Theta, X_1)$ the number of distinct directions chosen by the tree Θ to produce the leaf containing X_1 (upper bounded by $\log_2 n$). Then,

$$\begin{aligned}
\mathbb{P}(E) &\geq \mathbb{E}\left[\left(\frac{d - M(\Theta, X_1)}{d}\right)^{\log_2 n}\right] && (303)\\
&\geq \left(\frac{d - \log_2 n}{d}\right)^{\log_2 n} && (304)\\
&= \exp\left(\log_2 n \log\left(1 - \frac{\log_2 n}{d}\right)\right) && (305)\\
&\geq \exp\left(-\frac{\log_2^2 n}{d - \log_2 n}\right), && (306)
\end{aligned}$$

using, for all $x \in [0, 1)$, $\log(1 - x) \geq -x/(1 - x)$. The above probability tends to 1 as soon as $d \gg \log_2^2 n$. Then,

$$\begin{aligned}
&\mathbb{P}\left(X \in A_n(X_1, \Theta) \cap A_n(X_1, \Theta')\right) && (307)\\
&= \mathbb{P}\left(\{X \in A_n(X_1, \Theta) \cap A_n(X_1, \Theta')\} \cap E\right) + \mathbb{P}\left(\{X \in A_n(X_1, \Theta) \cap A_n(X_1, \Theta')\} \cap E^c\right)\\
&\leq \mathbb{P}\left(X \in A_n(X_1, \Theta) | X \in A_n(X_1, \Theta'), E\right) \mathbb{P}\left(X \in A_n(X_1, \Theta')\right) + \mathbb{P}\left(\{X \in A_n(X_1, \Theta)\} \cap E^c\right). && (308)
\end{aligned}$$

Applying Lemma B.9 (Line (241)) yields

$$\begin{aligned}\mathbb{P}\left(X \in A_n(X_1, \Theta')\right) &= \mathbb{E}\left[\mu\left(A_n(X_1, \Theta')\right]\right. && (309)\\ &= \mathbb{E}\left[\mathbb{E}\left[\mu(A_n(X_1, \Theta))|X_1, \boldsymbol{\delta}_1(X_1, \Theta), \ldots, \boldsymbol{\delta}_d(X_1, \Theta)\right]\right] && (310)\\ &\leq C_2 2^{-k} && (311)\end{aligned}$$

with $C_2 = 4\exp\left(\frac{5}{\sqrt{2}-1}\right)$. Moreover, conditional on E, $\{X \in A_n(X_1, \Theta)\}$ and $\{X \in A_n(X_1, \Theta')\}$ are independent as Θ and Θ' do not share any common direction on the path to X_1 and therefore the splits in Θ and Θ' are performed on independent sample components (by uniformity of X and X_1). Therefore, also by Lemma B.9,

$$\begin{aligned}\mathbb{P}\left(X \in A_n(X_1, \Theta)|X \in A_n(X_1, \Theta'), E\right) &= \mathbb{E}\left[\mu\left(A_n(X_1, \Theta)\right)\right] && (312)\\ &= \mathbb{E}\left[\mathbb{E}\left[\mu(A_n(X_1, \Theta))|X_1, \boldsymbol{\delta}_1(X_1, \Theta), \ldots, \boldsymbol{\delta}_d(X_1, \Theta)\right]\right] && (313)\\ &\leq C_2 2^{-k}. && (314)\end{aligned}$$

Similarly, the volume $\mu\left(A_n(X_1, \Theta)\right)$ is independent of the directions chosen to build the leaf, therefore

$$\begin{aligned}\mathbb{P}\left(\{X \in A_n(X_1, \Theta)\} \cap E^c\right) &= \mathbb{P}\left(X \in A_n(X_1, \Theta)\right)\mathbb{P}\left(E^c\right)\\ &\leq C_2 2^{-k}\left(1 - \exp\left(-\frac{\log_2^2 n}{d - \log_2 n}\right)\right).\end{aligned}$$

Overall,

$$\mathbb{P}\left(X \in A_n(X_1, \Theta) \cap A_n(X_1, \Theta')\right) \leq C_2^2 2^{-2k} + C_2 2^{-k}\left(1 - \exp\left(-\frac{\log_2^2 n}{d - \log_2 n}\right)\right)$$

and

$$V(f_{\infty,n}^{\text{MedRF}}) \leq C_2^2 n\sigma^2 2^{-2k} + n\sigma^2 C_2\left(1 - \exp\left(-\frac{\log_2^2 n}{d - \log_2 n}\right)\right)2^{-k}. \tag{315}$$

Since $k = \lfloor \log_2 n \rfloor$, we have $2^{-k} \leq 2/n$ and

$$V(f_{\infty,n}^{\text{MedRF}}) \leq \frac{4C_2^2\sigma^2}{n} + 2C_2\sigma^2\left(1 - e^{-\frac{\log_2^2 n}{d - \log_2 n}}\right). \tag{316}$$

□

B.5.2 Proof of Proposition 5.3 (Interpolation volume of Median RF)

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 $Z_1, \dots, Z_n$ be n i.i.d. random variables uniformly distributed over $[0,1]$. As in the infinite Median RF, the univariate trees, i.e., built by cutting along one direction only, appear almost surely. Then, the length of a leaf of such tree is bounded in expectation by $Z_{(k+1)} - Z_{(k-1)}$ where $Z_{(i)}$ indicates the i-the statistical order. Moreover, it is known that $Z_{(k)}$ follows a Beta distribution of parameters $(k, n-k+1)$. Therefore,

$$\mathbb{E}\left[Z_{(k+1)} - Z_{(k-1)}\right] = \frac{k+1}{n+1} - \frac{k-1}{n+1} \tag{317}$$

$$\leq \frac{2}{n}. \tag{318}$$

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.

B.6 Proofs of Section 6 (Interpolation volume of Breiman RF)

Proof of Proposition 6.1. Before diving into the computations, let us recall two facts about Breiman RF construction. First, in CART, each cut is made at the middle of two consecutive points in a given direction. Second, considering all univariate trees (trees whose splits are performed along one single direction), 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 around X_i as

$$\mathcal{A}_{min,X_i} := \bigcap_{M \in \mathbb{N}, \boldsymbol{\Theta}_M} \mathcal{A}_{X_i, \boldsymbol{\Theta}_M}. \tag{319}$$

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 6.

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 $\mathcal{A}_{X_i, \Theta_M} = \{x \in [0,1]^d, m_{M,n}(x, \boldsymbol{\Theta}_M) =$

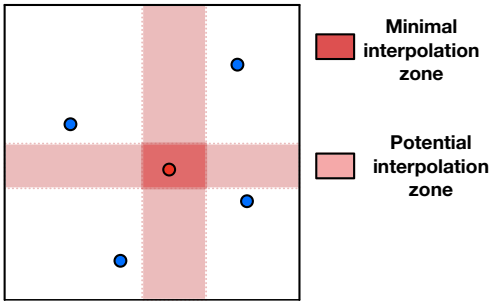

Figure 6: Different interpolation zones of a data point (in red).

$Y_i\}$. We begin with a one-dimensional analysis, and consider, without loss of generality, the first variable. We let $Z_1 := X_1^{(1)}, ..., Z_n := X_n^{(1)}$ the first components of the observations $X_1, \dots, X_n$. As $X_1, \dots, X_n$ are i.i.d. and follow a uniform distribution over $[0,1]^d$, $Z_1, ..., Z_n$ are i.i.d. and uniformly distributed on $[0,1]$. We consider the interpolation area at $x = Z_n$ and we reason conditional on Z_n in the following. The length (volume) of $\mathcal{A}_{min,x}$ 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,

$$\mu(A_{min,x}) = \frac{1}{2}\left(x - \max_{\{Z_i, Z_i<x\}\cup\{0\}} Z_i + \min_{\{Z_i, Z_i>x\}\cup\{1\}} Z_i - x\right). \tag{320}$$

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

$$\mathbb{P}\left(\frac{1}{2}\left(x - \max_{\{Z_i, Z_i<x\}\cup\{0\}} Z_i\right) \leq t \,\middle|\, x\right) \tag{321}$$

$$= 1 - \mathbb{P}\left(\max_{\{Z_i, Z_i<x\}\cup\{0\}} Z_i < x - 2t \,\middle|\, x\right) \tag{322}$$

$$= 1 - \mathbb{E}\left[\mathbb{E}\left[\mathbb{P}\left((Z_{i_1} < x-2t) \cap ... \cap (Z_{i_{N_x}} < x-2t)\middle| N_x, Z_{i_1} < x, ..., Z_{i_{N_x}} < x, x\right)\right] \middle| x\right] \tag{323}$$

$$= 1 - \mathbb{E}\left[\mathbb{P}\left(Z_1 < x - 2t | Z_1 \leq x, x\right)^{N_x} \middle| x\right] \tag{324}$$

$$= 1 - \sum_{k=0}^{n-1} \mathbb{P}\left(N_x = k | x\right) \mathbb{P}\left(Z_1 < x - 2t | Z_1 < x, x\right)^k \tag{325}$$

$$= 1 - \sum_{k=0}^{n-1} \mathbb{P}\left(N_x = k | x\right) \left(\frac{x-2t}{x}\right)^k \tag{326}$$

$$= 1 - \left((1-x) + x\left(\frac{x-2t}{x}\right)\right)^{n-1} \tag{327}$$

$$= 1 - (1-2t)^{n-1} \tag{328}$$

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 $t \geq 0$,

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

By symmetry,

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

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

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

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,

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

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

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

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.

□

C Experiments

For all experiments, we consider four different regression models, most of which have been already considered in [29]: 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:

- **Model 1:** $d = 2$, $Y = 2X_1^2 + \exp(-X_2^2)$
- **Model 2:** $d = 6$, $Y = X_1^2 + X_2^2 X_3 e^{-|X_4|} + X_5 - X_6 + \mathcal{N}(0, 0.5)$
- **Model 3:** $d = 8$, $Y = X_1X_2 + X_3^2 - X_4X_5 + X_6X_7 - X_8^2 + \mathcal{N}(0, 0.5)$
- **Model 4:** $d = 5$, $Y = 1/(1 + \exp(-10(\sum_{i=1}^{d} X_i - 1/2))) + \mathcal{N}(0, 0.05)$
- **Model 5**: $d = 4$, $Y = -\sin(2X_1X_2) + X_2^2 + X_3 - e^{X_4} + \mathcal{N}(0, 0.5)$
- **Model 6**: $d = 8$, $Y = \mathbb{1}_{\{X_1 \geq 0\}} + X_2^3 + \mathbb{1}_{\{X_3 + X_5 - X_6 - X_7 - X_8 \geq 1\}} + e^{-X_2^2} + \mathcal{N}(0, 0.5)$
- **Model 7**: $d = 4$, $Y = X_1 + 2(X_2 - 1)^2 + \frac{\sin(2\pi X_3)}{2 - \sin(2\pi X_3)} + 2\sin(2\pi X_4) + 2\cos(2\pi X_4) + 4\sin(2\pi X_4)^2 + 4\cos(2\pi X_4)^2 + \mathcal{N}(0, 0.5)$
- **Model 8**: $d = 4$, $Y = X_1 + 3X_2^2 - 2e^{X_3} + X_4$.

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. Experiments were run on 4 16-cores CPU and took at most a few hours to run.

C.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.
- 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.

C.1.1 Consistency of KeRF in the mean interpolation regime

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.

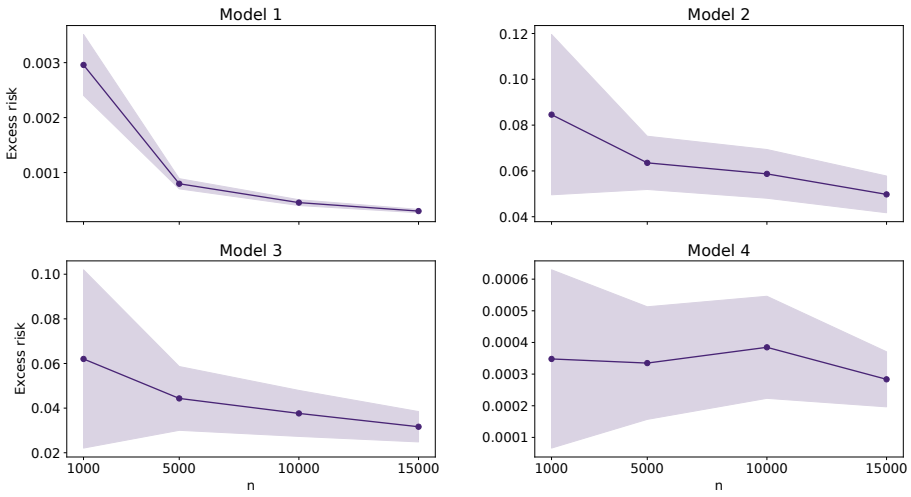

Figure 7: KeRF consistency results: excess risk w.r.t. sample sizes. For each sample size n, the experiment is repeated 30 times: we represent the mean over the 30 tries (bold line) and the mean $\pm$ std (filled zone).

Results On Figure 7, 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).

C.1.2 Consistency of Median RF in the interpolation regime

We analyze the empirical performances of Median RF in noiseless and noisy settings on the models specified above. 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 8 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.

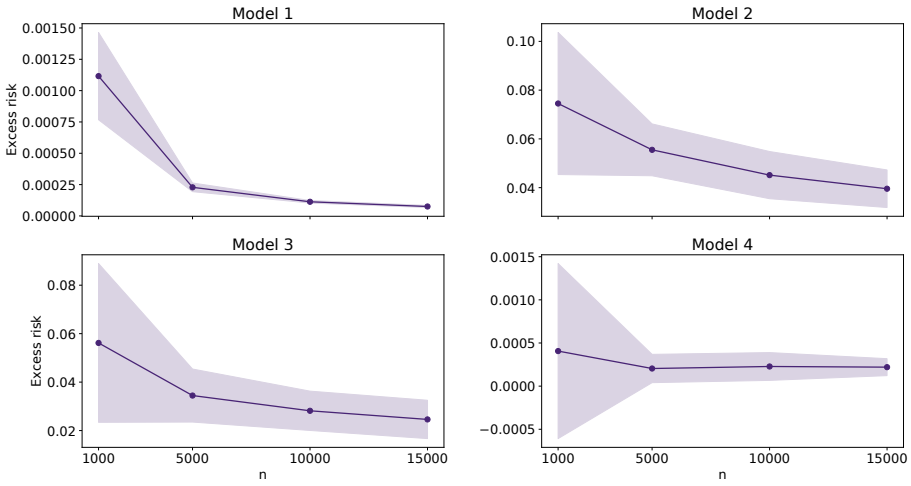

Figure 8: Consistency results for a Median RF with $M = 500$ trees: excess risk w.r.t. the sample size n. For each sample size, the experiment is repeated 30 times: we represent the mean over the 30 tries (bold line) and the mean $\pm$ std (filled zone).

C.1.3 Consistency of Breiman RF, additional models to Figure 1

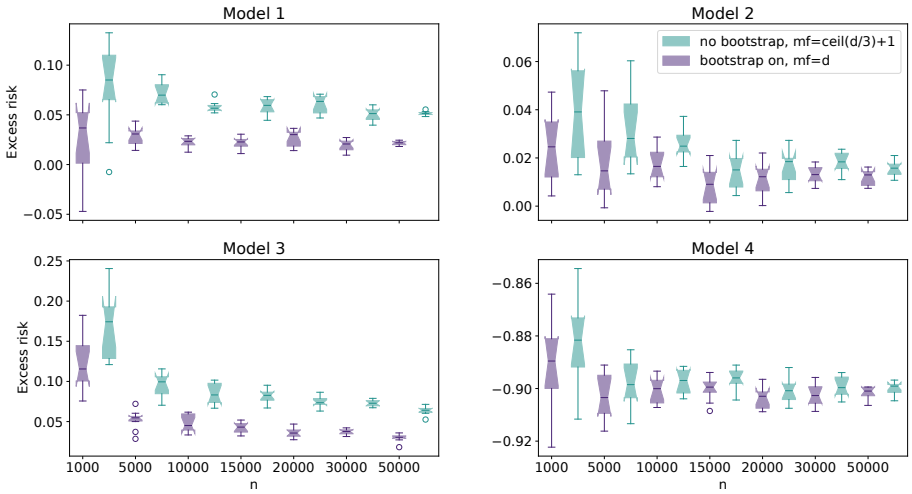

Figure 9: Consistency of Breiman RF: excess risk w.r.t the sample size n. RF parameters: 2000 trees, max-depth set to None, max-features= 1. Boxplots over 10 tries.

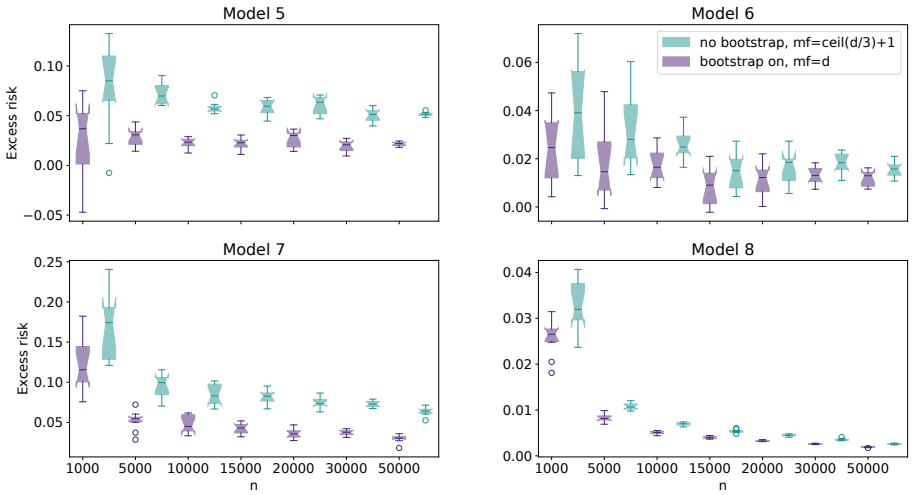

Figure 10: Consistency of Breiman RF: excess risk w.r.t the sample size n. RF parameters: 2000 trees, max-depth set to None, max-features= 1. Boxplots over 10 tries.

C.1.4 Consistency of Breiman RF with max-feature= 1

On Figure 11, 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.

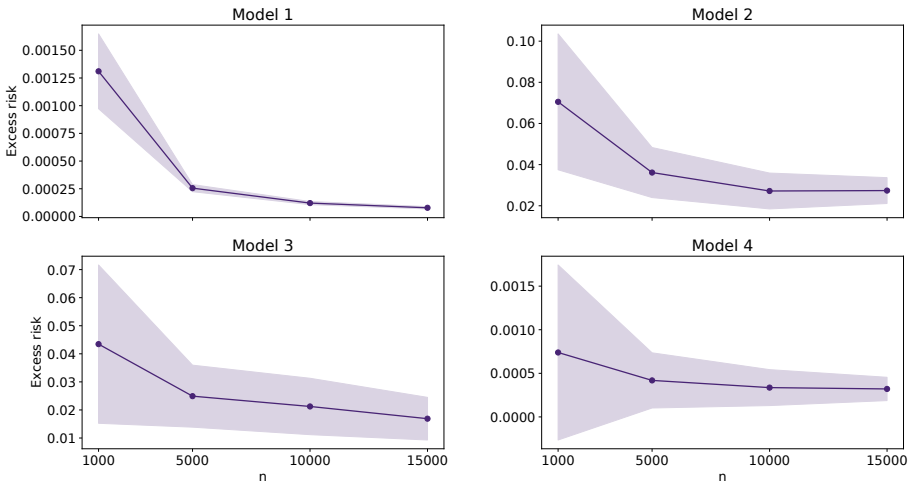

Figure 11: Consistency of Breiman RF: excess risk w.r.t sample size. RF parameters: 500 trees, max-depth set to None, max-features= 1, no bootstrap. Mean over 30 tries (doted line) and std (filled zone).

C.1.5 Decrease of the variance of the Breiman RF in a high-dimensional setting

Numerical experiments show the decrease of the variance of interpolating Breiman RF when d increases. The model involves no signal and only noise (with specified variance σ^2).

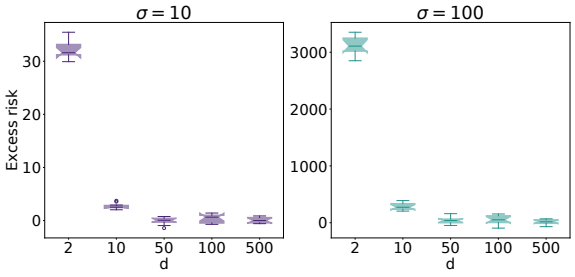

Figure 12: Decrease of the variance of an interpolating Breiman RF with `max-features=1` w.r.t. dimension d. 10 repetitions per boxplot, 5000 training points and 50000 testing points were used for each repetition. The Breiman RF contains 1000 trees.

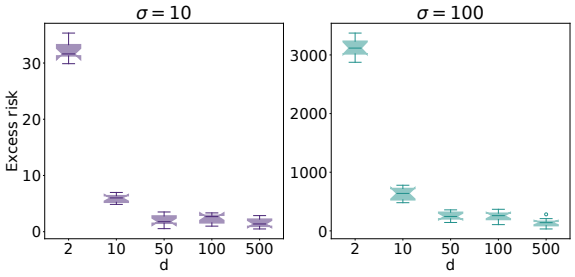

Figure 13: Decrease of the variance of an interpolating Breiman RF with `max-features=`$\lfloor d/3 \rfloor$ w.r.t. dimension d. 10 repetitions per boxplot, 5000 training points and 50000 testing points were used for each repetition. The Breiman RF contains 1000 trees.

C.1.6 Comparison of Breiman RF with and without bootstrap

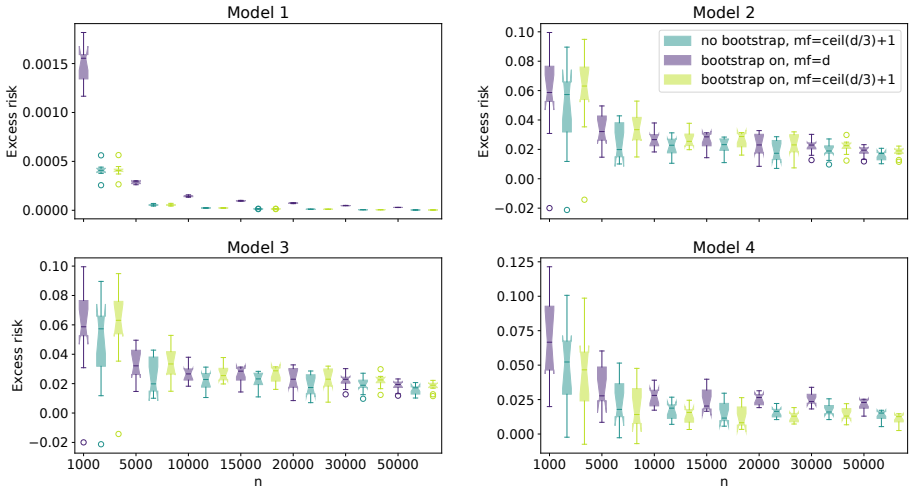

Figure 14: Consistency of Breiman RF: excess risk w.r.t sample size. RF parameters: 2000 trees, max-depth set to None, max-features= 1. Boxplots over 10 tries.

C.2 Interpolation experiments

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

We numerically evaluate the volume of the interpolation area of a Breiman RF (with 5000 trees, see Figure 17 in Appendix C.2 for details about this choice) when the sample size n increases.

In Figure 15, the volume of the minimal interpolation zone is shown to tend polynomially fast to 0

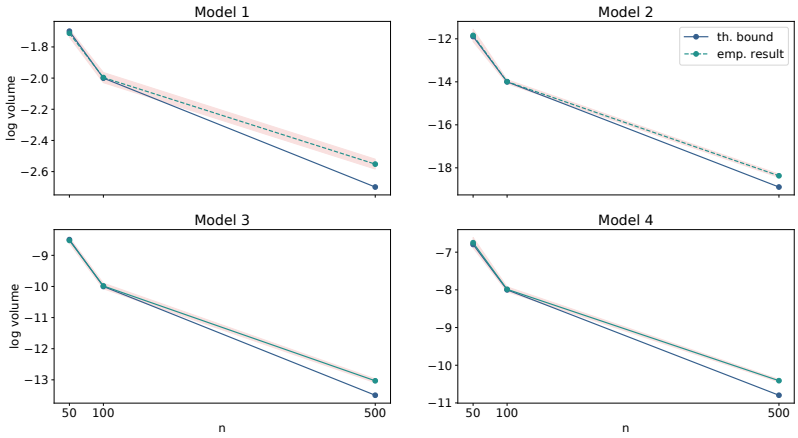

Figure 15: Log volume of the interpolation zone of a Breiman RF with 5000 Trees, max features set to 1, no bootstrap. Mean over 10 tries (red line) and mean ± std (filled zone). The theoretical bound (Proposition 6.1) is represented in green.

(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.

Increasing `max-feature` parameter We plot on Figure 16 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 15) 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.

C.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.

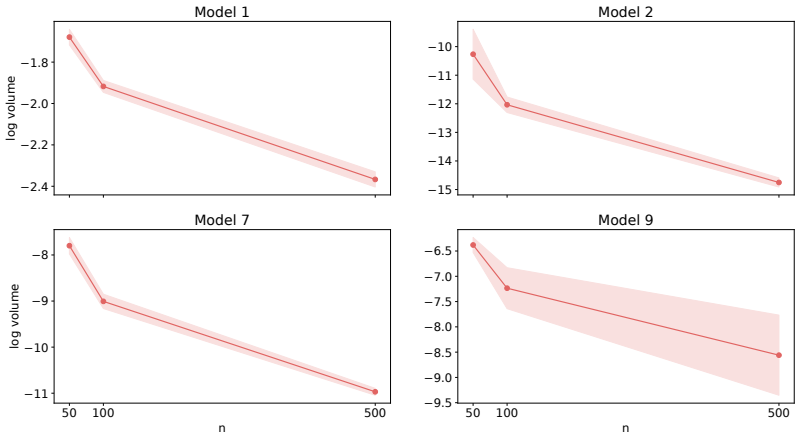

Figure 16: Log volume of Breiman RF interpolation zone w.r.t. sample size n. RF parameters: 500 trees, no bootstrap, max features $= \lceil d/3 \rceil$. Mean over 10 tries (bold line) and std (filled zone).

The numerical results in Figure 17 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.

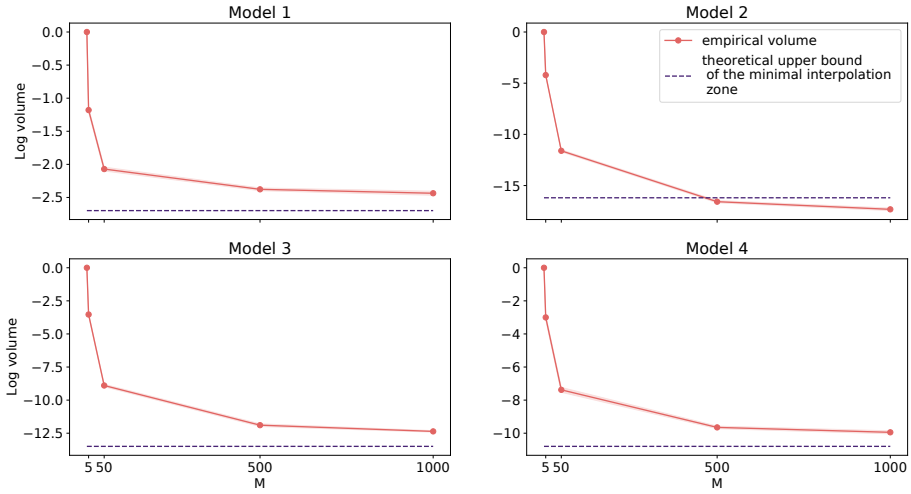

Figure 17: Log volume of Breiman RF interpolation zone w.r.t. the number M of trees. RF parameters: no bootstrap, max features $= 1$. Mean over 10 tries (bold line) and std (filled zone). Sample size $n = 500$.

C.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 18, we plot different quantiles of the above quantity as n varies.

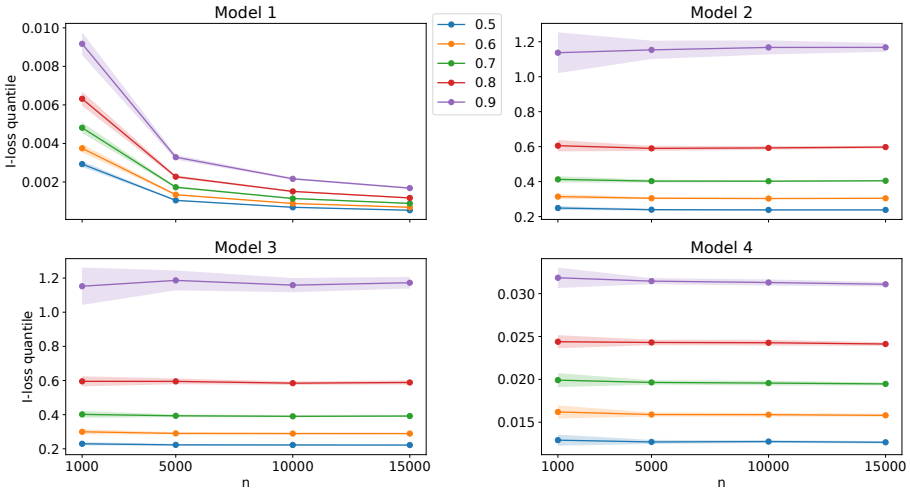

Figure 18: I_{loss} of a Breiman RF w.r.t sample size n. RF parameters: 500 trees, bootstrap on, max-features= d, max-depth set to None. Mean over 30 tries (doted lines) and std (filled zones).

For instance, if we take the 0.8-quantile in red on Figure 18 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).

On Figure 19, 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.

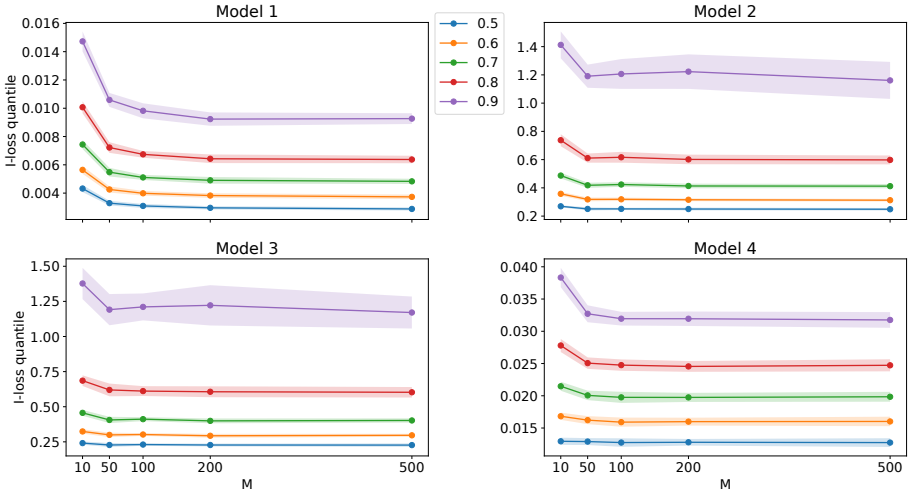

Figure 19: I_{loss} of a Breiman RF w.r.t number of trees. Parameters: bootstrap on, max-features= d, max-depth set to None. Sample size $n = 1000$. Mean over 30 tries (doted lines) and std (filled zones).