



# Convergence rate of the data-independent $P$ -greedy algorithm in kernel-based approximation

Gabriele Santin<sup>a</sup> · Bernard Haasdonk<sup>b</sup>

Communicated by A. De Rossi and E. Francomano

## Abstract

Kernel-based methods provide flexible and accurate algorithms for the reconstruction of functions from meshless samples. A major question in the use of such methods is the influence of the samples' locations on the behavior of the approximation, and feasible optimal strategies are not known for general problems. Nevertheless, efficient and greedy point-selection strategies are known. This paper gives a proof of the convergence rate of the data-independent  $P$ -greedy algorithm, based on the application of the convergence theory for greedy algorithms in reduced basis methods. The resulting rate of convergence is shown to be quasi-optimal in the case of kernels generating Sobolev spaces.

As a consequence, this convergence rate proves that, for kernels of Sobolev spaces, the points selected by the algorithm are asymptotically uniformly distributed, as conjectured in the paper where the algorithm has been introduced.

## 1 Introduction

We start by recalling some basic facts of kernel based approximation. Further details and a thorough treatment of the topic can be found e.g. in the monographs [20, 7, 2, 8].

On a compact set  $\Omega \subset \mathbb{R}^d$  we consider a continuous, symmetric and strictly positive definite kernel  $K : \Omega \times \Omega \rightarrow \mathbb{R}$ . Positive definiteness is understood in terms of the associated *kernel matrix*, i.e., for all  $n \in \mathbb{N}$  and  $\{x_1, \dots, x_n\} \subset \Omega$  pairwise distinct the kernel matrix  $A \in \mathbb{R}^{n \times n}$ ,  $A_{ij} := K(x_i, x_j)$ , is positive definite.

Associated with the kernel there is a uniquely defined *native space*  $\mathcal{H}_K(\Omega)$ , that is, the unique Hilbert space of functions from  $\Omega$  to  $\mathbb{R}$  in which  $K$  is the *reproducing kernel*, i.e.,

- (a)  $K(\cdot, x) \in \mathcal{H}_K(\Omega)$  for all  $x \in \Omega$ ,
- (b)  $(f, K(\cdot, x)) = f(x)$  for all  $f \in \mathcal{H}_K(\Omega)$ ,  $x \in \Omega$ .

We used here and we will use in the following the notation  $(\cdot, \cdot)$  and  $\|\cdot\|$ , without subscripts, for the inner product and norm of  $\mathcal{H}_K(\Omega)$ .

For any given finite set  $X_n := \{x_1, \dots, x_n\} \subset \Omega$  of  $n$  pairwise distinct points, the interpolation of a function  $f \in \mathcal{H}_K(\Omega)$  on  $X_n$  is well defined being the kernel strictly positive definite, and it coincides with the orthogonal projection  $\Pi_{V(X_n)}(f)$  of  $f$  into  $V(X_n)$ , where  $V(X_n) := \text{span} \{K(\cdot, x_k), 1 \leq k \leq n\}$  is the  $n$ -dimensional subspace of  $\mathcal{H}_K(\Omega)$  generated by the kernel translates on  $X_n$ . We will denote by  $|\cdot|$  the number of pairwise distinct elements of a finite set, i.e.,  $|X_n| := n$ .

Since  $\Pi_{V(X_n)}(f) \in V(X_n)$ , the interpolant is of the form

$$\Pi_{V(X_n)}(f) := \sum_{k=1}^n \alpha_k K(\cdot, x_k),$$

for some coefficients  $\{\alpha_k\}_{k=1}^n$ . To actually compute these, one imposes the interpolation conditions  $\Pi_{V(X_n)}(f)(x_i) = f(x_i)$ ,  $1 \leq i \leq n$ , which result in the linear system

$$A\alpha = b, \tag{1}$$

which has in fact a unique solution for all  $b \in \mathbb{R}^n$ ,  $b_i := f(x_i)$ ,  $A$  being positive definite.

A standard way to measure the interpolation error is by means of the *Power Function*  $P_{V(X_n)}$ , which is defined in a point  $x \in \Omega$  as the norm of the pointwise interpolation error at  $x$ , i.e.,

$$P_{V(X_n)}(x) := \sup_{f \in \mathcal{H}_K(\Omega), f \neq 0} \frac{|f(x) - \Pi_{V(X_n)}(f)(x)|}{\|f\|}, \tag{2}$$

<sup>a</sup>Institute for Applied Analysis and Numerical Simulation, University of Stuttgart, Germany, email: [santinge@mathematik.uni-stuttgart.de](mailto:santinge@mathematik.uni-stuttgart.de)

<sup>b</sup>Institute for Applied Analysis and Numerical Simulation, University of Stuttgart, Germany, email: [haasdonk@mathematik.uni-stuttgart.de](mailto:haasdonk@mathematik.uni-stuttgart.de)

and it is a continuous function on  $\Omega$ , vanishing only on  $X_n$ . Among other equivalent definitions of the Power Function (e.g., by considering a *cardinal basis*  $\{\ell_k\}_{k=1}^n$  of  $V(X_n)$ , i.e.,  $\ell_k(x_i) = \delta_{ki}$ ), the present one is easier to generalize to the setting considered in Section 3. From the definition, it is immediate to see that bounds on the maximal value of the Power Function in  $\Omega$  provide uniform bounds on the interpolation error as

$$\|f - \Pi_{V(X_n)}(f)\|_{L_\infty(\Omega)} \leq \|P_{V(X_n)}\|_{L_\infty(\Omega)} \|f\|, \quad f \in \mathcal{H}_K(\Omega). \quad (3)$$

It is thus of interest to find and characterize point sets  $X_n$  which guarantee a small value of  $\|P_{V(X_n)}\|_{L_\infty(\Omega)}$ , and the reason is twofold. If one is free to consider any point in  $\Omega$ , selecting good points means to construct an optimal or suboptimal discretization of the set with respect to kernel approximation. On the other hand, if a set of data points  $X_N \subset \Omega$  is provided (e.g., the location of the measurements coming from an application), it is often desirable to be able to select a subset  $X_n \subset X_N$ ,  $n \ll N$ , of the full data to reconstruct a sparse approximation of the unknown function, where sparsity is understood in the following sense. Selecting  $X_n \subset X_N$  means to solve the system (1) with respect to the submatrix defined by the small point set, or, equivalently, to compute an interpolant (or model of the data) given by an expansion of only  $n$  out of  $N$  kernel translates. Thus the interpolant can be represented as a linear combination of the kernel translates by means of a sparse coefficient vector, and this implies that its evaluation is cheaper and more suitable to be used as a surrogate model of the data. This concept of sparsity has not to be understood as the solution of a sparse linear system, since the solution is computed by selecting points, hence rows and columns, and solving the linear system for the small resulting submatrix, which is usually dense.

Although feasible selection criteria to construct an optimal set  $X_n$  are generally not known, different greedy techniques have been presented to construct *near optimal* points (see [17, 18, 3, 13, 12, 21]). They are based on the idea that it is possible to construct good sequences of nested sets of points starting from the empty set  $X_0 := \emptyset$ , and iteratively increasing the set as  $X_n := X_{n-1} \cup \{x_n\}$  by adding a new point chosen so that it maximizes a certain indicator. The resulting algorithms all share the same structure, while the choice of the point selection criteria is different.

Among various methods, we will concentrate here on the so-called *P-greedy* algorithm which has been introduced in [3]. It is a *data independent* algorithm, meaning that the selection of the points is made by only looking at  $K$  and  $\Omega$  (and possibly  $X_N$ ), but not at the samples of a particular function  $f \in \mathcal{H}_K(\Omega)$ , and it thus produces point sets which provide uniform approximation errors for *any* function  $f \in \mathcal{H}_K(\Omega)$ . To be more precise, the selection criterion picks at every iteration the point in  $\Omega \setminus X_{n-1}$  which maximizes the Power Function  $P_{V(X_{n-1})}$ , from which the name of the algorithm derives. By adding this point to the set  $X_{n-1}$ , the new Power Function  $P_{V(X_n)}$  vanishes at  $x_n$ , and indeed, as we will explain later,  $\|P_{V(X_n)}\|_{L_\infty(\Omega)} \leq \|P_{V(X_{n-1})}\|_{L_\infty(\Omega)}$ .

The goal of this paper is to prove that the points produced by this algorithm are indeed near-optimal, meaning that they have the same asymptotic decay of the best known, non greedy point distributions. In the paper [3], the authors considered the case of translational invariant and Fourier transformable kernels on domains satisfying an interior cone condition, for which the asymptotic decay of the Power Function obtained by certain point distributions is well understood. We remark that Radial Basis Functions are instances of such kernels. In this setting, in [3] the decay rate for the *P-greedy* algorithm has been shown to be as stated in the next Theorem, which is, up to our knowledge, the currently sharpest known convergence statement.

**Theorem 1.1.** *If  $\Omega$  is compact in  $\mathbb{R}^d$  and satisfies an interior cone condition, and  $K \in \mathcal{C}^2(\Omega_1 \times \Omega_1)$ , with  $\Omega \subset \Omega_1$ ,  $\Omega_1$  compact and convex, then the point sets  $\{X_n\}_n$  selected by the *P-greedy* algorithm have Power Functions such that, for any  $n \in \mathbb{N}$ ,*

$$\|P_{V(X_n)}\|_{L_\infty(\Omega)} \leq cn^{-\frac{1}{d}},$$

for a constant  $c$  not depending on  $n$ .

The proof of this theorem requires that  $K \in \mathcal{C}^2$  on a suitable set, and our bound indeed is similar to the present one under the same assumptions, while it will improve it when the additional smoothness of the kernel is taken into account. This refined error bound allows also to prove that the selected points, for certain kernels, are asymptotically uniformly distributed. We remark that other techniques are known to extract well distributed points from a set of scattered locations. They are designed using geometrical properties of the data sites, such as the thinning algorithms in [10, 5, 9] or error surrogates [6].

The paper is organized as follows. In Section 2 we review the known estimates on the decay of the Power Function and give further details on the *P-greedy* algorithm. Section 3 is devoted to provide a connection between Kolmogorov widths and maximization of the Power Function. This connection allows to employ the theory of [1, 4] in Section 4 to prove the main results of this paper. Finally, in Section 5 we present some numerical experiments which verify the expected rates of convergence.

*Remark 1.* We remark that, although our analysis is presented for the reconstruction of scalar-valued functions, it applies also to the vector-valued case when using product spaces: namely, as pointed out in [21], for  $q \geq 1$  it is possible to use kernel methods to reconstruct functions  $f : \Omega \rightarrow \mathbb{R}^q$  simply by considering  $q$  copies of  $\mathcal{H}_K(\Omega)$ , i.e., the product space

$$\mathcal{H}_K(\Omega)^q := \{f : \Omega \rightarrow \mathbb{R}^q, f_j \in \mathcal{H}_K(\Omega)\}$$

equipped with the inner product

$$(f, g)_q := \sum_{j=1}^q (f_j, g_j),$$

where, to avoid having  $q$  different expansions, one for each component, we can make the further assumption that a unique subspace  $V(X_n)$  is used for every component. In this context, the present discussion on the *P-greedy* algorithm is directly applicable without modifications.

## 2 Power Function and the P-greedy algorithm

To assess the convergence rate of the  $P$ -greedy algorithm, we compare it with the known estimates on the decay of the Power Function. The following bounds apply to the notable case of translational invariant kernels, for which the behavior of the Power Function is well understood.

To be more precise, we assume from now on that there exists a function  $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$  such that  $K(x, y) := \Phi(x - y)$ , and that  $\Phi$  has a continuous Fourier transform  $\hat{\Phi}$  on  $\mathbb{R}^d$ . We further assume that  $\Omega$  satisfies an interior cone condition. Under these assumptions, the decay of  $\|P_{V(X_n)}\|_{L_\infty(\Omega)}$  can be related to the smoothness of  $\Phi$  (hence of  $K$ ) and to the *fill distance*

$$h_{X_n, \Omega} := \sup_{x \in \Omega} \min_{x_j \in X_n} \|x - x_j\|_2,$$

where  $\|\cdot\|_2$  is the Euclidean norm on  $\mathbb{R}^d$ . The next theorem summarizes such estimates (see [16]). We remark that the two cases (a) and (b) are substantially different. The first one regards kernels for which there exist  $c_\Phi, C_\Phi > 0$  and  $\beta \in \mathbb{N}$ ,  $\beta > d/2$ , such that

$$c_\Phi (1 + \|\omega\|_2^2)^{-\beta} \leq \hat{\Phi}(\omega) \leq C_\Phi (1 + \|\omega\|_2^2)^{-\beta},$$

shortly  $\hat{\Phi}(\omega) \sim (1 + \|\omega\|_2^2)^{-\beta}$ , in which case  $K \in C^\beta$  and  $\mathcal{H}_K(\mathbb{R}^d)$  is norm equivalent to the Sobolev space  $W_2^\beta(\mathbb{R}^d)$ . The second one applies to kernels of infinite smoothness, such as the Gaussian kernel. We will use the notion of kernels of *finite or infinite smoothness* to indicate precisely these two cases.

**Theorem 2.1.** *Under the assumptions on  $K$  and  $\Omega$  as above, for suitable constants  $\hat{c}_1, \hat{c}_2, \hat{c}_3$  not depending on  $X_n$  we have the following cases.*

(a) *If  $K$  has finite smoothness  $\beta \in \mathbb{N}$ ,*

$$\|P_{V(X_n)}\|_{L_\infty(\Omega)} \leq \hat{c}_1 h_{X_n, \Omega}^{\beta-d/2}.$$

(b) *If  $K$  is infinitely smooth,*

$$\|P_{V(X_n)}\|_{L_\infty(\Omega)} \leq \hat{c}_2 \exp(-\hat{c}_3/h_{X_n, \Omega}).$$

In particular, one can look at *asymptotically uniformly distributed points* in  $\Omega$ , i.e., sequences  $\{X_n\}_n$  of points such that  $h_{X_n, \Omega} \leq cn^{-1/d}$ , for a constant  $c \in \mathbb{R}$  not depending on  $n$ . The above estimates can then be written only in terms of  $n$ .

**Corollary 2.2.** *In the same setting as in Theorem 2.1, there exists a sequence  $\{X_n\}_n$  of points in  $\Omega$  and constants  $c_1, c_2, c_3$ , so that for  $n \in \mathbb{N}$  the Power Function behaves as follows.*

(i) *If  $K$  has finite smoothness  $\beta \in \mathbb{N}$ ,*

$$\|P_{V(X_n)}\|_{L_\infty(\Omega)} \leq c_1 n^{-\frac{\beta}{d} + \frac{1}{2}}.$$

(ii) *If  $K$  is infinitely smooth,*

$$\|P_{V(X_n)}\|_{L_\infty(\Omega)} \leq c_2 \exp(-c_3 n^{1/d}).$$

To refer to the convergence of the Power Function as  $n$  increases, in the above rates, we will concisely write

$$\|P_{V(X_n)}\|_{L_\infty(\Omega)} \leq \gamma_n \quad \text{with} \quad \lim_{n \rightarrow \infty} \gamma_n = 0.$$

### 2.1 The $P$ -greedy algorithm

We describe here in more detail the structure of the algorithm, and provide some aspects of its implementation. The algorithm starts with an empty set  $X_0 := \emptyset$  and with the zero subspace  $V(X_0) := \{0\}$ , and it constructs a sequence of nested point sets

$$X_0 \subset X_1 \subset \dots \subset X_n \subset \dots \subset \Omega,$$

by sequentially adding a new point, i.e.,  $X_n := X_{n-1} \cup \{x_n\}$ . A sequence of nested linear subspaces

$$V(X_0) \subset V(X_1) \subset \dots \subset V(X_n) \subset \dots \subset \mathcal{H}_K(\Omega),$$

is associated to the point sets, and for each of them a Power Function  $P_{V(X_n)}$  can be defined. For  $n = 0$ , Definition (2) gives  $P_{V(X_0)} := \sqrt{K(x, x)}$ , since

$$|f(x) - \Pi_{V(X_0)}(f)(x)| = |f(x)| = |(f, K(\cdot, x))| \leq \|K(\cdot, x)\| \|f\| = \sqrt{K(x, x)} \|f\|,$$

and equality is obtained for  $f := K(\cdot, x)$ .

The points are chosen by picking the current maximum on  $\Omega \setminus X_n$  of the the  $n$ -th Power Function, i.e.,

$$\begin{aligned} x_1 &:= \arg \max_{x \in \Omega} P_{V(X_0)}(x) = \sqrt{K(x, x)}, \\ x_n &:= \arg \max_{x \in \Omega \setminus X_{n-1}} P_{V(X_{n-1})}(x). \end{aligned}$$

In particular, the choice of the first point is arbitrary for a translational invariant kernel, and in general all other points are not uniquely defined, being the not necessarily unique maxima of the Power Function.

This  $P$ -greedy algorithm has an efficient implementation in terms of the *Newton basis* (see [12]), which allows to easily deal with nested subspaces and the corresponding orthogonal projections. Namely, assuming to have a sequence  $\{X_n\}_n$  of nested point sets, the construction of the Newton basis is a Gram-Schmidt procedure over the set of the kernel translates at these points, and the resulting set of functions  $\{v_k\}_{k=1}^n$  is indeed an orthonormal basis of  $V(X_n)$ , with the further property that  $\text{span}\{v_k, 1 \leq k \leq n\} = V(X_n)$ . In particular, the basis does not need to be recomputed when a new point is added. We remark that this construction can be efficiently implemented by a matrix-free partial Cholesky factorization of the kernel matrix (i.e., only one column at a time is computed), with the pivoting rule given by the present selection criteria (see [13]).

As mentioned in Section 1, the  $P$ -greedy selection strategy guarantees that the Power Function decreases. To prove this fact, we first recall the following characterization of the Power Function, which we prove for completeness.

**Lemma 2.3.** *For any subspace  $V(X_n) \subset \mathcal{H}_K(\Omega)$  and  $x \in \Omega$ , the Power Function has the representation*

$$P_{V(X_n)}(x) = \|K(\cdot, x) - \Pi_{V(X_n)}(K(\cdot, x))\|. \quad (4)$$

*Proof.* Using (2), it suffices to consider the case  $f \in \mathcal{H}_K(\Omega)$ ,  $\|f\| = 1$ . We consider an orthonormal basis  $\{v_k\}_k$  of  $V(X_n)$  and define here  $v_x := K(\cdot, x)$  for simplicity of notation. The interpolation error for  $f$ , measured at  $x \in \Omega$ , is

$$\begin{aligned} f(x) - \Pi_{V(X_n)}(f)(x) &= (v_x, f) - \left( v_x, \sum_{k=1}^n (f, v_k) v_k \right) \\ &= (v_x, f) - \sum_{k=1}^n (f, v_k) (v_x, v_k) = (v_x, f) - \left( \sum_{k=1}^n (v_x, v_k) v_k, f \right) \\ &\leq \left\| v_x - \sum_{k=1}^n (v_x, v_k) v_k \right\| \|f\| = \|v_x - \Pi_{V(X_n)}(v_x)\| \|f\|, \end{aligned}$$

thus  $P_{V(X_n)}(x) \leq \|v_x - \Pi_{V(X_n)}(v_x)\|$ , and the equality is actually reached by taking

$$f_x := \frac{v_x - \Pi_{V(X_n)}(v_x)}{\|v_x - \Pi_{V(X_n)}(v_x)\|}.$$

□

It is then clear that, for any orthonormal basis  $\{v_k\}_{k=1}^n$  of  $V(X_n)$ , we have

$$P_{V(X_n)}(x)^2 = \|K(\cdot, x) - \Pi_{V(X_n)}(K(\cdot, x))\|^2 = K(x, x) - \sum_{k=1}^n v_k(x)^2, \quad (5)$$

and in particular, by using the Newton basis as an orthonormal basis,

$$P_{V(X_n)}(x)^2 = P_{V(X_{n-1})}(x)^2 - v_n(x)^2. \quad (6)$$

This means that the Power Function is decreasing whatever the choice of  $x_n \in \Omega \setminus X_n$  is, i.e., we have

$$\|P_{V(X_n)}\|_{L_\infty(\Omega)} \leq \|P_{V(X_{n-1})}\|_{L_\infty(\Omega)}.$$

### 3 Power Function and Kolmogorov width

We can now provide a connection between the Power Function and the *Kolmogorov  $n$ -width* of a particular compact subset of  $\mathcal{H}_K(\Omega)$ . Recall that for a subset  $\mathcal{V} \subset \mathcal{H}_K(\Omega)$  the Kolmogorov  $n$ -width of  $\mathcal{V}$  in the Hilbert space  $\mathcal{H}_K(\Omega)$  is defined as (see e.g. [14])

$$d_n(\mathcal{V}, \mathcal{H}_K(\Omega)) := \inf_{\substack{V_n \subset \mathcal{H}_K(\Omega) \\ \dim(V_n) = n}} \sup_{f \in \mathcal{V}} \|f - \Pi_{V_n}(f)\| = \inf_{\substack{V_n \subset \mathcal{H}_K(\Omega) \\ \dim(V_n) = n}} E(\mathcal{V}, V_n),$$

where the term  $E(\mathcal{V}, V_n)$  represents the worst-case error in approximating elements of  $\mathcal{V}$  by means of elements of the linear subspace  $V_n$ . One has  $d_n(\mathcal{V}, \mathcal{H}_K(\Omega)) \leq \sup_{f \in \mathcal{V}} \|f\|$ , and in particular we allow  $d_n(\mathcal{V}, \mathcal{H}_K(\Omega)) = +\infty$  for unbounded sets.

In order to analyze the connection between  $P_{V_n}$  and  $d_n$ , we recall that a generalized interpolation operator can be defined for any  $n$ -dimensional linear subspace  $V_n$  of  $\mathcal{H}_K(\Omega)$ , not necessarily in the form  $V(X_n)$ , simply by considering the orthogonal projection operator  $\Pi_{V_n} : \Omega \rightarrow V_n$  as a generalized interpolation operator. A generalized Power Function can be defined also in this case by directly using the definition (2) (see [15]), and Lemma 2.3 still holds with the same proof, i.e.,

$$P_{V_n}(x) = \|K(\cdot, x) - \Pi_{V_n}(K(\cdot, x))\| \quad \text{for all } x \in \Omega. \quad (7)$$

With this characterization at hand, it comes easy to provide a connection with Kolmogorov widths. Namely, for any subset  $\Omega_h \subseteq \Omega$ , we can define the subset  $\mathcal{V}(\Omega_h) := \{K(\cdot, x), x \in \Omega_h\} \subset \mathcal{H}_K(\Omega)$ . Thanks to (5) and Lemma 2.3, it is clear that

$$E(\mathcal{V}(\Omega_h), V_n) = \sup_{f \in \mathcal{V}(\Omega_h)} \|f - \Pi_{V_n}(f)\| = \sup_{x \in \Omega_h} P_{V_n}(x) = \|P_{V_n}\|_{L_\infty(\Omega_h)}. \quad (8)$$

We have then the following.

**Lemma 3.1.** Let  $\Omega_h \subseteq \Omega$ . If there exist point sets  $\{X_n\}_n \subset \Omega$ , each of  $n$  pairwise distinct points, and a sequence  $\{\gamma_n\}_n \subset \mathbb{R}$  such that  $\|P_{V(X_n)}\|_{L_\infty(\Omega_h)} \leq \gamma_n$ , then

$$d_n(\mathcal{V}(\Omega_h), \mathcal{H}_K(\Omega)) \leq \gamma_n, \quad (9)$$

and  $\mathcal{V}(\Omega_h)$  is compact in  $\mathcal{H}_K(\Omega)$  if  $\lim_{n \rightarrow \infty} \gamma_n = 0$ . In particular, in the setting of Corollary 2.2,

(a) if  $K$  has finite smoothness  $\beta \in \mathbb{N}$ ,

$$d_n(\mathcal{V}(\Omega_h), \mathcal{H}_K(\Omega)) \leq c_1 n^{-\frac{\beta}{d} + \frac{1}{2}},$$

(b) if  $K$  is infinitely smooth,

$$d_n(\mathcal{V}(\Omega_h), \mathcal{H}_K(\Omega)) \leq c_2 \exp(-c_3 n^{1/d}),$$

and in both cases  $\mathcal{V}(\Omega_h)$  is compact in  $\mathcal{H}_K(\Omega)$ .

*Proof.* From (8), and from the definition of the Kolmogorov width, one has

$$\begin{aligned} d_n(\mathcal{V}(\Omega_h), \mathcal{H}_K(\Omega)) &= \inf_{\substack{V_n \subset \mathcal{H}_K(\Omega) \\ \dim(V_n) = n}} \|P_{V_n}\|_{L_\infty(\Omega_h)} \leq \inf_{\substack{X_n \subset \Omega \\ |X_n| = n}} \|P_{V(X_n)}\|_{L_\infty(\Omega_h)} \\ &\leq \inf_{\substack{X_n \subset \Omega \\ |X_n| = n}} \|P_{V(X_n)}\|_{L_\infty(\Omega)}, \end{aligned}$$

where the first inequality follows from restricting the set over which the infimum is computed, and the second one by considering the  $L_\infty$ -norm over the larger set  $\Omega \supseteq \Omega_h$ .

Now, for any  $X_n \subset \Omega$  with  $|X_n| = n$ ,  $\|P_{V(X_n)}\|_{L_\infty(\Omega)}$  is an upper bound on the last term of the above inequalities, being  $X_n$  non necessarily optimal. In particular this holds for the sequence of points  $\{X_n\}_n$  of the statement, with Power Functions bounded by a sequence  $\{\gamma_n\}_n$ , which proves (9). Moreover, according to [14, Prop. 1.2], a set  $\mathcal{V} \subset \mathcal{H}_K(\Omega)$  is compact if and only if it is bounded and  $d_n(\mathcal{V}, \mathcal{H}_K(\Omega)) \rightarrow 0$  as  $n \rightarrow \infty$ . But this is the case for  $\mathcal{V} := \mathcal{V}(\Omega_h)$  whenever  $\lim_{n \rightarrow \infty} \gamma_n = 0$ , since

$$\sup \{\|f\|, f \in \mathcal{V}(\Omega_h)\} = \sup \{\sqrt{K(x, x)}, x \in \Omega_h\} = \sqrt{\Phi(0)} \in \mathbb{R}.$$

In particular, by using the rates of convergence of Corollary 2.2 one gets the estimates (a) and (b) for different kernel smoothness.  $\square$

*Remark 2.* It is clear that this result holds for  $\Omega_h = \Omega$ , and this is indeed the most interesting case. Nevertheless, in actual computations one has generally never access to  $\Omega$ , but only to a subset  $\Omega_h$ , being it an arbitrary discretization required for numerically representing the continuous set, or a large set of data  $\Omega_h = X_N$  coming from an application. In this case, also the optimization required by the greedy algorithm is performed on  $\Omega_h$ , and not on  $\Omega$ . By explicitly considering this restricted set in the above Kolmogorov width, we will be able to give exact bounds on the convergence of the  $P$ -greedy algorithm when executed over  $\Omega_h$ , as will be explained in the next Section.

## 4 Convergence rate of the $P$ -greedy algorithm

The discussion of the previous Section is what we need to provide a connection to the theory of greedy algorithms developed in the papers [1, 4]. Indeed, the  $P$ -greedy algorithm can be rewritten in terms of the so-called *strong* greedy algorithm of these papers as follows.

We consider a target compact set  $\mathcal{V}(\Omega_h) \subset \mathcal{H}_K(\Omega)$ , and, for  $n \geq 1$ , we select a sequence of functions  $\{f_k\}_k \subset \mathcal{V}(\Omega_h)$  such that  $\text{span}\{f_k, 1 \leq k \leq n\}$  is an approximation of  $\mathcal{V}$ . The first element  $f_1$  is defined as

$$f_1 := \arg \max_{f \in \mathcal{V}(\Omega_h)} \|f\| = \arg \max_{x \in \Omega_h} \sqrt{K(x, x)}.$$

Assuming  $f_1, \dots, f_{n-1}$  has been selected and  $V_{n-1} := \text{span}\{f_1, \dots, f_{n-1}\} = \text{span}\{K(\cdot, x_k), x_k \in X_{n-1}\}$ , the next element is

$$f_n := \arg \max_{f \in \mathcal{V}(\Omega_h)} E(f, V_{n-1}) = \arg \max_{x \in \Omega_h} P_{V_{n-1}}(x) = \arg \max_{x \in \Omega_h \setminus X_{n-1}} P_{V_{n-1}}(x),$$

where we used in the last step the fact that  $P_{V(X_{n-1})} = 0$  on  $X_{n-1}$ . It is clear that the present algorithm is exactly the  $P$ -greedy algorithm. Observe also that the orthonormal system  $\{f_k^*\}_k$  obtained in the cited papers by Gram-Schmidt orthogonalization of  $\{f_n\}_n$  is precisely the Newton basis  $\{v_n\}_n$ .

In this case, thanks to the compactness of  $\mathcal{V}(\Omega_h)$ , we can use the estimates of [4, Corollary 3.3], which are in fact bounds on  $\max_{f \in \mathcal{V}(\Omega_h)} E(f, V_n)$ , i.e., on  $\|P_{V(X_{n-1})}\|_{L_\infty(\Omega_h)}$ , in terms of  $d_n(\mathcal{V}(\Omega_h), \mathcal{H}_K(\Omega))$ . In our case they read as follows.

**Theorem 4.1.** Assume  $K, \Omega$  satisfy the hypothesis of Corollary 2.2. The  $P$ -greedy algorithm applied to  $\Omega_h \subseteq \Omega$  gives point sets  $X_n \subseteq \Omega_h$  with the following decay of the Power Function.

(a) If  $K$  has finite smoothness  $\beta \in \mathbb{N}$ ,

$$\|P_{V(X_n)}\|_{L_\infty(\Omega_h)} \leq \hat{c}_1 n^{-\frac{\beta}{d} + \frac{1}{2}}.$$

(b) If  $K$  has infinitely many smooth derivatives,

$$\|P_{V(X_n)}\|_{L_\infty(\Omega_h)} \leq \hat{c}_2 \exp(-\hat{c}_3 n^{1/d}).$$

The constants  $\hat{c}_1, \hat{c}_2, \hat{c}_3$  do not depend on  $n$  and can be computed as

$$\hat{c}_1 := c_1 2^{\frac{5\beta}{d} - \frac{3}{2}}, \quad \hat{c}_2 := \sqrt{2c_2}, \quad \hat{c}_3 := 2^{-1 - \frac{\beta}{d}} c_3.$$

*Remark 3.* In the case (a) of the above theorem, something more can be deduced on the quality of the approximation provided by the  $P$ -greedy algorithm. Indeed, in this case the native space on  $\Omega = \mathbb{R}^d$  is norm-equivalent to the Sobolev space  $W_2^\beta(\Omega)$  and in these spaces the behavior of the best approximation is well understood. Indeed, denoting as  $B_1 \subset \mathcal{H}_K(\Omega)$  the unit ball in the native space and by  $\Pi_{L_2, V_n}$  the  $L_2(\Omega)$ -orthogonal projection into a linear subspace  $V_n \subset L_2(\Omega)$ , we can consider the Kolmogorov width

$$d_n(B_1, L_2(\Omega)) := \inf_{V_n \subset L_2(\Omega)} \sup_{f \in B_1} \|f - \Pi_{L_2, V_n}(f)\|_{L_2(\Omega)},$$

which is known to behave (see [11]) as

$$cn^{-\beta/d} \leq d_n(B_1, L_2(\Omega)) \leq Cn^{-\beta/d}, \quad c, C > 0.$$

Moreover, it has been proven in [18] that the same rate (in fact precisely the same value) can be obtained by considering subspaces  $V_n \subset \mathcal{H}_K(\Omega)$  and the  $\mathcal{H}_K(\Omega)$ -orthogonal projection  $\Pi_{V_n}$  (the same one we used so far in this paper), i.e.,

$$\kappa_n(B_1, L_2(\Omega)) := \inf_{V_n \subset \mathcal{H}_K(\Omega)} \sup_{f \in B_1} \|f - \Pi_{V_n}(f)\|_{L_2(\Omega)} = d_n(B_1, L_2(\Omega)).$$

Unfortunately, the above infimum is reached by considering a subspace generated by eigenfunctions of a particular integral operator, which are not known in general (see e.g. [15]). Nevertheless, again in the paper [18] it has been observed that standard kernel-based approximation can reach almost the same asymptotic order of convergence in a bounded set  $\Omega \subset \mathbb{R}^d$ . Indeed, by considering an asymptotically uniformly distributed point sequence  $\{X_n\}_n \subset \Omega$ , Corollary 2.2 and the error bound (3) give

$$\begin{aligned} \sup_{f \in B_1} \|f - \Pi_{V(X_n)}(f)\|_{L_2(\Omega)} &\leq \sup_{f \in B_1} \|f\| \|P_{V(X_n)}\|_{L_2(\Omega)} \\ &\leq \text{meas}(\Omega)^{1/2} \|P_{V(X_n)}\|_{L_\infty(\Omega)} \leq \text{meas}(\Omega)^{1/2} c_1 n^{-\frac{\beta}{d} + \frac{1}{2}}, \end{aligned}$$

where  $\text{meas}(\cdot)$  is the Lebesgue measure.

Thanks to Theorem 4.1, this asymptotically quasi-optimal rate of convergence in Sobolev spaces can be reached also by greedy techniques. Moreover, we will see in Section 5 that the actual convergence of the  $P$ -greedy algorithm seems to be in fact of rate  $n^{-\beta/d}$ , and not only  $n^{-\beta/d+1/2}$  as proven here.

#### 4.1 Distribution of the selected points

The previous result has also some consequence on the distribution of the points selected by the  $P$ -greedy algorithm. When the algorithm was introduced in [3], the authors noticed that the points were placed in an asymptotically uniform way inside  $\Omega$ , and they also proved the following result.

**Theorem 4.2.** *Assume  $K$  and  $\Omega$  satisfy the same assumptions as in Theorem 2.1, with  $\hat{\Phi}(\omega) \sim (1 + \|\omega\|_2^2)^\beta$ ,  $\beta > d/2$ . Then for any  $\alpha > \beta$ , there exists a constant  $M_\alpha > 0$  such that, if  $\varepsilon_n > 0$  and  $X_n \subset \Omega$  satisfy*

$$\|f - \Pi_{V(X_n)}\|_{L_\infty(\Omega)} \leq \varepsilon_n \|f\| \quad \text{for all } f \in \mathcal{H}_K(\Omega),$$

then

$$h_{X_n, \Omega} \leq M_\alpha \varepsilon_n^{1/(\alpha-d/2)}.$$

Unfortunately, the rate of convergence of Theorem 1.1 was not enough to conclude that the points are asymptotically uniformly distributed, which is instead possible with the bounds of Theorem 4.1.

**Corollary 4.3.** *Under the same assumptions of the previous Theorem, there exists a constant  $c > 0$  such that, for any  $n \in \mathbb{N}$ , the sets  $\{X_n\}_n$  selected by the  $P$ -greedy algorithm satisfy*

$$h_{X_n, \Omega} \leq cn^{-\frac{1}{d}(1-\varepsilon)},$$

for any  $\varepsilon \in (0, 1)$ , where  $c$  is independent of  $n$ .

*Proof.* Under these assumptions and from Theorem 4.1 we have  $\varepsilon \leq \hat{c}_1 n^{-\frac{\beta}{d} + \frac{1}{2}}$ . Theorem 4.2 then implies that, for all  $\alpha > \beta$ ,

$$h_{X_n, \Omega} \leq M_\alpha \left( \hat{c}_1 n^{-\frac{\beta}{d} + \frac{1}{2}} \right)^{\frac{1}{\alpha-d/2}},$$

and for any  $\varepsilon \in (0, 1)$  there is an  $\alpha > \beta$  such that the exponent can be written as follows

$$\left( -\frac{\beta}{d} + \frac{1}{2} \right) \left( \frac{1}{\alpha-d/2} \right) = -\frac{1}{d} \left( \frac{\beta-d/2}{\alpha-d/2} \right) = -\frac{1}{d} (1-\varepsilon).$$

□

We remark that the above result does not apply in the case of infinitely smooth kernels. On one side, the proof of Theorem 4.2 uses tools which are related to Sobolev spaces, hence to kernels of finite smoothness. On the other hand, one could not expect a decay of the fill distance with exponential speed with respect to the number of points. Nevertheless, it is plausible to expect that also for kernels of this kind an algebraic convergence of the fill distance is possible, even if it is not clear with what rate.

## 5 Numerical experiments

We test in this Section the theoretical rates obtained in Theorem 4.1 for kernels of different smoothness and in different space dimensions.

In order to ensure the validity of the hypothesis on  $\Omega$ , in all the following experiments we consider as a base domain the unit ball  $\Omega := \{x \in \mathbb{R}^d, \|x\|_2 \leq 1\}$ , for  $d = 1, 2, 3$ . Furthermore, to implement numerical calculations  $\Omega$  is represented by a discretization  $\Omega_h \subset \Omega$ , obtained by intersecting a uniform grid in  $[-1, 1]^d$  with the unit ball. The grids have respectively  $10^4$  ( $d = 1$ ),  $114^2$  ( $d = 2$ ),  $28^3$  ( $d = 3$ ) points, so that the resulting number of points of  $\Omega_h$  is approximately  $10^4$ . The point selection, and both the computation of the supremum norm and of the fill distance are performed on this discretized set. We point out that this choice of  $\Omega_h$  is somehow arbitrary, but it is justified in view of Remark 2.

As kernels we consider radial basis functions which satisfy the requirements of the convergence results, namely the Gaussian kernel  $G$  defined by  $\Phi(r) := \exp(-(\varepsilon r)^2)$ , as an infinitely smooth kernel, and the Wendland kernels  $W_{\beta,d}$  for  $\beta = 2, 3$ , as kernels of finite smoothness  $\beta$  (see [19]). We consider unscaled version of the kernels, i.e., in all the experiments the shape parameter  $\varepsilon$  is fixed to the value  $\varepsilon = 1$ .

The  $P$ -greedy algorithm is applied via a matrix-free implementation of the Newton basis, based on [12, 13]. The code can be found on the website of G. Santin<sup>1</sup>. The algorithm is stopped by means of a tolerance of  $\tau = 10^{-15}$  on the maximal value of the square of the Power Function on  $\Omega_h$ , or a maximum expansion size of  $n = 1000$ . We remark that the present implementation actually computes the square of the Power Function via the formula (6), so numerical cancellation can happen when  $\|P_{V(x_n)}\|_{L^\infty(\Omega_h)}^2$  is close to the machine precision. We remark also that for some class of kernels it is possible to employ a more stable and accurate computation method for the Power Function (see [8, Section 14.1.1]), even if it is not clear if and how it applies to an iterative computation like the present one.

The numerical decay rate of the Power Function for the Gaussian kernel are presented in Figure 1, and the experiments confirm the expected decay rates of Theorem 4.1. The coefficients  $\hat{c}_2, \hat{c}_3$  are estimated numerically, and are reported in Table 1. They are computed by linear least-squares approximation of the logarithm of the maxima of the Power Function, for values of  $n$  as in Figure 1.

	$d = 1$	$d = 2$	$d = 3$
$\hat{c}_2$	26.94	148.87	580.67
$\hat{c}_3$	1.22	1.80	2.31

**Table 1:** Estimated coefficients for the decay rate of the Power Function with the Gaussian kernel.

Figure 2 shows the results of the same experiment for the Wendland kernels. Here we can observe that the theoretical rate of Theorem 4.1 seems to be not sharp, and instead the rate of Remark 3 seems to be valid. We report both rates in the figure, computed with scaling coefficients as in Table 2. These results could be an insight into the optimality of kernel methods in Sobolev spaces, where the optimal decay rate can be reached also by greedy methods.

	$d = 1$	$d = 2$	$d = 3$		$d = 1$	$d = 2$	$d = 3$
$\beta = 2$	0.003	0.01	0.02	$\beta = 2$	0.08	0.34	0.49
$\beta = 3$	0.03	0.02	0.02	$\beta = 3$	0.32	0.52	0.67

**Table 2:** Estimated coefficient  $\hat{c}_1$  for the decay rate of the Power Function with the Wendland kernels for the theoretical rate of convergence (left) and the modified rate of convergence (right).

In the same setting, also the fill distance of the selected points is computed. The results are shown in Figure 3, and they confirm the decay rate expected from Corollary 4.3. Also in this case the theoretical rate is scaled by a positive coefficient. Observe that in this case the use of a discretized set  $\Omega_h$  in place of  $\Omega$  influences the results of the computations.

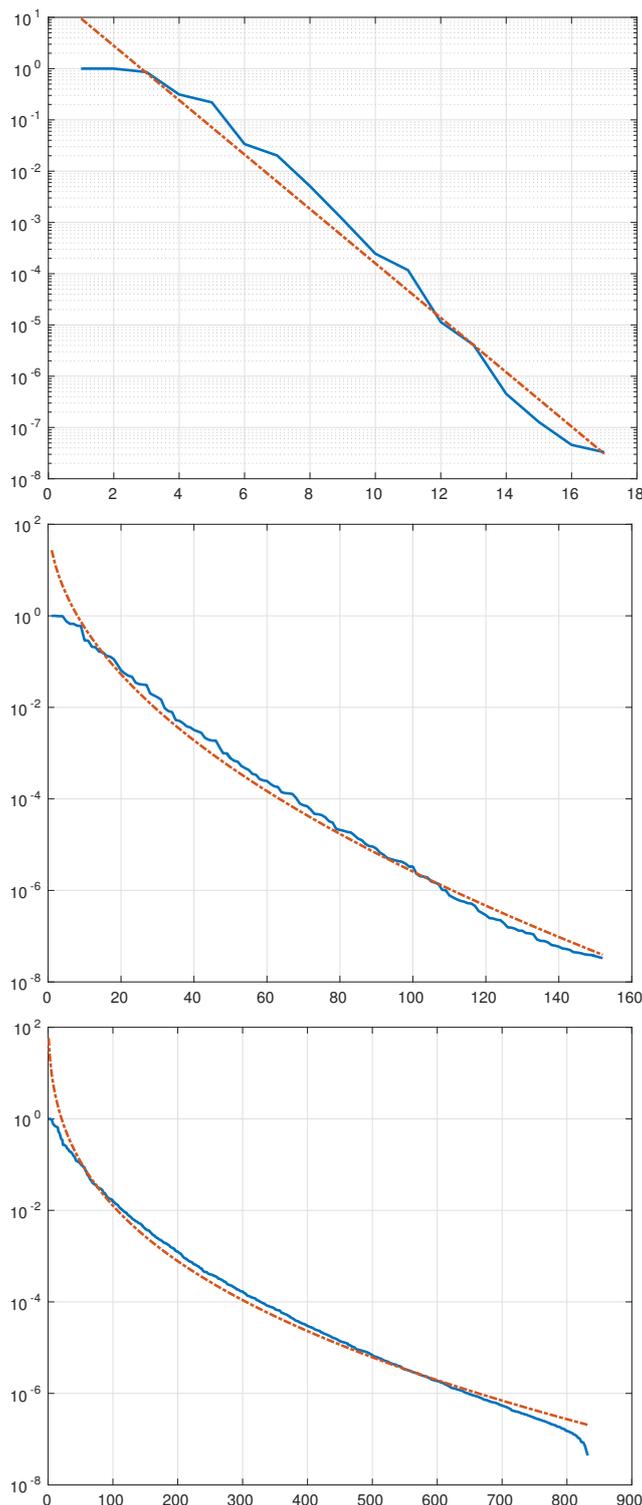
Finally, we show an example demonstrating the quasi-uniformity of the selected points on a non trivial domain. To this end, we run the algorithm with the Wendland kernel  $W_{2,2}$  and  $\varepsilon = 1$  on the bivariate domain  $\Omega$  defined as the union of three disks centered in  $(0.4, 0.2)$ ,  $(-0.3, -0.2)$ ,  $(-0.4, 0.6)$  and with radii  $r = 0.6, 0.5, 0.4$ , respectively. The greedy optimization is performed on the set  $\Omega_h$  of 55237 points obtained by intersecting a uniform grid in  $[-1, 1]^2$  with  $\Omega$ . The first 1000 points selected by the algorithm are shown in Figure 4.

## 6 Conclusion and open problems

In this paper we proved that the data-independent  $P$ -greedy algorithm is able to approximate functions from the native space of a given kernel with a rate of convergence which is asymptotically equivalent to the best known, non greedy decay rate. This rate is shown to be even asymptotically quasi-optimal in the case of kernels of finite smoothness generating Sobolev spaces. Furthermore, in this case the points selected by the algorithm are proven to be asymptotically uniformly distributed in the domain.

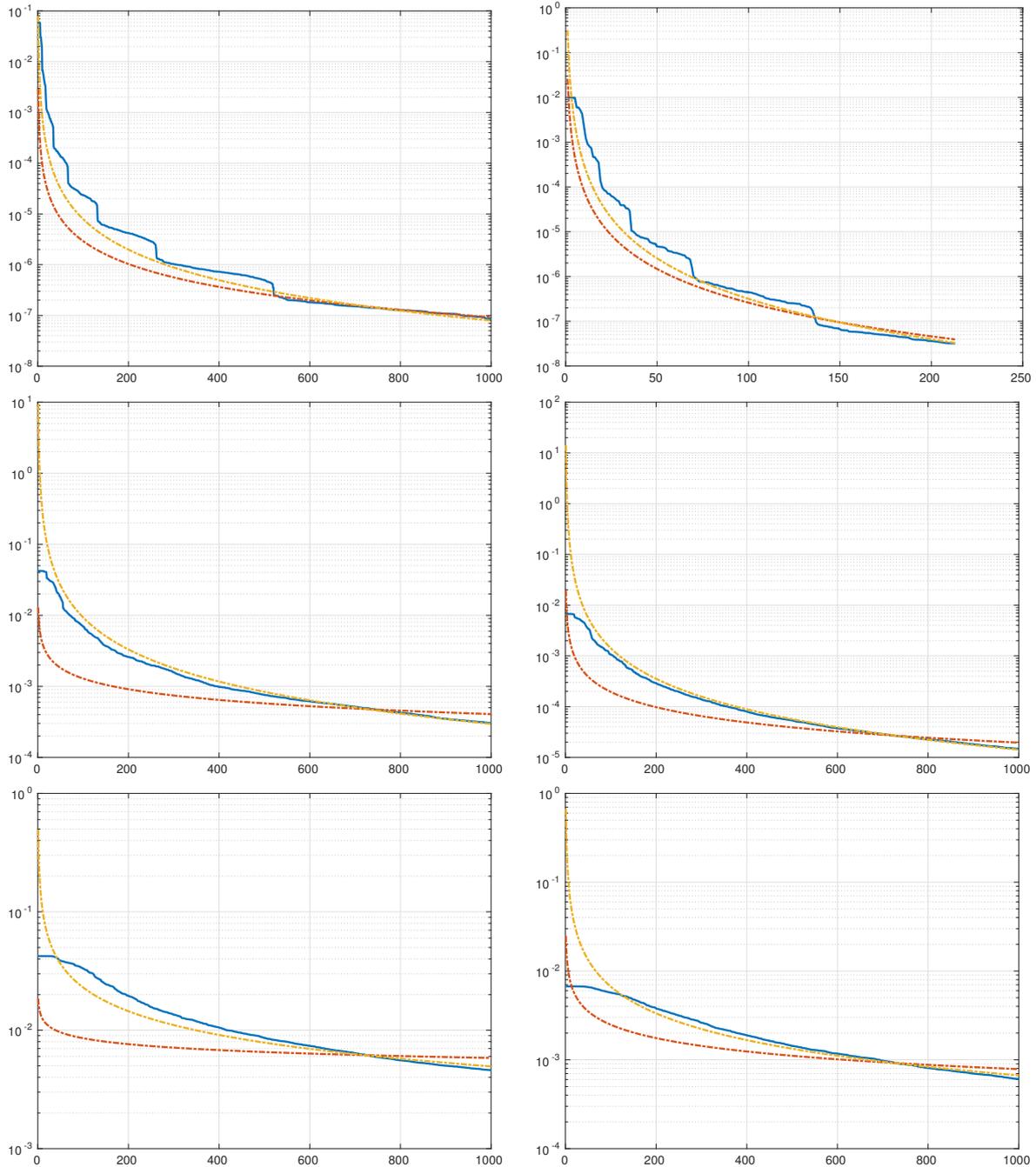
Some interesting questions remain open. First, only in the case of kernels of finite smoothness the points selected by the  $P$ -greedy algorithm can be shown to be asymptotically uniformly distributed. In the case of infinitely smooth kernels, on the contrary, no information can be deduced. Nevertheless, one could expect that a proper combination of algebraic convergence rates can be used to deduce similar properties also in the case of infinitely smooth kernels.

<sup>1</sup><http://www.mathematik.uni-stuttgart.de/fak8/ians/lehrstuhl/agh/orga/people/santin/index.html>



**Figure 1:** Expected theoretical rate of convergence (dotted red lines) and computed decay of the Power Function for the  $P$ -greedy algorithm (solid blue lines), with the setting described in Section 5 and the Gaussian kernel. From top to bottom:  $d = 1, 2, 3$ .

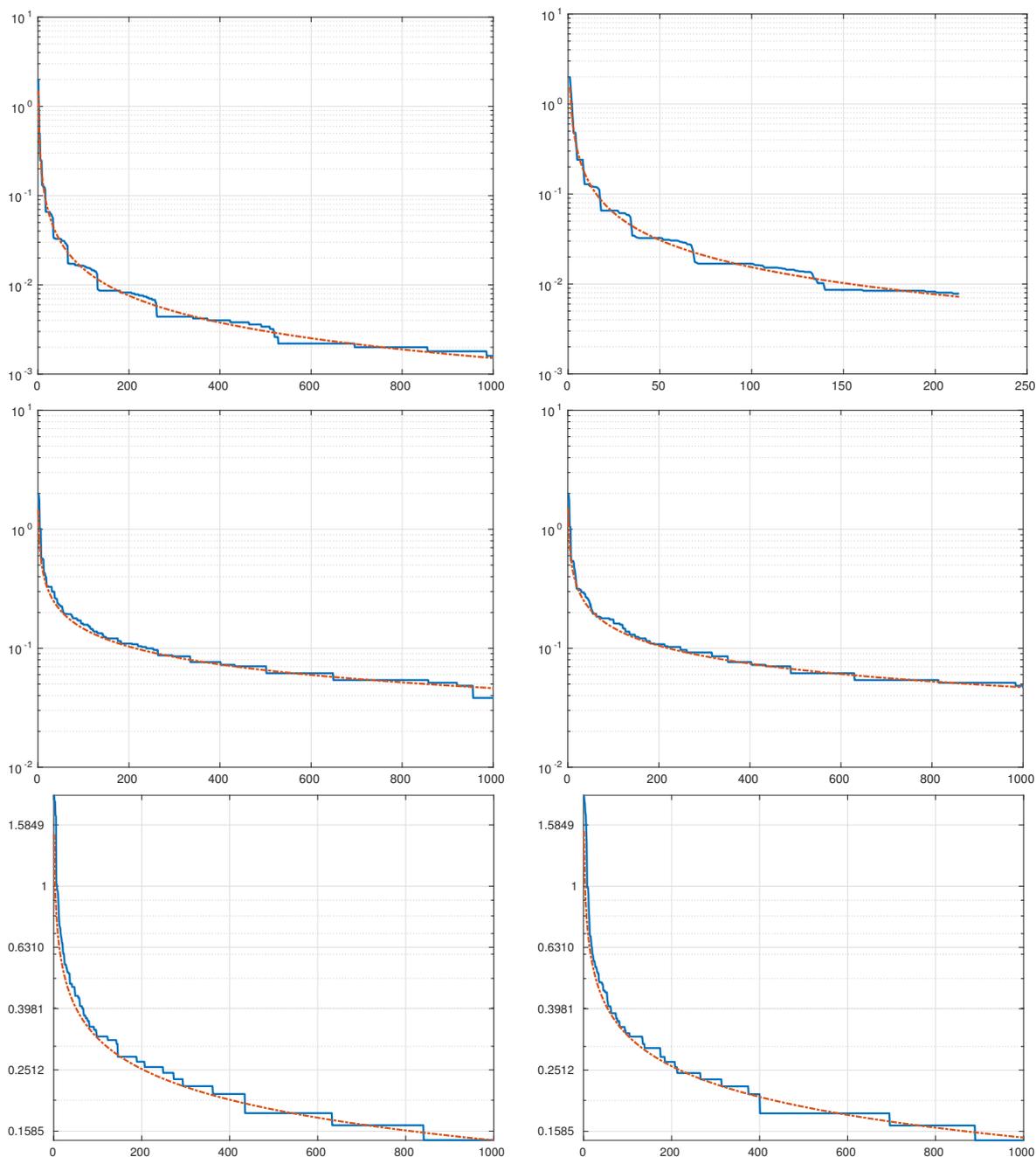
Furthermore, the asymptotically optimal convergence rate of the  $P$ -greedy algorithm leads to questions on the convergence rate of different greedy selection techniques. Indeed, the data-dependent  $f$ -greedy [17, 18] and  $f/P$ -greedy algorithms [12] (i.e., the points are selected to approximate one single function  $f$ , using its evaluations and the Power Function) should provide equal or even better approximation, while at present the known convergence rates are still non completely satisfactory. In particular, the convergence of the  $f$ -greedy procedure is proven to be optimal only in the one-dimensional case  $\Omega \subset \mathbb{R}$ , while in the general



**Figure 2:** Expected theoretical rate of convergence (dotted red lines), improved rate of convergence (dotted yellow lines), and computed decay of the Power Function for the  $P$ -greedy algorithm (solid blue lines), with the setting described in Section 5 and the Wendland kernels, with  $d = 1, 2, 3$  (from top to bottom) and  $\beta = 2, 3$  (from left to right).

case only an order  $n^{-1}$  is known (see [12]), and  $f/P$ -greedy has a proven convergence order of  $n^{-1/2}$ , even if the bound in dimension-independent ([21]). Further work is required on this side.

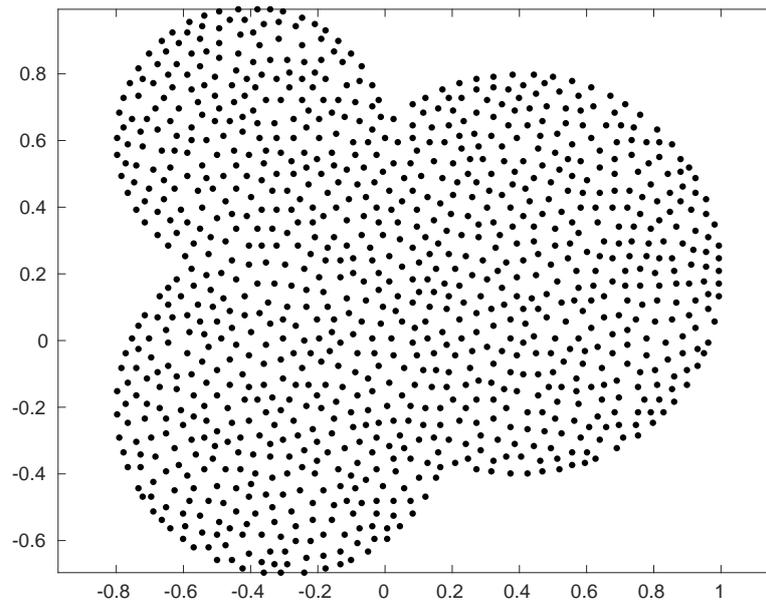
**Acknowledgements:** We thank Dominik Wittwar for fruitful discussions. Further, we thank the anonymous referees for many valuable comments on the first version of the manuscript. Both authors would like to thank the German Research Foundation (DFG) for financial support within the Cluster of Excellence in Simulation Technology (EXC 310/1) at the University of Stuttgart.



**Figure 3:** Expected decay of the fill distance (red dotted lines) and computed decay (solid blue lines), with the setting described in Section 5 for the Wendland kernels with  $\beta = 2, 3$  (from left to right) and  $d = 1, 2, 3$  (from top to bottom).

## References

- [1] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk. Convergence rates for greedy algorithms in reduced basis methods. *SIAM J. Math. Anal.*, 43(3):1457–1472, 2011.
- [2] M. D. Buhmann. *Radial basis functions: theory and implementations*, volume 12 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2003.
- [3] S. De Marchi, R. Schaback, and H. Wendland. Near-optimal data-independent point locations for radial basis function interpolation. *Adv. Comput. Math.*, 23(3):317–330, 2005.
- [4] R. DeVore, G. Petrova, and P. Wojtaszczyk. Greedy algorithms for reduced bases in Banach spaces. *Constr. Approx.*, 37(3):455–466, 2013.
- [5] N. Dyn, M. Floater, and A. Iske. Adaptive thinning for bivariate scattered data. *J. Comput. Appl. Math.*, 145(2):505 – 517, 2002.
- [6] N. Dyn and P. Kozlov. Adaptive thinning of centers for approximation of a large data set by radial functions. *J. Approx. Theory*, 200:153 – 169, 2015.



**Figure 4:** Points selected by the algorithm on a test domain given by the union of three disks.

- [7] G. E. Fasshauer. *Meshfree Approximation Methods with MATLAB*, volume 6 of *Interdisciplinary Mathematical Sciences*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2007. With 1 CD-ROM (Windows, Macintosh and UNIX).
- [8] G. E. Fasshauer and M. McCourt. *Kernel-Based Approximation Methods Using MATLAB*, volume 19 of *Interdisciplinary Mathematical Sciences*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2015.
- [9] M. S. Floater and A. Iske. Thinning algorithms for scattered data interpolation. *BIT Numerical Mathematics*, 38(4):705–720, 1998.
- [10] A. Iske. Optimal distribution of centers for radial basis function methods. Technical report, Technische Universität München, Report TUM-M0003, 2000.
- [11] J. W. Jerome. On  $n$ -widths in Sobolev spaces and applications to elliptic boundary value problems. *J. Math. Anal. Appl.*, 29:201–215, 1970.
- [12] S. Müller and R. Schaback. A Newton basis for kernel spaces. *J. Approx. Theory*, 161(2):645–655, 2009.
- [13] M. Pazouki and R. Schaback. Bases for kernel-based spaces. *J. Comput. Appl. Math.*, 236(4):575 – 588, 2011.
- [14] A. Pinkus.  *$n$ -Widths in Approximation Theory*, volume 7 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1985.
- [15] G. Santin and R. Schaback. Approximation of eigenfunctions in kernel-based spaces. *Adv. Comput. Math.*, 42(4):973–993, 2016.
- [16] R. Schaback. Error estimates and condition numbers for radial basis function interpolation. *Adv. Comput. Math.*, 3(3):251–264, 1995.
- [17] R. Schaback and H. Wendland. Adaptive greedy techniques for approximate solution of large RBF systems. *Numer. Algorithms*, 24(3):239–254, 2000.
- [18] R. Schaback and H. Wendland. Approximation by positive definite kernels. In M. Buhmann and D. Mache, editors, *Advanced Problems in Constructive Approximation*, volume 142 of *International Series in Numerical Mathematics*, pages 203–221, 2002.
- [19] H. Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.*, 4(1):389–396, 1995.
- [20] H. Wendland. *Scattered Data Approximation*, volume 17 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2005.
- [21] D. Wirtz and B. Haasdonk. A vectorial kernel orthogonal greedy algorithm. *Dolomites Res. Notes Approx.*, 6:83–100, 2013. Proceedings of DWCAA12.