Research article

Improved convergence rates for some kernel random forest algorithms

Iakovidis Isidoros* and Nicola Arcozzi

Dipartimento di Matematica, Università di Bologna, Bologna 40126, Italy

* **Correspondence:** Email: isidoros.iakovidis2@unibo.it.

Abstract: Random forests are notable learning algorithms introduced by Breiman in 2001. They are widely used for classification and regression tasks and their mathematical properties are under ongoing research. We consider a specific class of random forest algorithms related to kernel methods, the so-called Kernel Random Forests (KeRF). In particular, we investigate thoroughly two explicit algorithms, designed independently of the data set, the centered KeRF and the uniform KeRF. In the present article, we provide an improvement in the rate of convergence for both algorithms and we explore the related reproducing kernel Hilbert space defined by the explicit kernel of the centered random forest.

Keywords: nonparametric statistics; kernel random forests; reproducing kernel spaces

1. Introduction

Random forests are a class of non-parametric statistic machine learning algorithms used for regression and classification tasks. Random forest algorithms have the capability to perform sparse tasks with high accuracy in high dimensions, avoiding overfitting. In particular, random forests are considered to be among the most accurate learning algorithm classes for general tasks. They are routinely used in many fields including bio-informatics [15], economics [30], biology [10], and linguistics [18].

The most widely used random forest algorithm was introduced by Breiman [13], who was inspired by the work on random subspaces of Ho [19], and geometrical feature selection of Amit and Geman [2]. In Breiman's random forest, the trees are grown based on the classification and regression trees (CART) procedure, where both splitting directions and training sets are randomized. However, despite the few parameters that need to be tuned [16, 21], their mathematical properties are still areas of active research [6, 22]. A significant distinction among the class of random forest algorithms consists in the way each individual tree is constructed, and, in particular, the dependence of each tree on the data set. Some of the researchers consider random forests designed independently from the data set [7, 14, 25].

In 2012, Biau [5] studied a random forest model proposed by Breiman, where the construction is

independent of the data set, called in literature *centered random forest*. In [5] an upper bound on the rate of consistency of the algorithm and its adaption to sparsity were proven. More precisely, about the first item, for a data set of n samples in a space of dimension d, the convergence rate was $O\left(n^{-\frac{1}{d\frac{4}{3}\log 2+1}}\right)$.

In 2021, Klusowski [20] improved the rate of convergence to $O\left((n\log^{\frac{d-1}{2}}n)^{-\left(\frac{1+\delta}{d\log 2+1}\right)}\right)$, where δ is a positive constant that depends on the dimension of the feature space d and converges to zero as d approaches infinity. In addition, in the same paper, Klusowski proved that the rate of convergence of the algorithm is sharp, although it fails to reach the minimax rate of consistency over the class of the Lipschitz functions $O\left(n^{\frac{-2}{d+2}}\right)$ [29]. There is also important work on the consistency of algorithms that depend on data [23, 27, 28]. For a comprehensive overview of both theoretical and practical aspects of the random forests, see [8], which surveys the subject up to 2016.

An important tool for algorithmically manipulating random forests is through kernel methods. Breiman [11] observed the connection between kernel theory and random forests by showing the equivalence between tree construction and kernel action. Later this was formalized by Geurts et al. in [17]. In the same direction Scornet in [26] defined Kernel Random Forest (KeRF) by modifying the original algorithm, and providing theoretical and practical results. In particular, in his important work, Scornet provided explicit kernels for some generalizations of algorithms, their rate of consistency, and comparisons with the corresponding random forests. Furthermore, Arnould et al. [3] investigated the trade-off between interpolation of several random forest algorithms and their consistency results.

In the first part of the paper, we provide the notation and the definitions of the centered and uniform random forests and their corresponding kernel-related formulations. In addition, we improve the rate of consistency for the centered KeRF algorithm. Let $k \geq 1$ be the depth of the trees used to estimate the target variable Y (see Section 2 for definitions and notation).

Theorem 1. *Suppose that $X = (X_1, \ldots, X_d)$ and Y are related by $Y = m(X) + \epsilon$ where: ϵ is a zero mean Gaussian noise with finite variance independent of X, X is uniformly distributed in $[0, 1]^d$, and m is a regression function, which we assume to be Lipschitz. Then, there exists a constant $\tilde{C}$ such that, for every $n > 1$ and for every $x \in [0, 1]^d$,*

$$\mathbb{E}(\tilde{m}_{\infty,n}^{Cen}(x) - m(x))^2 \leq \tilde{C}n^{-\left(\frac{1}{1+d\log 2}\right)}(\log n).$$

Here, $m(x) = \mathbb{E}[Y|X = x]$ is the predicted value of Y for $X = x \in [0, 1]^d$, while $\tilde{m}_{\infty,n}^{Cen}(x)$ is the estimate for m provided by the kernel version of the centered random forest algorithm.

Similarly, with $\tilde{m}_{\infty,n}^{Un}(x)$ playing for the uniform KeRF algorithm the role $\tilde{m}_{\infty,n}^{Cen}(x)$ had above, we have:

Theorem 2. *Let X, Y, m, and ϵ be as in Theorem 1, with $Y = m(X) + \epsilon$. Then there exists a constant $\tilde{C}$ such that for every $n > 1$, for every $x \in [0, 1]^d$*

$$\mathbb{E}(\tilde{m}_{\infty,n}^{Un}(x) - m(x))^2 \leq \tilde{C}n^{-\left(\frac{1}{1+\frac{3}{2}d\log 2}\right)}(\log n).$$

Moreover, in Section 4, we provide numerical examples and experiments concerning the tuning parameter k, which is the tree depth of the two kernel-based random forest algorithms, by comparing the L_2-error for different values and under specific assumptions on the data set.

In the final part of the article, we consider the reproducing kernel K used in the centered KeRF algorithm *per se*. It is rewarding looking at it as defined on the finite Abelian group $\mathbb{Z}_2^{kd}$, where, as above, d is the dimension of the vector X and k is the depth of the tree. By using elementary Fourier analysis on groups, we obtain several equivalent expressions for K and its group transform, we characterize the functions belonging to the corresponding Reproducing Kernel Hilbert Space (RKHS) H_K, we derive results on multipliers, and we obtain bounds for the dimension of H_K, which is much smaller than what one might expect.

2. Notation

A usual problem in machine learning is, based on n observations of a random vector $(X, Y) \in \mathcal{X} \times \mathbb{R} \subseteq \mathbb{R}^d \times \mathbb{R}$, to estimate the function $m(x) = \mathbb{E}(Y|X = x)$. In classification problems, Y ranges over a finite set. In particular we assume that we are given a training sample $\mathcal{D}_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ of independent random variables, where $X_i \in [0, 1]^d$ for every $i = 1, ..., n$ and $Y \in \mathbb{R}$ with a shared joint distribution $\mathbb{P}_{X,Y}$. The goal is using the data set to construct an estimate $m_n : \mathcal{X} \subseteq [0, 1]^d \to \mathbb{R}$ of the function m. Our convergence rate requires an a priori assumption on the regularity of the function m. Following [26], we suppose that m belongs to the class of L Lipschitz functions,

$$|m(x) - m(x')| \le L \cdot \|x - x'\|.$$

Here, as is [26], we consider on $\mathbb{R}^d$ the distance

$$\|x - x'\| = \sum_{j=1}^{d} |x_j - x'_j|.$$

2.1. The random forest algorithm

Next, we provide the general random forest framework by defining firstly the notion of a random tree. Additionally, we present two specific variations of the original random forest algorithm, namely, the centered and uniform random forest algorithms.

Let's assume Θ_i for $i = 1, ..., M$ is a collection of independent random variables, distributed as Θ. The random variables Θ_i correspond to sample the training set or select the positions for splitting. The detailed construction in the case of the centered random forest is performed in Appendix.

Definition 1. *For the j-th tree in the forest, the predicted value x will be denoted by*

$$m_{n,\Theta_j,\mathcal{D}_n}(x) = \sum_{i=1}^{n} \frac{\mathbb{1}_{X_i \in A_{n,\Theta_j,\mathcal{D}_n}(x)} Y_i}{N_{n,\Theta_j,\mathcal{D}_n}(x)}.$$

Where $A_{n,\Theta_j,\mathcal{D}_n}(x)$ is the cell containing x and $N_{n,\Theta_j,\mathcal{D}_n}(x)$ is the number of points that fall into the cell that x belongs to.

For a fixed value of $x \in [0, 1]^d$, the value of the tree is the empirical expectation of Y in the unique cell containing x; which is, this is the hope, a good guess for the target value corresponding to x.

A *random forest* is a finite collection (average) of independent, finite random trees $\{\Theta_1, \ldots, \Theta_M\}$.

Definition 2. *The finite M forest is*

$$m_{M,n}(x) = \frac{1}{M} \sum_{j=1}^{M} m_{n,\Theta_j,\mathcal{D}_n}(x).$$

From a modeling point of view, we let $M \to \infty$ and consider the infinite forest estimate

$$m_{\infty,n,\mathcal{D}_n}(x) = \mathbb{E}_\Theta(m_{n,\Theta,\mathcal{D}_n}(x)).$$

The convergence holds almost surely by the law of the large numbers conditionally on $\mathcal{D}_n$ [12] and [25, Theorem 3.1].

2.2. Kernel random forest algorithm

In 2016, Scornet [26] introduced kernel methods in the random forest world (KeRF), producing a kernel-based algorithm, together with estimates on how this compares with the old one, described above.

To understand the intuition behind KeRF construction, we reformulate the random forest algorithm. For all $x \in [0,1]^d$,

$$m_{M,n}(x) = \frac{1}{M} \sum_{j=1}^{M} \left(\sum_{i=1}^{n} \frac{\mathbb{1}_{X_i \in A_{n,\Theta_j,\mathcal{D}_n}(x)} Y_i}{N_{n,\Theta_j,\mathcal{D}_n}(x)} \right).$$

Therefore we can define the weights of every observation Y_i as

$$W_{i,j,n}(x) = \frac{\mathbb{1}_{X_i \in A_{n,\Theta_j,\mathcal{D}_n}(x)}}{N_{n,\Theta_j,\mathcal{D}_n}(x)}.$$

Hence it is clear that the value of weights changes significantly concerning the number of points in each cell. A way to overcome this nuisance is by simultaneously considering all tree cells containing x, as the tree is randomly picked in the forest.

For all $x \in [0,1]^d$,

$$\tilde{m}_{M,n,\Theta_1,\Theta_2,\dots,\Theta_M}(x) = \frac{1}{\sum_{j=1}^{M} N_{n,\Theta_j}(x)} \sum_{j=1}^{M} \sum_{i=1}^{n} Y_i \mathbb{1}_{X_i \in A_{n,\Theta_j}(x)}.$$

This way, empty cells do not affect the computation of the prediction function of the algorithm.

It is proven in [26], that this representation has indeed a kernel representation.

Proposition 1. *[26, Proposition 1] For all $x \in [0,1]^d$ almost surely, it holds*

$$\tilde{m}_{M,n,\Theta_1,\Theta_2,\dots,\Theta_M}(x) = \frac{\sum_{i=1}^{n} K_{M,n}(x,X_i) Y_i}{\sum_{i=1}^{n} K_{M,n}(x,X_i)},$$

where

$$K_{M,n}(x,z) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}_{x \in A_{n,\Theta_i,\mathcal{D}_n}(z)}$$

is the proximity function of the M forest.

The *infinite random forest* arises as the number of trees tends to infinity.

Definition 3. *The infinite KeRF is defined as:*

$$\tilde{m}_{\infty,n}(x) = \lim_{M\to\infty} \tilde{m}_{M,n}(x, \Theta_1, \Theta_2, ..., \Theta_M).$$

The extension of the kernel follows also in the infinite random forest.

Proposition 2. *[26, Proposition 2] Almost surely for all $x, y \in [0.1]^d$*

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

where

$$K_n(x, y) = \mathbb{P}_\Theta(x \in A_n(y, \Theta)),$$

where the left-hand side is the probability that x and y belong to the same cell in the infinite forest.

2.3. The centered random forest vs centered KeRF, and the uniform random forest vs uniform KeRF

In this paper, we say that an estimator function m_n of m is consistent if the following L_2−type of convergence holds,

$$\mathbb{E}(m_n(x) - m(x))^2 \to 0,$$

as $n \to \infty$.

In the centered and uniform forest algorithms, the way the data set $\mathcal{D}_n$ is partitioned is independent of the data set itself.

2.3.1. The centered random forest/centered KeRF

The centered forest is designed as follows:

1) Fix $k \in \mathbb{N}$.
2) At each node of each individual tree choose a coordinate uniformly from $\{1, 2, ..d\}$.
3) Split the node at the midpoint of the interval of the selected coordinate.

Repeat steps 2) and 3) k times. At the end, we have 2^k leaves, or cells. A toy example of this iterative process for $k = 1, 2$ is in Figures 1 and 2. Our estimation at a point x is achieved by averaging the Y_i corresponding to the X_i in the cell containing x.

Scornet in [26] introduced the corresponding kernel-based centered random forest providing explicitly the proximity kernel function.
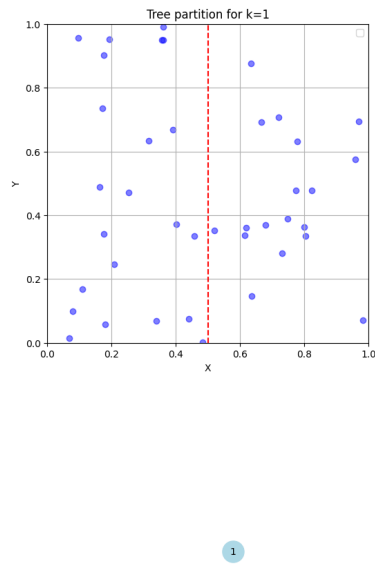
Figure 1. Centered algorithm with tree level $k = 1$ with the convention that 1 corresponds to x axis and 2 to the y axis.
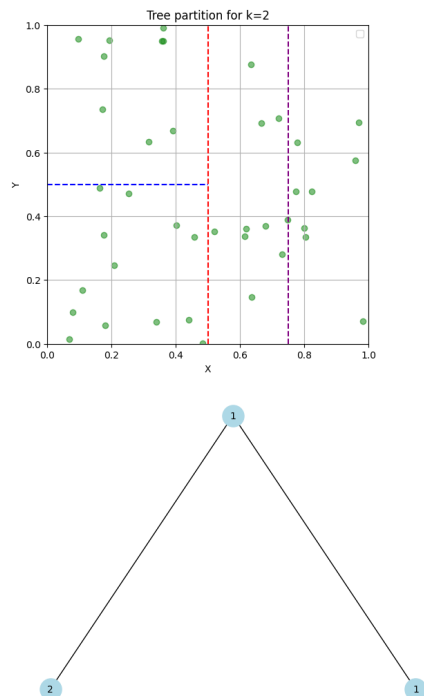
Figure 2. Centered algorithm with tree level $k = 1$ with the convention that 1 corresponds to x axis and 2 to the y axis.

Proposition 3. *A centered random forest kernel with $k \in \mathbb{N}$ parameter has the following multinomial expression [26, Proposition 5],*

$$K_k^{Cen}(x, z) = \sum_{\sum_{j=1}^d k_j = k} \frac{k!}{k_1! \ldots k_d!} \left(\frac{1}{d}\right)^k \prod_{j=1}^d \mathbb{1}_{\left[2^{k_j} x_j\right] = \left[2^{k_j} z_j\right]},$$

where K_k^{Cen} is the kernel of the corresponding centered random forest.

2.3.2. The uniform random forest/kernel random forest

Uniform random forest was introduced by Biau et al. [7] and is another toy model of Breinman's random forest as a centered random forest. The algorithm forms a partition in $[0, 1]^d$ as follows:

1) Fix $k \in \mathbb{N}$.
2) At each node of each individual tree choose a coordinate uniformly from $\{1, 2, ..d\}$.
3) The splitting is performed uniformly on the side of the cell of the selected coordinate.

Repeat steps 2) and 3) k times. At the end, we have 2^k leaves. Our final estimation at a point x is achieved by averaging the Y_i corresponding to the X_i in the cell x.

Again Scornet in [26, Proposition 6] proved the corresponding kernel-based uniform random forest.

Proposition 4. *The corresponding proximity kernel for the uniform KeRF for parameter $k \in \mathbb{N}$ and $x \in [0, 1]^d$ has the following form:*

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

with the convention that

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

and by continuity we can extend the kernel also for zero components of the vector.

Unfortunately, it is very hard to obtain a general formula for $K^{Un}(x, y)$ but we consider instead a translation invariant KeRF uniform forest:

$$m_{\infty,n}^{Un}(x) = \frac{\sum_{i=1}^n Y_i K_k^{Un}(0, |X_i - x|)}{\sum_{i=1}^n K_k^{Un}(0, |X_i - x|)}.$$

3. Proofs of the main theorems

In this section, after providing some measure concentration type results [9], we improve the rate of consistency of the centered KeRF algorithm. The following lemmata will provide inequalities to derive upper bounds for averages of iid random variables. Lacking a reference, for completeness, we provide detailed proofs of these lemmata. Moreover, we assume for this article that all random variables are real-valued and $\|X\|_{L_p} := (\mathbb{E}|X|^p)^{\frac{1}{p}}$ and $\|X\|_\infty := \inf\{t : P(|X| \le t) = 1\}$.

Lemma 1. *Let $X_1, ..., X_n$ be a sequence of real independent and identically distributed random variables with $\mathbb{E}(X_i) = 0$. Assuming also that there is a uniform bound for the L_1-norm and the supremum norm i.e., $\mathbb{E}(|X_i|) \leq C$, $\|X_i\|_\infty \leq CM$ for every $i = 1, ..., n$. Then for every $t \in (0, 1)$*

$$\mathbb{P}(\{\frac{|\sum_{i=1}^n X_i|}{n} \geq t\}) \leq 2e^{-\tilde{C}_C \frac{t^2 n}{M}}$$

for some positive constant $\tilde{C}_C$ that depends only on C.

Proof. $\forall x \in [0, 1]$ one has that $e^x \leq 1 + x + x^2$. By using the hypothesis for every $\lambda \leq \frac{1}{CM}$,

$$e^{\lambda X_i} \leq 1 + \lambda X_i + (\lambda X_i)^2 \quad \Rightarrow$$
$$\mathbb{E}e^{\lambda X_i} \leq 1 + \lambda^2 \mathbb{E}(X_i)^2$$
$$\leq 1 + \lambda^2 \|X_i\|_1 \|X_i\|_\infty$$
$$\leq 1 + \lambda^2 C^2 M$$
$$\leq e^{\lambda^2 C^2 M}.$$

By the independence of the random variables X_i,

$$\mathbb{E}e^{\sum_{i=1}^n \lambda X_i} = \prod_{i=1}^n \mathbb{E}e^{\lambda X_i}$$
$$\leq \prod_{i=1}^n e^{\lambda^2 C^2 M} = e^{n\lambda^2 C^2 M}.$$

Therefore, by Markov inequality

$$\mathbb{P}(\{\frac{\sum_{i=1}^n X_i}{n} \geq t\}) \leq e^{-\lambda t n} \mathbb{E}e^{\sum_{i=1}^n \lambda X_i} \leq e^{-\lambda t n} e^{n\lambda^2 C^2 M} = e^{n\lambda^2 C^2 M - \lambda t n}.$$

Finally if $C \geq \frac{1}{4}$ we choose, $\lambda = \frac{t}{2C^2 M}$, otherwise for $\lambda = \frac{t}{16CM}$

$$\mathbb{P}(\{\frac{\sum_{i=1}^n X_i}{n} \geq t\}) \leq e^{-\tilde{C}_C \frac{t^2 n}{M}}.$$

By replacing X_i with $-X_i$ we conclude the proof. □

Lemma 2. *Let $X_1, ..., X_n$ be a non-negative sequence of independent and identically distributed random variables with $\mathbb{E}(X_i) \leq 2$, $\|X_i\|_\infty \leq M$ for every $i = 1, ..., n$. Let also a sequence of independent random variables ϵ_i following normal distribution with zero mean and finite variance σ^2, for every $i = 1, ..., n$. We assume also that ϵ_i are independent from X_i for every $i = 1, ..., n$.*

Then for every $t \in (0, 1)$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n |\epsilon_i X_i| \geq t\right) \leq 2\exp\left(-Ct^2 \frac{n}{M}\right)$$

with the positive constant C depending only on σ.

Proof.

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i X_i \geq t\Big) = \mathbb{P}\Big(\exp\Big(\frac{\lambda}{n}\sum_{i=1}^{n}\epsilon_i X_i \geq \exp(\lambda t)\Big) \quad \text{(for a positive } \lambda)$$

$$\leq \exp(-\lambda t)\mathbb{E}\exp\Big(\frac{\lambda}{n}\sum_{i=1}^{n}\epsilon_i X_i\Big) \quad \text{(by Chebyshev's inequality)}$$

$$= \exp(-\lambda t)\prod_{i=1}^{n}\mathbb{E}\exp\Big(\frac{\lambda}{n}\epsilon_i X_i\Big) \quad \text{(by independence)}$$

$$= \exp(-\lambda t)\prod_{i=1}^{n}\Big(1 + \sum_{k=2}^{\infty}\frac{\lambda^k \mathbb{E}X_i^k \mathbb{E}\epsilon_i^k}{n^k k!}\Big)$$

$$\leq \exp(-\lambda t)\prod_{i=1}^{n}\Big(1 + \frac{2}{M}\sum_{k=2}^{\infty}\frac{\lambda^k M^k \mathbb{E}\epsilon_i^k}{n^k k!}\Big)$$

$$= \exp(-\lambda t)\prod_{i=1}^{n}\Big(1 + \frac{2}{M}\Big(\mathbb{E}\exp\Big(\frac{\lambda M}{n}\epsilon_i\Big) - 1\Big)\Big)$$

$$\leq \exp(-\lambda t)\prod_{i=1}^{n}\Big(1 + \frac{2}{M}\Big(\exp\Big(\frac{\lambda^2\sigma^2 M^2}{n^2}\Big) - 1\Big)\Big)$$

$$= \exp(-\lambda t)\exp\Big(\sum_{i=1}^{n}\Big(\log\Big(1 + \frac{2}{M}\Big(\exp\Big(\frac{\lambda^2\sigma^2 M^2}{n^2}\Big) - 1\Big)\Big)\Big)\Big)$$

$$\leq \exp(-\lambda t)\exp\Big(\sum_{i=1}^{n}\frac{2}{M}\Big(\exp\Big(\frac{\lambda^2\sigma^2 M^2}{n^2}\Big) - 1\Big)\Big)$$

$$\leq \exp(-\lambda t)\exp\Big(\frac{2n}{M}\Big(\exp\Big(\frac{\lambda^2\sigma^2 M^2}{n^2}\Big) - 1\Big)\Big)$$

$$\leq \exp(-\lambda t)\exp\Big(\frac{2n}{M}\Big(2\frac{\lambda^2\sigma^2 M^2}{n^2}\Big)\Big) \quad \text{(for } \lambda \leq \frac{n}{\sigma M})$$

$$= \exp\Big(-\lambda t + \frac{4M}{n}\lambda^2\sigma^2\Big).$$

Finally, we select $\lambda = \frac{tn}{8M\sigma^2}$, when $\sigma \geq \frac{1}{8}$ and $\lambda = \frac{tn}{M\sigma}$, when $\sigma \leq \frac{1}{8}$

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i X_i \geq t\Big) \leq \exp\Big(-C\frac{t^2 n}{M}\Big).$$

Replacing X_i with $-X_i$ we conclude the proof. □

Theorem 3. *$Y = m(X) + \epsilon$ where ϵ is a zero mean Gaussian noise with finite variance independent of X. Assuming also that X is uniformly distributed in $[0,1]^d$ and m is a Lipschitz function. Then there exists exists a constant $\tilde{C}$ such that for every $n > 1$, for every $x \in [0,1]^d$*

$$\mathbb{E}(\tilde{m}_{\infty,n}^{Cen}(x) - m(x))^2 \leq \tilde{C}n^{-\left(\frac{1}{1+d\log 2}\right)}(\log n).$$

Proof. Following the notation in [26], let $x \in [0, 1]^d$, $\|m\|_\infty = \sup_{x \in [0,1]^d} |m(x)|$, and by the construction of the algorithm

$$\tilde{m}_{n,\infty}^{Cen}(x) = \frac{\sum_{i=1}^n Y_i K_k(x, X_i)}{\sum_{i=1}^n K_k(x, X_i)}.$$

Let

$$A_n(x) = \frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i K_k(x, X_i) - \mathbb{E}(Y K_k(x, X))}{\mathbb{E}(K_k(x, X))} \right),$$

$$B_n(x) = \frac{1}{n} \sum_{i=1}^n \left(\frac{K_k(x, X_i) - \mathbb{E}(K_k(x, X))}{\mathbb{E}(K_k(x, X))} \right),$$

and

$$M_n(x) = \frac{\mathbb{E}(Y K_k(x, X))}{\mathbb{E}(K_k(x, X))}.$$

Hence, we can reformulate the estimator as

$$\tilde{m}_{n,\infty}^{Cen}(x) = \frac{M_n(x) + A_n(x)}{B_n(x) + 1}.$$

Let $t \in (0, \frac{1}{2})$ and the event $C_t(x)$ where $\{A_n(x), B_n(x) \le t\}$.

$$\mathbb{E}(\tilde{m}_{n,\infty}^{cc}(x) - m(x))^2 = \mathbb{E}(\tilde{m}_{n,\infty}^{cc}(x) - m(x))^2 \mathbb{1}_{C_t(x)} + \mathbb{E}(\tilde{m}_{n,\infty}^{cc}(x) - m(x))^2 \mathbb{1}_{C_t^c(x)}$$

$$\le \mathbb{E}(\tilde{m}_{n,\infty}^{cc}(x) - m(x))^2 \mathbb{1}_{C_t^c(x)} + c_1 \left(1 - \frac{1}{2d} \right)^{2k} + c_2 t^2$$

where the last inequality was obtained in [26, p.1496]. Moreover, in [26],

$$\mathbb{E}(\tilde{m}_{n,\infty}^{cc}(x) - m(x))^2 \mathbb{1}_{C_t^c(x)} \le c_3 (\log n)(\mathbb{P}(C_t^c(x)))^{\frac{1}{2}}.$$

In order to find the rate of consistency we need a bound for the probability $\mathbb{P}(C_t^c(x))$. Obviously,

$$\mathbb{P}(C_t^c(x)) \le \mathbb{P}(|A_n(x)| > t) + \mathbb{P}(|B_n(x)| > t).$$

We will work separately to obtain an upper bound for both probabilities.

Proposition 5. *Let*

$$\tilde{X}_i = \frac{K_k(x, X_i)}{\mathbb{E}(K_k(x, X))} - 1,$$

a sequence of iid random variables. Then for any $t \in (0, 1)$,

$$\mathbb{P}(\{ \frac{|\sum_{i=1}^n \tilde{X}_i|}{n} \ge t \}) = \mathbb{P}(|B_n(x)| \ge t) \le 2e^{-\tilde{C}_1 \frac{t^2 n}{2^k}}$$

for some positive constant $\tilde{C}_1$.

Proof. It is easy to verify that $\mathbb{E}\tilde{X}_i = 0$, and

$$|\tilde{X}_i| = |\frac{K_k(x, X_i)}{\mathbb{E}(K_k(x, X))} - 1| \le \frac{K_k(x, X_i)}{\mathbb{E}(K_k(x, X))} + 1,$$

hence, $\mathbb{E}|\tilde{X}_i| \le 2$.

Finally,

$$\|\tilde{X}_i\|_\infty = \sup\{|\tilde{X}_i|\} = \sup\{|\frac{K_k(x, X_i)}{\mathbb{E}(K_k(x, X))} - 1|\} \le \frac{1}{\mathbb{E}(K_k(x, X))} \sup K_k(x, X_i) + 1 \le 2^k + 1 \le 2^{k+1}.$$

By Lemma 1,

$$\mathbb{P}(\{\frac{|\sum_{i=1}^n \tilde{X}_i|}{n} \ge t\}) = \mathbb{P}(|B_n(x)| \ge t) \le 2e^{-\tilde{C}_1 \frac{t^2 n}{2^k}}.$$

□

We need a bound for $\mathbb{P}(|A_n(x)| > t)$ where,

$$A_n(x) = \frac{1}{n} \sum_{i=1}^n (\frac{Y_i K_k(x, X_i) - \mathbb{E}(Y K_k(x, X))}{\mathbb{E}(K_k(x, X))}).$$

Proposition 6. *Let*

$$\tilde{Z}_i = \frac{Y_i K_k(x, X_i) - \mathbb{E}(Y K_k(x, X))}{\mathbb{E}(K_k(x, X))}$$

for $i = 1, ..., n$, then for every $t \in (0, 1)$,

$$\mathbb{P}(\{\frac{|\sum_{i=1}^n \tilde{Z}_i|}{n} \ge t\}) = \mathbb{P}(|A_n(x)| \ge t) \le 4e^{-C \frac{t^2 n}{2^k}},$$

for some constant C depending only on $\sigma, \|m\|_\infty$.

Proof.

$$A_n(x) = \frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i K_k(x, X_i) - \mathbb{E}(Y K_k(x, X))}{\mathbb{E}(K_k(x, X))}\right)$$

$$= \frac{1}{n} \sum_{i=1}^n \left(\frac{m(X_i)K_k(x, X_i) - \mathbb{E}(m(X)K_k(x, X))}{\mathbb{E}(K_k(x, X))}\right) + \frac{1}{n} \sum_{i=1}^n \left(\frac{\epsilon_i K_k(x, X_i) - \mathbb{E}(\epsilon K_k(x, X))}{\mathbb{E}(K_k(x, X))}\right)$$

$$= \frac{1}{n} \sum_{i=1}^n \left(\frac{m(X_i)K_k(x, X_i) - \mathbb{E}(m(X)K_k(x, X))}{\mathbb{E}(K_k(x, X))}\right) + \frac{1}{n} \sum_{i=1}^n \left(\frac{\epsilon_i K_k(x, X_i)}{\mathbb{E}(K_k(x, X))}\right).$$

Therefore,

$$\mathbb{P}(|A_n(x)| \ge t) \le \mathbb{P}\left(\left|\frac{2}{n} \sum_{i=1}^n \frac{m(X_i)K_k(x, X_i) - \mathbb{E}(m(X)K_k(x, X))}{\mathbb{E}(K_k(x, X))}\right| \ge t\right)$$

$$+ \mathbb{P}\left(\left|\frac{2}{n} \sum_{i=1}^n \frac{\epsilon_i K_k(x, X_i)}{\mathbb{E}(K_k(x, X))}\right| \ge t\right).$$

Let

$$Z_i = \frac{2(m(X_i)K_k(x, X_i) - \mathbb{E}(m(X)K_k(x, X)))}{\mathbb{E}(K_k(x, X))},$$

a sequence of iid random variables. It is easy to verify that $\tilde{Z}_i$ are centered and

$$|\tilde{Z}_i| = |\frac{m(X_i)K_k(x, X_i) - \mathbb{E}(m(X)K_k(x, X))}{\mathbb{E}(K_k(x, X))}| \leq 2\|m\|_\infty \frac{K_k(x, X_i) + \mathbb{E}(K_k(x, X))}{\mathbb{E}(K_k(x, X))}.$$

Hence, $\mathbb{E}|Z_i| \leq 4\|m\|_\infty$. Finally,

$$\begin{aligned}
\|Z_i\|_\infty &= \sup\{|Z_i|\} \\
&= 2\sup\{|\frac{m(X_i)K_k(x, X_i) - \mathbb{E}(m(X)K_k(x, X))}{\mathbb{E}(K_k(x, X))}|\} \\
&\leq 2\|m\|_\infty(2^k + 1) \\
&\leq 4\|m\|_\infty 2^k.
\end{aligned}$$

By Lemma 1,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\frac{m(X_i)K_k(x, X_i) - \mathbb{E}(m(X)K_k(x, X))}{\mathbb{E}(K_k(x, X))}\right| \geq t\right) \leq 2e^{-C\frac{nt^2}{2^k}}.$$

Furthermore, let

$$\tilde{W}_i = \frac{2\epsilon_i K_k(x, X_i)}{\mathbb{E}(K_k(x, X))}$$

for $i = 1, ..., n$, a sequence of independent and identically distributed random variables. We can verify that for every for $i = 1, ..., n$:

$$\mathbb{E}\left(\frac{2K_k(x, X_i)}{\mathbb{E}(K_k(x, X))}\right) \leq 2.$$

Finally,

$$\sup\left\{\left|\frac{2K_k(x, X_i)}{\mathbb{E}(K_k(x, X))}\right|\right\} \leq \frac{2}{\mathbb{E}(K_k(x, X))}\sup\{K_k(x, X_i)\} \leq 2^{k+1}.$$

By Lemma 2 it is clear,

$$\mathbb{P}\left(\left|\frac{2}{n}\sum_{i=1}^{n}\frac{\epsilon_i K_k(x, X_i)}{\mathbb{E}(K_k(x, X))}\right| \geq t\right) \leq 2e^{-C_2\frac{nt^2}{2^k}}.$$

We conclude the proposition by observing

$$\mathbb{P}(|A_n(x)| \geq t) \leq 4e^{-\min\{C_2, C\}\frac{nt^2}{2^k}}.$$

$\square$

Finally, let us compute the rate of consistency of the algorithm-centered KeRF. By Propositions 5 and 6, one has that

$$\left(\mathbb{P}(C_t^c(x))\right)^{\frac{1}{2}} \leq \left(\mathbb{P}(|A_n(x)| > t) + \mathbb{P}(|B_n(x)| > t)\right)^{\frac{1}{2}} \leq c_3 e^{-c_4\frac{nt^2}{2^k}},$$

for some constants c_3, c_4 independent of k and n.

Thus,

$$\mathbb{E}(\tilde{m}_{\infty,n} - m(x))^2 \leq c_1 \left(1 - \frac{1}{2d}\right)^{2k} + c_2 t^2 + c_3 \log n e^{-c_4 t^2 \frac{n}{2^k}}.$$

We compute the minimum of the right-hand side of the inequality for $t \in (0, 1)$,

$$2c_2 t - 2t c_4 \log n c_3 \frac{n}{2^k} e^{-c_4 t^2 \frac{n}{2^k}} = 0 \quad \Rightarrow$$

$$e^{-c_4 t^2 \frac{n}{2^k}} = \frac{c_2}{c_3 c_4} \frac{2^k}{n \log n} \quad \text{and}$$

$$t^2 = \frac{1}{c_4} \frac{2^k}{n} \log \left(\frac{c_2}{c_3 c_4} \frac{n \log n}{2^k}\right).$$

Hence, the inequality becomes

$$\mathbb{E}(\tilde{m}_{\infty,n} - m(x))^2 \leq c_1 \left(1 - \frac{1}{2d}\right)^{2k} + c_2 \frac{1}{c_4} \frac{2^k}{n} \log \left(\frac{c_2}{c_3 c_4} \frac{n \log n}{2^k}\right) + c_3 \log n \frac{c_2}{c_3 c_4} \frac{2^k}{n \log n}$$

$$= c_1 \left(1 - \frac{1}{2d}\right)^{2k} + c_2 \frac{1}{c_4} \frac{2^k}{n} \log \left(\frac{c_2}{c_3 c_4} \frac{n \log n}{2^k} e^{\frac{c_2}{c_4}}\right).$$

For every $\epsilon_n \in (0, 2]$ it holds, $\log x \leq \frac{1}{\epsilon_n} x^{\epsilon_n}$. Then one has that

$$\mathbb{E}(\tilde{m}_{\infty,n} - m(x))^2 \leq c_1 \left(1 - \frac{1}{2d}\right)^{2k} + \frac{c_2 (e^{\frac{c_2}{c_4}} \frac{c_2}{c_3 c_4})^n}{c_4 \epsilon_n} \left(\frac{2^k}{n} (\log n)^{\frac{\epsilon_n}{1-\epsilon_n}}\right)^{1-\epsilon_n}.$$

We pick

$$k = c(d) \log_2 \frac{n}{(\log n)^{\frac{\epsilon_n}{1-\epsilon_n}}},$$

thus,

$$\frac{c_2 (e^{\frac{c_2}{c_4}} \frac{c_2}{c_3 c_4})^n}{c_4 \epsilon_n} \left(\frac{2^k}{n} (\log n)^{\frac{\epsilon_n}{1-\epsilon_n}}\right)^{1-\epsilon_n} \leq \frac{c'}{\epsilon_n} n^{(c(d)-1)(1-\epsilon_n)} \log n^{\epsilon_n (1-c(d))},$$

for a constant c' independent of n and,

$$c_1 \left(1 - \frac{1}{2d}\right)^{2k} = c_1 \left(1 - \frac{1}{2d}\right)^{2(c(d) \log_2 \frac{n}{(\log n)^{\frac{\epsilon_n}{1-\epsilon_n}}})}$$

$$= c_1 2^{2c(d) \log_2 \left(1 - \frac{1}{2d}\right) \log_2 \frac{n}{(\log n)^{\frac{\epsilon_n}{1-\epsilon_n}}}}$$

$$= c_1 n^{2c(d) \log_2 \left(1 - \frac{1}{2d}\right)} \frac{1}{(\log n)^{c(d) \frac{2\epsilon_n}{1-\epsilon_n} \log_2 \left(1 - \frac{1}{2d}\right)}}.$$

Therefore,

$$c(d) = \frac{\epsilon_n - 1}{2 \log_2 \left(1 - \frac{1}{2d}\right) - (1 - \epsilon_n)}.$$

Finally,

$$
c_1 n^{2c(d)\log_2\left(1-\frac{1}{2d}\right)} \frac{1}{(\log n)^{c(d)\frac{2\epsilon_n}{1-\epsilon_n}\log_2\left(1-\frac{1}{2d}\right)}}
$$

$$
= c_1 n^{\frac{2(\epsilon_n-1)}{2\log_2\left(1-\frac{1}{2d}\right)-(1-\epsilon_n)}\log_2\left(1-\frac{1}{2d}\right)} \times \frac{1}{(\log n)^{\frac{2(\epsilon_n-1)}{2\log_2\left(1-\frac{1}{2d}\right)-(1-\epsilon_n)}\frac{2\epsilon_n}{1-\epsilon_n}\log_2\left(1-\frac{1}{2d}\right)}}
$$

$$
= c_1 n^{\frac{2(\epsilon_n-1)}{2\left(\frac{-\frac{1}{2d}}{\log 2}\right)-(1-\epsilon_n)}\left(\frac{-\frac{1}{2d}}{\log 2}\right)} \times \frac{1}{(\log n)^{\frac{2(\epsilon_n-1)}{2\log_2\left(1-\frac{1}{2d}\right)-(1-\epsilon_n)}\frac{2\epsilon_n}{1-\epsilon_n}\log_2\left(1-\frac{1}{2d}\right)}}
$$

$$
= c_1 n^{-\left(\frac{1-\epsilon_n}{1+(1-\epsilon_n)d\log 2}\right)}(\log n)^{\left(\frac{\epsilon_n}{1+d\log 2(1-\epsilon_n)}\right)}
$$

and, for the second term, with the same arguments

$$
\frac{\tilde{c}}{\epsilon_n} n^{(c(d)-1)(1-\epsilon_n)}\log n^{\epsilon_n(1-c(d))} = \frac{\tilde{c}}{\epsilon_n} n^{-\left(\frac{1-\epsilon_n}{1+(1-\epsilon_n)d\log 2}\right)}(\log n)^{\left(\frac{\epsilon_n}{1+d\log 2(1-\epsilon_n)}\right)}
$$

for a constant $\tilde{c}$ independent of ϵ_n, hence,

$$
\mathbb{E}(\tilde{m}_{\infty,n}^{Cen}(x) - m(x))^2 \leq \frac{C}{\epsilon_n} n^{-\left(\frac{1-\epsilon_n}{1+(1-\epsilon_n)d\log 2}\right)}(\log n)^{\left(\frac{\epsilon_n}{1+d\log 2(1-\epsilon_n)}\right)},
$$

and consequently,

$$
\frac{C}{\epsilon_n} n^{-\left(\frac{1-\epsilon_n}{1+(1-\epsilon_n)d\log 2}\right)}(\log n)^{\left(\frac{\epsilon_n}{1+d\log 2(1-\epsilon_n)}\right)}
$$

$$
= \frac{C}{\epsilon_n} n^{-\left(\frac{1}{1+d\log 2}\right)}(\log n)^{\left(\frac{\epsilon_n}{1+d\log 2(1-\epsilon_n)}\right)} \times n^{\left(\frac{\epsilon_n}{(1+d\log 2)(1+(1-\epsilon_n))d\log 2}\right)}
$$

$$
\leq \frac{C}{\epsilon_n} n^{-\left(\frac{1}{1+d\log 2}\right)}(\log n)^{\left(\frac{\epsilon_n}{d\log 2(1-\epsilon_n)}\right)} \times (\log n)^{\frac{\log n}{\log\log n}\left(\frac{\epsilon_n}{(d\log 2)^2(1-\epsilon_n)}\right)}.
$$

Finally we finish the proof by selecting $\epsilon_n = \frac{1}{\log n}$, and

$$
\mathbb{E}(\tilde{m}_{\infty,n}^{Cen}(x) - m(x))^2 \leq \tilde{C} n^{-\left(\frac{1}{1+d\log 2}\right)}(\log n).
$$

$\square$

Theorem 4. *$Y = m(X) + \epsilon$ where ϵ is a zero mean Gaussian noise with finite variance independent of X. Assuming also that X is uniformly distributed in $[0,1]^d$ and m is a Lipschitz function. Providing $k \to \infty$, there exists a constant $\tilde{C}$ such that for every $n > 1$, for every $x \in [0,1]^d$*

$$
\mathbb{E}(\tilde{m}_{\infty,n}^{Un}(x) - m(x))^2 \leq \tilde{C} n^{-\left(\frac{1}{1+\frac{3}{2}d\log 2}\right)}(\log n).
$$

Proof. By arguing with the same reasoning as the proof of the centered random forest we can verify that

$$
\left(\mathbb{P}(C_t^c(x))\right)^{\frac{1}{2}} \leq \left(\mathbb{P}(|A_n(x)| > t) + \mathbb{P}(|B_n(x)| > t)\right)^{\frac{1}{2}} \leq c_3 e^{-c_4\frac{nt^2}{2^k}}
$$

for some constants c_3, c_4 independent of k and n. The rate of consistency for the Uniform KeRF is the minimum of the right hand in the inequality in terms of n

$$\mathbb{E}(\tilde{m}_{\infty,n}^{Un} - m(x))^2 \leq c_1 \left(1 - \frac{1}{3d}\right)^{2k} + c_2 t^2 + c_3 \log n e^{-c_4 t^2 \frac{n}{2^k}}.$$

We compute the minimum of the right-hand side of the inequality for $t \in (0, 1)$,

$$2c_2 t - 2t c_4 \log n c_3 \frac{n}{2^k} e^{-c_4 t^2 \frac{n}{2^k}} = 0 \quad \Rightarrow$$

$$e^{-c_4 t^2 \frac{n}{2^k}} = \frac{c_2}{c_3 c_4} \frac{2^k}{n \log n} \quad \text{and}$$

$$t^2 = \frac{1}{c_4} \frac{2^k}{n} \log \left(\frac{c_2}{c_3 c_4} \frac{n \log n}{2^k}\right).$$

Hence, the inequality becomes

$$\mathbb{E}(\tilde{m}_{\infty,n}^{Un}(x) - m(x))^2 \leq c_1 \left(1 - \frac{1}{3d}\right)^{2k} + c_2 \frac{1}{c_4} \frac{2^k}{n} \log \left(\frac{c_2}{c_3 c_4} \frac{n \log n}{2^k}\right)$$

$$+ c_3 \log n \frac{c_2}{c_3 c_4} \frac{2^k}{n \log n}$$

$$= c_1 \left(1 - \frac{1}{3d}\right)^{2k} + c_2 \frac{1}{c_4} \frac{2^k}{n} \log \left(\frac{c_2}{c_3 c_4} \frac{n \log n}{2^k} e^{\frac{c_2}{c_4}}\right).$$

For every $\epsilon_n \in (0, 2]$ it holds, $\log x \leq \frac{1}{\epsilon_n} x^{\epsilon_n}$. Then one has that,

$$\mathbb{E}(\tilde{m}_{\infty,n}^{Un} - m(x))^2 \leq c_1 \left(1 - \frac{1}{3d}\right)^{2k} + \frac{c_2 (e^{\frac{c_2}{c_4}} \frac{c_2}{c_3 c_4})^n}{c_4 \epsilon_n} \left(\frac{2^k}{n} (\log n)^{\frac{\epsilon_n}{1-\epsilon_n}}\right)^{1-\epsilon_n}.$$

We pick

$$k = c(d) \log_2 \frac{n}{(\log n)^{\frac{\epsilon_n}{1-\epsilon_n}}}.$$

Therefore,

$$\frac{c_2 (e^{\frac{c_2}{c_4}} \frac{c_2}{c_3 c_4})^n}{c_4 \epsilon_n} \left(\frac{2^k}{n} (\log n)^{\frac{\epsilon_n}{1-\epsilon_n}}\right)^{1-\epsilon_n} \leq \frac{c'}{\epsilon_n} n^{(c(d)-1)(1-\epsilon_n)} \log n^{\epsilon_n(1-c(d))},$$

for a constant c' independent of n and,

$$c_1 \left(1 - \frac{1}{3d}\right)^{2k} = c_1 \left(1 - \frac{1}{3d}\right)^{2c(d) \log_2 \frac{n}{(\log n)^{\frac{\epsilon_n}{1-\epsilon_n}}}}$$

$$= c_1 2^{2c(d) \log_2 \left(1 - \frac{1}{3d}\right) \log_2 \frac{n}{(\log n)^{\frac{\epsilon_n}{1-\epsilon_n}}}}$$

$$= c_1 n^{2c(d) \log_2 \left(1 - \frac{1}{3d}\right)} \frac{1}{(\log n)^{c(d) \frac{2\epsilon_n}{1-\epsilon_n} \log_2 \left(1 - \frac{1}{3d}\right)}}.$$

Therefore,

$$c(d) = \frac{\epsilon_n - 1}{2\log_2\left(1 - \frac{1}{3d}\right) - (1 - \epsilon_n)}.$$

Finally,

$$c_1 n^{2c(d)\log_2\left(1-\frac{1}{3d}\right)} \frac{1}{(\log n)^{c(d)\frac{2\epsilon_n}{1-\epsilon_n}\log_2\left(1-\frac{1}{3d}\right)}}$$

$$= c_1 n^{\frac{2(\epsilon_n-1)}{2\log_2\left(1-\frac{1}{3d}\right)-(1-\epsilon_n)}\log_2\left(1-\frac{1}{3d}\right)} \times \frac{1}{(\log n)^{\frac{2(\epsilon_n-1)}{2\log_2\left(1-\frac{1}{3d}\right)-(1-\epsilon_n)}\frac{2\epsilon_n}{1-\epsilon_n}\log_2\left(1-\frac{1}{3d}\right)}}$$

$$= c_1 n^{\frac{2(\epsilon_n-1)}{2\left(\frac{-\frac{1}{3d}}{\log 2}\right)-(1-\epsilon_n)}\left(\frac{-\frac{1}{3d}}{\log 2}\right)} \times \frac{1}{(\log n)^{\frac{2(\epsilon_n-1)}{2\left(\frac{-\frac{1}{3d}}{\log 2}\right)-(1-\epsilon_n)}\frac{2\epsilon_n}{1-\epsilon_n}\left(\frac{-\frac{1}{3d}}{\log 2}\right)}}$$

$$= n^{-\left(\frac{2(1-\epsilon_n)}{1+(1-\epsilon_n)d\log 2}\right)} \frac{1}{(\log n)^{\frac{2\epsilon_n}{-2+3d\log 2(\epsilon_n-1)}}}$$

$$= n^{-\left(\frac{2(1-\epsilon_n)}{2+(1-\epsilon_n)3d\log 2}\right)}(\log n)^{\left(\frac{2\epsilon_n}{2+3d\log 2(1-\epsilon_n)}\right)}$$

and, for the second term, with the same arguments

$$\frac{\tilde{c}}{\epsilon_n} n^{(c(d)-1)(1-\epsilon_n)}\log n^{\epsilon_n(1-c(d))} = \frac{\tilde{c}}{\epsilon_n} n^{-\left(\frac{1-\epsilon_n}{1+(1-\epsilon_n)\frac{3}{2}d\log 2}\right)}(\log n)^{\left(\frac{\epsilon_n}{1+d\frac{3}{2}\log 2(1-\epsilon_n)}\right)},$$

for a constant $\tilde{c}$ independent of ϵ_n hence,

$$\mathbb{E}(\tilde{m}_{\infty,n}^{Un}(x) - m(x))^2 \le \frac{C}{\epsilon_n} n^{-\left(\frac{1-\epsilon_n}{1+(1-\epsilon_n)\frac{3}{2}d\log 2}\right)}(\log n)^{\left(\frac{\epsilon_n}{1+\frac{3}{2}d\log 2(1-\epsilon_n)}\right)},$$

and consequently,

$$\frac{C}{\epsilon_n} n^{-\left(\frac{1-\epsilon_n}{1+(1-\epsilon_n)\frac{3}{2}d\log 2}\right)}(\log n)^{\left(\frac{\epsilon_n}{1+\frac{3}{2}d\log 2(1-\epsilon_n)}\right)}$$

$$= \frac{C}{\epsilon_n} n^{-\left(\frac{1}{1+\frac{3}{2}d\log 2}\right)}(\log n)^{\left(\frac{\epsilon_n}{1+\frac{3}{2}d\log 2(1-\epsilon_n)}\right)} \times n^{\left(\frac{\epsilon_n}{(1+\frac{3}{2}d\log 2)(1+(1-\epsilon_n))d\log 2}\right)}$$

$$\le \frac{C}{\epsilon_n} n^{-\left(\frac{1}{1+\frac{3}{2}d\log 2}\right)}(\log n)^{\left(\frac{\epsilon_n}{\frac{3}{2}d\log 2(1-\epsilon_n)}\right)} \times (\log n)^{\frac{\log n}{\log\log n}\left(\frac{\epsilon_n}{(\frac{3}{2}d\log 2)^2(1-\epsilon_n)}\right)}.$$

Finally we finish the proof by selecting $\epsilon_n = \frac{1}{\log n}$, and

$$\mathbb{E}(\tilde{m}_{\infty,n}^{Un}(x) - m(x))^2 \le \tilde{C} n^{-\left(\frac{1}{1+\frac{3}{2}d\log 2}\right)}(\log n).$$

$\square$

4. Plots and experiments

In the following section, we summarize the rates of convergence for the centered KeRF and the uniform KeRF, and we compare them with the minimax rate of convergence over the class of the Lipschitz functions [29]. In addition, we provide numerical simulations where we compare the L_2−error for different choices of the tree depth. All experiments performed with the software Python (https://www.python.org/), mainly with the numpy library, where random sets uniformly distributed in $[0, 1]^d$ have been created, for various examples for the dimension d and the function Y. For every experiment the set is divided into a training set (80 %) and a testing set (20 %); then the L_2−error ($\sum_{X_i \in \text{test set}} (\tilde{m}(X_i) - Y_i)^2$) and the standard deviation of the error is computed.

For the centered KeRF we compare three different values of tree depth, which from theory provide different convergence rates. First, the choice of k in [26, Theorem 1] that provides the previous convergence rate; second, the selection of k as it was delivered from the Theorem 1; and, third, the case where the estimator, in probability, interpolates the data set, but the known convergence rate is slow [3, Theorem 4.1], $O(\log n^{-\frac{d-5}{6}})$ for the dimension of the feature space $d > 5$.

For the uniform KeRF, we compare the two values of tree depth as they were derived from [26] and Theorem 2 nevertheless, it is not known if the uniform-KeRF algorithm converges when our estimator function interpolates the data set. Of course, in practice, since real data might violate the assumptions of the theorems, one should try cross-validation for tuning the parameters of the algorithms.

Comparing the rates of consistency for centered KeRF and the depth of the corresponding trees:

- Scornet in [26, Theorem 1] rate of convergence:

$$n^{-(\frac{1}{d \log 2 + 3})}(\log n)^2, \quad \text{and } k = \lceil \frac{1}{\log 2 + \frac{3}{d}} \log \frac{n}{\log n^2} \rceil.$$

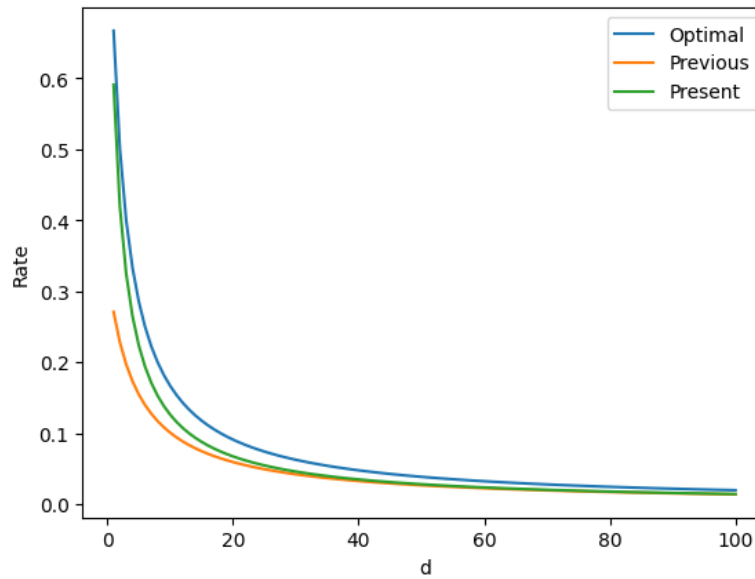- New rate of convergence:

$$n^{-(\frac{1}{1 + d \log 2})}(\log n), \quad \text{and } k = \lceil \frac{\frac{1}{\log n} - 1}{2 \log_2(1 - \frac{1}{2d}) - (1 - \frac{1}{\log n})} \log_2 \frac{n}{(\log n)^{1 - \frac{1}{\log n}}} \rceil.$$

- Minimax [29] rate of consistency over the class of Lipschitz functions: $n^{\frac{-2}{d+2}}$ functions.

Thus, theoretically, the improved rate of consistency is achieved when trees grow at a deeper level compared with the parameter selection in [26, Theorem 1].

As it is evident from Figure 3, the improvement in the convergence rate is more significant in the low dimensional feature space scenarios. The constant $\tilde{C} = \tilde{C}(d)$ of Theorem 1 depends on the dimension d of the space. The convergence rates in the literature do not try to have a good estimate for $\tilde{C}$, and they are significant for fixed values of d only.

Figure 3. Plot of the exponents of n, for the previous rate of convergence for the centered KeRF algorithm, the new rate of convergence, and the optimal over the class of the Lipschitz functions.

Finally, we note that Klusowski's rate of convergence in [20], $O\left((n \log^{\frac{d-1}{2}} n)^{-\left(\frac{1+\delta}{d \log 2+1}\right)}\right)$, where δ is a positive constant that depends on the dimension of the feature space d and converges to zero as d approaches infinity, is sharp and better than the one in Theorem 1 $O\left(n^{-\left(\frac{1}{1+d \log 2}\right)}(\log n)\right)$ for small values of d. For large values of n and d the two estimates are essentially the same, but for now, we do not know if in general, the rate of convergence of the centered KeRF is not improvable.

Comparing the rates of convergence for uniform KeRF and the depth of the corresponding trees:

- Scornet in [26, Theorem 2] rate of convergence:

$$n^{-\left(\frac{2}{3d log 2+6}\right)}(\log n)^2, \text{ and } k = \lceil \frac{1}{\log 2 + \frac{3}{d}} \log \frac{n}{\log n^2} \rceil.$$
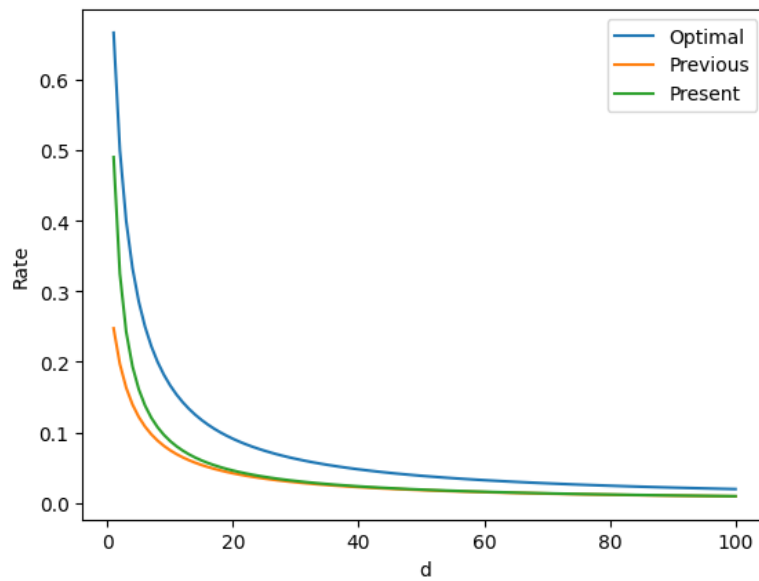
- New rate of convergence:

$$n^{-\left(\frac{2}{3d \log 2+2}\right)}(\log n), \text{ and } k = \lceil \frac{\frac{1}{\log n} - 1}{2 \log_2(1 - \frac{1}{3d}) - (1 - \frac{1}{\log n})} \log_2 \frac{n^{\frac{1}{\log n}}}{(\log n)^{1 - \frac{1}{\log n}}} \rceil.$$

- Minimax [29] rate of convergence for the consistency over the class of Lipschitz functions: $n^{\frac{-2}{d+2}}$ functions.

Thus, theoretically, as in the case of centered random KeRF, the improved rate of consistency is achieved when trees grow at a deeper level compared with the parameter selection in [26, Theorem 2].

The same considerations on the dependence of the constant $\tilde{C}$ on d we made for the centered KeRF hold in the uniform KeRF case, as it is evident in Figure 4. Moreover, as of now, it is still unknown to us if the convergence rate of the uniform KeRF can be improved.

Figure 4. Plot of the exponents of n, for the previous rate of convergence for the uniform KeRF algorithm, the new rate of convergence, and the optimal over the class of the Lipschitz functions.

Finally, numerical simulations of the L_2−error of the centered KeRF-approximation for three different values of k in Figure 5 are reported with the standard deviation of the errors in Figure 6. In Appendix, more simulations and plots for different target functions and for both algorithms are illustrated.
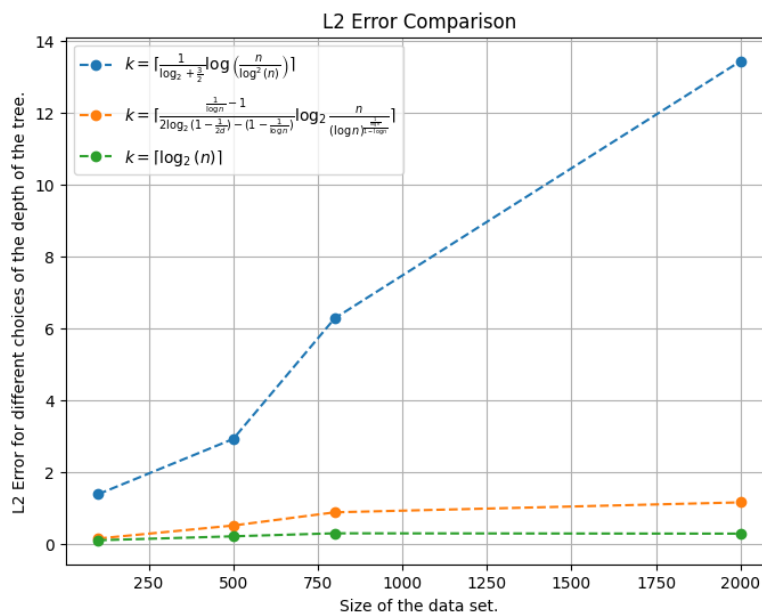
Figure 5. Plot of the L_2−error of the centered KeRF-approximation for three different values of k for the function $Y = X_1^2 + e^{-X_2^2} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \frac{1}{2})$, against different data set size.
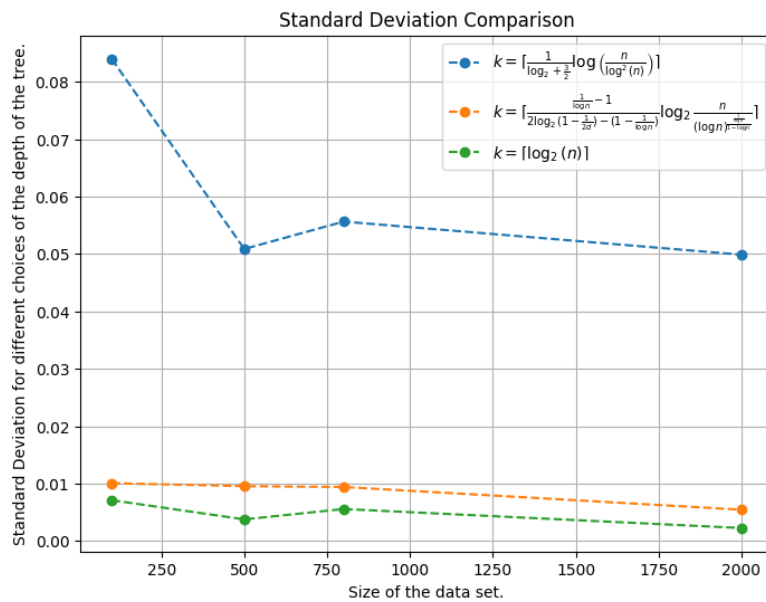
Figure 6. Plot of the standard deviation of errors, for the centered KeRF-approximation for three different values of k of the function $Y = X_1^2 + e^{-X_2^2} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \frac{1}{2})$, against different data set size.

5. Conclusions

In this paper, we have obtained improved rates of convergence for two kernel-based random forests, the centered KeRF and the uniform KeRF. In addition, we provided numerical simulations for both algorithms concerning the parameters of the methods. Finally, we explored the reproducing kernel Hilbert space related to the centered KeRF providing bounds for the dimension of the aforementioned space.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. J. Agler, J. E. McCarthy, *Pick interpolation and Hilbert function spaces*, 2002.

2. Y. Amit, D. Geman, Shape quantization and recognition with randomized trees, *Neural Comput.*, **9** (1997), 1545–1588. https://doi.org/10.1162/neco.1997.9.7.1545

3. L. Arnould, C. Boyer, E. Scornet, Is interpolation benign for random forest regression? *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, **206** (2023), 5493–5548.

4. N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.*, **68** (1950), 337–404.

5. G. Biau, Analysis of a random forests model, *J. Mach. Learn. Res.*, **13** (2012), 1063–1095.

6. G. Biau, L. Devroye, On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification, *J. Multivariate Anal.*, **101** (2010), 2499–2518. https://doi.org/10.1016/j.jmva.2010.06.019

7. G. Biau, L. Devroye, G. Lugosi, Consistency of random forests and other averaging classifiers, *J. Mach. Learn. Res.*, **9** (2008), 2015–2033.

8. G. Biau, E. Scornet, A random forest guided tour, *TEST*, **25** (2016), 197–227. https://doi.org/10.1007/s11749-016-0481-7

9. S. Boucheron, G. Lugosi, P. Massart, *Concentration inequalities: a nonasymptotic theory of independence*, Oxford University Press, 2013.

10. A. Boulesteix, S. Janitza, J. Kruppa, I. R. König, Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics, *Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery*, **2** (2012), 493–507. https://doi.org/10.1002/widm.1072

11. L. Breiman, *Some infinity theory for predictor ensembles*, Technical Report 579, Statistics Dept. UCB, 2000.

12. L. Breiman, Random forests, *Mach. Learn.*, **45** (2001), 5–32. https://doi.org/10.1023/A:1010933404324

13. L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, *Classification and regression trees*, 1 Ed., New York: Chapman and Hall/CRC, 1984. https://doi.org/10.1201/9781315139470

14. M. Denil, D. Matheson, N. Freitas, Consistency of online random forests, *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, Georgia, 2013, 1256–1264.

15. D. J. Dittman, T. M. Khoshgoftaar, A. Napolitano, The effect of data sampling when using random forest on imbalanced bioinformatics data, *2015 IEEE International Conference on Information Reuse and Integration*, IEEE, 2015, 457–463. https://doi.org/10.1109/IRI.2015.76

16. R. Genuer, J. M. Poggi, C. Tuleau, Random forests: some methodological insights, *arXiv*, 2008. https://doi.org/10.48550/arXiv.0811.3619

17. P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, *Mach. Learn.*, **63** (2006), 3–42. https://doi.org/10.1007/s10994-006-6226-1

18. S. T. Gries, On classification trees and random forests in corpus linguistics: some words of caution and suggestions for improvement, *Corpus Linguist. Ling. Theory*, **16** (2019), 617–647. https://doi.org/10.1515/cllt-2018-0078

19. T. K. Ho, Random decision forests, *Proceedings of 3rd International Conference on Document Analysis and Recognition*, IEEE, **1** (1995), 278–282. https://doi.org/10.1109/ICDAR.1995.598994

20. J. M. Klusowski, Sharp analysis of a simple model for random forests, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, **130** (2021), 757–765.

21. A. Liaw, M. Wiener, Classification and regression by randomForest, *R News*, **2** (2002), 18–22.

22. Y. Lin, Y. Jeon, Random forests and adaptive nearest neighbors, *J. Amer. Stat. Assoc.*, **101** (2006), 578–590. https://doi.org/10.1198/016214505000001230

23. L. K. Mentch, G. Hooker, Ensemble trees and CLTs: statistical inference for supervised learning, *arXiv*, 2014. https://doi.org/10.48550/arXiv.1404.6473

24. W. Rudin, *Fourier analysis on groups*, Courier Dover Publications, 2017.

25. E. Scornet, On the asymptotics of random forests, *J. Multivariate Anal.*, **146** (2016), 72–83. https://doi.org/10.1016/j.jmva.2015.06.009

26. E. Scornet, Random forests and kernel methods, *IEEE Trans. Inf. Theory*, **62** (2016), 1485–1500. https://doi.org/10.1109/TIT.2016.2514489

27. E. Scornet, G. Biau, J. P. Vert, Consistency of random forests, *Ann. Statist.*, **43** (2015), 1716–1741. https://doi.org/10.1214/15-AOS1321

28. S. Wager, Asymptotic theory for random forests, *arXiv*, 2014. https://doi.org/10.48550/arXiv.1405.0352

29. Y. Yang, A. Barron, Information-theoretic determination of minimax rates of convergence, *Ann. Statist.*, **27** (1999), 1564–1599.

30. J. Yoon, Forecasting of real GDP growth using machine learning models: gradient boosting and random forest approach, *Comput. Econ.*, **57** (2021), 247–265. https://doi.org/10.1007/s10614-020-10054-w

Appendix

In this last section, we provide more experiments of low dimensional regression examples with additive noise, regarding the centered and the uniform KeRF. In particular, we calculate and compare the L_2−errors and the standard deviations against different sample sizes for different values of the parameter k of the estimator. Moreover, in the following subsection, we study the corresponding RKHS produced by the kernel

$$K_k^{Cen}(x,z) = \sum_{\sum_{j=1}^d k_j = k} \frac{k!}{k_1!...k_d!} \left(\frac{1}{d}\right)^k \prod_{j=1}^d \mathbb{1}_{\left[2^{k_j} x_j\right] = \left[2^{k_j} z_j\right]},$$

defined in the abelian group $\mathbb{Z}_2^{kd}$. To conclude we recall some notation for finite abelian groups, necessary to define the aforementioned RKHS and estimate its dimension.

Appendix A. Analysis of the kernel

We provide here an alternative description of the centered random forest algorithm, where the dyadic tiling of the hypercube motivates us to define the kernel in the abelian group $\mathbb{Z}_2^{kd}$. First, we define a *Random Tree* Θ. Start with a random variable Θ^0, uniformly distributed in $\{1, \ldots, d\}$, and split

$$I := [0, 1]^d = I_0^{\Theta^0} \cup I_1^{\Theta^0},$$

where

$$I_l^{\Theta_0} = [0, 1] \times \cdots \times [l/2, (l + 1)/2] \times \ldots [0, 1],$$

where for $l = 0, 1$ the splitting was performed in the Θ^0-th coordinate. Choose then random variables $\Theta^{1,l}$ ($l = 0, 1$), distributed as Θ^0, and split each

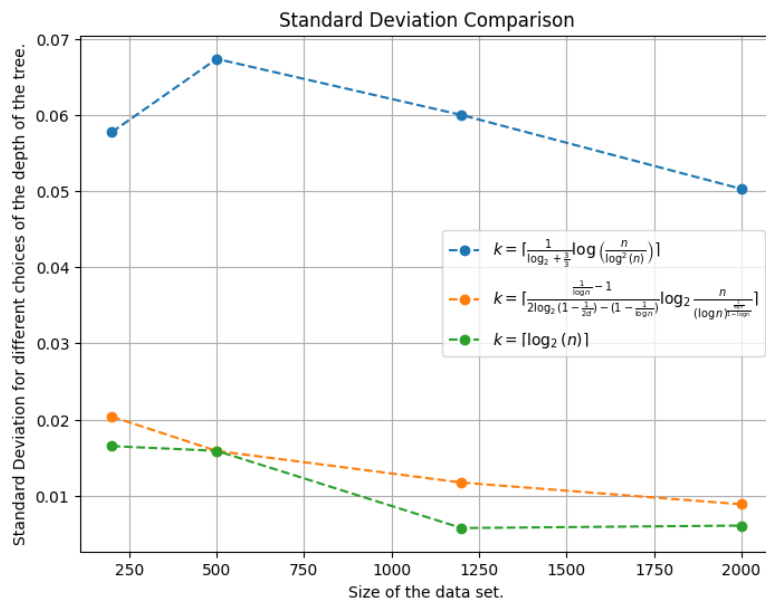$$I_l^{\Theta^0} = I_{l,0}^{\Theta^0, \Theta^1} \cup I_{l,1}^{\Theta^0, \Theta^1},$$

where, as before, the splitting is performed at the Θ_1-th coordinate, and $I_{l,0}^{\Theta^0, \Theta^1}$ is the lower half of $I_l^{\Theta^0}$. Iterate the same procedure k times. In order to do that, we need random variables $\Theta^{j;\eta_0,\ldots,\eta_j}$, with $\eta_l \in \{1, \ldots, d\}$ and $j = 1, \ldots, k$. We assume that all such random variables are independent. It is useful think of $\Theta = \{\Theta^{j;\eta_0,\ldots,\eta_j}\}$ as indexed by a d-adic tree, and, in fact, we refer to Θ as a random tree in $[0, 1]^d$. We call *cells*, or *leaves*, each of the 2^k rectangles into which $[0, 1]^d$ is split at the end of the kth subdivision.

On Figure A1 we see the L_2−error of the centered KeRF-approximation for a three dimensional feature space and on Figure A2 the standard deviation of the errors.

Figures A3 and A4 show the L_2−error and the standard deviation for the uniform KeRF in a two dimensional feature space and Figures A5 and A6 present a three dimensional example respectively.
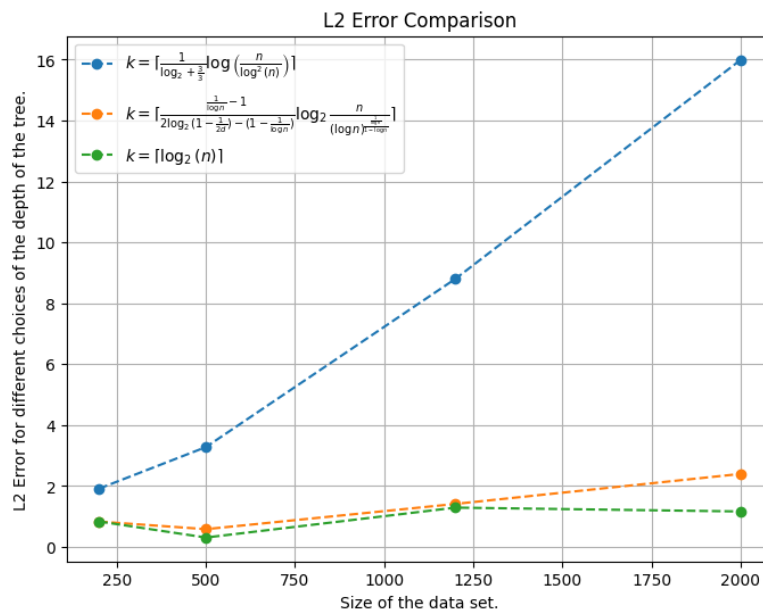
Figure A1. Plot of the L_2−error of the centered KeRF-approximation for three different values of k for the function $Y = X_1^2 + \frac{1}{e^{X_2^2 + X_3^2}} + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.5)$, against different data set size.
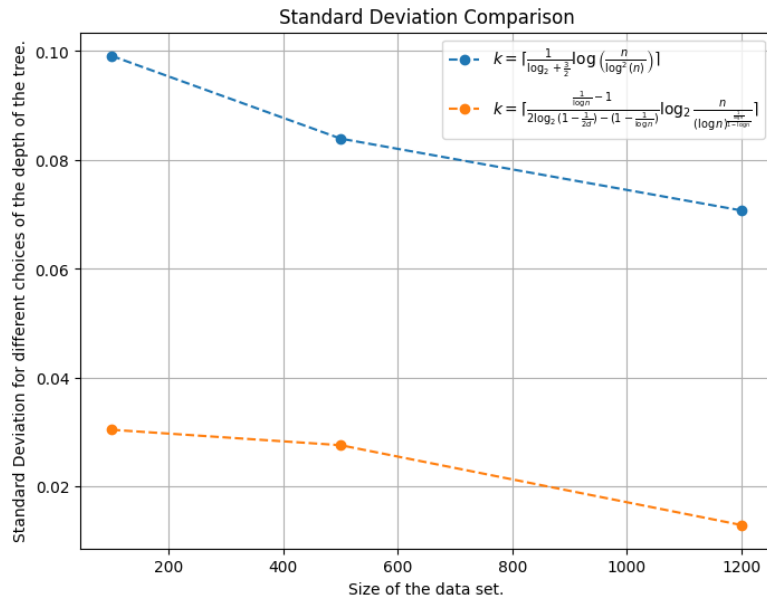
Figure A2. Plot of the standard deviation of errors, of the centered KeRF-approximation for three different values of k of the function $Y = X_1^2 + \frac{1}{e^{X_2^2} + e^{X_3^2}} + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.5)$, against different data set size.
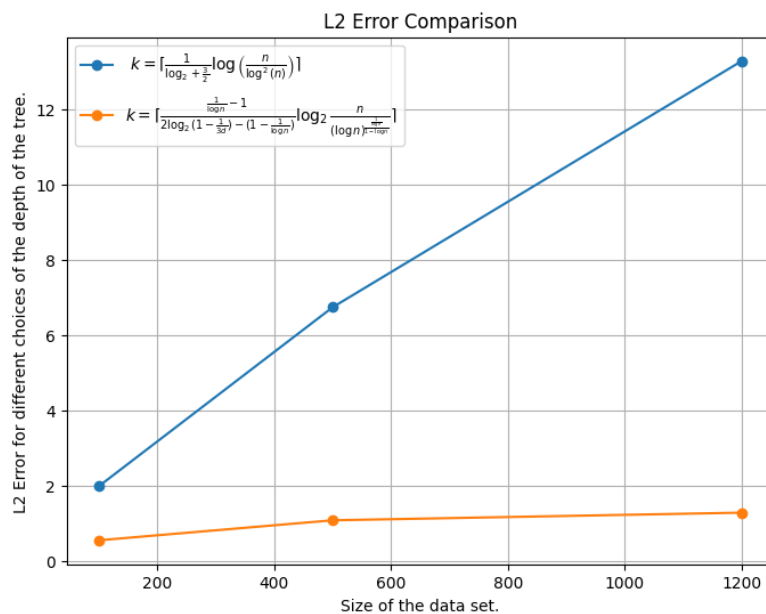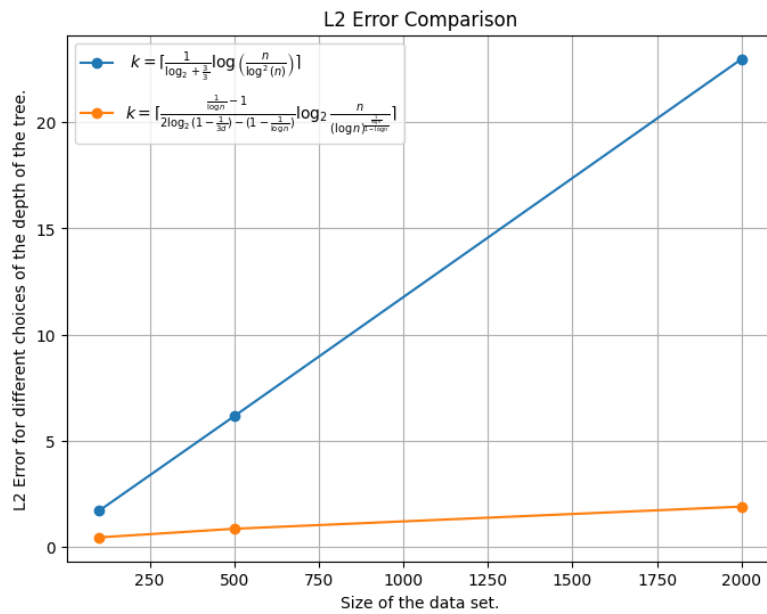

Figure A3. Plot of the L_2-error of the uniform KeRF-approximation for two different values of k for the function $Y = X_1^2 + e^{-X_2^2} + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.5)$, against different data set size.
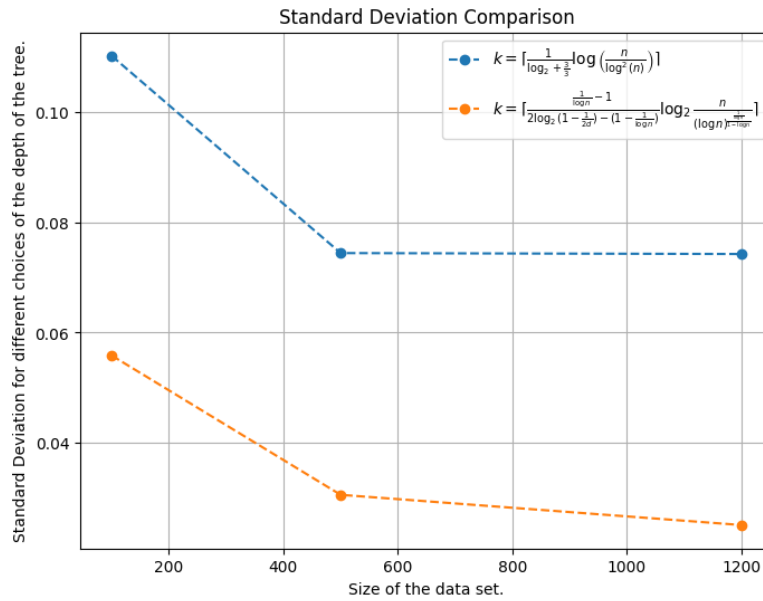
Figure A4. Plot of the standard deviation of errors, of the uniform KeRF-approximation for two different values of k for the function $Y = X_1^2 + e^{-X_2^2} + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.5)$, against different data set size.

Figure A5. Plot of the L_2-error of the uniform KeRF-approximation for two different values of k for the function $Y = X_1^2 + \frac{1}{(e^{X_3^2} + e^{X_2^2})} + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.5)$, against different data set size.

Figure A6. Plot of the standard deviation of the errors, of the uniform KeRF-approximation for two different values of k for the function $Y = X_1^2 + \frac{1}{(e^{X_3^2} + e^{X_2^2})} + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.5)$, against different data set size.

Appendix B. A reproducing kernel from the centered KeRF

B.1. The Fourier analysis of the kernel

The kernel in the centered KeRF defines a RKHS H_K structure on a set Γ having 2^{kd} points [4]. In fact, Γ has a group structure, and Fourier analysis can be used. Much research is done in RKHS theory, and in this section (see [1]), we study the structure of H_K in itself. A priori, H_K might have any dimension less or equal to $\#\Gamma$. We show in fact that its dimension is much lower than that, a fact which is somehow surprising, and we believe it is interesting in itself. Furthermore, we prove that there are no nonconstant multipliers in the space H_K. For completeness we provide definitions and notation on Fourier analysis on Abelian groups in Appendix C.

We identify every real number $x \in [0, 1]$ with its binary expression $x = 0.x_1x_2x_3\ldots$ with $x_i \in \{0, 1\}$ for $i \in \mathbb{N}$.

Here we consider the group

$$G = \mathbb{Z}_2^{kd} \ni x = (x_i^j)_{\substack{i=1,\ldots,k \\ j=1,\ldots,d}} = (x^1|x^2|\ldots|x^d) = \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_k \end{pmatrix}. \tag{B.1}$$

The kernel $K : \Gamma \times \Gamma \to \mathbb{C}$ corresponding to the kernel K_k^{cen} is,

$$K(a,b) \;=\; \sum_{\substack{l\in\mathbb{N}^d \\ |l|=k}} \frac{1}{d^k}\binom{k}{l}\prod_{j=1}^{d}\chi\left(a_1^j=b_1^j,\ldots,a_{k_j}^j=b_{k_j}^j\right)$$

$$=\; \sum_{\substack{l\in\mathbb{N}^d \\ |l|=k}} \frac{1}{d^k}\binom{k}{l}\prod_{j=1}^{d}\prod_{i=1}^{k_j}\chi\left(a_i^j=b_i^j\right)$$

$$=\; \varphi(a-b), \tag{B.2}$$

where $\binom{k}{l}$ is the multidimensional binomial coefficient, χ_E is the characteristic function of the set E, and $a-b$ is the difference in the group $\mathbb{Z}_2^{kd}$. Incidentally, (B.2) shows that the kernel K can be viewed as a convolution kernel on the appropriate group structure. For the last equality, we consider the binary representation of a number in $(0,1]$ whose digits are not definitely vanishing. The fact that 0 does not have such representation is irrelevant since the probability that one of the coordinates of the data vanishes is zero.

We now compute the anti-Fourier transform $\mu=\check\varphi$. We know that $\sharp(\Gamma)=2^{kd}$, and that the characters of $\mathbb{Z}_2^{kd}$ have the form

$$\gamma_a(x),\; x\in\mathbb{Z}_2^{kd},\; a\in\widehat{\mathbb{Z}_2^{kd}},\; a\cdot x = a_1^1 x_1^1+\cdots+a_k^d x_k^d. \tag{B.3}$$

Hence,

$$2^{kd}p^n\mu(x) \;=\; d^k\sum_{a\in\Gamma}\varphi(a)\gamma_a(x)$$

$$=\; d^k\sum_{a\in\Gamma}\varphi(a)(-1)^{a\cdot x}$$

$$=\; \sum_{a\in\Gamma}\sum_{\substack{l\in\mathbb{N}^d \\ |l|=k}}\binom{k}{l}\prod_{j=1}^{d}\prod_{i=1}^{k_j}\chi\left(a_i^j=0\right)(-1)^{a\cdot x}$$

$$=\; \sum_{a\in\Gamma}\sum_{\substack{l\in\mathbb{N}^d \\ |l|=k}}\binom{k}{l}\prod_{j=1}^{d}(-1)^{\tilde a_j^{k_j}\cdot\tilde x_j^{k_j}}\prod_{i=1}^{k_j}\left[\chi\left(a_i^j=0\right)(-1)^{a_i^j x_i^j}\right]$$

where $\tilde a_j^{k_j}=\begin{pmatrix}a_{k_j+1}^j\\ \ldots\\ a_n^j\end{pmatrix}$ is the lower, $(k-k_j)$-dimensional part of the column a^j,

$$=\; \sum_{\substack{l\in\mathbb{N}^d \\ |l|=k}}\binom{k}{l}\prod_{j=1}^{d}(-1)^{\tilde a_j^{k_j}\cdot\tilde x_j^{k_j}}\prod_{i=1}^{k_j}\chi\left(a_i^j=0\right)$$

$$=\; \sum_{\substack{l\in\mathbb{N}^d \\ |l|=k}}\binom{k}{l}\sum_{\substack{a\in\Gamma \\ a_1^1=\ldots a_{k_1}^1=a_1^2=\cdots=a_{k_d}^d=0}}\prod_{j=1}^{d}(-1)^{\tilde a_j^{k_j}\cdot\tilde x_j^{k_j}}. \tag{B.4}$$

The last expression vanishes exactly when for all l, there are some $1\le j\le d$, and some $k_j+1\le i\le k$ such that $x_i^j=1$, due to the presence of the factor $(-1)^{a_i^j x_i^j}=(-1)^{a_i^j}$ which takes values ± 1 on summands having, two by two, the same absolute values.

If, on the contrary, there is l such that for all $1 \le j \le d$, and $k_j + 1 \le i \le k$, we have that $x_i^j = 0$, then $\mu(x) \ne 0$. Since $|l| = k$ and there are kd binary digits involved in the expression of x, the latter occurs exactly when the binary matrix representing x has a large lower region in which all entries are 0. More precisely, the number of vanishing entries must be at least

$$(k - k_1) + \cdots + (k - k_p) = (d - 1)k. \tag{B.5}$$

The number $N(d, k)$ of such matrices is the dimension of H_K, the Hilbert space having K as a reproducing kernel.

Next, we prove some estimates for the dimension of the RKHS.

We summarize the main items in the following statement.

Theorem B.1. *Let $K : \Gamma \times \Gamma \to \mathbb{C}$ be the kernel in (B.2), $K(a, b) = \varphi(a - b)$, and let*

$$E_K = supp(\check{\varphi}) \in K. \tag{B.6}$$

Then,

(i) As a linear space, $H_K = L_{E_K}$, where

$$
\begin{aligned}
E_K &= \{x = (x^1 | \dots | x^d) : \ x_i^j = 0 \ for \ k_j + 1 \le i \le k, \ for \ some \ l \\
&= (k_1, \dots, k_d) \in \mathbb{N}^d \ with \ k_1 + \cdots + k_d = k\};
\end{aligned} \tag{B.7}
$$

(ii) For $x \in E_K$,

$$\check{\varphi}(x) = \frac{1}{2^k d^k} \sum_{\substack{l \in \mathbb{N}^d, \ |l| = k \\ x_i^j = 0 \ for \ k_j + 1 \le i \le k}} \binom{k}{l}. \tag{B.8}$$

To obtain the expression on (B.8), we used the fact that

$$\sharp\{a : \ a_1^1 = \dots a_{k_1}^1 = a_1^2 = \cdots = a_{k_p}^p = 0\} = 2^{(d-1)k}.$$

B.2. Some properties of H_K

Linear relations. Among all functions $\psi : \Gamma \to \mathbb{C}$, those belonging to H_K (i.e., those belonging to L_{E_K}) are characterized by a set of linear equations,

$$0 = 2^{np} p^n \mu(x) = \sum_{\substack{k \in \mathbb{N}^p, \ |k| = n \\ x_i^j = 0 \ for \ k_j + 1 \le i \le n}} \binom{n}{k} \text{ for } x \notin E_K. \tag{B.9}$$

Multipliers. A *multiplier* of H_K is a function $m : \Gamma \to \mathbb{C}$ such that $m\psi \in H_K$ whenever $\psi \in H_K$.

Proposition B.1. *The space H_K has no nonconstant multiplier.*

In particular, it does not enjoy the *complete Pick property*, which has been subject of intensive research for the past twenty-five years [1].

Proof. The space H_K coincides as a linear space with L_{E_K}. Let $\Lambda_E = \check{L}_E$, which is spanned by $\{\delta_x : x \in E\}$. Observe that, since $0 = (0| \dots |0) \in E_K$, the constant functions belong to H_K, hence, any multiplier m of H_K lies in H_K; $m = m \cdot 1 \in H_K$.

Suppose m is not constant. Then, $\check{m}(a) \neq 0$ for some $a \in E_K$, $a \neq 0$. Let a be an element in E_K such that $\check{m}(a) \neq 0$. Since $H_K \ni m \cdot \widehat{\delta_x}$ for all x in E_K, and $m \cdot \widehat{\delta_x} = \widehat{\check{m} * \delta_x}$, we have that the support of $\check{m} * \delta_x$ lies in H_K. Now,

$$\check{m} * \delta_x(y) = \check{m}(x - y),$$

hence, we have that, for any x in E_K, $y = x - a$ lies in E_K as well. This forces $a = 0$, hence m to be constant. $\qquad\square$

B.3. Bounds for dimension and generating functions

Theorem B.2. *For fixed $d \geq 1$, we have the estimates:*

$$dim(H_K) \sim \frac{2^{k-d+1}k^{d-1}}{(d-1)!}, \; hence \; \frac{dim(H_K)}{2^{kd}} \sim \frac{k^{d-1}}{2^{d-1}(d-1)!2^{k(d-1)}}. \tag{B.10}$$

Proof. Let $l_1, l_2, ..., l_d$ such that

$$0 \leq l_1 + l_2 + ... + l_d = m \leq k$$

where m is a parameter and let also $\lambda = |j : l_j \geq 1| = |\{\text{stop 1-digits}\}| = |\{\text{back-entries}\}|$ where $|\cdot|$ denotes the size of the sets, and of course we have that $0 \leq m \leq k$ and $0 \leq \lambda \leq d, m$. Goal to obtain a bound for

$$N(k, d) = \sum_{m=0}^{k} \sum_{\lambda=0}^{d \wedge m} 2^{m-\lambda} \binom{d}{\lambda} |\{(l_1, l_2, ..., l_d) : l_1 + l_2 + ... + l_d = m|$$

and

$$|\{j : l_j = 1\} = 1|\}.$$

Let $A(m, \lambda)$ the m-th coefficient of x in the polynomial

$$\begin{aligned}
(x + x^2 + ...x^m + ...)^\lambda &= (x(1 + x + x^2 + ...)^\lambda \\
&= (x^\lambda(1 + x + ...)^\lambda) \\
&= \frac{x^\lambda}{(1-x)^\lambda}
\end{aligned}$$

and $2^m A(m, \lambda)$ is the m-th coefficient of x, for the fraction $\frac{(2x)^\lambda}{(1-2x)^\lambda}$, therefore $2^{m-\lambda}A(m, \lambda)$ is the m-th coefficient of $\frac{x^\lambda}{(1-2x)^\lambda}$. Let's see the first sum, $B(m, d)$ is the m-th coefficient of x:

$$\begin{aligned}
\sum_{\lambda=0}^{d \wedge m} \binom{d}{\lambda} 2^{m-\lambda} A(m, \lambda) &= \sum_{\lambda=0}^{d \wedge m} \binom{d}{\lambda} (\frac{x}{1-2x})^\lambda \\
&= (1 + \frac{x}{1-2x})^d \\
&= (\frac{1-x}{1-2x})^d.
\end{aligned}$$

Again by the same combinatoric argument we are looking the k-th coefficient of the function

$$f(x) = \frac{1}{1-x}\left(\frac{1-x}{1-2x}\right)^d.$$

Back to the estimate, Let a_k the coefficient of the power series centered at $z = 0$.

$$\max_{|z|=r} |f(z)| = \max_{|z|=r} \left|\frac{(1-z)^{d-1}}{(1-2z)^d}\right| = \max_{|z|=r} \left|\frac{1}{1-z}\left(\frac{1-z}{1-2z}\right)^d\right| \le 2 \max_{\theta \in (-\pi,\pi)} \left|\frac{1-re^{i\theta}}{1-2re^{i\theta}}\right|^d.$$

After some calculations since r is fixed one has that the maximum is achieved for $\theta = 0$. So

$$\max_{|z|=r} |f(z)| \le 2\left(\frac{1-r}{1-2r}\right)^d.$$

Our estimation finally becomes:

$$\begin{aligned}
|a_k| &\le \frac{2\left(\frac{1-r}{1-2r}\right)^d}{r^k} \\
&= \frac{2(1-r)^d}{r^k(1-2r)^k} \\
&= k^d 2^k \left(\frac{1}{2} + \frac{1}{2k}\right)^d, \qquad \left(\text{since,} \quad r = \frac{1}{2}\left(1 - \frac{1}{k}\right)\right) \\
&= k^d\left(1 + \frac{1}{k}\right)^d 2^{k-d}.
\end{aligned}$$

Thus an estimate for the dimension of H_K is

$$\frac{|a_k|}{2^{kd}} \lesssim \frac{k^d\left(1 + \frac{1}{k}\right)^d 2^{k(1-d)}}{2^d}.$$

Another estimate about the dimension of H_K. For $f(z) = \sum_{n=0}^{\infty} a_n z^n$, we have

$$|a_n| \le \frac{\max\{|f\left(re^{it}\right)| : \; |t| \le \pi\}}{r^n}.$$

Consider the function

$$f(z) = \frac{(1-z)^{d-1}}{(1-2z)^d}$$

in $|z| < 1/2$ and let $r = \frac{1-1/k}{2}$. Then,

$$\begin{aligned}
|a_k| &\le \frac{(3/2)^{d-1}}{(1/k)^d(1-1/k)^k 2^{-k}} \\
&\le (3/2)^{d-1} 2^k e k^d.
\end{aligned}$$

Thus,

$$\frac{|a_k|}{2^{kd}} \lesssim \frac{k^d(3/2)^d}{2^{k(d-1)}}.$$

Recursively working out the generating function one gets the estimates in (B.10). $\qquad\square$

Appendix C. Some results from Fourier analysis on finite groups

Following the notation of [24], we recall here the basic notions of Fourier theory for a finite, abelian group G, which we employed above. Here, G is endowed with its counting measure. The *dual group* $\Gamma = \widehat{G}$ of G is populated by labels a for homomorphisms $\gamma_a : G \to \mathbb{T} = \{e^{it} : t \in \mathbb{R}\}$. Given a function $f : G \to \mathbb{C}$, its *Fourier transform* $\widehat{f} : \Gamma \to \mathbb{C}$ is defined as

$$\widehat{f}(a) = \sum_{x \in G} f(x)\overline{\gamma_a(x)}. \tag{C.1}$$

We make Γ into a (finite), additive group by setting

$$\gamma_{a+b} = \gamma_a \cdot \gamma_b, \text{ and } \gamma_x(a) := \gamma_a(x).$$

It turns out that they have the same number of elements, $\sharp(G) = \sharp(\Gamma)$. Some basic properties are:

$$
\begin{aligned}
f(x) &= \frac{1}{\sharp(\Gamma)} \sum_{a \in \Gamma} \widehat{f}(a)\gamma_a(x) \text{ (inverse Fourier transform)} \\
\sum_{x \in G} |f(x)|^2 &= \frac{1}{\sharp(\Gamma)} \sum_{a \in \Gamma} |\widehat{f}(a)|^2 \text{ (Plancherel)} \\
\widehat{f * g} &= \widehat{f} \cdot \widehat{g},
\end{aligned}
$$

where

$$(f * g)(x) = \sum_{y \in G} f(x - y)g(y). \tag{C.2}$$

We write

$$\check{\varphi}(x) = \sharp(\Gamma)^{-1} \sum_{a \in \Gamma} \varphi(a)\gamma_a(x), \text{ so that } \widehat{\check{\varphi}} = \varphi. \tag{C.3}$$

The unit element of convolution in G is δ_0.

In the other direction, for $\varphi, \psi : \Gamma \to \mathbb{C}$ we define

$$(\varphi * \psi)(a) = \frac{1}{\sharp(\Gamma)} \sum_{b \in \Gamma} \varphi(a - b)\psi(b), \tag{C.4}$$

and similarly to above, $\widehat{\check{\varphi}\check{\psi}} = \varphi * \psi$. The unit element on convolution in Γ is $\sharp(\Gamma)\delta_0$.

A function φ on Γ is *positive definite* if

$$\sum_{a,b \in \Gamma}^{n} c(a)\overline{c(b)}\varphi(b - a) \geq 0.$$

Theorem C.1. *[Bochner's theorem] A function $\varphi : \Gamma \to \mathbb{C}$ is positive definite if and only if there exists $\mu : G \to \mathbb{R}_+$ such that $\varphi = \widehat{\mu}$.*

The theorem holds in great generality, and its proof in the finite group case is elementary. We include it because it highlights the relationship between the measure μ on G and the positive definite function (the kernel) φ.

Proof. If

$$\sharp(\Gamma)^{-2} \sum_{a,b\in\Gamma} \widehat{\mu}(b-a)c(a)\overline{c(b)} \;=\; \sum_{x\in G} \sharp(\Gamma)^{-2} \sum_{a,b\in\Gamma} \mu(x)\overline{\gamma_{b-a}(x)}c(a)\overline{c(b)}$$

$$= \sum_{x\in G} \sharp(\Gamma)^{-2} \sum_{a,b\in\Gamma} \mu(x)\overline{\gamma_b(x)c(b)}\gamma_a(x)c(a)$$

$$= \sum_{x\in G} \mu(x)\left|\sharp(\Gamma)^{-1}\sum_{a\in\Gamma} c(a)\gamma_a(x)\right|^2$$

$$= \sum_{x\in G} \mu(x)\,|\check{c}(x)|^2 \geq 0. \tag{C.5}$$

Only if, since for all b in Γ,

$$\mu(x)\sharp(\Gamma) \;=\; \sum_{a\in\Gamma} \varphi(a)\gamma_x(a)$$

$$= \sum_{a\in\Gamma} \varphi(a-b)\gamma_x(a-b)$$

$$= \sum_{a\in\Gamma} \varphi(a-b)\gamma_x(a)\overline{\gamma_x(b)}, \tag{C.6}$$

we have

$$\mu(x)\sharp(\Gamma)^2 = \sum_{a,b\in\Gamma} \varphi(a-b)\gamma_x(a)\overline{\gamma_x(b)} \geq 0, \tag{C.7}$$

by the assumption. $\qquad\square$

We now come to reproducing kernels on Γ which are based on positive definite functions $\varphi : \Gamma \to \mathbb{R}_+$. Set

$$K(a,b) = \varphi(a-b) = K_b(a), \; K : \Gamma \times \Gamma \to \mathbb{C}, \tag{C.8}$$

and set

$$H_K = \operatorname{span}\{K_b : \; b \in \Gamma\} \ni \sum_{b\in\Gamma} c(b)K_b, \tag{C.9}$$

where H_K is the Hilbert space having K as reproducing kernel. We wish to have a more precise understanding of it.

We start by expressing the norm of an element on H_K is several equivalent ways,

$$\left\|\sum_{b\in\Gamma} c(b)K_b\right\|_{H_K}^2 \;=\; \sum_{a,b\in\Gamma} \overline{c(a)}c(b)\langle K_b, K_a\rangle$$

$$= \sum_{a,b\in\Gamma} \overline{c(a)}c(b)K(a,b) = \sum_{a,b\in\Gamma} \overline{c(a)}c(b)\widehat{\mu}(a-b)$$

$$= \sum_{a,b\in\Gamma} \overline{c(a)}c(b) \sum_{x\in G} \mu(x)\gamma_{b-a}(x)$$

$$= \sum_{x\in G} \mu(x) \sum_{a,b\in\Gamma} \overline{c(a)}c(b)\gamma_b(x)\overline{\gamma_a(x)}$$

$$= \sum_{x\in G} \mu(x) \left|\sum_{b\in\Gamma} c(b)\gamma_b(x)\right|^2$$

$$
\begin{aligned}
&= \sharp(\Gamma)^2 \sum_{x \in G} \mu(x) \, |\check{c}(x)|^2 \\
&= \sharp(\Gamma)^2 \sum_{x \in G} \left| \mu(x)^{1/2} \check{c}(x) \right|^2 .
\end{aligned}
\tag{C.10}
$$

In other terms,

$$
\sharp(\Gamma)^{-1} \sum_{b \in \Gamma} c(b) K_b \mapsto \check{c}
\tag{C.11}
$$

is an isometry of H_K onto $L^2(\mu)$. This will become important later, when we verify that for our kernels $\operatorname{supp}(\mu)$ is sparse in G. In fact, $\dim(H_K) = \sharp(\operatorname{supp}(\mu))$.

Corollary C.1. *As a linear space, H_K is determined by $\operatorname{supp}(\mu)$:*

$$
\psi \in H_K \text{ if and only if } \operatorname{supp}(\check{\psi}) \subseteq \operatorname{supp}(\mu).
$$

Let $E \subseteq G$. We denote

$$
L_E = \{ G \xrightarrow{\psi} \mathbb{C} : \ \operatorname{supp}(\check{\psi}) \subseteq E \}.
\tag{C.12}
$$

Next, we look for the natural orthonormal system provided by the Fourier isometry (C.11). Fr $x \in G$, let $\check{c}_x = \mu(x)^{-1/2} \delta_x$: $\{ \check{c}_x : \ x \in E := \operatorname{supp}(\mu) \}$ is a orthonormal system for $L^2(\mu)$, and so $\{ e_x : \ x \in E \}$ is an orthonormal basis for H_K, where

$$
c_x(b) = \sum_{y \in G} \mu(x)^{-1/2} \delta_x(y) \overline{\gamma_b(y)} = \mu(x)^{-1/2} \overline{\gamma_b(x)},
\tag{C.13}
$$

and

$$
\begin{aligned}
e_x(a) &= \sharp(\Gamma)^{-1} \sum_{b \in \Gamma} c_x(b) K_b(a) \\
&= \frac{\mu(x)^{-1/2}}{\sharp(\Gamma)} \sum_{b \in \Gamma} K_b(a) \overline{\gamma_b(x)} \\
&= \frac{\mu(x)^{-1/2}}{\sharp(\Gamma)} \sum_{b \in \Gamma} \varphi(a - b) \overline{\gamma_b(x)} \\
&= \frac{\mu(x)^{-1/2}}{\sharp(\Gamma)} \sum_{b \in \Gamma} \varphi(a - b) \gamma_{a-b}(x) \overline{\gamma_a(x)} \\
&= \mu(x)^{-1/2} \mu(x) \overline{\gamma_a(x)} \\
&= \mu(x)^{1/2} \overline{\gamma_a(x)}.
\end{aligned}
\tag{C.14}
$$

Let's verify that we obtain the reproducing kernel from the o.n.b. as expected,

$$
\begin{aligned}
\sum_{x \in \Gamma} e_x(a) \overline{e_x(b)} &= \sum_{x \in \Gamma} \mu(x) \overline{\gamma_x(a)} \gamma_x(b) \\
&= \sum_{x \in \Gamma} \mu(x) \overline{\gamma_x(a - b)} \\
&= \widehat{\mu}(a - b) \\
&= \varphi(a - b).
\end{aligned}
\tag{C.15}
$$

Remark C.1. *Since any finite, abelian group can be written as the direct product of cyclic groups,*

$$G = \bigoplus_{l=1}^{L} \mathbb{Z}_{m_l},$$ (C.16)

its dual Γ can be written in the same way, because $\widehat{\mathbb{Z}_m} \equiv \mathbb{Z}_m$. From the Fourier point of view, the only difference is that, if on G we consider the counting measure, then on Γ we consider normalized counting measure, as we did above.