Random forests and kernel methods

Erwan Scornet

HAL Id: hal-01116158
https://hal.science/hal-01116158v2

Submitted on 16 Sep 2015

Random forests and kernel methods

Erwan Scornet
Sorbonne Universités, UPMC Univ Paris 06, F-75005, Paris, France
erwan.scornet@upmc.fr

Abstract

Random forests are ensemble methods which grow trees as base learners and combine their predictions by averaging. Random forests are known for their good practical performance, particularly in high-dimensional settings. On the theoretical side, several studies highlight the potentially fruitful connection between random forests and kernel methods. In this paper, we work out in full details this connection. In particular, we show that by slightly modifying their definition, random forests can be rewritten as kernel methods (called KeRF for Kernel based on Random Forests) which are more interpretable and easier to analyze. Explicit expressions of KeRF estimates for some specific random forest models are given, together with upper bounds on their rate of consistency. We also show empirically that KeRF estimates compare favourably to random forest estimates.

Index Terms — Random forests, randomization, consistency, rate of consistency, kernel methods.

2010 Mathematics Subject Classification: 62G05, 62G20.

1 Introduction

Random forests are a class of learning algorithms used to solve pattern recognition problems. As ensemble methods, they grow many trees as base learners and aggregate them to predict. Growing many different trees from a single data set requires to randomize the tree building process by, for example, sampling the data set. Thus, there exists a variety of random forests, depending on how trees are built and how the randomness is introduced in the tree building process.

One of the most popular random forests is that of Breiman (2001) which grows trees based on CART procedure (Classification and Regression Trees, Breiman et al., 1984) and randomizes both the training set and the splitting directions. Breiman's (2001) random forests have been under active investigation during the last decade mainly because of their good practical performance and their ability to handle high dimensional data sets. Moreover, they are easy to run since they only depend on few parameters which are easily tunable (Liaw and Wiener, 2002; Genuer et al., 2008). They are acknowledged to be state-of-the-art methods in fields such as genomics (Qi, 2012) and pattern recognition (Rogez et al., 2008), just to name a few.

However, even if random forests are known to perform well in many contexts, little is known about their mathematical properties. Indeed, most authors study forests whose construction does not depend on the data set. Although, consistency of such simplified models has been addressed in the literature (e.g., Biau

et al., 2008; Ishwaran and Kogalur, 2010; Denil et al., 2013), these results do not adapt to Breiman's forests whose construction strongly depends on the whole training set. The latest attempts to study the original algorithm are by Mentch and Hooker (2014) and Wager (2014) who prove its asymptotic normality or by Scornet et al. (2014) who prove its consistency under appropriate assumptions.

Despite these works, several properties of random forests still remain unexplained. A promising way for understanding their complex mechanisms is to study the connection between forests and kernel estimates, that is estimates m_n which take the form

$$m_n(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K_k(\mathbf{X}_i, \mathbf{x})}{\sum_{i=1}^n K_k(\mathbf{X}_i, \mathbf{x})}, \tag{1}$$

where $\{(\mathbf{X}_i, Y_i) : 1 \le i \le n\}$ is the training set, $(K_k)_k$ is a sequence of kernel functions, and k ($k \in \mathbb{N}$) is a parameter to be tuned. Unlike the most used Nadaraya-Watson kernels (Nadaraya, 1964; Watson, 1964) which satisfy a homogeneous property of the form $K_h(\mathbf{X}_i, \mathbf{x}) = K((\mathbf{x} - \mathbf{X}_i)/h)$, kernels K_k are not necessarily of this form. Therefore, the analysis of kernel estimates defined by (1) turns out to be more complicated and cannot be based on general results regarding Nadaraya-Watson kernels.

Breiman (2000) was the first to notice the link between forest and kernel methods, a link which was later formalized by Geurts et al. (2006). On the practical side, Davies and Ghahramani (2014) highlight the fact that a specific kernel based on random forests can empirically outperform state-of-the-art kernel methods. Another approach is taken by Lin and Jeon (2006) who establish the connection between random forests and adaptive nearest neighbor, implying that random forests can be seen as adaptive kernel estimates (see also Biau and Devroye, 2010). The latest study is by Arlot and Genuer (2014) who show that a specific random forest can be written as a kernel estimate and who exhibit rates of consistency. However, despite these works, the literature is relatively sparse regarding the link between forests and kernel methods.

Our objective in the present paper is to prove that a slight modification of random forest procedures have explicit and simple interpretations in terms of kernel methods. Thus, the resulting kernel based on random forest (called KeRF in the rest of the paper) estimates are more amenable to mathematical analysis. They also appear to be empirically as accurate as random forest estimates. To theoretically support these results, we also make explicit the expression of some KeRF. We prove upper bounds on their rates of consistency, which compare favorably to the existing ones.

The paper is organized as follows. Section 2 is devoted to notations and to the definition of KeRF estimates. The link between KeRF estimates and random forest estimates is made explicit in Section 3. In Section 4, two KeRF estimates are presented and their consistency is proved along with their rate of consistency. Section 5 contains experiments that highlight the good performance of KeRF compared to their random forests counterparts. Proofs are postponed to Section 6.

2 Notations and first definitions

2.1 Notations

Throughout the paper, we assume to be given a training sample $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ of $[0,1]^d \times \mathbb{R}$-valued independent random variables distributed as the independent prototype pair $(\mathbf{X}, Y)$, where $\mathbb{E}[Y^2] < \infty$. We aim at predicting the response Y, associated with the random variable $\mathbf{X}$, by estimating the regression function $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. In this context, we use infinite random forests (see the definition below) to build an estimate $m_{\infty,n} : [0,1]^d \to \mathbb{R}$ of m, based on the data set $\mathcal{D}_n$.

A random forest is a collection of M randomized regression trees (for an overview on tree construction, see e.g., Chapter 20 in Györfi et al., 2002). For the j-th tree in the family, the predicted value at point $\mathbf{x}$ is denoted by $m_n(\mathbf{x}, \Theta_j)$, where $\Theta_1, \ldots, \Theta_M$ are independent random variables, distributed as a generic random variable Θ, independent of the sample $\mathcal{D}_n$. This random variable can be used to sample the training set or to select the candidate directions or positions for splitting. The trees are combined to form the finite forest estimate

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) = \frac{1}{M} \sum_{j=1}^{M} m_n(\mathbf{x}, \Theta_j). \tag{2}$$

By the law of large numbers, for all $\mathbf{x} \in [0,1]^d$, almost surely, the finite forest estimate tends to the infinite forest estimate

$$m_{\infty,n}(\mathbf{x}) = \mathbb{E}_\Theta \left[m_n(\mathbf{x}, \Theta) \right],$$

where $\mathbb{E}_\Theta$ denotes the expectation with respect to Θ, conditionally on $\mathcal{D}_n$.

As mentioned above, there is a large variety of forests, depending on how trees are grown and how the random variable Θ influences the tree construction. For instance, tree construction can be independent of $\mathcal{D}_n$ (Biau, 2012). On the other hand, it can depend only on the $\mathbf{X}_i$'s (Biau et al., 2008) or on the whole training set (Cutler and Zhao, 2001; Geurts et al., 2006; Zhu et al., 2012). Throughout the paper, we use three important types of random forests to exemplify our results: Breiman's, centred and uniform forests. In Breiman's original procedure, splits are performed to minimize the variances within the two resulting cells. The algorithm stops when each cell contains less than a small pre-specified number of points (typically between 1 and 5; see Breiman, 2001, for details). Centred forests are a simpler procedure which, at each node, uniformly select a coordinate among $\{1, \ldots, d\}$ and performs splits at the center of the cell along the pre-chosen coordinate. The algorithm stops when a full binary tree of level k is built (that is, each cell is cut exactly k times), where $k \in \mathbb{N}$ is a parameter of the algorithm (see Breiman, 2004, for details on the procedure). Uniform forests are quite similar to centred forests except that once a split direction is chosen, the split is drawn uniformly on the side of the cell, along the preselected coordinate (see, e.g., Arlot and Genuer, 2014).

2.2 Kernel based on random forests (KeRF)

To be more specific, random forest estimates satisfy, for all $\mathbf{x} \in [0,1]^d$,

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) = \frac{1}{M} \sum_{j=1}^{M} \left(\sum_{i=1}^{n} \frac{Y_i \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)}}{N_n(\mathbf{x}, \Theta_j)} \right),$$

where $A_n(\mathbf{x}, \Theta_j)$ is the cell containing $\mathbf{x}$, designed with randomness Θ_j and data set $\mathcal{D}_n$, and

$$N_n(\mathbf{x}, \Theta_j) = \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)}$$

is the number of data points falling in $A_n(\mathbf{x}, \Theta_j)$. Note that, the weights $W_{i,j,n}(\mathbf{x})$ of each observation Y_i defined by

$$W_{i,j,n}(\mathbf{x}) = \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)}}{N_n(\mathbf{x}, \Theta_j)}$$

depend on the number of observations $N_n(\mathbf{x}, \Theta_j)$. Thus the contributions of observations that are in cells with a high density of data points are smaller than that of observations which belong to less populated cells. This is particularly true for non adaptive forests (i.e., forests built independently of data) since the number of observations in each cell cannot be controlled. Giving important weights to observations that are in low-density cells can potentially lead to rough estimates. Indeed, as an extreme example, trees of non adaptive forests can contain empty cells which leads to a substantial misestimation (since the prediction in empty cells is set, by default, to zero).

In order to improve the random forest methods and compensate the misestimation induced by random forest weights, a natural idea is to consider KeRF estimates defined, for all $\mathbf{x} \in [0,1]^d$, by

$$\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) = \frac{1}{\sum_{j=1}^{M} N_n(\mathbf{x}, \Theta_j)} \sum_{j=1}^{M} \sum_{i=1}^{n} Y_i \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)}. \qquad (3)$$

Note that $\tilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)$ is equal to the mean of the Y_i's falling in the cells containing $\mathbf{x}$ in the forest. Thus, each observation is weighted by the number of times it appears in the trees of the forests. Consequently, in this setting, an empty cell does not contribute to the prediction.

The proximity between KeRF estimates $\widetilde{m}_{M,n}$ and random forest estimates will be thoroughly discussed in Section 3. As for now, we focus on (3) and start by proving that it is indeed a kernel estimate whose expression is given by Proposition 1.

Proposition 1. *Almost surely, for all $\mathbf{x} \in [0,1]^d$, we have*

$$\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) = \frac{\sum_{i=1}^{n} Y_i K_{M,n}(\mathbf{x}, \mathbf{X}_i)}{\sum_{\ell=1}^{n} K_{M,n}(\mathbf{x}, \mathbf{X}_\ell)}, \qquad (4)$$

where

$$K_{M,n}(\mathbf{x}, \mathbf{z}) = \frac{1}{M} \sum_{j=1}^{M} \mathbb{1}_{\mathbf{z} \in A_n(\mathbf{x}, \Theta_j)}. \tag{5}$$

We call $K_{M,n}$ the connection function of the M finite forest.

Proposition 1 states that KeRF estimates have a more interpretable form than random forest estimates since their kernels are the connection functions of the forests. This connection function can be seen as a geometrical characteristic of the cells in the random forest. Indeed, fixing $\mathbf{X}_i$, the quantity $K_{M,n}(\mathbf{x}, \mathbf{X}_i)$ is nothing but the empirical probability that $\mathbf{X}_i$ and $\mathbf{x}$ are connected (i.e. in the same cell) in the M finite random forest. Thus, the connection function is a natural way to build kernel functions from random forests, a fact that had already been noticed by Breiman (2001). Note that these kernel functions have the nice property of being positive semi-definite, as proved by Davies and Ghahramani (2014).

A natural question is to ask what happens to KeRF estimates when the number of trees M goes to infinity. To this aim, we define infinite KeRF estimates $\widetilde{m}_{\infty,n}$ by, for all $\mathbf{x}$,

$$\widetilde{m}_{\infty,n}(\mathbf{x}) = \lim_{M \to \infty} \widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M). \tag{6}$$

In addition, we say that an infinite random forest is discrete (resp. continuous) if its connection function K_n is piecewise constant (resp. continuous). For example, Breiman forests and centred forests are discrete but uniform forests are continuous. Denote by $\mathbb{P}_\Theta$ the probability with respect to Θ, conditionally on $\mathcal{D}_n$. Proposition 2 extends the results of Proposition 1 to the case of infinite KeRF estimates.

Proposition 2. *Consider an infinite discrete or continuous forest. Then, almost surely, for all $\mathbf{x}, \mathbf{z} \in [0,1]^d$,*

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

where

$$K_n(\mathbf{x}, \mathbf{z}) = \mathbb{P}_\Theta \left[\mathbf{z} \in A_n(\mathbf{x}, \Theta) \right].$$

We call K_n the connection function of the infinite random forest. Thus, for all $\mathbf{x} \in [0,1]^d$, one has

$$\widetilde{m}_{\infty,n}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K_n(\mathbf{x}, \mathbf{X}_i)}{\sum_{\ell=1}^{n} K_n(\mathbf{x}, \mathbf{X}_\ell)}.$$

This lemma shows that infinite KeRF estimates are kernel estimates with kernel function equal to K_n. Observing that $K_n(\mathbf{x}, \mathbf{z})$ is the probability that $\mathbf{x}$ and $\mathbf{z}$ are connected in the infinite forest, the function K_n characterizes the shape of the cells in the infinite random forest.

Now that we know the expression of KeRF estimates, we are ready to study how close this approximation is to random forest estimates. This link will be further work out in Section 4 for centred and uniform KeRF and empirically studied in Section 5.

3 Relation between KeRF and random forests

In this section, we investigate in which cases KeRF and forest estimates are close to each other. To achieve this goal, we will need the following assumption.

(H1) *Fix* $\mathbf{x} \in [0,1]^d$, *and assume that* $Y \geq 0$ *a.s.. Then, one of the following two conditions holds:*

> **(H1.1)** *There exist sequences* $(a_n), (b_n)$ *such that, a.s.,*
>
> $$a_n \leq N_n(\mathbf{x}, \Theta) \leq b_n.$$

> **(H1.2)** *There exist sequences* $(\varepsilon_n), (a_n), (b_n)$ *such that, a.s.,*
>
> - $1 \leq a_n \leq \mathbb{E}_\Theta [N_n(\mathbf{x}, \Theta)] \leq b_n$,
> - $\mathbb{P}_\Theta \left[a_n \leq N_n(\mathbf{x}, \Theta) \leq b_n \right] \geq 1 - \varepsilon_n$.

(H1) assumes that the number of points in every cell of the forest can be bounded from above and below. **(H1.1)** holds for finite forests for which the number of points in each cell is controlled almost surely. Typically, **(H1.1)** is verified for adaptive random forests, if the stopping rule is properly chosen. On the other hand, **(H1.2)** holds for infinite forests. Note that the first condition $\mathbb{E}_\Theta [N_n(\mathbf{x}, \Theta)] \geq 1$ in **(H1.2)** is technical and is true if the level of each tree is tuned appropriately. Several random forests which satisfy **(H1)** are discussed below.

Proposition 3 states that finite forest estimate $m_{M,n}$ and finite KeRF estimate $\widetilde{m}_{M,n}$ are close to each other assuming that **(H1.1)** holds.

Proposition 3. *Assume that* **(H1.1)** *is satisfied. Thus, almost surely,*

$$\left| \frac{m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)}{\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)} - 1 \right| \leq \frac{b_n - a_n}{a_n},$$

with the convention that $0/0 = 1$.

Since KeRF estimates are kernel estimates of the form (1), Proposition 3 stresses that random forests are close to kernel estimates if the number of points in each cell is controlled. As highlighted by the following discussion, the assumptions of Proposition 3 are satisfied for some types of random forests.

Centred random forests of level k**.** For this model, whenever $\mathbf{X}$ is uniformly distributed over $[0,1]^d$, each cell has a Lebesgue-measure of 2^{-k}. Thus, fixing $\mathbf{x} \in [0,1]^d$, according to the law of the iterated logarithm, for all n large enough, almost surely,

$$\left| N_n(\mathbf{x}, \Theta) - \frac{n}{2^k} \right| \leq \frac{\sqrt{2n \log \log n}}{2}.$$

Consequently, **(H1.1)** is satisfied for $a_n = n2^{-k} - \sqrt{2n \log \log n}/2$ and $b_n = n2^{-k} + \sqrt{2n \log \log n}/2$. This yields, according to Proposition 3, almost surely,

$$\left| \frac{m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)}{\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)} - 1 \right| \leq \frac{\sqrt{2n \log \log n}}{n2^{-k} - \sqrt{2n \log \log n}/2}.$$

Thus, choosing for example $k = (\log_2 n)/3$, centred KeRF estimates are asymptotically equivalent to centred forest estimates as $n \to \infty$. The previous inequality can be extended to the case where $\mathbf{X}$ has a density f satisfying $c \leq f \leq C$, for some constants $0 < c < C < \infty$. In that case, almost surely,

$$\left| \frac{m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)}{\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)} - 1 \right| \leq \frac{\sqrt{2n \log \log n} + (C - c)n/2^k}{nc2^{-k} - \sqrt{2n \log \log n}/2}.$$

However, the right-hand term does not tend to zero as $n \to \infty$, meaning that the uniform assumption on $\mathbf{X}$ is crucial to prove the asymptotic equivalence of $m_{M,n}$ and $\widetilde{m}_{M,n}$ in the case of centred forests.

Breiman's forests. Each leaf in Breiman's trees contains a small number of points (typically between 1 and 5). Thus, if each cell contains exactly one point (default settings in classification problems), **(H1.1)** holds with $a_n = b_n = 1$. Thus, according to Proposition 3, almost surely,

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) = \widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M).$$

More generally, if the number of observations in each cell varies between 1 and 5, one can set $a_n = 1$ and $b_n = 5$. Thus, still by Proposition 3, almost surely,

$$\left| \frac{m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)}{\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)} - 1 \right| \leq 4.$$

Median forests of level k. In this model, each cell of each tree is split at the empirical median of the observations belonging to the cell. The process is repeated until every cell is cut exactly k times (where $k \in \mathbb{N}$ is a parameter chosen by the user). Thus, each cell contains the same number of points ± 2 (see, e.g., Biau and Devroye, 2013, for details), and, according to Proposition 3, almost surely,

$$\left| \frac{m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)}{\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)} - 1 \right| \leq \frac{2}{a_n}.$$

Consequently, if the level k of each tree is chosen such that $a_n \to \infty$ as $n \to \infty$, median KeRF estimates are equivalent to median forest estimates.

The following lemma extends Proposition 3 to infinite KeRF and forest estimates.

Proposition 4. *Assume that* **(H1.2)** *is satisfied. Thus, almost surely,*

$$|m_{\infty,n}(\mathbf{x}) - \widetilde{m}_{\infty,n}(\mathbf{x})| \leq \frac{b_n - a_n}{a_n} \widetilde{m}_{\infty,n}(\mathbf{x}) + n\varepsilon_n \left(\max_{1 \leq i \leq n} Y_i \right).$$

Considering inequalities provided in Proposition 4, we see that infinite KeRF estimates are close to infinite random forest estimates if the number of observations in each cell is bounded (via a_n and b_n).

It is worth noticing that controlling the number of observations in each cell while obtaining a simple partition shape is difficult to achieve. On the one hand, if the tree construction depends on the training set, the algorithm can be stopped when each leaf contains exactly one point and thus KeRF estimate is equal to random forest estimate. However, in that case, the probability $K_n(\mathbf{x}, \mathbf{z})$ is very difficult to express since the geometry of each tree partitioning strongly depends on the training set. On the other hand, if the tree construction is independent of the training set, the probability $K_n(\mathbf{x}, \mathbf{z})$ can be made explicit in some cases, for example for centred forests (see Section 5). However, the number of points in each cell is difficult to control (every leaf cannot contain exactly one point with a non-adaptive cutting strategy) and thus KeRF estimate can be far away from random forest estimate. Consequently, one cannot deduce an explicit expression for random forest estimates from the explicit expression of KeRF estimates.

4   Two particular KeRF estimates

According to Proposition 2, infinite KeRF estimate $\widetilde{m}_{\infty,n}$ depends only on the connection function K_n via the following equation

$$\widetilde{m}_{\infty,n}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K_n(\mathbf{x}, \mathbf{X}_i)}{\sum_{\ell=1}^{n} K_n(\mathbf{x}, \mathbf{X}_\ell)}. \tag{7}$$

To take one step further into the understanding of KeRF, we study in this section the connection function of two specific infinite random forests. We focus on infinite KeRF estimates for two reasons. Firstly, the expressions of infinite KeRF estimates are more amenable to mathematical analysis since they do not depend on the particular trees used to build the forest. Secondly, the prediction accuracy of infinite random forests is known to be better than that of finite random forests (see, e.g., Scornet, 2014). Therefore infinite KeRF estimates are likely to be more accurate than finite KeRF estimates.

Practically, both infinite KeRF estimates and infinite random forest estimates can only be approximated by Monte Carlo simulations. Here, we show that centred KeRF estimates have an explicit expression, that is their connection function can be made explicit. Thus, infinite centred KeRF estimates and infinite uniform KeRF estimates (up to an approximation detailed below) can be directly computed using equation (7).

Centred KeRF   As seen above, the construction of centred KeRF of level k is the same as for centred forests of level k except that predictions are made according to equation (3). Centred random forests are closely related to Breiman's forests in a linear regression framework. Indeed, in this context, splits that are performed at a low level of the trees are roughly located at the middle of each cell. In that case, Breiman's forests and centred forests are close to each other, which justifies the interest for these simplified models, and thus for centred KeRF.

8

In the sequel, the connection function of the centred random forest of level k is denoted by K_k^{cc}. This notation is justified by the fact that the construction of centred KeRF estimates depends only on the size of the training set through the choice of k.

Proposition 5. *Let $k \in \mathbb{N}$ and consider an infinite centred random forest of level k. Then, for all $\mathbf{x}, \mathbf{z} \in [0, 1]^d$,*

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

Note that ties are broken by imposing that cells are of the form $\prod_{i=1}^{d} A_i$ where the A_i are equal to $]a_i, b_i]$ or $[0, b_i]$, for all $0 < a_i < b_i \leq 1$. Figure 1 shows a graphical representation of the function f defined as

$$f_k : \quad [0, 1] \times [0, 1] \quad \to \quad [0, 1]$$
$$\mathbf{z} = (z_1, z_2) \quad \mapsto \quad K_k^{cc}\left((\tfrac{1}{2}, \tfrac{1}{2}), \mathbf{z}\right).$$
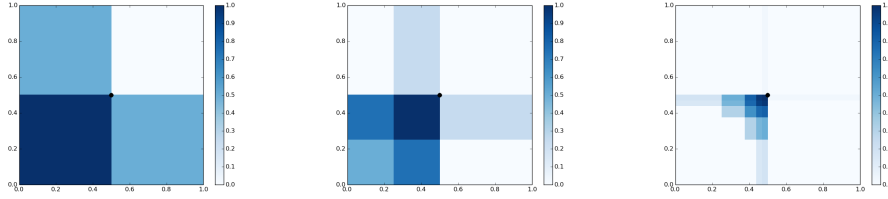

Figure 1: Representations of f_1, f_2 and f_5 in $[0, 1]^2$

Denote by $\widetilde{m}_{\infty,n}^{cc}$ the infinite centred KeRF estimate, associated with the connection function K_k^{cc}, defined as

$$\widetilde{m}_{\infty,n}^{cc}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K_k^{cc}(\mathbf{x}, \mathbf{X}_i)}{\sum_{\ell=1}^{n} K_k^{cc}(\mathbf{x}, \mathbf{X}_\ell)}.$$

To pursue the analysis of $\widetilde{m}_{\infty,n}^{cc}$, we will need the following assumption on the regression model.

(H2) *One has*

$$Y = m(\mathbf{X}) + \varepsilon,$$

where ε is a centred Gaussian noise, independent of $\mathbf{X}$, with finite variance $\sigma^2 < \infty$. Moreover, $\mathbf{X}$ is uniformly distributed on $[0, 1]^d$ and m is Lipschitz.

Our theorem states that infinite centred KeRF estimates are consistent whenever **(H2)** holds. Moreover, it provides an upper bound on the rate of consistency of centred KeRF.

Theorem 4.1. *Assume that* **(H2)** *is satisfied. Then, providing $k \to \infty$ and $n/2^k \to \infty$, there exists a constant $C_1 > 0$ such that, for all $n > 1$, and for all $\mathbf{x} \in [0,1]^d$,*

$$\mathbb{E}\left[\widetilde{m}_{\infty,n}^{cc}(\mathbf{x}) - m(\mathbf{x})\right]^2 \leq C_1 n^{-1/(3+d\log 2)}(\log n)^2.$$

Observe that centred KeRF estimates fail to reach minimax rate of consistency $n^{-2/(d+2)}$ over the class of Lipschitz functions. A similar upper bound on the rate of consistency $n^{-3/4d\log 2+3}$ of centred random forests was obtained by Biau (2012). It is worth noticing that, for all $d \geq 9$, the upper bound on the rate of centred KeRF is sharper than that of centred random forests. This theoretical result supports the fact that KeRF procedure has a better performance compared to centred random forests. This will be supported by simulations in Section 5 (see Figure 5)

Uniform KeRF   Recall that the infinite uniform KeRF estimates of level k are the same as infinite uniform forest of level k except that predictions are computed according to equation (3). Uniform random forests, first studied by Biau et al. (2008), remain under active investigation. They are a nice modelling of Breiman forests, since with no a priori on the split location, we can consider that splits are drawn uniformly on the cell edges. Other related versions of these forests have been thoroughly investigated by Arlot and Genuer (2014) who compare the bias of a single tree to that of the whole forest.

As for the connection function of centred random forests, we use the notational convention K_k^{uf} to denote the connection function of uniform random forests of level k.

Proposition 6. *Let $k \in \mathbb{N}$ and consider an infinite uniform random forest of level k. Then, for all $\mathbf{x} \in [0,1]^d$,*

$$K_k^{uf}(\mathbf{0}, \mathbf{x}) = \sum_{\substack{k_1,\ldots,k_d \\ \sum_{\ell=1}^d k_\ell = k}} \frac{k!}{k_1! \ldots k_d!} \left(\frac{1}{d}\right)^k \prod_{m=1}^d \left(1 - x_m \sum_{j=0}^{k_m-1} \frac{(-\ln x_m)^j}{j!}\right),$$

with the convention $\sum_{j=0}^{-1} \frac{(-\ln x_m)^j}{j!} = 0$.

Proposition 6 gives the explicit expression of $K_k^{uf}(\mathbf{0}, \mathbf{x})$. Figure 2 shows a representation of the functions f_1, f_2 and f_5 defined as

$$\begin{aligned} f_k: \quad [0,1] \times [0,1] \quad &\to \quad [0,1] \\ \mathbf{z} = (z_1, z_2) \quad &\mapsto \quad K_k^{uf}\left(\mathbf{0}, \left|\mathbf{z} - (\tfrac{1}{2}, \tfrac{1}{2})\right|\right), \end{aligned}$$

where $|\mathbf{z} - \mathbf{x}| = (|z_1 - x_1|, \ldots, |z_d - x_d|)$.

Unfortunately, the general expression of the connection function $K_k^{uf}(\mathbf{x}, \mathbf{z})$ is difficult to obtain. Indeed, for $d = 1$, cuts are performed along a single axis, but the probability of connection between two points x and z does not depend only upon the distance $|z - x|$ but rather on the positions x and z, as stressed in the following Lemma.
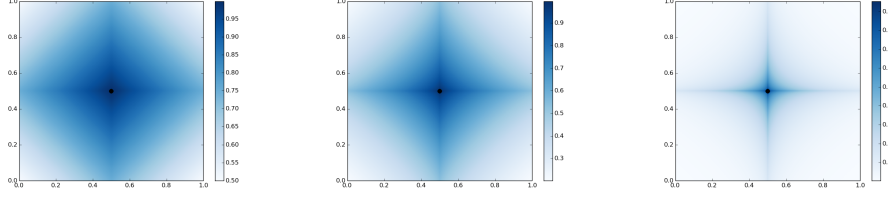
Figure 2: Representations of f_1, f_2 and f_5 in dimension two

Lemma 1. *Let $x, z \in [0, 1]$. Then,*

$$K_1^{uf}(x, z) = 1 - |z - x|,$$

$$K_2^{uf}(x, z) = 1 - |z - x| + |z - x| \log \left(\frac{z}{1 - x} \right).$$

A natural way to deal with this difficulty is to replace the connection function K_k^{uf} by the function $(\mathbf{x}, \mathbf{z}) \to K_k^{uf}(\mathbf{0}, |\mathbf{z} - \mathbf{x}|)$. Indeed, this is a simple manner to build an invariant-by-translation version of the uniform kernel K_k^{uf}. The extensive simulations in Section 5 support the fact that estimates of the form (7) built with these two kernels have similar prediction accuracy. As for infinite centred KeRF estimates, we denote by $\widetilde{m}_{\infty,n}^{uf}$ the infinite uniform KeRF estimates but built with the invariant-by-translation version of K_k^{uf}, namely

$$\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K_k^{uf}(\mathbf{0}, |\mathbf{X}_i - \mathbf{x}|)}{\sum_{\ell=1}^{n} K_k^{uf}(\mathbf{0}, |\mathbf{X}_\ell - \mathbf{x}|)}.$$

Our last theorem states the consistency of infinite uniform KeRF estimates along with an upper bound on their rate of consistency.

Theorem 4.2. *Assume that* **(H2)** *is satisfied. Then, providing $k \to \infty$ and $n/2^k \to \infty$, there exists a constant $C_1 > 0$ such that, for all $n > 1$ and for all $\mathbf{x} \in [0, 1]^d$,*

$$\mathbb{E} \left[\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x}) \right]^2 \leq C_1 n^{-2/(6+3d \log 2)} (\log n)^2.$$

As for centred KeRF estimates, the rate of consistency does not reach the minimax rate on the class of Lipschitz functions, and is actually worse than that of centred KeRF estimates, whatever the dimension d is. Besides, centred KeRF estimates have better performance than uniform KeRF estimates and this will be highlighted by simulations (Section 5).

Although centred and uniform KeRF estimates are kernel estimates of the form (1), the usual tools used to prove consistency and to find rate of consistency of kernel methods cannot be applied here (see, e.g., Chapter 5 in Györfi et al., 2002). Indeed, the support of $\mathbf{z} \mapsto K_k^{cc}(\mathbf{x}, \mathbf{z})$ and that of $\mathbf{z} \mapsto K_k^{uf}(\mathbf{0}, |\mathbf{z} - \mathbf{x}|)$ cannot be contained in a ball centred on $\mathbf{x}$, whose diameter tends to zero (see Figure 1 and 2). The proof of Theorem 4.1 and 4.2 are then based on the previous work of Greblicki et al. (1984) who proved the consistency of kernels with unbounded support. In particular, we use their bias/variance decomposition of kernel estimates to exhibit upper bounds on the rate of consistency.

11

5 Experiments

Practically speaking, Breiman's random forests are among the most widely used forest algorithms. Thus a natural question is to know whether Breiman KeRF compare favourably to Breiman's forests. In fact, as seen above, the two algorithms coincide whenever Breiman's forests are fully grown. But this is not always the case since by default, each cell of Breiman's forests contain between 1 and 5 observations.

We start this section by comparing Breiman KeRF and Breiman's forest estimates for various regression models described below. Some of these models are toy models (**Model 1, 5-8**). **Model 2** can be found in van der Laan et al. (2007) and **Models 3-4** are presented in Meier et al. (2009). For all regression frameworks, we consider covariates $\mathbf{X} = (X_1, \ldots, X_d)$ that are uniformly distributed over $[0,1]^d$. We also let $\widetilde{X}_i = 2(X_i - 0.5)$ for $1 \leq i \leq d$.

Model 1: $n = 800, d = 50, Y = \widetilde{X}_1^2 + \exp(-\widetilde{X}_2^2)$

Model 2: $n = 600, d = 100, Y = \widetilde{X}_1 \widetilde{X}_2 + \widetilde{X}_3^2 - \widetilde{X}_4 \widetilde{X}_7 + \widetilde{X}_8 \widetilde{X}_{10} - \widetilde{X}_6^2 + \mathcal{N}(0, 0.5)$

Model 3: $n = 600, d = 100, Y = -\sin(2\widetilde{X}_1) + \widetilde{X}_2^2 + \widetilde{X}_3 - \exp(-\widetilde{X}_4) + \mathcal{N}(0, 0.5)$

Model 4: $n = 600, d = 100, Y = \widetilde{X}_1 + (2\widetilde{X}_2 - 1)^2 + \sin(2\pi\widetilde{X}_3)/(2 - \sin(2\pi\widetilde{X}_3)) + \sin(2\pi\widetilde{X}_4) + 2\cos(2\pi\widetilde{X}_4) + 3\sin^2(2\pi\widetilde{X}_4) + 4\cos^2(2\pi\widetilde{X}_4) + \mathcal{N}(0, 0.5)$

Model 5: $n = 700, d = 20, Y = \mathbb{1}_{\widetilde{X}_1 > 0} + \widetilde{X}_2^3 + \mathbb{1}_{\widetilde{X}_4 + \widetilde{X}_6 - \widetilde{X}_8 - \widetilde{X}_9 > 1 + \widetilde{X}_{10}} + \exp(-\widetilde{X}_2^2) + \mathcal{N}(0, 0.5)$

Model 6: $n = 500, d = 30, Y = \sum_{k=1}^{10} \mathbb{1}_{\widetilde{X}_k^3 < 0} - \mathbb{1}_{\mathcal{N}(0,1) > 1.25}$

Model 7: $n = 600, d = 300, Y = \widetilde{X}_1^2 + \widetilde{X}_2^2 \widetilde{X}_3 \exp(-|\widetilde{X}_4|) + \widetilde{X}_6 - \widetilde{X}_8 + \mathcal{N}(0, 0.5)$

Model 8: $n = 500, d = 1000, Y = \widetilde{X}_1 + 3\widetilde{X}_3^2 - 2\exp(-\widetilde{X}_5) + \widetilde{X}_6$

All numerical implementations have been performed using the free Python software, available online at https://www.python.org/. For each experiment, the data set is divided into a training set (80% of the data set) and a test set (the remaining 20%). Then, the empirical risk ($\mathbb{L}^2$ error) is evaluated on the test set.

To start with, Figure 3 depicts the empirical risk of Breiman's forests and Breiman KeRF estimates for two regression models (the conclusions are similar for the remaining regression models). Default settings were used for Breiman's forests (`minsamplessplit = 2`, `maxfeatures = 0.333`) and for Breiman KeRF, except that we did not bootstrap the data set. Figure 3 puts in evidence that Breiman KeRF estimates behave similarly (in terms of empirical risk) to Breiman forest estimates. It is also interesting to note that bootstrapping the data set does not change the performance of the two algorithms.

Figure 4 (resp. Figure 5) shows the risk of uniform (resp. centred) KeRF estimates compared to the risk of uniform (resp. centred) forest estimates (only
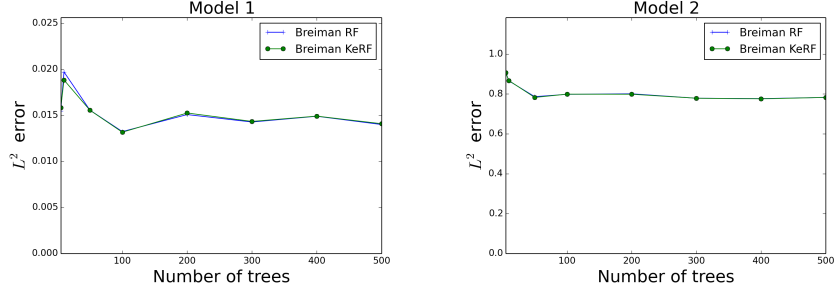
Figure 3: Empirical risks of Breiman KeRF estimates and Breiman forest estimates.

two models shown). In these two experiments, uniform and centred forests and their KeRF counterparts have been grown in such a way that each tree is a complete binary tree of level $k = \lfloor \log_2 n \rfloor$. Thus, in that case, each cell contains on average $n/2^k \simeq 1$ observation. Once again, the main message of Figure 4 is that the uniform KeRF accuracy is close to the uniform forest accuracy.
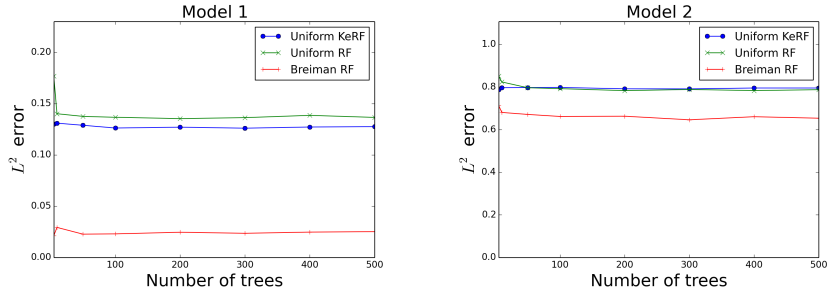

Figure 4: Empirical risks of uniform KeRF and uniform forest.

On the other hand, it turns out that the performance of centred KeRF and centred forests are not similar (Figure 5). In fact, centred KeRF estimates are either comparable to centred forest estimates (as, for example, in **Model 2**), or have a better accuracy (as, for example, in **Model 1**). A possible explanation for this phenomenon is that centred forests are non-adaptive in the sense that their construction does not depend on the data set. Therefore, each tree is likely to contain cells with unbalanced number of data points, which can result in random forest misestimation. This undesirable effect vanishes using KeRF methods since they assign the same weights to each observation.

The same series of experiments were conducted, but using bootstrap for computing both KeRF and random forest estimates. The general finding is that the results are similar—Figure 6 and 7 depict the accuracy of corresponding algorithms for a selected choice of regression frameworks.

An important aspect of infinite centred and uniform KeRF is that they can be explicitly computed (see Proposition 5 and 6). Thus, we have plotted in Figure
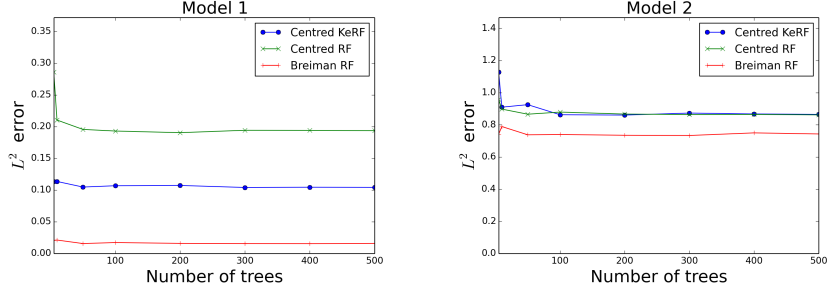
13

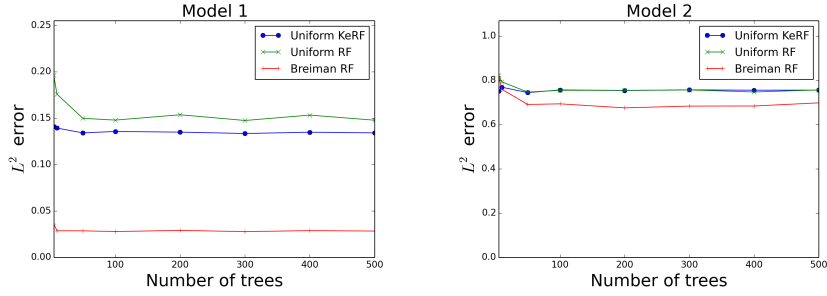Figure 5: Empirical risks of centred KeRF and centred forest.

Figure 6: Empirical risks of uniform KeRF and uniform forest (with bootstrap).
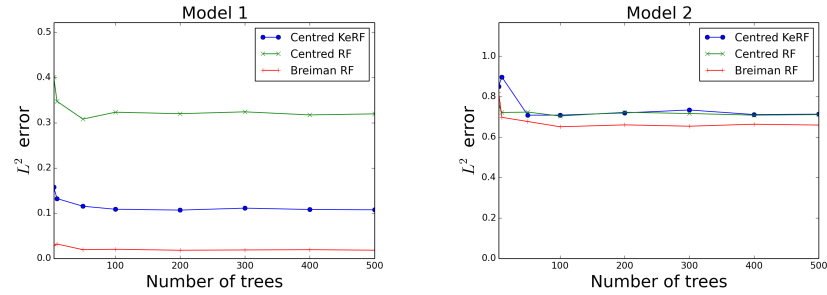

Figure 7: Empirical risks of centred KeRF and centred forests (with bootstrap).

8 the empirical risk of both finite and infinite centred KeRF estimates for some examples (for $n = 100$ and $d = 10$). We clearly see in this figure that the accuracy of finite centred KeRF tends to the accuracy of infinite centred KeRF as M tends to infinity. This corroborates Proposition 2.

The same comments hold for uniform KeRF (see Figure 9). Note however that, in that case, the proximity between finite uniform KeRF and infinite uniform KeRF estimate strengthens the approximation that has been made on infinite uniform KeRF in Section 4.

The computation time for finite KeRF estimate is very acceptable for finite
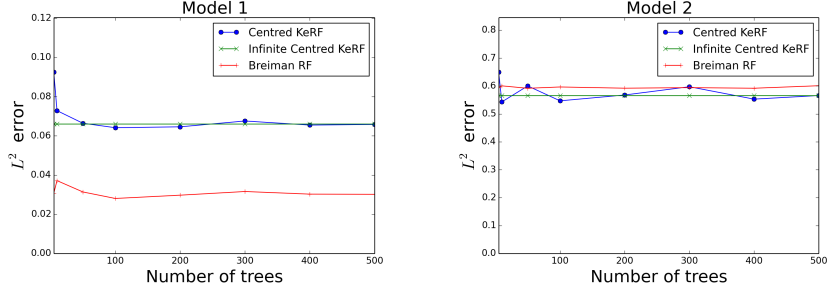
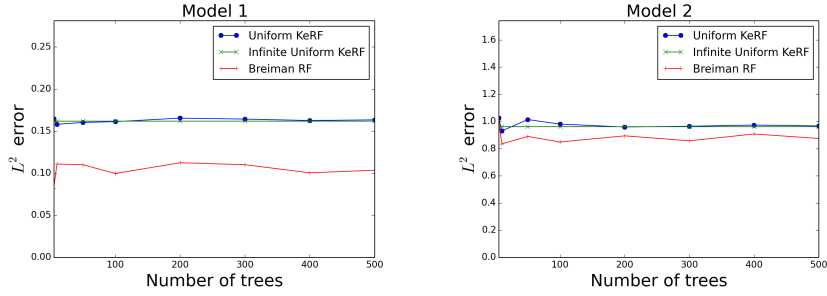Figure 8: Risks of finite and infinite centred KeRF.

Figure 9: Risks of finite and infinite uniform KeRF.

KeRF and similar to that of random forest (Figure 3-5). However, the story is different for infinite KeRF estimates. In fact, KeRF estimates can only be evaluated for low dimensional data sets and small sample sizes. To see this, just note that the explicit formulation of KeRF involves a multinomial distribution (Proposition 5 and 6). Each evaluation of the multinomial creates computational burden when the dimensions (d and n) of the problems increases. For example, in Figure 8 and 9, the computation time needed to compute infinite KeRF estimates ranges between thirty minutes to 3 hours. As a matter of fact, infinite KeRF methods should be seen as theoretical tools rather than a practical substitute for random forests.

6 Proofs

Proof of Proposition 1. By definition,

$$
\begin{aligned}
\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) &= \frac{1}{\sum_{j=1}^{M} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)}} \sum_{j=1}^{M} \sum_{i=1}^{n} Y_i \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)} \\
&= \frac{M}{\sum_{j=1}^{M} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)}} \sum_{i=1}^{n} Y_i K_{M,n}(\mathbf{x}, \mathbf{X}_i).
\end{aligned}
$$

15

Finally, observe that

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

which concludes the proof. □

Proof of Proposition 2. We prove the result for $d = 2$. The other cases can be treated similarly. For the moment, we assume the random forest to be continuous. Recall that, for all $\mathbf{x}, \mathbf{z} \in [0,1]^2$, and for all $M \in \mathbb{N}$,

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

According to the strong law of large numbers, almost surely, for all $\mathbf{x}, \mathbf{z} \in \mathbb{Q}^2\cap[0,1]^2$

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

Set $\varepsilon > 0$ and $\mathbf{x},\mathbf{z} \in [0,1]^2$ where $\mathbf{x} = (x^{(1)}, x^{(2)})$ and $\mathbf{z} = (z^{(1)}, z^{(2)})$. Assume, without loss of generality, that $x^{(1)} < z^{(1)}$ and $x^{(2)} < z^{(2)}$. Let

$$A_{\mathbf{x}} = \{\mathbf{u} \in [0,1]^2, u^{(1)} \le x^{(1)} \text{ and } u^{(2)} \le x^{(2)}\},$$
$$\text{and } A_{\mathbf{z}} = \{\mathbf{u} \in [0,1]^2, u^{(1)} \ge z^{(1)} \text{ and } u^{(2)} \ge z^{(2)}\}.$$

Choose $\mathbf{x}_1 \in A_{\mathbf{x}} \cap \mathbb{Q}^2$ (resp. $\mathbf{z}_2 \in A_{\mathbf{z}} \cap \mathbb{Q}^2$) and take $\mathbf{x}_2 \in [0,1]^2 \cap \mathbb{Q}^2$ (resp. $\mathbf{z}_1 \in [0,1]^2 \cap \mathbb{Q}^2$) such that $x_1^{(1)} \le x^{(1)} \le x_2^{(1)}$ and $x_1^{(2)} \le x^{(2)} \le x_2^{(2)}$ (resp. $z_1^{(1)} \le z^{(1)} \le z_2^{(1)}$ and $z_1^{(2)} \le z^{(2)} \le z_2^{(2)}$, see Figure 10).

Figure 10: Respective positions of $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{z}, \mathbf{z}_1, \mathbf{z}_2$

Observe that, because of the continuity of K_n, one can choose $\mathbf{x}_1, \mathbf{x}_2$ close enough to $\mathbf{x}$ and $\mathbf{z}_2, \mathbf{z}_1$ close enough to $\mathbf{z}$ such that

$$|K_n(\mathbf{x}_2, \mathbf{x}_1) - 1| \le \varepsilon,$$
$$|K_n(\mathbf{z}_1, \mathbf{z}_2) - 1| \le \varepsilon,$$
$$\text{and } |K_n(\mathbf{x}_1, \mathbf{z}_2) - K_n(\mathbf{x}, \mathbf{z})| \le \varepsilon.$$

16

Bounding the difference between $K_{M,n}$ and K_n, we have

$$
\begin{aligned}
|K_{M,n}(\mathbf{x}, \mathbf{z}) - K_n(\mathbf{x}, \mathbf{z})| \leq\; & |K_{M,n}(\mathbf{x}, \mathbf{z}) - K_{M,n}(\mathbf{x}_1, \mathbf{z}_2)| \\
& + |K_{M,n}(\mathbf{x}_1, \mathbf{z}_2) - K_n(\mathbf{x}_1, \mathbf{z}_2)| \\
& + |K_n(\mathbf{x}_1, \mathbf{z}_2) - K_n(\mathbf{x}, \mathbf{z})|. \tag{8}
\end{aligned}
$$

To simplify notation, we let $\mathbf{x} \overset{\Theta_m}{\leftrightarrow} \mathbf{z}$ be the event where $\mathbf{x}$ and $\mathbf{z}$ are in the same cell in the tree built with randomness Θ_j and dataset $\mathcal{D}_n$. We also let $\mathbf{x} \overset{\Theta_m}{\nleftrightarrow} \mathbf{z}$ be the complement event of $\mathbf{x} \overset{\Theta_m}{\leftrightarrow} \mathbf{z}$. Accordingly, the first term on the right side in equation (8) is bounded above by

$$
\begin{aligned}
|K_{M,n}(\mathbf{x}, \mathbf{z}) - K_{M,n}(\mathbf{x}_1, \mathbf{z}_2)| &\leq \frac{1}{M} \sum_{m=1}^{M} \left| \mathbb{1}_{\mathbf{x} \overset{\Theta_m}{\leftrightarrow} \mathbf{z}} - \mathbb{1}_{\mathbf{x}_1 \overset{\Theta_m}{\leftrightarrow} \mathbf{z}_2} \right| \\
&\leq \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\mathbf{x}_1 \overset{\Theta_m}{\nleftrightarrow} \mathbf{x}} + \mathbb{1}_{\mathbf{z}_2 \overset{\Theta_m}{\nleftrightarrow} \mathbf{z}} \\
&\quad \text{(given the positions of } \mathbf{x}, \mathbf{x}_1, \mathbf{z}, \mathbf{z}_2) \\
&\leq \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\mathbf{x}_1 \overset{\Theta_m}{\nleftrightarrow} \mathbf{x}_2} + \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\mathbf{z}_2 \overset{\Theta_m}{\nleftrightarrow} \mathbf{z}_1}, \tag{9}
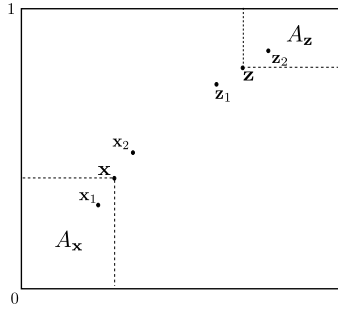\end{aligned}
$$

given the respective positions of $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{z}, \mathbf{z}_1, \mathbf{z}_2$. But, since $\mathbf{x}_2, \mathbf{z}_1, \mathbf{x}_1, \mathbf{z}_2 \in \mathbb{Q}^2 \cap [0,1]^2$, we deduce from inequation (9) that, for all M large enough,

$$
|K_{M,n}(\mathbf{x}, \mathbf{z}) - K_{M,n}(\mathbf{x}_1, \mathbf{z}_2)| \leq 1 - K_n(\mathbf{x}_2, \mathbf{x}_1) + 1 - K_n(\mathbf{z}_1, \mathbf{z}_2) + 2\varepsilon.
$$

Combining the last inequality with equation (8), we obtain, for all M large enough,

$$
\begin{aligned}
|K_{M,n}(\mathbf{x}, \mathbf{z}) - K_n(\mathbf{x}, \mathbf{z})| \leq\; & 1 - K_n(\mathbf{x}_2, \mathbf{x}_1) + 1 - K_n(\mathbf{z}_1, \mathbf{z}_2) \\
& + |K_{M,n}(\mathbf{x}_1, \mathbf{z}_2) - K_n(\mathbf{x}_1, \mathbf{z}_2)| \\
& + |K_n(\mathbf{x}_1, \mathbf{z}_2) - K_n(\mathbf{x}, \mathbf{z})| + 2\varepsilon \\
\leq\; & 6\varepsilon.
\end{aligned}
$$

Consequently, for any continuous random forest, almost surely, for all $\mathbf{x}, \mathbf{z} \in [0,1]^2$,

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

The proof can be easily adapted to the case of discrete random forests. Thus, this complete the first part of the proof. Next, observe that

$$
\lim_{M \to \infty} \frac{\sum_{i=1}^{n} Y_i K_{M,n}(\mathbf{x}, \mathbf{X}_i)}{\sum_{j=1}^{n} K_{M,n}(\mathbf{x}, \mathbf{X}_j)} = \frac{\sum_{i=1}^{n} Y_i K_n(\mathbf{x}, \mathbf{X}_i)}{\sum_{j=1}^{n} K_n(\mathbf{x}, \mathbf{X}_j)},
$$

for all $\mathbf{x}$ satisfying $\sum_{j=1}^{n} K_n(\mathbf{x}, \mathbf{X}_j) \neq 0$. Thus, almost surely for those $\mathbf{x}$,

$$
\lim_{M \to \infty} \widetilde{m}_{M,n}(\mathbf{x}) = \widetilde{m}_{\infty,n}(\mathbf{x}). \tag{10}
$$

Now, if there exists any $\mathbf{x}$ such that $\sum_{j=1}^{n} K_n(\mathbf{x}, \mathbf{X}_j) = 0$, then $\mathbf{x}$ is not connected with any data points in any tree of the forest. In that case, $\sum_{j=1}^{n} K_{M,n}(\mathbf{x}, \mathbf{X}_j) = 0$ and, by convention, $\widetilde{m}_{\infty,n}(\mathbf{x}) = \widetilde{m}_{M,n}(\mathbf{x}) = 0$. Finally, formula (10) holds for all $\mathbf{x} \in [0,1]^2$. $\qquad\square$

Proof of Proposition 3. Fix $\mathbf{x} \in [0,1]^d$ and assume that, a.s., $Y \geq 0$. By assumption **(H1.1)**, there exist sequences $(a_n), (b_n)$ such that, almost surely,

$$a_n \leq N_n(\mathbf{x}, \Theta) \leq b_n.$$

To simplify notation, we let $\bar{N}_{M,n}(\mathbf{x}, \Theta) = \frac{1}{M} \sum_{j=1}^{M} N_n(\mathbf{x}, \Theta_j)$. Thus, almost surely,

$$
\begin{aligned}
|m_{M,n}(\mathbf{x}) - \widetilde{m}_{M,n}(\mathbf{x})| &= \left| \sum_{i=1}^{n} Y_i \left(\frac{1}{M} \sum_{m=1}^{M} \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_m)}}{N_n(\mathbf{x}, \Theta_m)} \right) \right. \\
&\quad \left. - \sum_{i=1}^{n} Y_i \left(\frac{1}{M} \sum_{m=1}^{M} \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_m)}}{\bar{N}_{M,n}(\mathbf{x})} \right) \right| \\
&\leq \frac{1}{M} \sum_{i=1}^{n} Y_i \sum_{m=1}^{M} \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_m)}}{\bar{N}_{M,n}(\mathbf{x})} \times \left| \frac{\bar{N}_{M,n}(\mathbf{x})}{N_n(\mathbf{x}, \Theta_m)} - 1 \right| \\
&\leq \frac{b_n - a_n}{a_n} \widetilde{m}_{M,n}(\mathbf{x}).
\end{aligned}
$$

$\square$

Proof of Proposition 4. Fix $\mathbf{x} \in [0,1]^d$ and assume that, almost surely, $Y \geq 0$. By assumption **(H1.2)**, there exist sequences (a_n), (b_n), (ε_n) such that, letting A be the event where

$$a_n \leq N_n(\mathbf{x}, \Theta) \leq b_n,$$

we have, almost surely,

$$\mathbb{P}_{\Theta}[A] \geq 1 - \varepsilon_n \quad \text{and} \quad 1 \leq a_n \leq \mathbb{E}_{\Theta}[N_n(\mathbf{x}, \Theta)] \leq b_n.$$

Therefore, a.s.,

$$
\begin{aligned}
&|m_{\infty,n}(\mathbf{x}) - \widetilde{m}_{\infty,n}(\mathbf{x})| \\
&= \left| \sum_{i=1}^{n} Y_i \mathbb{E}_{\Theta} \left[\frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta)}}{N_n(\mathbf{x}, \Theta)} \right] - \sum_{i=1}^{n} Y_i \mathbb{E}_{\Theta} \left[\frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta)}}{\mathbb{E}_{\Theta}[N_n(\mathbf{x}, \Theta)]} \right] \right| \\
&= \left| \sum_{i=1}^{n} Y_i \mathbb{E}_{\Theta} \left[\frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta)}}{\mathbb{E}_{\Theta}[N_n(\mathbf{x}, \Theta)]} \left(\frac{\mathbb{E}_{\Theta}[N_n(\mathbf{x}, \Theta)]}{N_n(\mathbf{x}, \Theta)} - 1 \right) (\mathbb{1}_A + \mathbb{1}_{A^c}) \right] \right| \\
&\leq \frac{b_n - a_n}{a_n} \widetilde{m}_{\infty,n}(\mathbf{x}) + \left(\max_{1 \leq i \leq n} Y_i \right) \mathbb{E}_{\Theta} \left[\left| 1 - \frac{N_n(\mathbf{x}, \Theta)}{\mathbb{E}_{\Theta}[N_n(\mathbf{x}, \Theta)]} \right| \mathbb{1}_{A^c} \right] \\
&\leq \frac{b_n - a_n}{a_n} \widetilde{m}_{\infty,n}(\mathbf{x}) + n \left(\max_{1 \leq i \leq n} Y_i \right) \mathbb{P}[A^c].
\end{aligned}
$$

Consequently, almost surely,

$$|m_{\infty,n}(\mathbf{x}) - \widetilde{m}_{\infty,n}(\mathbf{x})| \leq \frac{b_n - a_n}{a_n} \widetilde{m}_{\infty,n}(\mathbf{x}) + n\varepsilon_n \left(\max_{1 \leq i \leq n} Y_i \right).$$

$\square$

Proof of Proposition 5. Assume for the moment that $d = 1$. Take $x, z \in [0, 1]$ and assume, without loss of generality, that $x \leq z$. Then the probability that x and z be in the same cell, after k cuts, is equal to

$$\mathbb{1}_{\lceil 2^{k_j} x_j \rceil = \lceil 2^{k_j} z_j \rceil}.$$

To prove the result in the multivariate case, take $\mathbf{x}, \mathbf{z} \in [0, 1]^d$. Since cuts are independent, the probability that $\mathbf{x}$ and $\mathbf{z}$ are in the same cell after k cuts is given by the following multinomial

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

$\square$

To prove Theorem 4.1, we need to control the bias of the centred KeRF estimate, which is done in Theorem 6.1.

Theorem 6.1. *Let f be a L-Lipschitz function. Then, for all k,*

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

Proof of Theorem 6.1. Let $\mathbf{x} \in [0, 1]^d$ and $k \in \mathbb{N}$. Take f a L-Lipschitz function. In the rest of the proof, for clarity reasons, we use the notation $\mathrm{d}\mathbf{z}$ instead of $\mathrm{d}z_1 \ldots \mathrm{d}z_d$. Thus,

$$\left| \frac{\int_{[0,1]^d} K_k^{cc}(\mathbf{x}, \mathbf{z}) f(\mathbf{z}) \mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K_k^{cc}(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z}} - f(\mathbf{x}) \right| \leq \frac{\int_{[0,1]^d} K_k^{cc}(\mathbf{x}, \mathbf{z}) |f(\mathbf{z}) - f(\mathbf{x})| \mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K_k^{cc}(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z}}.$$

Note that,

$$\int_{[0,1]^d} K_k^{cc}(\mathbf{x}, \mathbf{z}) |f(\mathbf{z}) - f(\mathbf{x})| \mathrm{d}\mathbf{z}$$

$$\leq L \sum_{\ell=1}^d \int_{[0,1]^d} K_k^{cc}(\mathbf{x}, \mathbf{z}) |z_\ell - x_\ell| \mathrm{d}\mathbf{z}$$

$$\leq L \sum_{\ell=1}^d \int_{[0,1]^d} \sum_{\substack{k_1, \ldots, k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1! \ldots k_d!} \left(\frac{1}{d}\right)^k \prod_{m \neq \ell} \int_0^1 K_{k_m}^{cc}(x_m, z_m) \mathrm{d}z_m$$

$$\times \int_0^1 K_{k_\ell}^{cc}(x_\ell, z_\ell) |z_\ell - x_\ell| \mathrm{d}z_\ell. \tag{11}$$

The last integral is upper bounded by

$$
\int_{[0,1]} K^{cc}_{k_\ell}(x_\ell, z_\ell)|x_\ell - z_\ell|\mathrm{d}z_\ell = \int_{[0,1]} \mathbb{1}_{\lceil 2^{k_\ell} x_\ell \rceil = \lceil 2^{k_\ell} z_\ell \rceil}|x_\ell - z_\ell|\mathrm{d}z_\ell
$$

$$
\leq \left(\frac{1}{2}\right)^{k_\ell} \int_{[0,1]} \mathbb{1}_{\lceil 2^{k_\ell} x_\ell \rceil = \lceil 2^{k_\ell} z_\ell \rceil}\mathrm{d}z_\ell
$$

$$
\leq \left(\frac{1}{2}\right)^{k_\ell} \int_{[0,1]} K^{cc}_{k_\ell}(x_\ell, z_\ell)\mathrm{d}z_\ell. \tag{12}
$$

Therefore, combining inequalities (11) and (12), we obtain,

$$
\int_{[0,1]^d} K^{cc}_{k}(\mathbf{x}, \mathbf{z})|f(\mathbf{z}) - f(\mathbf{x})|\mathrm{d}\mathbf{z}
$$

$$
\leq L \sum_{\ell=1}^{d} \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^{d} k_j = k}} \frac{k!}{k_1!\ldots k_d!} \left(\frac{1}{2}\right)^{k_\ell} \left(\frac{1}{d}\right)^{k} \prod_{m=1}^{d} \int_{0}^{1} K^{cc}_{k_m}(x_m, z_m)\mathrm{d}z_m
$$

$$
\leq L \left(\frac{1}{d}\right)^{k} \sum_{\ell=1}^{d} \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^{d} k_j = k}} \frac{k!}{k_1!\ldots k_d!} \left(\frac{1}{2}\right)^{k_\ell + k}, \tag{13}
$$

since, simple calculations show that, for all $x_m \in [0,1]$ and for all $k_m \in \mathbb{N}$,

$$
\int_{0}^{1} K^{cc}_{k_m}(x_m, z_m)\mathrm{d}z_m = \int_{[0,1]} \mathbb{1}_{\lceil 2^{k_m} x_m \rceil = \lceil 2^{k_m} z_m \rceil}dz_m = \left(\frac{1}{2}\right)^{k_m}. \tag{14}
$$

Consequently, we get from inequality (13) that

$$
\frac{\int_{[0,1]^d} K^{cc}_{k}(\mathbf{x}, \mathbf{z})|f(\mathbf{z}) - f(\mathbf{x})|\mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K^{cc}_{k}(\mathbf{x}, \mathbf{z})\mathrm{d}\mathbf{z}} \leq L \left(\frac{1}{d}\right)^{k} \sum_{\ell=1}^{d} \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^{d} k_j = k}} \frac{k!}{k_1!\ldots k_d!} \left(\frac{1}{2}\right)^{k_\ell}.
$$

Taking the first term of the sum, we obtain

$$
\left(\frac{1}{d}\right)^{k} \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^{d} k_j = k}} \frac{k!}{k_1!\ldots k_d!} \left(\frac{1}{2}\right)^{k_1} = \sum_{k_1=0}^{k} \left(\frac{1}{2d}\right)^{k_1} \left(1 - \frac{1}{d}\right)^{k-k_1} \frac{k!}{k_1!(k-k_1)!}
$$

$$
\leq \left(1 - \frac{1}{2d}\right)^{k}.
$$

Finally,

$$
\frac{\int_{[0,1]^d} K^{cc}_{k}(\mathbf{x}, \mathbf{z})|f(\mathbf{z}) - f(\mathbf{x})|\mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K^{cc}_{k}(\mathbf{x}, \mathbf{z})\mathrm{d}\mathbf{z}} \leq Ld \left(1 - \frac{1}{2d}\right)^{k}.
$$

$\square$

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

$$\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K^{cc}_k(\mathbf{x}, \mathbf{X}_i)}{\sum_{i=1}^n K^{cc}_k(\mathbf{x}, \mathbf{X}_i)}.$$

Thus, letting

$$A_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i K^{cc}_k(\mathbf{x}, \mathbf{X}_i)}{\mathbb{E}\left[K^{cc}_k(\mathbf{x}, \mathbf{X})\right]} - \frac{\mathbb{E}\left[Y K^{cc}_k(\mathbf{x}, \mathbf{X})\right]}{\mathbb{E}\left[K^{cc}_k(\mathbf{x}, \mathbf{X})\right]} \right),$$

$$B_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \left(\frac{K^{cc}_k(\mathbf{x}, \mathbf{X}_i)}{\mathbb{E}\left[K^{cc}_k(\mathbf{x}, \mathbf{X})\right]} - 1 \right),$$

and $M_n(\mathbf{x}) = \frac{\mathbb{E}\left[Y K^{cc}_k(\mathbf{x}, \mathbf{X})\right]}{\mathbb{E}\left[K^{cc}_k(\mathbf{x}, \mathbf{X})\right]}$,

the estimate $\widetilde{m}^{cc}_{\infty,n}(\mathbf{x})$ can be rewritten as

$$\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) = \frac{M_n(\mathbf{x}) + A_n(\mathbf{x})}{1 + B_n(\mathbf{x})},$$

which leads to

$$\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) - m(\mathbf{x}) = \frac{M_n(\mathbf{x}) - m(\mathbf{x}) + A_n(\mathbf{x}) - B_n(\mathbf{x})m(\mathbf{x})}{1 + B_n(\mathbf{x})}.$$

According to Theorem 6.1, we have

$$\begin{aligned}
|M_n(\mathbf{x}) - m(\mathbf{x})| &= \left| \frac{\mathbb{E}\left[m(\mathbf{X}) K^{cc}_k(\mathbf{x}, \mathbf{X})\right]}{\mathbb{E}\left[K^{cc}_k(\mathbf{x}, \mathbf{X})\right]} + \frac{\mathbb{E}\left[\varepsilon K^{cc}_k(\mathbf{x}, \mathbf{X})\right]}{\mathbb{E}\left[K^{cc}_k(\mathbf{x}, \mathbf{X})\right]} - m(\mathbf{x}) \right| \\
&\leq \left| \frac{\mathbb{E}\left[m(\mathbf{X}) K^{cc}_k(\mathbf{x}, \mathbf{X})\right]}{\mathbb{E}\left[K^{cc}_k(\mathbf{x}, \mathbf{X})\right]} - m(\mathbf{x}) \right| \\
&\leq C_1 \left(1 - \frac{1}{2d} \right)^k,
\end{aligned}$$

where $C_1 = Ld$. Take $\alpha \in]0, 1/2]$. Let $\mathcal{C}_\alpha(\mathbf{x})$ be the event on which $\{|A_n(\mathbf{x})|, |B_n(\mathbf{x})| \leq \alpha\}$. On the event $\mathcal{C}_\alpha(\mathbf{x})$, we have

$$\begin{aligned}
|\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})|^2 &\leq 8|M_n(\mathbf{x}) - m(\mathbf{x})|^2 + 8|A_n(\mathbf{x}) - B_n(\mathbf{x})m(\mathbf{x})|^2 \\
&\leq 8C_1^2 \left(1 - \frac{1}{2d} \right)^{2k} + 8\alpha^2 (1 + \|m\|_\infty)^2.
\end{aligned}$$

Thus,

$$\mathbb{E}[|\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha(\mathbf{x})}] \leq 8C_1^2 \left(1 - \frac{1}{2d} \right)^{2k} + 8\alpha^2 (1 + \|m\|_\infty)^2. \quad (15)$$

Consequently, to find an upper bound on the rate of consistency of $\widetilde{m}^{cc}_{\infty,n}$, we

just need to upper bound

$$\mathbb{E}\Big[|\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}^c_\alpha(\mathbf{x})}\Big] \leq \mathbb{E}\Big[|\max_{1 \leq i \leq n} Y_i + m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}^c_\alpha(\mathbf{x})}\Big]$$

(since $\widetilde{m}^{cc}_{\infty,n}$ is a local averaging estimate)

$$\leq \mathbb{E}\Big[|2\|m\|_\infty + \max_{1 \leq i \leq n} \varepsilon_i|^2 \mathbb{1}_{\mathcal{C}^c_\alpha(\mathbf{x})}\Big]$$

$$\leq \left(\mathbb{E}\left[2\|m\|_\infty + \max_{1 \leq i \leq n} \varepsilon_i\right]^4 \mathbb{P}\left[\mathcal{C}^c_\alpha(\mathbf{x})\right]\right)^{1/2}$$

(by Cauchy-Schwarz inequality)

$$\leq \left(\left(16\|m\|^4_\infty + 8\mathbb{E}\Big[\max_{1 \leq i \leq n} \varepsilon_i\Big]^4\right) \mathbb{P}\left[\mathcal{C}^c_\alpha(\mathbf{x})\right]\right)^{1/2}.$$

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

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

Thus, there exists C_2 such that, for all $n > 1$,

$$\mathbb{E}\Big[|\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}^c_\alpha(\mathbf{x})}\Big] \leq C_2(\log n)(\mathbb{P}\left[\mathcal{C}^c_\alpha(\mathbf{x})\right])^{1/2}. \tag{16}$$

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

$$\mathbb{P}\left[|A_n(\mathbf{x})| > \alpha\right] \leq \frac{1}{n\alpha^2}\mathbb{E}\left[\frac{Y K^{cc}_k(\mathbf{x}, \mathbf{X})}{\mathbb{E}\left[K^{cc}_k(\mathbf{x}, \mathbf{X})\right]} - \frac{\mathbb{E}\left[Y K^{cc}_k(\mathbf{x}, \mathbf{X})\right]}{\mathbb{E}\left[K^{cc}_k(\mathbf{x}, \mathbf{X})\right]}\right]^2$$

$$\leq \frac{1}{n\alpha^2}\frac{1}{(\mathbb{E}\left[K^{cc}_k(\mathbf{x}, \mathbf{X})\right])^2}\mathbb{E}\left[Y^2 K^{cc}_k(\mathbf{x}, \mathbf{X})^2\right]$$

$$\leq \frac{2}{n\alpha^2}\frac{1}{(\mathbb{E}\left[K^{cc}_k(\mathbf{x}, \mathbf{X})\right])^2}\left(\mathbb{E}\left[m(\mathbf{X})^2 K^{cc}_k(\mathbf{x}, \mathbf{X})^2\right]\right.$$

$$\left. + \mathbb{E}\left[\varepsilon^2 K^{cc}_k(\mathbf{x}, \mathbf{X})^2\right]\right)$$

$$\leq \frac{2(\|m\|^2_\infty + \sigma^2)}{n\alpha^2}\frac{\mathbb{E}\left[K^{cc}_k(\mathbf{x}, \mathbf{X})\right]}{(\mathbb{E}\left[K^{cc}_k(\mathbf{x}, \mathbf{X})\right])^2}$$

(since $\sup_{\mathbf{x},\mathbf{z} \in [0,1]^d} K^{cc}_k(\mathbf{x}, \mathbf{z}) \leq 1$)

$$\leq \frac{2M^2_1}{\alpha^2}\frac{2^k}{n}$$

(according to inequality (14)),

where $M^2_1 = \|m\|^2_\infty + \sigma^2$. Meanwhile with respect to $B_n(\mathbf{x})$, we obtain, still by

Chebyshev's inequality,

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

$$\leq \frac{1}{n\alpha^2} \frac{1}{\mathbb{E}\left[K_k^{cc}(\mathbf{x}, \mathbf{X})\right]}$$

$$\left(\text{since} \sup_{\mathbf{x}, \mathbf{z} \in [0,1]^d} K_k^{cc}(\mathbf{x}, \mathbf{z}) \leq 1\right)$$

$$\leq \frac{2^k}{n\alpha^2}.$$

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

$$\mathbb{P}\big[\mathcal{C}_\alpha(\mathbf{x})\big] \geq 1 - \mathbb{P}\big(|A_n(\mathbf{x})| \geq \alpha\big) - \mathbb{P}\big(|B_n(\mathbf{x})| \geq \alpha\big)$$

$$\geq 1 - \frac{2^k}{n} \frac{2M_1^2}{\alpha^2} - \frac{2^k}{n\alpha^2}$$

$$\geq 1 - \frac{2^k(2M_1^2 + 1)}{n\alpha^2}.$$

Consequently, according to inequality (16), we obtain

$$\mathbb{E}\left[|\widetilde{m}_{\infty,n}^{cc}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(\mathbf{x})}\right] \leq C_2(\log n)\left(\frac{2^k(2M_1^2 + 1)}{n\alpha^2}\right)^{1/2}.$$

Then using inequality (15),

$$\mathbb{E}\left[\widetilde{m}_{\infty,n}^{cc}(\mathbf{x}) - m(\mathbf{x})\right]^2$$

$$\leq \mathbb{E}\left[|\widetilde{m}_{\infty,n}^{cc}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha(\mathbf{x})}\right] + \mathbb{E}\left[|\widetilde{m}_{\infty,n}^{cc}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(\mathbf{x})}\right]$$

$$\leq 8C_1^2 \left(1 - \frac{1}{2d}\right)^{2k} + 8\alpha^2(1 + \|m\|_\infty)^2 + C_2(\log n)\left(\frac{2^k(2M_1^2 + 1)}{n\alpha^2}\right)^{1/2}$$

Optimizing the right hand side in α, we get

$$\mathbb{E}\left[\widetilde{m}_{\infty,n}^{cc}(\mathbf{x}) - m(\mathbf{x})\right]^2 \leq 8C_1^2 \left(1 - \frac{1}{2d}\right)^{2k} + C_3 \left(\frac{(\log n)^2 2^k}{n}\right)^{1/3},$$

for some constant $C_3 > 0$. The last expression is minimized for

$$k = C_4 + \frac{1}{\log 2 + \frac{3}{d}} \log\left(\frac{n}{(\log n)^2}\right),$$

where $C_4 = \left(\frac{1}{d} + \frac{\log 2}{3}\right)^{-1} \log\left(\frac{C_3 d \log 2}{24 C_1^2}\right)$. Consequently, there exists a constant C_5 such that, for all $n > 1$,

$$\mathbb{E}\left[\widetilde{m}_{\infty,n}^{cc}(\mathbf{x}) - m(\mathbf{x})\right]^2 \leq C_5 n^{-\frac{1}{d\log 2 + 3}} (\log n)^2.$$

$\square$

Proof of Lemma 1. Let $x, z \in [0, 1]$ such that $x < z$. The first statement comes from the fact that splits are drawn uniformly over $[0, 1]$. To address the second one, denote by Z_1 (resp. Z_2) the position of the first (resp. second) split used to build the cell containing $[x, z]$. Observe that, given $Z_1 = z_1$, Z_2 is uniformly distributed over $[z_1, 1]$ (resp. $[0, z_1]$) if $z_1 \leq x$ (resp. $z_1 \geq z$). Thus, we have

$$K_2^{uf}(x, z) = \int_{z_1=0}^{x} \left(\int_{z_2=z_1}^{x} \frac{1}{1-z_1} \mathrm{d}z_1 \mathrm{d}z_2 + \int_{z_2=z}^{1} \frac{1}{1-z_1} \mathrm{d}z_1 \mathrm{d}z_2 \right)$$
$$+ \int_{z_1=z}^{1} \left(\int_{z_2=0}^{x} \frac{1}{1-z_1} \mathrm{d}z_1 \mathrm{d}z_2 + \int_{z_2=z}^{z_1} \frac{1}{1-z_1} \mathrm{d}z_1 \mathrm{d}z_2 \right).$$

The first term takes the form

$$\int_0^x \frac{1}{z_1} \left(\int_{z_1}^x \mathrm{d}z_2 \right) \mathrm{d}z_1 = \int_0^x \frac{x - z_1}{1 - z_1} \mathrm{d}z_1$$
$$= x - (1 - x) \log(1 - x).$$

Similarly, one has

$$\int_0^x \int_z^1 \frac{1}{1-z_1} \mathrm{d}z_1 \mathrm{d}z_2 = (1 - z) \log(1 - x),$$
$$\int_z^1 \int_z^{z_1} \frac{1}{z_1} \mathrm{d}z_1 \mathrm{d}z_2 = (1 - z) + z \log z,$$
$$\int_z^1 \int_0^x \frac{1}{z_1} \mathrm{d}z_1 \mathrm{d}z_2 = -x \log z.$$

Consequently,

$$K_2^{uf}(x, z) = x - (1 - x) \log(1 - x) + (1 - z) \log(1 - x)$$
$$- x \log z + (1 - z) + z \log z$$
$$= 1 - (z - x) + (z - x) \log \left(\frac{z}{1 - x} \right).$$

$\square$

Proof of Proposition 6. The result is proved in Technical Proposition 2 in Scornet (2014).

$\square$

To prove Theorem 4.2, we need to control the bias of uniform KeRF estimates, which is done in Theorem 6.2.

Theorem 6.2. *Let f be a L-Lipschitz function. Then, for all k,*

$$\sup_{\mathbf{x} \in [0,1]^d} \left| \frac{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z} - \mathbf{x}|) f(\mathbf{z}) \mathrm{d}z_1 \ldots \mathrm{d}z_d}{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z} - \mathbf{x}|) \mathrm{d}z_1 \ldots \mathrm{d}z_d} - f(\mathbf{x}) \right| \leq \frac{Ld2^{2d+1}}{3} \left(1 - \frac{1}{3d} \right)^k.$$

Proof of Theorem 6.2. Let $\mathbf{x} \in [0,1]^d$ and $k \in \mathbb{N}$. Let f be a L-Lipschitz function. In the rest of the proof, for clarity reasons, we use the notation $\mathrm{d}\mathbf{z}$ instead of $\mathrm{d}z_1 \ldots \mathrm{d}z_d$. Thus,

$$\left| \frac{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z} - \mathbf{x}|) f(\mathbf{z}) \mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z} - \mathbf{x}|) \mathrm{d}\mathbf{z}} - f(\mathbf{x}) \right| \le \frac{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z} - \mathbf{x}|) |f(\mathbf{z}) - f(\mathbf{x})| \mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z} - \mathbf{x}|) \mathrm{d}\mathbf{z}}.$$

Note that,

$$\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z} - \mathbf{x}|) |f(\mathbf{z}) - f(\mathbf{x})| \mathrm{d}\mathbf{z}$$

$$\le L \sum_{\ell=1}^d \int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z} - \mathbf{x}|) |z_\ell - x_\ell| \mathrm{d}\mathbf{z}$$

$$\le L \sum_{\ell=1}^d \int_{[0,1]^d} \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1! \ldots k_d!} \left(\frac{1}{d}\right)^k \prod_{m=1}^d K_{k_m}^{uf}(0, |z_m - x_m|) |z_\ell - x_\ell| \mathrm{d}\mathbf{z}$$

$$\le L \sum_{\ell=1}^d \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1! \ldots k_d!} \left(\frac{1}{d}\right)^k \prod_{m \ne \ell} \int_0^1 K_{k_m}^{uf}(0, |z_m - x_m|) \mathrm{d}z_m$$

$$\times \int_0^1 K_{k_\ell}^{uf}(0, |z_\ell - x_\ell|) |z_\ell - x_\ell| \mathrm{d}z_\ell$$

$$\le L \sum_{\ell=1}^d \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1! \ldots k_d!} \left(\frac{2}{3}\right)^{k_l + 1} \left(\frac{1}{d}\right)^k \prod_{m=1}^d \int_0^1 K_{k_m}^{uf}(0, |z_m - x_m|) \mathrm{d}z_m$$

(according to the second statement of Lemma 2, see below)

$$\le \frac{L}{2^{k-d}} \sum_{\ell=1}^d \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1! \ldots k_d!} \left(\frac{2}{3}\right)^{k_\ell + 1} \left(\frac{1}{d}\right)^k, \tag{17}$$

according to the first statement of Lemma 2. Still by Lemma 2 and using inequality (17), we have,

$$\frac{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z} - \mathbf{x}|) |f(\mathbf{z}) - f(\mathbf{x})| \mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z} - \mathbf{x}|) \mathrm{d}\mathbf{z}}$$

$$\le \frac{L 2^{2d+1}}{3} \sum_{\ell=1}^d \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1! \ldots k_d!} \left(\frac{2}{3}\right)^{k_\ell} \left(\frac{1}{d}\right)^k.$$

Taking the first term of the sum, we obtain

$$\sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1! \ldots k_d!} \left(\frac{2}{3}\right)^{k_1} \left(\frac{1}{d}\right)^k = \sum_{k_1=0}^k \left(\frac{2}{3d}\right)^{k_1} \left(1 - \frac{1}{d}\right)^{k-k_1} \frac{k!}{k_1!(k-k_1)!}$$

$$\leq \left(1 - \frac{1}{3d}\right)^k.$$

Finally,

$$\frac{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|)|f(\mathbf{z}) - f(\mathbf{x})|\mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|)\mathrm{d}\mathbf{z}} \leq \frac{L2^{2d+1}}{3}\left(1 - \frac{1}{3d}\right)^k.$$

$\square$

Proof of Theorem 4.2. Let $\mathbf{x} \in [0,1]^d$, $\|m\|_\infty = \sup_{\mathbf{x}\in[0,1]^d} |m(\mathbf{x})|$ and recall that

$$m_{\infty,n}^{uf}(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K_k^{uf}(\mathbf{0}, |\mathbf{X}_i - \mathbf{x}|)}{\sum_{i=1}^n K_k^{uf}(\mathbf{0}, |\mathbf{X}_i - \mathbf{x}|)}.$$

Thus, letting

$$A_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \left(\frac{Y_i K_k^{uf}(\mathbf{0}, |\mathbf{X}_i - \mathbf{x}|)}{\mathbb{E}\left[K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)\right]} - \frac{\mathbb{E}\left[Y K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)\right]}{\mathbb{E}\left[K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)\right]}\right),$$

$$B_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \left(\frac{K_k^{uf}(\mathbf{0}, |\mathbf{X}_i - \mathbf{x}|)}{\mathbb{E}\left[K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)\right]} - 1\right),$$

and $M_n(\mathbf{x}) = \dfrac{\mathbb{E}\left[Y K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)\right]}{\mathbb{E}\left[K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)\right]},$

the estimate $m_{\infty,n}^{uf}(\mathbf{x})$ can be rewritten as

$$m_{\infty,n}^{uf}(\mathbf{x}) = \frac{M_n(\mathbf{x}) + A_n(\mathbf{x})}{1 + B_n(\mathbf{x})},$$

which leads to

$$m_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x}) = \frac{M_n(\mathbf{x}) - m(\mathbf{x}) + A_n(\mathbf{x}) - B_n(\mathbf{x})m(\mathbf{x})}{1 + B_n(\mathbf{x})}.$$

Note that, according to Theorem 6.2, we have

$$|M_n(\mathbf{x}) - m(\mathbf{x})| = \left|\frac{\mathbb{E}[m(\mathbf{X})K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]} + \frac{\mathbb{E}[\varepsilon K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]} - m(\mathbf{x})\right|$$

$$\leq \left|\frac{\mathbb{E}[m(\mathbf{X})K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]} - m(\mathbf{x})\right|$$

$$\leq C_1\left(1 - \frac{1}{3d}\right)^k,$$

26

where $C_1 = L2^{2d+1}/3$. Take $\alpha \in]0, 1/2]$. Let $\mathcal{C}_\alpha(\mathbf{x})$ be the event on which $\{|A_n(\mathbf{x})|, |B_n(\mathbf{x})| \leq \alpha\}$. On the event $\mathcal{C}_\alpha(\mathbf{x})$, we have

$$|m_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})|^2 \leq 8|M_n(\mathbf{x}) - m(\mathbf{x})|^2 + 8|A_n(\mathbf{x}) - B_n(\mathbf{x})m(\mathbf{x})|^2$$
$$\leq 8C_1^2 \left(1 - \frac{1}{3d}\right)^{2k} + 8\alpha^2(1 + \|m\|_\infty)^2.$$

Thus,

$$\mathbb{E}[|m_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha(\mathbf{x})}] \leq 8C_1^2 \left(1 - \frac{1}{3d}\right)^{2k} + 8\alpha^2(1 + \|m\|_\infty)^2. \quad (18)$$

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

$$\mathbb{E}\left[|\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(\mathbf{x})}\right] \leq \mathbb{E}\left[|\max_{1 \leq i \leq n} Y_i + m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(\mathbf{x})}\right]$$

$$\text{(since } \widetilde{m}_{\infty,n}^{uf} \text{ is a local averaging estimate)}$$

$$\leq \mathbb{E}\left[|2\|m\|_\infty + \max_{1 \leq i \leq n} \varepsilon_i|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(\mathbf{x})}\right]$$

$$\leq \left(\mathbb{E}\left[2\|m\|_\infty + \max_{1 \leq i \leq n} \varepsilon_i\right]^4 \mathbb{P}\left[\mathcal{C}_\alpha^c(\mathbf{x})\right]\right)^{1/2}$$

$$\text{(by Cauchy-Schwarz inequality)}$$

$$\leq \left(\left(16\|m\|_\infty^4 + 8\mathbb{E}\left[\max_{1 \leq i \leq n} \varepsilon_i\right]^4\right) \mathbb{P}\left[\mathcal{C}_\alpha^c(\mathbf{x})\right]\right)^{1/2}.$$

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

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

Thus, there exists C_2 such that, for all $n > 1$,

$$\mathbb{E}\left[|\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(\mathbf{x})}\right] \leq C_2(\log n)(\mathbb{P}\left[\mathcal{C}_\alpha^c(\mathbf{x})\right])^{1/2}. \quad (19)$$

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

$$\mathbb{P}\left[|A_n(\mathbf{x})| > \alpha\right] \leq \frac{1}{n\alpha^2} \mathbb{E}\left[\frac{YK_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)]} - \frac{\mathbb{E}[YK_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)]}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)]}\right]^2$$

$$\leq \frac{1}{n\alpha^2} \frac{1}{(\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)])^2} \mathbb{E}\left[Y^2 K_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)^2\right]$$

$$\leq \frac{2}{n\alpha^2} \frac{1}{(\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)])^2} \left(\mathbb{E}[m(\mathbf{X})^2 K_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)^2]\right.$$

$$\left. + \mathbb{E}[\varepsilon^2 K_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)^2]\right),$$

which leads to

$$\mathbb{P}\big[|A_n(\mathbf{x})| > \alpha\big] \leq \frac{2(\|m\|_\infty^2 + \sigma^2)}{n\alpha^2} \frac{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)]}{(\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)])^2}$$

$$(\text{since} \sup_{\mathbf{x}, \mathbf{z} \in [0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z} - \mathbf{x}|) \leq 1)$$

$$\leq \frac{M_1^2}{\alpha^2} \frac{2^k}{n}$$

(according to the first statement of Lemma 2),

where $M_1^2 = 2^{d+1}(\|m\|_\infty^2 + \sigma^2)$. Meanwhile with respect to $B_n(\mathbf{x})$, we have, still by Chebyshev's inequality,

$$\mathbb{P}\big[|B_n(\mathbf{x})| > \alpha\big] \leq \frac{1}{n\alpha^2} \mathbb{E}\left[\frac{K_k^{uf}(\mathbf{0}, |\mathbf{X}_i - \mathbf{x}|)}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)]}\right]^2$$

$$\leq \frac{1}{n\alpha^2} \frac{1}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)]}$$

$$\leq \frac{2^{k+d}}{n\alpha^2}.$$

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

$$\mathbb{P}\big[\mathcal{C}_\alpha(\mathbf{x})\big] \geq 1 - \mathbb{P}\big(|A_n(\mathbf{x})| \geq \alpha\big) - \mathbb{P}\big(|B_n(\mathbf{x})| \geq \alpha\big)$$

$$\geq 1 - \frac{2^k}{n} \frac{M_1^2}{\alpha^2} - \frac{2^{k+d}}{n\alpha^2}$$

$$\geq 1 - \frac{2^k(M_1^2 + 2^d)}{n\alpha^2}.$$

Consequently, according to inequality (19), we obtain

$$\mathbb{E}\Big[|\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(\mathbf{x})}\Big] \leq C_2(\log n)\left(\frac{2^k(M_1^2 + 2^d)}{n\alpha^2}\right)^{1/2}.$$

Then using inequality (18),

$$\mathbb{E}\Big[\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})\Big]^2$$

$$\leq \mathbb{E}\Big[|\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha(\mathbf{x})}\Big] + \mathbb{E}\Big[|\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(\mathbf{x})}\Big]$$

$$\leq 8C_1^2\left(1 - \frac{1}{3d}\right)^{2k} + 8\alpha^2(1 + \|m\|_\infty)^2 + C_2(\log n)\left(\frac{2^k(M_1^2 + 2^d)}{n\alpha^2}\right)^{1/2}.$$

Optimizing the right hand side in α, we get

$$\mathbb{E}\Big[\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})\Big]^2 \leq 8C_1^2\left(1 - \frac{1}{3d}\right)^{2k} + C_3\left(\frac{(\log n)^2 2^k}{n}\right)^{1/3},$$

for some constant $C_3 > 0$. The last expression is minimized for

$$k = C_4 + \frac{1}{\log 2 + \frac{2}{d}} \log\left(\frac{n}{(\log n)^2}\right),$$

28

where $C_4 = -3\left(\log 2 + \frac{2}{d}\right)^{-1} \log\left(\frac{dC_3 \log 2}{16C_1^2}\right)$. Thus, there exists a constant $C_5 > 0$ such that, for all $n > 1$,

$$\mathbb{E}\left[\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})\right]^2 \leq Cn^{-2/(6+3d\log 2)}(\log n)^2.$$

$\square$

Lemma 2. *For all $k \in \mathbb{N}$ and $x \in [0,1]$,*

(i)

$$\left(\frac{1}{2}\right)^{k_l+1} \leq \int_0^1 K_{k_l}^{uf}(0, |z_l - x_l|)dz \leq \left(\frac{1}{2}\right)^{k_l-1}.$$

(ii)

$$\int_{[0,1]} K_{k_l}^{uf}(0, |z_l - x_l|)|x_l - z_l|dz_l \leq \left(\frac{2}{3}\right)^{k_l+1} \int_{[0,1]} K_{k_l}^{uf}(0, |z_l - x_l|)dz_l.$$

Proof of Lemma 2. Let $k_l \in \mathbb{N}$ and $x_l \in [0,1]$. We start by proving (i). According to Proposition 6, the connection function of uniform random forests of level k_l takes the form

$$
\begin{aligned}
\int_{[0,1]} K_{k_l}^{uf}(0, |z_l - x_l|)dz_l &= \int_{-\log x_l}^{\infty} e^{-2u} \sum_{j=k_l}^{\infty} \frac{u^j}{j!} du + \int_{-\log(1-x_l)}^{\infty} e^{-2u} \sum_{j=k_l}^{\infty} \frac{u^j}{j!} du \\
&= \sum_{j=k_l}^{\infty} \left(\frac{1}{2}\right)^{j+1} \int_{-2\log x_l}^{\infty} e^{-u} \frac{u^j}{j!} du \\
&+ \sum_{j=k_l}^{\infty} \left(\frac{1}{2}\right)^{j+1} \int_{-2\log(1-x_l)}^{\infty} e^{-u} \frac{u^j}{j!} du \\
&= \sum_{j=k_l}^{\infty} \left(\frac{1}{2}\right)^{j+1} x_l^2 \sum_{i=0}^{j} \frac{(-2\log x_l)^i}{i!} \\
&+ \sum_{j=k_l}^{\infty} \left(\frac{1}{2}\right)^{j+1} (1-x_l)^2 \sum_{i=0}^{j} \frac{(-2\log(1-x_l))^i}{i!}.
\end{aligned}
$$

Therefore,

$$\int_{[0,1]} K_{k_l}^{uf}(0, |z_l - x_l|)dz_l \leq \left(\frac{1}{2}\right)^{k_l-1},$$

and

$$\int_{[0,1]} K_{k_l}^{uf}(0, |z_l - x_l|)dz_l \geq \left(x_l^2 + (1-x_l)^2\right) \left(\frac{1}{2}\right)^{k_l} \geq \left(\frac{1}{2}\right)^{k_l+1}.$$

Regarding the second statement of Lemma 2, we have

$$\int_{[0,1]} K_{k_l}^{uf}(0,|z_l - x_l|)|x_l - z_l|\mathrm{d}z_l$$

$$= \int_{[0,1]} |x_l - z_l|^2 \sum_{j=k_l}^{\infty} \frac{(-\log|x_l - z_l|)^j}{j!}\mathrm{d}z_l$$

$$= \int_{z_l \leq x_l} (x_l - z_l)^2 \sum_{j=k_l}^{\infty} \frac{(-\log|x_l - z_l|)^j}{j!}\mathrm{d}z_l$$

$$\quad + \int_{z_l > x_l} (z_l - x_l)^2 \sum_{j=k_l}^{\infty} \frac{(-\log|x_l - z_l|)^j}{j!}\mathrm{d}z_l$$

$$= \int_{[0,x_l]} v^2 \sum_{j=k_l}^{\infty} \frac{(-\log v)^j}{j!}dv + \int_{[0,1-x_l]} u^2 \sum_{j=k_l}^{\infty} \frac{(-\log u)^j}{j!}du$$

$$= \int_{-\log(x_l)}^{\infty} e^{-3w} \sum_{j=k_l}^{\infty} \frac{w^j}{j!}\mathrm{d}w + \int_{-\log(1-x_l)}^{\infty} e^{-3w} \sum_{j=k_l}^{\infty} \frac{w^j}{j!}\mathrm{d}w$$

$$= \frac{2}{3}\int_{-3\log(x_l)/2}^{\infty} e^{-2w} \sum_{j=k_l}^{\infty} \frac{(2v/3)^j}{j!}dv + \frac{2}{3}\int_{-3\log(1-x_l)/2}^{\infty} e^{-2v} \sum_{j=k_l}^{\infty} \frac{(2v/3)^j}{j!}dv$$

$$\leq \left(\frac{2}{3}\right)^{k_l+1} \left(\int_{-\log(x_l)}^{\infty} e^{-2w} \sum_{j=k_l}^{\infty} \frac{v^j}{j!}dv + \int_{-\log(1-x_l)}^{\infty} e^{-2v} \sum_{j=k_l}^{\infty} \frac{v^j}{j!}dv\right)$$

$$\leq \left(\frac{2}{3}\right)^{k_l+1} \int_{[0,1]} K_{k_l}^{uf}(0,|z_l - x_l|)\mathrm{d}z_l.$$

$\square$

7 Acknowledgments

References

S. Arlot and R. Genuer. Analysis of purely random forests bias. arXiv:1407.3939, 2014.

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

G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101:2499–2518, 2010.

G. Biau and L. Devroye. Cellular tree classifiers. *Electronic Journal of Statistics*, 7:1875–1912, 2013.

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

L. Breiman. *Some infinity theory for predictor ensembles*. Technical Report 577, UC Berkeley, 2000.

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

L. Breiman. *Consistency for a simple model of random forests*. Technical Report 670, UC Berkeley, 2004.

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

A. Cutler and G. Zhao. Pert - perfect random tree ensembles. *Computing Science and Statistics*, 33:490–497, 2001.

A. Davies and Z. Ghahramani. The random forest kernel and other kernels for big data from random partitions. arXiv:1402.4293, 2014.

M. Denil, D. Matheson, and N. de Freitas. Consistency of online random forests. In *Proceedings of the ICML Conference*, 2013. arXiv:1302.4853.

R. Genuer, J.-M. Poggi, and C. Tuleau. Random forests: some methodological insights. arXiv:0811.3619, 2008.

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

W. Greblicki, A. Krzyzak, and M. Pawlak. Distribution-free pointwise consistency of kernel regression estimate. *The Annals of Statistics*, pages 1570–1575, 1984.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

H. Ishwaran and U.B. Kogalur. Consistency of random survival forests. *Statistics & Probability Letters*, 80:1056–1064, 2010.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R news*, 2:18–22, 2002.

Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101:578–590, 2006.

L. Meier, S. Van de Geer, and P. Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37:3779–3821, 2009.

L. Mentch and G. Hooker. Ensemble trees and clts: Statistical inference for supervised learning. arXiv:1404.6473, 2014.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9:141–142, 1964.

Y. Qi. *Ensemble Machine Learning*, chapter Random forest for bioinformatics, pages 307–323. Springer, 2012.

G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. H. Torr. Randomized trees for human pose detection. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

E. Scornet. On the asymptotics of random forests. arXiv:1409.2090, 2014.

E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests. arXiv:1405.2881, 2014.

M. van der Laan, E.C. Polley, and A.E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6, 2007.

S. Wager. Asymptotic theory for random forests. arXiv:1405.0352, 2014.

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

R. Zhu, D. Zeng, and M.R. Kosorok. *Reinforcement learning trees*. Technical Report, University of North Carolina, 2012.