Institut für Numerische Simulation
Rheinische Friedrich-Wilhelms-Universität Bonn
Wegelerstraße 6 • 53115 Bonn • Germany
phone +49 228 73-3427 • fax +49 228 73-7527
www.ins.uni-bonn.de

C. Rieger

Incremental Kernel Based Approximations for Bayesian Inverse Problems

INS Preprint No. 1807

May 2018

INCREMENTAL KERNEL BASED APPROXIMATIONS FOR BAYESIAN INVERSE PROBLEMS

CHRISTIAN RIEGER

ZUSAMMENFASSUNG. We provide an interpretation for the covariance of the predictive process of Bayesian Gaussian process regression as reproducing kernel of a subset of the Cameron Martin space of the prior. We demonstrate that this deterministic viewpoint enables us to relate particular greedy methods using that subset kernel to instances of powerful low-rank matrix approximation techniques such as *adaptive cross approximation* or *pivoted Cholesky decomposition*. In particular, we can show convergence results for such algorithms which appear to be novel in the case of finitely smooth kernels.

Moreover, we consider the inverse problem to reconstruct a parametrized diffusion coefficient from point evaluations of the solution to a diffusion equation with that parametrized coefficient. To this end, we present a Gaussian process regression based approach to approximate the observation operator. The error estimates for this approximation methods are capable to take deterministic model errors explicitly into account. Finally, we show how the findings about incremental low rank approximations can be applied to these reconstruction problems.

1. INTRODUCTION

Gaussians process regression is an important tool to reconstruct a function from finitely many data. The Gaussian process has many different interpretations. We outline connections between Gaussian process regression and numerical approximation methods in reproducing kernel Hilbert spaces (RKHS). Here, we especially focus on the deterministic interpretation of the predictive process and its covariance. It turns out that the covariance of the predictive process is nothing but the so-called power kernel from [10]. The power kernel construction naturally shows up in the *adaptive cross approximation* method, though this powerful method can be formulated in much more generality, using e.g. non-symmetric kernels, see [1]. Algorithmically both methods can be related to the variants of the Cholesky decomposition, see [6]. These findings allow very efficient low-rank approximations of the covariance of the predictive process. In particular, the interpretation in terms of reproducing kernel methods allows to derive an error analysis for the case of finitely smooth covariance functions.

As an application, we apply the general framework of Gaussian process regression to an inverse problem. The inverse problem is to reconstruct a parametrized diffusion coefficient from point evaluations of solutions to the stationary diffusion equation. Here, we follow [15], where a very general theory to bound the reconstruction error for the diffusion coefficients is presented. The difference to the approach in [15] is that we do not assume to have arbitrarily precise solutions to the differential equation. Of course, this is a constructed model problem. If we had all data available

Date: 8. Mai 2018.

to compute the solution to the differential equation, there is no need to recover the diffusion coefficient. So our analysis might be directly applied if the solution to a differential equation is stored as data but the information about the diffusion coefficients are lost. We, however, treat this numerical model error as a prototype for more general types of model errors. Such a model error might be the wrong differential equation to describe the physical processes accurately, or many more. We are convinced that many of those model errors are deterministic in nature and hence, we present an analysis for this case.
The necessity to include the discretization error of the solution to the differential equation is also discussed in a statistical framework in [2]. We pursue a similar approach as we also employ a kernel method to solve the differential equation. As already encountered in Section 2, those kernel methods can be seen as Gaussian process regression. We, however, employ a deterministic a priori error analysis as developed in [4]. To our mind, the deterministic error model is more realistic in the context of model errors. To account for such a deterministic error, we employ techniques from machine learning. Moreover, we can use techniques from low-rank matrix approximation methods to distribute the points according to the covariance of the predictive distribution function in the spirit of [5, 7]. We develop a new error analysis for the inverse reconstruction problem as outlined above. Moreover, we discuss how techniques from adaptive cross approximation can be applied in such problems especially in the design of new observation points.
The remainder of the manuscript is organized as follows: In Section 2, we recall basic facts from the theory of Gaussian process regression and reproducing kernel Hilbert space methods. Moreover, we present a new interpretation of the adaptive cross approximation in the setting of reproducing kernel Hilbert space methods. In Section 3, we present the inverse problem to determine the parametrized diffusion coefficient from point evaluations of the stationary diffusion equation with that coefficient. In Section 4, we present a numerical scheme to solve the differential equation and present its error analysis. Finally, we present an error analysis for the inverse problem in Section 5. We end with some concluding remarks in Section 6.

2. Gaussian process regression and adaptive cross approximation

Here, we follow [13, Chapter 2.2]. Let $\mathbb{X} \subset \mathbb{R}^d$. We consider a Gaussian process

$$f \sim \nu_0 = \mathcal{N}(0, \mathcal{C}) \quad \text{or in other notation} \quad f \sim \mathrm{GP}(0, \kappa).$$

Let $\Xi_N = \{\xi_1, \dots, \xi_N\} \subset \mathbb{X}$ be a discrete set. The goal is to extract information about the function f from finitely many point values. We get for the distribution

$$f|_{\Xi_N} \sim \mathrm{GP}(\mathbf{0}, S_{\Xi_N} S_{\Xi_N}^\star).$$

In order to formulate the problem more precisely, we need to specify the prior or the Gaussian measure. We consider a symmetric positive definite kernel function $\kappa : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$. Then, there is a uniquely defined Hilbert space $\mathcal{H}_\kappa(\mathbb{X}) \subset \{f : \mathbb{X} \to \mathbb{R}\}$ which is called reproducing kernel Hilbert space (RKHS) and satisfies

$$\begin{aligned} \kappa_x &:= \kappa(\cdot, x) \in \mathcal{H}_\kappa(\mathbb{X}) \quad \text{for all } x \in \mathbb{X} \\ f(x) &= (f, \kappa_x)_{\mathcal{H}_\kappa(\mathbb{X})} \quad \text{for all} f \in \mathcal{H}_\kappa(\mathbb{X}) \text{ and for all } x \in \mathbb{X}. \end{aligned}$$

The integral operator

$$\mathcal{C}_\kappa : L_2(\mathbb{X}) \to L_2(\mathbb{X}), \quad f \mapsto \mathcal{C}_\kappa(f) = \int_{\mathbb{X}} \kappa(\cdot, y) f(y) dy \tag{1}$$

gives rise to a (centered) Gaussian measure $\nu_0 = \mathcal{N}(0, \mathcal{C}_\kappa)$ on $L_2(\mathbb{X})$. Furthermore we have

$$\mathrm{Range}(\mathcal{C}_\kappa^{\frac{1}{2}}) = \mathcal{H}_\kappa(\mathbb{X}) \subset L_2(\mathbb{X}),$$

i.e., the RKHS is the Cameron Martin space on ν_0. This means that we choose ν_0 as prior. We consider the case that $\nu_0(C(\mathbb{X})) = 1$ which requires some regularity of the kernel $\kappa : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ and is in our setting usually related to Sobolev embeddings. We introduce the maps

$$\begin{aligned} S_{\Xi_N} &: \mathcal{H}_\kappa(\mathbb{X}) \to \mathbb{R}^N, \quad f \mapsto (f(\xi_1), \dots, f(\xi_N))^\top \\ S^\star_{\Xi_N} &: \mathbb{R}^N \to \mathcal{H}_\kappa(\mathbb{X}), \quad \boldsymbol{c} \mapsto \sum_{\xi_j \in \Xi_N} c_j \kappa_{\xi_j}. \end{aligned}$$

The notation is justified since we have

$$(\boldsymbol{d}, S_{\Xi_N} f)_{\ell_2(\Xi_N)} = \sum_{j=1}^N d_j f(\xi_j) = \left(f, S^\star_{\Xi_N} d\right)_{\mathcal{H}_\kappa(\mathbb{X})} \quad \text{for all } f \in \mathcal{H}_\kappa(\mathbb{X}) \text{ and all } \boldsymbol{d} \in \mathbb{R}^N.$$

We use the notation

$$\mathcal{I}_{\Xi_N} : \mathcal{H}_\kappa(\mathbb{X}) \to \mathcal{H}_\kappa(\mathbb{X}), \quad f \mapsto \mathcal{I}_{\Xi_N; f}(\cdot) = \left(S^\star_{\Xi_N} \left(S_{\Xi_N} S^\star_{\Xi_N} \right)^{-1} S_{\Xi_N}(f) \right)(\cdot).$$

Moreover, we observe that

$$S_{\Xi_N} S^\star_{\Xi_N} = \boldsymbol{K}_{\Xi_N, \Xi_N} = \begin{pmatrix} \kappa(\xi_1, \xi_1) & \dots & \kappa(\xi_1, \xi_N) \\ \vdots & \ddots & \vdots \\ \kappa(\xi_N, \xi_1) & \dots & \kappa(\xi_N, \xi_N) \end{pmatrix} \in \mathbb{R}^{N \times N}$$

is a symmetric and positive definite matrix. We introduce the space

$$\ker(S_{\Xi_N}) = \{ f \in \mathcal{H}_\kappa(\mathbb{X}) \ : \ S_{\Xi_N}(f) = \mathbf{0} \} \subset \mathcal{H}_\kappa(\mathbb{X})$$

and note that its reproducing kernel is given by [10, Section 5]

$$\kappa_{\Xi_N} : \mathbb{X} \times \mathbb{X} \to \mathbb{R}, \quad (x, y) \mapsto \left(\left(\mathrm{Id} - \mathcal{I}_{\Xi_N} \right)^{(1)} \left(\mathrm{Id} - \mathcal{I}_{\Xi_N} \right)^{(2)} \kappa(\cdot, \cdot) \right)(x, y), \tag{2}$$

where the superscript denotes to which argument the function is applied. Following [10], we call the kernel (2) the *power kernel*. The terminology is motivated by the identity

$$\kappa_{\Xi_N}(x, x) = P^2_{\Xi_N}(x) = \mathrm{dist}^2_{\mathcal{H}^*_\kappa(\mathbb{X})} \left(\delta_x, \{ \delta_\xi \ : \ \xi \in \Xi_N \} \right), \tag{3}$$

where P_{Ξ_N} is usually called the *power function*. Then, we have the following representation

$$f(\cdot) | S_{\Xi_N}(f) \sim \mathrm{GP} \left(\mathcal{I}_{\Xi_N; f}(\cdot), \kappa_{\Xi_N}(\cdot, \cdot) \right)$$

for the predictive process, see [13, Eq. (2.19)] or [15, Eq. (3.6)].

2.1. **Recursive interpolation.** We assume that the data points are not given as a single set of points but that they can be viewed as a data stream. We denote by

$$\Xi_N = \{\xi_1, \dots, \xi_{N(\ell)}\} = \Xi_{N-1} \cup \{\xi_N\} \quad \text{and} \quad \Xi_0 = \emptyset.$$

The first observation is that such a data stream fits the asymptotic nature of most error estimates better than a single set of points. The perspective of incremental approximations can also be employed in a low-rank approximation of the kernel matrix. Following [10, Section 5], we have the following recursion for the power kernel (2)

$$\begin{aligned} &\kappa_{\Xi_0,\Xi_0}(x,y) = \kappa(x,y) \quad \text{and} \\ &\kappa_{\Xi_{N+1},\Xi_{N+1}}(x,y) = \kappa_{\Xi_N,\Xi_N}(x,y) - \frac{\kappa_{\Xi_N,\Xi_N}(x,\xi_{N+1})\kappa_{\Xi_N,\Xi_N}(\xi_{N+1},y)}{\kappa_{\Xi_N,\Xi_N}(\xi_{N+1},\xi_{N+1})} \quad \text{for all } x,y \in \mathbb{X}. \end{aligned} \tag{4}$$

Such a recursion appears also for the symmetric kernel case in the adaptive cross approximation, c.f. [1, Eq 16]. Following [10, Section 5], we consider also the matrix version of that recursion. We define the kernel matrix for the n-th power kernel defined in (4) with $n \leq N$

$$\mathbb{K}_{n;\Xi_N,\Xi_N} = \begin{pmatrix} \kappa_{\Xi_n,\Xi_n}(\xi_1,\xi_1) & \dots & \kappa_{\Xi_n,\Xi_n}(\xi_1,\xi_N) \\ \vdots & \ddots & \vdots \\ \kappa_{\Xi_n,\Xi_n}(\xi_N,\xi_1) & \dots & \kappa_{\Xi_n,\Xi_n}(\xi_N,\xi_N) \end{pmatrix} \in \mathbb{R}^{N\times N},$$

evaluated at the point set Ξ_N. This is the matrix C_j from [10, Eq. (5)]. As outlined there, this matrix is the n-th iterate of the *inline Cholesky decomposition*

$$\begin{aligned} &\mathbb{K}_{0;\Xi_N,\Xi_N} = \boldsymbol{K}_{\Xi_N,\Xi_N} \quad \text{and} \\ &\mathbb{K}_{n;\Xi_N,\Xi_N} = \mathbb{K}_{n-1;\Xi_N,\Xi_N} - \frac{1}{\mathbb{K}_{n-1;\Xi_N,\Xi_N}(n,n)} \mathbb{K}_{n-1;\Xi_N,\Xi_N}(:,n) \otimes \mathbb{K}_{n-1;\Xi_N,\Xi_N}(n,:), \end{aligned} \tag{5}$$

using the same MATLAB inspired notation as in [10, Eq. (6)]. Algorithmically, this relates to the recursion of the adaptive cross approximation as in [1, Section 3.1] for symmetric kernels. The iteration (5) yields a Cholesky decomposition of the kernel matrix, i.e.,

$$\boldsymbol{K}_{\Xi_N,\Xi_N} = \mathbb{K}_{N;\Xi_N,\Xi_N} \mathbb{K}^{\top}_{N;\Xi_N,\Xi_N}.$$

As outlined in [12, Section 6], there is a relation between matrix factorizations and special bases, for instance a Cholesky decomposition of the kernel matrix leads to a $(\cdot,\cdot)_{\mathcal{H}_\kappa(\mathbb{X})}$-orthogonal basis. In order to describe this, we define the vector-field

$$\boldsymbol{\kappa}_{\Xi_N} := S^{(2)}_{\Xi_N} \kappa(\cdot,\cdot) = (\kappa(\cdot,\xi_1), \dots, \kappa(\cdot,\xi_N))^{\top}$$

and construct a new *Netwon–type* basis as

$$(\psi_{1;\Xi_N}, \dots, \psi_{N;\Xi_N}) = S^{(2)}_{\Xi_N} \kappa(\cdot,\cdot) \cdot \left(\mathbb{K}^{\top}_{N;\Xi_N,\Xi_N}\right)^{-1}, \tag{6}$$

where $\mathbb{K}^{\top}_{N;\Xi_N,\Xi_N}$ is by construction via the Cholesky decomposition an upper triangular matrix. We have that

$$\operatorname{span}\{\psi_{1;\Xi_N}, \dots, \psi_{N;\Xi_N}\} = \operatorname{span}\{\kappa_{\xi_1}, \dots, \kappa_{\xi_N}\} =: V_{\Xi_N}.$$

The basis ψ_j was already discussed in [11] with a different normalization. The main feature of this basis is that 'old' basis elements remain unchanged if a new point is added. This makes this new basis especially favorable in sequential applications. A related basis change is also described in [1, Lemma 4], with the choices $\hat{\ell}_i = \kappa_{\xi_i}$ and $C = \mathbb{K}_{N;\Xi_N,\Xi_N}^\top$.

2.2. **Low-rank approximation of kernel matrices.** We consider a very fine discrete set of points $\Upsilon_{N_\infty} \subset \mathbb{X}$ with $N_\infty < \infty$. In theory, it would be possible to build the matrix $\boldsymbol{K}_{\Upsilon_{N_\infty},\Upsilon_{N_\infty}} \in \mathbb{R}^{N_\infty \times N_\infty}$ but N_∞ is assumed to be so large that direct inversion of this matrix should be avoided. Low-rank approximation can provide reliable approximations to that matrix which are cheaper to store and allow for faster matrix algorithms. To this end, we fix a precision $\varepsilon_{\text{chol}} > 0$ and try to find a rank $N_{\varepsilon_{\text{chol}}}$ and a matrix $\boldsymbol{A} \in \mathbb{R}^{N_\infty \times N_\infty}$ such that

$$\left\| \boldsymbol{K}_{\Upsilon_{N_\infty},\Upsilon_{N_\infty}} - \boldsymbol{A} \right\| \leq \varepsilon_{\text{chol}}$$

and $\operatorname{Rank}(\boldsymbol{A}) \leq N_{\varepsilon_{\text{chol}}}$. We can use [12, Eq. (20)] to get

$$\kappa(x,y) = \lim_{h_{\Xi_N,\mathbb{X}} \to 0} \sum_{j=1}^{N} \psi_{j;\Xi_N}(x)\psi_{j;\Xi_N}(y), \tag{7}$$

where the limit means that the point set Ξ_N needs to get dense in $\mathbb{X}$ which in turn necessarily implies $N \to \infty$. Moreover, we have the identity (c.f. [12, p. 584])

$$\kappa(x,y) - \kappa_{\Xi_N,\Xi_N}(x,y) = \sum_{j=1}^{N} \psi_{j;\Xi_N}(x)\psi_{j;\Xi_N}(y).$$

The representation of the kernel (7) leads to

$$\boldsymbol{K}_{\Upsilon_{N_\infty},\Upsilon_{N_\infty}} = \lim_{h_{\Xi_N,\mathbb{X}} \to 0} \sum_{j=1}^{N} S_{\Upsilon_{N_\infty}} \psi_{j;\Xi_N} \otimes S_{\Upsilon_{N_\infty}} \psi_{j;\Xi_N}. \tag{8}$$

Such a series expansion motivates a choice for the matrix $\boldsymbol{A}$ as

$$\boldsymbol{A} = \boldsymbol{A}_{\Xi_N} = \sum_{j=1}^{N} S_{\Upsilon_{N_\infty}} \psi_{j;\Xi_N} \otimes S_{\Upsilon_{N_\infty}} \psi_{j;\Xi_N}. \tag{9}$$

The important point which remains open is how to construct the point sets Ξ_n iteratively to achieve a small ϵ_{chol}. Both [1] and [6] provide efficient algorithms to compute such a low rank approximation. As already mentioned in [6, Section 3], both approaches are similar if a *total pivoting* strategy is used in [1]. We will present an other kernel-based derivation for such algorithms. In particular, the kernel based approach also allows for an error analysis in the case of finitely smooth kernels.

Kernel-based power greedy method. For any discrete set $\Xi \subset \mathbb{X}$, there is an error formula for kernel-based interpolation

$$|(f - \mathcal{I}_\Xi(f))(x)| \leq P_\Xi(x) \|f\|_{\mathcal{H}_\kappa(\mathbb{X})} \quad \text{for all } f \in \mathcal{H}_\kappa(\mathbb{X}),$$

using the power function from (3), see [16]. The important feature of this error estimate is that the power function does not depend on the specific function f. Hence, such a error estimate is universal for the whole function space $\mathcal{H}_\kappa(\mathbb{X})$. By definition, we observe

$$\kappa_\Xi(\xi,\xi) = P_\Xi^2(\xi) = 0 \quad \text{for all } \xi \in \Xi.$$

The *power greedy methods* is now defined as follows: Start with a point $\xi_1 \in \mathbb{X}$ and set $\Xi_1 = \{\xi_1\}$. Define iteratively

$$\xi_{n+1} := \arg\max_{x\in\mathbb{X}} P_{\Xi_n}^2(x) \quad \text{and set} \quad \Xi_{n+1} = \Xi_n \cup \{\xi_{n+1}\}. \tag{10}$$

The rationale behind this algorithm is to try to reduce the high values of the power function as quickly as possible and hence to make the power function globally small. For the convergence analysis we refer to [14] for the case of finitely smooth kernels and to [6] for infinitely smooth kernels. As finitely smooth reproducing kernels, we will consider *radial basis functions* with a specific decay of their Fourier transforms. Precisely, we consider translation-invariant kernels Φ such that there is a ϕ satisfying

$$\Phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, x \mapsto \Phi(x-y) = \phi(\|x-y\|_2),$$

where $\phi : \mathbb{R} \to \mathbb{R}$. The decay condition on the Fourier transform reads: there are constants $0 < c_\phi, C_\phi < \infty$ such that for $\lfloor \tau \rfloor > \frac{d}{2}$ it holds that

$$c_\Phi \left(1+\|\omega\|_2^2\right)^{\tau} \le \hat{\Phi}(\omega) \le C_\Phi \left(1+\|\omega\|_2^2\right)^{\tau} \quad \text{for all } \omega \in \mathbb{R}^d. \tag{11}$$

We assume also $\phi \in C^2(\mathbb{R})$. This is for instance satisfied for certain Wendland kernels, see [16, proof of Theorem 11.17]. For those kernels, it turns out that

$$\mathcal{H}_\Phi(\mathbb{R}^d) \cong W_2^\tau(\mathbb{R}^d).$$

Moreover, we obtain [16, Corollary 10.48] that

$$\mathcal{H}_\Phi(\mathbb{X}) \cong W_2^\tau(\mathbb{X})$$

if $\mathbb{X} \subset \mathbb{R}^d$ has a Lipschitz boundary and we use a suitable norm which is equivalent to the standard norm on the Sobolev space. Furthermore, the covariance operator (1) behaves in the assumed way [16, Lemma 10.27].

Convergence of P-greedy algorithm for finitely smooth kernels. The algorithm (10) is numerically infeasible since the maximum over the whole set $\mathbb{X}$ is not computable. Following [14, Section 4], we replace the set $\mathbb{X}$ by the finite set $\Upsilon_{N_\infty} = \{\upsilon_1, \dots, \upsilon_{N_\infty}\}$. The algorithm then reads

$$\begin{aligned} \xi_1 &:= \arg\max_{x\in\Upsilon_{N_\infty}} \kappa(x,x) \quad \text{and} \quad \Xi_1 = \{\xi_1\} \\ \xi_{n+1} &:= \arg\max_{x\in\Upsilon_{N_\infty}\setminus\Xi_n} P_{\Xi_n}^2(x) \quad \text{and} \quad \Xi_{n+1} = \Xi_n \cup \{\xi_{n+1}\}. \end{aligned} \tag{12}$$

Then, we have by [14, Theorem 4.1]

$$\|P_{\Xi_n}\|_{\ell_\infty(\Upsilon_{N_\infty})} \le C_P n^{-\frac{\tau}{d}+\frac{1}{2}}. \tag{13}$$

This result is enough to get an error estimate for the low-rank matrix approximation, where the error is measured in the Frobenius norm.

Theorem 2.1. *Let $\Xi \subset \mathbb{R}^d$ satisfy an interior cone condition and let $\Phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be of the form (11). Then, there is a constant $c_\epsilon > 0$ such that for all sets Ξ_n created by the discrete power greedy method (12), the matrix $\boldsymbol{A}_{\Xi_n}$ satisfies the estimate*

$$\left\| \boldsymbol{K}_{\Upsilon_{N_\infty}, \Upsilon_{N_\infty}} - \boldsymbol{A}_{\Xi_n} \right\|_F \leq C_P N_\infty n^{-\frac{\tau}{d}+\frac{1}{2}}.$$

Proof. We observe using (8) and (9) that

$$\boldsymbol{K}_{\Upsilon_{N_\infty}, \Upsilon_{N_\infty}} - \boldsymbol{A}_{\Xi_N} = \begin{pmatrix} \kappa_{\Xi_n,\Xi_n}(\upsilon_1, \upsilon_1) & \dots & \kappa_{\Xi_n,\Xi_n}(\upsilon_1, \upsilon_{N_\infty}) \\ \vdots & \ddots & \vdots \\ \kappa_{\Xi_n,\Xi_n}(\upsilon_{N_\infty}, \upsilon_1) & \dots & \kappa_{\Xi_n,\Xi_n}(\upsilon_{N_\infty}, \upsilon_{N_\infty}) \end{pmatrix} \in \mathbb{R}^{N_\infty \times N_\infty}. \tag{14}$$

This error representation (14) implies an error bound

$$\left\| \boldsymbol{K}_{\Upsilon_{N_\infty}, \Upsilon_{N_\infty}} - \boldsymbol{A}_{\Xi_N} \right\|_F = \left(\sum_{i=1}^{N_\infty} \sum_{j=1}^{N_\infty} \kappa^2_{\Xi_n,\Xi_n}(\upsilon_i, \upsilon_j) \right)^{\frac{1}{2}} \leq N_\infty \left\| P_{\Xi_n} \right\|_{\ell_\infty(\Upsilon_{N_\infty})}.$$

Now, (13) yields the claim. □

In order to derive kernel-based error estimates, we also have to study the distribution of the point sets Ξ_n. In particular, we have to bound $h_{\Xi_n,\mathbb{X}}$ in terms of n. To this end, we will provide a new proof for the corollary [14, Corollary 4.3] in order to make the dependence on the finite set Υ_{N_∞} more explicitly. In particular, we need to make sure that the finite set of points Υ_{N_∞} is fine enough to yield

$$\left\| P_{\Xi_n} \right\|_{\ell_\infty(\Upsilon_{N_\infty})} \leq \left\| P_{\Xi_n} \right\|_{L_\infty(\mathbb{X})} \leq C \left\| P_{\Xi_n} \right\|_{\ell_\infty(\Upsilon_{N_\infty})}$$

for a suitable constant $C > 0$.

Theorem 2.2. *Let $\mathbb{X} \subset \mathbb{R}^d$ satisfy an interior cone condition and let $\Phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be of the form (11). Let $h_{\Upsilon_{N_\infty},\mathbb{X}} \leq h_0$, where h_0 will be defined in (17). Then, for every $\epsilon \in (0, 1)$ there is a constant $c_\epsilon > 0$ such that for all sets Ξ_n created by the discrete power greedy method (12), we have that there is a constant $\theta = \tau - \frac{d}{2} > 0$ such that*

$$h_{\Xi_n,\mathbb{X}} := \sup_{x \in \mathbb{X}} \min_{\xi \in \Xi_n} \|x - \xi\|_2 \leq c_\epsilon n^{-\frac{\theta}{\theta+\epsilon}\frac{1}{d}} \quad \textit{for all}\,. \tag{15}$$

Proof. We mainly follow [14, Corollary 4.3]. Here, we have to use an approximation argument in order to apply [9, Theorem 3.1]. Precisely, [9, Theorem 3.1] provides for every $\epsilon > 0$ the existence of a constant $C(\epsilon)$ such that

$$h_{\Xi_n,\mathbb{X}} \leq C(\epsilon) \left\| P_{\Xi_n} \right\|^{\frac{1}{\tau-\frac{d}{2}+\epsilon}}_{L_\infty(\mathbb{X})}.$$

In order to use (13), we have to estimate $\|P_{\Xi_n}\|_{L_\infty(\mathbb{X})}$ in terms of $\|P_{\Xi_n}\|_{\ell_\infty(\Upsilon_{N_\infty})}$. From the condition (11), we have $\lfloor \tau \rfloor > \frac{d}{2}$. Here, we use [16, Theorem 2.6], to get

$$\|P_{\Xi_n}\|_{L_\infty(\mathbb{X})} \leq C_S \left(h_{\Upsilon_{N_\infty},\mathbb{X}}^{1-\frac{d}{2d}} \|P_{\Xi_n}\|_{W_{2d}^1(\mathbb{X})} + \|P_{\Xi_n}\|_{\ell_\infty(\Upsilon_{N_\infty})} \right). \tag{16}$$

We use [9, Lemma 4.2] to get

$$|\partial_k P_{\Xi_n}(x)| \leq 2P_{\Xi_n}(x)\sqrt{\partial_k^{(1)}\partial_k^{(2)}\kappa(x,x)} \quad \text{for all } x \in \mathbb{X} \text{ and } 1 \leq k \leq d,$$

where the integrability follows from the assumed regularity of the kernel. Hence, we obtain

$$\|P_{\Xi_n}\|_{W_{2d}^1(\mathbb{X})} \leq C(\kappa, d, \mathbb{X}) \|P_{\Xi_n}\|_{L_\infty(\mathbb{X})}.$$

This implies that if $h_{\Upsilon_{N_\infty},\mathbb{X}} \leq h_0$ with

$$h_0 \leq (2C_S C(\kappa, d, \mathbb{X}))^{-2} \tag{17}$$

we have by (16)

$$\|P_{\Xi_n}\|_{L_\infty(\mathbb{X})} \leq 2C_S \|P_{\Xi_n}\|_{\ell_\infty(\Upsilon_{N_\infty})} \leq 2C_S C_P n^{-\frac{\tau}{d}+\frac{1}{2}}$$

where the last inequality is given by (13). Hence, we obtain

$$h_{\Xi_n,\mathbb{X}} \leq C(\epsilon) \|P_{\Xi_n}\|_{L_\infty(\mathbb{X})}^{\frac{1}{\tau-\frac{d}{2}+\epsilon}} \leq C(\epsilon) \left(2C_S C_P n^{-\frac{\tau}{d}+\frac{1}{2}}\right)^{\frac{1}{\tau-\frac{d}{2}+\epsilon}}.$$

Manipulating the exponent yields

$$\left(-\frac{\tau}{d}+\frac{1}{2}\right)\frac{1}{\tau-\frac{d}{2}+\epsilon} = -\frac{1}{d}\frac{\tau-\frac{d}{2}}{\tau-\frac{d}{2}+\epsilon}.$$

Setting now $\theta = \tau - \frac{d}{2}$ finishes the proof. □

The proof also shows the following corollary which is a stronger version of (13).

Corollary 2.3. *Let $\mathbb{X} \subset \mathbb{R}^d$ satisfy an interior cone condition and let $\Phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be of the form (11). Let $h_{\Upsilon_{N_\infty},\mathbb{X}} \leq h_0$, where h_0 is defined in (17). Then, for ever $\epsilon \in (0,1)$ there is a constant $c_\epsilon > 0$ such that for all sets Ξ_n created by the discrete power greedy method (12), we have that*

$$\|P_{\Xi_n}\|_{L_\infty(\mathbb{X})} \leq c_\epsilon n^{-\frac{\tau}{d}+\frac{1}{2}}$$

Summarizing, we get results on the distribution of the set of points created by the power greedy method.

3. INVERSE MODEL PROBLEM

Let $D \subset \mathbb{R}^d$ be an open bounded set with boundary ∂D. Let $\tau \in \mathbb{N}$ with $\tau > \frac{d}{2} + 2$. We will assume $\partial D \in C^{\tau+2}$. Let $\phi_k \in C^{\tau+1}(\bar{D})$ for $1 \leq k \leq K$ be given functions with $\|\phi_k\|_{L_\infty(D)} = 1$ for all $1 \leq k \leq K$. Let

$$\mathcal{A} : I^K := [-1,1]^K \to C^{\tau+1}(\bar{D}), \quad \boldsymbol{c} = (c_1, \ldots, c_d) \mapsto \mathcal{A}(\boldsymbol{c}, x) = 1 + \sum_{k=1}^{K} \frac{c_k}{2K} \phi_k.$$

We directly observe that

$$A_{\min} = \frac{1}{2} \leq \mathcal{A}(\boldsymbol{c}, x) \leq A_{\max} = \frac{3}{2} \quad \text{for all } \boldsymbol{c} \in I^K \text{ and all } x \in D. \tag{18}$$

We consider the second order elliptic boundary value problem for given $g \in W_2^{\tau-2}(D)$ with

$$\begin{cases} L_{\boldsymbol{c}}(u) := -\nabla \cdot (\mathcal{A}(\boldsymbol{c}, x) \mathbf{Id}_{d\times d} \nabla u(\boldsymbol{c}, x)) = g(x) & \text{for all } x \in D \\ u(\boldsymbol{c}, x) = 0 & \text{for all } x \in \partial D \end{cases} \tag{19}$$

and for fixed parameter $\boldsymbol{c} \in I^K$. We introduce the operator

$$\mathcal{L}_{D;\boldsymbol{c}} : W_2^\tau(D) \to W_2^{\tau-2}(D) \times W_2^{\tau-\frac{1}{2}}(\partial D), \quad u \mapsto (L_{\boldsymbol{c}} u, \gamma_0(u)), \tag{20}$$

where γ_0 is the usual trace map. Elliptic regularity theory [3, Sec. 6.3, Thm. 5] implies that $\mathcal{L}_{D;\boldsymbol{c}}$ given as in (20) is continuously invertible, i.e., there is a map

$$\mathcal{L}_{D;\boldsymbol{c}}^{-1} : W_2^{\tau-2}(D) \times W_2^{\tau-\frac{1}{2}}(\partial D) \to W_2^\tau(D), \quad (g, 0) \mapsto u = \mathcal{L}_{D;\boldsymbol{c}}^{-1}(g, 0), \tag{21}$$

and we have

$$\|u\|_{W_2^\tau(D)} = \left\| \mathcal{L}_{D;\boldsymbol{c}}^{-1}(g, 0) \right\|_{W_2^\tau(D)} \leq C(A_{\min}, A_{\max}) \|g\|_{W_2^{\tau-2}(D)},$$

where we used to notation $C(A_{\min}, A_{\max})$ to denote that the constant depends on $\boldsymbol{c}$ only via the upper and lower bounds in (18). The assumption on the smoothness of D also ensures the existence of a bounded extension operator

$$\mathcal{E}_D : W_2^\tau(D) \to W_2^\tau(\mathbb{R}^d)$$

such that

$$\mathcal{E}_D u|_D = u \quad \text{and} \quad \|\mathcal{E}_D u\|_{W_2^\tau(\mathbb{R}^d)} \leq \|\mathcal{E}_D\| \, \|u\|_{W_2^\tau(D)} \quad \text{for all } u \in W_2^\tau(D).$$

Following [15], we consider the mapping

$$G : I^K \to W_2^\tau(D), \quad \boldsymbol{c} \mapsto u = \mathcal{L}_{D;\boldsymbol{c}}^{-1}(g, 0),$$

where $u \in W_2^\tau(D)$ is a solution to (19). Furthermore, we introduce a sampling operator associated with a discrete set $X_N := \{x_1, \ldots, x_N\} \subset D \subset \mathbb{R}^d$ as

$$S_{X_N} : W_2^\tau(\mathbb{R}^d) \to \mathbb{R}^N,$$

which is well-defined due to the condition $\tau > \frac{d}{2}$. The concatenation of the operators

$$\mathcal{G} := S_{X_N} \circ \mathcal{E}_D \circ G : I^K \to \mathbb{R}^N, \quad \boldsymbol{c} \mapsto S_{X_N}(\mathcal{E}_D(u)) = S_{X_N}(\mathcal{E}_D \circ \mathcal{L}_{D;\boldsymbol{c}}^{-1}(g,0)) \tag{22}$$

is called the observation operator in [15]. As in [15], we consider the associated inverse problem to find $\boldsymbol{c} \in I^K$ from data

$$\boldsymbol{y}_m = \mathcal{G}(\boldsymbol{c}_m) + \boldsymbol{\epsilon}_{m;L} \quad \text{for } 1 \leq m \leq M, \tag{23}$$

where $\boldsymbol{\epsilon}_{m;L}$ denotes the numerical model error. Since we cannot evaluate the observation operator for all parameters $\boldsymbol{c} \in I^K$. We define the set $C_M := \{c_1, \dots, c_M\} \subset I^K$. Hence, we have only the finitely many values

$$S_{C_M}(\hat{\mathcal{G}}) = \left(\hat{\mathcal{G}}(\boldsymbol{c}_1), \dots, \hat{\mathcal{G}}(\boldsymbol{c}_M)\right)^\top \in \mathbb{R}^{N \times M}, \tag{24}$$

where $\hat{\mathcal{G}} \approx \mathcal{G}$ is a suitable approximation to the observation operator, at hand to gain information on the continuous observation operator $\mathcal{G} : I^K \to \mathbb{R}^N$. The strategy of [15] is to assume a prior measure μ_0 such that $\boldsymbol{c}$ is distributed according to μ_0. The information, we would like to infer about the parameters is the posterior distribution $\mu^{\boldsymbol{y}}$ which is the distribution of the conditioned random variable $\boldsymbol{c}|\boldsymbol{y}$. Bayes' theorem in this situation reads as follows, see[15, Proposition 2.1].

Proposition 3.1. *Given that $\mathcal{G} : I^K \to \mathbb{R}^N$ is continuous and we have $\mu_0(I^K) = 1$. Then the distribution of the conditioned random variable $\boldsymbol{c}|\boldsymbol{y}$ is absolutely continuous with respect to the prior μ_0 and hence the Radon-Nikodým derivative exists and is given as*

$$\frac{d\mu^{\boldsymbol{y}}}{d\mu_0}(\boldsymbol{c}) = \frac{1}{\mathbb{E}_{\mu_0}\left[\exp\left(-\|\boldsymbol{y} - \mathcal{G}(\boldsymbol{c})\|_2\right)\right]} \exp\left(-\|\boldsymbol{y} - \mathcal{G}(\boldsymbol{c})\|_2\right) \tag{25}$$

Following [15], we construct an approximation to the measure (25) by approximation to the observation operator (22). We construct an approximation

$$\mathcal{G}_{N,M} \approx \mathcal{G} \tag{26}$$

and use this to construct the posterior measure as

$$\frac{d\mu^{\boldsymbol{y}}}{d\mu_0}(\boldsymbol{c}) \approx \frac{1}{\mathbb{E}_{\mu_0}\left[\exp\left(-\|\boldsymbol{y} - \mathcal{G}(\boldsymbol{c})\|_2\right)\right]} \exp\left(-\|\boldsymbol{y} - \mathcal{G}_{N,M}(\boldsymbol{c})\|_2\right) \tag{27}$$

The problem is that $\mathcal{G}(\boldsymbol{c})$ involves two approximation processes. The first approximation stems from the use of only finitely many data C_M and the second approximation process involves the solution of the system of equations (19) for a fixed parameter.

4. Approximation of the forward problem – numerical model error

The forward problem consists in solving the system of equations (19) for a given $\boldsymbol{c} \in I^K$. As outlined in [2] there is an uncertainty on the solution $u \in W_2^\tau(D)$ due to the fact that most numerical schemes use only a finite number of point-values of g for the reconstruction of $u = \mathcal{L}_{D;\boldsymbol{c}}^{-1}(g,0)$. We treat this uncertainty, in contrast to [2], as a deterministic (numerical) model error. We note as already mentioned in the introduction that this numerical model error serves as prototype for more general model errors.

Following [4], we apply a symmetric kernel–based collocation approach. This collocation method is a special case of generalized kernel-based interpolation. In an abstract framework the generalized kernel-based interpolation can be described as follows. Let $\mathcal{H}_\kappa(\mathbb{X})$ be an RKHS with kernel κ. Consider a set of linear functionals

$$\Lambda_N := \{\lambda_1, \dots, \lambda_N\} \subset \mathcal{H}_\kappa^\star(\mathbb{X}),$$

where we assume that $\dim \operatorname{span} \Lambda_N = N$. Given a vector $\boldsymbol{y} \in \mathbb{R}^N$, the generalized interpolation problem can be formulated as an optimization problem to find

$$s_{\boldsymbol{y};\Lambda_N} := \arg\min \left\{ \|s\|_{\mathcal{H}_\kappa(\mathbb{X})} \ : \ \lambda_k(s) = y_k \text{ for } 1 \le k \le N \right\}. \tag{28}$$

The solution to this problem turns out to be contained in a finite dimensional space

$$s_{\boldsymbol{y};\Lambda_N} \in V_{\kappa;\Lambda_N} = \operatorname{span} \left\{ \lambda_j^{(1)} \kappa(\cdot,\cdot) \ : \ \lambda_j \in \Lambda_N \right\}.$$

The coefficients can be computed by

$$s_{\boldsymbol{y};\Lambda_N} = \sum_{j=1}^N \alpha_j \lambda_j^{(1)} \kappa(\cdot,\cdot), \text{ where } \begin{pmatrix} \lambda_1^{(1)}\lambda_1^{(2)}\kappa(\cdot,\cdot) & \dots & \lambda_1^{(1)}\lambda_N^{(2)}\kappa(\cdot,\cdot) \\ \vdots & \ddots & \vdots \\ \lambda_N^{(1)}\lambda_1^{(2)}\kappa(\cdot,\cdot) & \dots & \lambda_N^{(1)}\lambda_N^{(2)}\kappa(\cdot,\cdot) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}.$$

In order to apply this to a kernel-based collocation method for the numerical solution to (19), we consider a union of two collocation sets $Z_L = Z_{L_1} \cup Z_{L_2}$ with $Z_{L_1} := \{z_1, \dots, z_{L_1}\} \subset D$ and $Z_{L_2} := \{z_{L_1+1}, \dots, z_{L_2}\} \subset \partial D$. We define

$$\Lambda_L := \Lambda_{L_1} \cup \Lambda_{L_2} = \left\{ \delta_{z_j} \circ L_{\boldsymbol{c}} \ : \ z_j \in Z_{L_1} \right\} \cup \left\{ \delta_{z_j} \ : \ z_j \in Z_{L_2} \right\} \subset W_2^{-\tau}(\mathbb{R}^d).$$

The data vector is given by

$$\boldsymbol{y} = (g(z_1), \dots, g(z_{L_1}), 0, \dots, 0)^\top \in \mathbb{R}^L = \mathbb{R}^{L_1+L_2}. \tag{29}$$

Moreover, we assume that there are open sets $V_k \subset \mathbb{R}^d$ such that

$$\partial D \subset \bigcup_{k=1}^{n_{\partial D}} V_k, \quad C^{\tau+2} \ni \phi_k : B(\mathbf{0},1) \xrightarrow{\simeq} V_k$$

and we denote by w_k a subordinate partition of unity which allows to write for $1 \le p < \infty$

$$\|u\|_{W_p^\tau(\partial D)}^p := \sum_{k=1}^{n_{\partial D}} \|(u \cdot w_k) \circ \phi_k\|_{W_p^\tau(B)}^p$$

and the usual modifications for $p = \infty$. The discretization parameters are defined for a discrete set of point $P = \{p_1, \dots, p_{\#P}\} \subset A \subset \mathbb{R}^d$ as in (15), i.e.,

$$h_{P,A} := \sup_{\boldsymbol{a} \in A} \min_{\boldsymbol{p} \in P} \|\boldsymbol{a} - \boldsymbol{p}\|_2 .$$

Hence, we can directly use this definition for $h_{Z_{L_1},D}$. For the points on the boundary, we define

$$h_{Z_{L_2},\partial D} := \max_{k=1}^{n_{\partial D}} h_{\phi_k^{-1}(Z_{L_2} \cap V_k), B(\mathbf{0},1)}.$$

Since the atlas is assumed to be fixed, the dependence on the particular atlas does not matter. Then, we can use [4, Corollary 3.13].

Proposition 4.1. *Let $\mathcal{L}^{-1}_{D;\boldsymbol{c}}(g,0) \in W_2^\tau(D)$ be the solution to (19) and set $\hat{\mathcal{L}}^{-1}_{D;\boldsymbol{c}_m;L}(g,0) := s_{\boldsymbol{y};\Lambda_L}$ with the notation from (28). Then, there is a constant $C > 0$ such that*

$$\begin{aligned}\left\|\mathcal{L}^{-1}_{D;\boldsymbol{c}_m}(g,0)) - \hat{\mathcal{L}}^{-1}_{D;\boldsymbol{c}_m;L}(g,0))\right\|_{L_\infty(D)} &\leq C\left(h_{Z_{L_1},D}^{\tau-2-\frac{d}{2}} + h_{Z_{L_2},\partial D}^{\tau-\frac{1+(d-1)}{2}}\right)\|u\|_{W_2^\tau(D)}\\ &\leq CC(A_{\min},A_{\max})\left(h_{Z_{L_1},D}^{\tau-2-\frac{d}{2}} + h_{Z_{L_2},\partial D}^{\tau-\frac{d}{2}}\right)\|g\|_{W_2^{\tau-2}(D)} \qquad (30)\end{aligned}$$

for all $\boldsymbol{c} \in I^K$.

So far, (30) is an error estimate for the numerical solution of a differential equation.

Corollary 4.2. *We define a perturbed observation operator*

$$\widehat{\mathcal{G}} : I^K \to \mathbb{R}^N, \quad \boldsymbol{c} \mapsto S_{X_N}(\mathcal{E}_D \circ \hat{\mathcal{L}}^{-1}_{D;\boldsymbol{c}}(g,0)).$$

Then, we obtain an error bound for the difference of the perturbed observation operator and its unperturbed counterpart (22)

$$\left\|\widehat{\mathcal{G}}(\boldsymbol{c}) - \mathcal{G}(\boldsymbol{c})\right\|_{\ell_\infty(\mathbb{R}^N)} \leq \epsilon_L := CC(A_{\min},A_{\max})\left(h_{Z_{L_1},D}^{\tau-2-\frac{d}{2}} + h_{Z_{L_2},\partial D}^{\tau-\frac{d}{2}}\right)\|g\|_{W_2^{\tau-2}(D)} \qquad (31)$$

for all $\boldsymbol{c} \in I^K$.

Proof. We directly observe for $\epsilon_{m,L}$ given in (23)

$$\|\boldsymbol{\epsilon}_{m;L}\|_{\ell_\infty(X_N)} := \left\|S_{X_N} \circ \mathcal{E}_D\left(\mathcal{L}^{-1}_{D;\boldsymbol{c}_m}(g,0) - \hat{\mathcal{L}}^{-1}_{D;\boldsymbol{c}_m;L}(g,0))\right)\right\|_{\ell_\infty(X_N)} \leq \epsilon_L,$$

which is a uniform bound for $\boldsymbol{c}_m \in C_M \subset I^K$. □

5. Error Analysis for the inverse problem

Here, we mainly follow [15]. The main goal of Bayesian inverse problems is to reconstruct the posterior distribution (25).

5.1. Reconstruction of observation operator.

In this section, we consider the reconstruction of the observation operator (22) $\mathcal{G} : I^K \to \mathbb{R}^N$ given polluted evaluations at the points $C_M \subset I^K$. As a first step, we consider a fixed set of points and later study the incremental construction of the points sets. We recall the structure of the data $\boldsymbol{y}_m$ for $1 \leq m \leq M$

$$\begin{aligned}\boldsymbol{y}_m &= \mathcal{G}(\boldsymbol{c}_m) = S_{X_N}(\mathcal{L}^{-1}_{D;\boldsymbol{c}_m}(g,0))\\ &= S_{X_N}(\hat{\mathcal{L}}^{-1}_{D;\boldsymbol{c}_m;L}(g,0)) + S_{X_N}(\hat{\mathcal{L}}^{-1}_{D;\boldsymbol{c}_m;L}(g,0) - \mathcal{L}^{-1}_{D;\boldsymbol{c}_m}(g,0))\\ &=: \boldsymbol{y}_m + \boldsymbol{\epsilon}_{m;L} = \bar{\boldsymbol{y}}_m,\end{aligned}$$

where $\bar{\boldsymbol{y}}_m = S_{X_N} \circ \mathcal{E}_D \circ \hat{\mathcal{L}}^{-1}_{D;\boldsymbol{c}_m;L}(g,0))$ are the available observations which are polluted by the numerical model error given in (31).

Since $\mathcal{G} : I^K \to \mathbb{R}^N$, we write

$$\mathcal{G}(\boldsymbol{c}) = (\mathcal{G}_1(\boldsymbol{c}), \dots, \mathcal{G}_N(\boldsymbol{c}))^\top$$

and also

$$S_{C_M}(\mathcal{G}) = \begin{pmatrix} y_{1;1} & \dots & y_{1:M} \\ \vdots & \ddots & \vdots \\ y_{1;N} & \dots & y_{N;M} \end{pmatrix} \approx \bar{\boldsymbol{Y}} = \begin{pmatrix} \bar{y}_{1;1} & \dots & \bar{y}_{1:M} \\ \vdots & \ddots & \vdots \\ \bar{y}_{1;N} & \dots & \bar{y}_{N;M} \end{pmatrix} \in \mathbb{R}^{N \times M} \tag{32}$$

and treat each component separately. Since we have a (vector-valued) reconstruction problem for $\mathcal{G}$, we choose a regression approach instead of the interpolation approach of [15]. We assume that for some $\lfloor \sigma \rfloor > \frac{d}{2}$

$$\mathcal{G} \in W_2^\sigma(I^K)$$

and again $W_2^\sigma(\mathbb{R}^d)$ is an RKHS with reproducing kernel

$$K^{(\sigma)} : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}$$

which satisfies the conditions of (11) with $\tau = \sigma$ and $d = K$. We consider the functional

$$J_{\bar{\boldsymbol{y}},\alpha,n} : W_2^\tau(I^K) \to \mathbb{R}, \quad J_{\bar{\boldsymbol{y}},\alpha,n}(s) := \sum_{m=1}^{M} (s(\boldsymbol{c}_m) - \bar{y}_{n;m})^2 + \alpha_n \|s\|^2_{W_2^\sigma(I^K)}. \tag{33}$$

The usual representer's theorem yields that an optimum exists and we define

$$s_n^\star := \arg \min_{s \in W_2^\sigma(\mathbb{R}^d)} J_{\bar{\boldsymbol{y}},\alpha,n}(s) \in \operatorname{span}\left\{ K^{(\sigma)}(\cdot, \boldsymbol{c}_1), \dots, K^{(\sigma)}(\cdot, \boldsymbol{c}_M) \right\}. \tag{34}$$

Let $\hat{\boldsymbol{s}}_n^\star \in \mathbb{R}^M$ denote the coefficients of $s_n^\star$, i.e.,

$$s_n^\star = \sum_{m=1}^{M} \hat{s}_{n;m}^\star K^{(\sigma)}(\cdot, \boldsymbol{c}_m).$$

We consider the symmetric positive definite matrix

$$\boldsymbol{K}^{(\sigma)}_{C_M,C_M} := \begin{pmatrix} K^{(\sigma)}(\boldsymbol{c}_1, \boldsymbol{c}_1) & \dots & K^{(\sigma)}(\boldsymbol{c}_1, \boldsymbol{c}_M) \\ \vdots & \ddots & \vdots \\ K^{(\sigma)}(\boldsymbol{c}_M, \boldsymbol{c}_1) & \dots & K^{(\sigma)}(\boldsymbol{c}_M, \boldsymbol{c}_M) \end{pmatrix} \in \mathbb{R}^{M \times M}.$$

The coefficients can be obtained by solving the linear system

$$\left(\boldsymbol{K}^{(\sigma)}_{C_M,C_M} + \alpha_n \mathbf{Id}_{M \times M} \right) \hat{\boldsymbol{s}}_n^\star = \bar{\boldsymbol{y}}_{n;1:M} = \bar{\boldsymbol{Y}}^\top \boldsymbol{e}_n^{(N)}, \tag{35}$$

where $\boldsymbol{e}_n^{(N)}$ is the n-th unit vector in $\mathbb{R}^N$. Note at this point, that the optimization problem (34) can be seen as a penalty methods for the optimization problem (28). For the error analysis, we follow [16].

Theorem 5.1. *For the error between the observation operator (22) and the reconstructed observation operator using the data (32) we obtain for $\mathcal{G}_{n;N,M} = s_n^\star$ defined in (34) the error estimate*

$$\|\mathcal{G}_n - \mathcal{G}_{n;N,M}\|_{L_2(I^K)} \lesssim h_{C_M,I^K}^{\frac{K}{2}} \left(\sqrt{M}\epsilon_L + \sqrt{\alpha_n}\,\|\mathcal{G}_n\|_{W_2^\sigma(I^K)}\right) + h_{C_M,I^K}^{\sigma} \left(\frac{\sqrt{M}}{\sqrt{\alpha_n}}\epsilon_L + \|\mathcal{G}_n\|_{W_2^\sigma(I^K)}\right),$$

where $\mathcal{G}_{n;N,M}$ denotes the n-th component of $\mathcal{G}_{N,M}$. For quasi-uniform sets of points, the error estimate reduces to

(36)
$$\|\mathcal{G}_n - \mathcal{G}_{n;N,M}\|_{L_2(I^K)} \lesssim \left(1 + h_{C_M,I^K}^{\sigma-\frac{K}{2}} \frac{1}{\sqrt{\alpha_n}}\right)\epsilon_L + \left(h_{C_M,I^K}^{\frac{K}{2}}\sqrt{\alpha_n} + h_{C_M,I^K}^{\sigma}\right)\|\mathcal{G}_n\|_{W_2^\sigma(I^K)}.$$

Proof. From the definition of the optimization functional (33), we can conclude that

$$\begin{aligned}
&\sum_{m=1}^M \left(s_n^\star(\boldsymbol{c}_m) - \mathcal{G}_n(\boldsymbol{c}_m)\right)^2 = \sum_{m=1}^M \left(s_n^\star(\boldsymbol{c}_m) - \hat{y}_{n;m}\right)^2 \\
&\le 2\sum_{m=1}^M \left(s_n^\star(\boldsymbol{c}_m) - \bar{y}_{n;m}\right)^2 + 2\sum_{m=1}^M \left|\boldsymbol{\epsilon}_{m;L}\cdot \boldsymbol{e}_n^{(N)}\right|^2 \le 2J_{\bar{\boldsymbol{y}},\alpha,n}(s_n^\star) + 2M\epsilon_L^2 \\
&\le 2J_{\bar{\boldsymbol{y}},\alpha,n}(\mathcal{G}_n) + 2M\epsilon_L^2 \le 2\alpha_n\,\|\mathcal{G}_n\|_{W_2^\sigma(I^K)} + 2M\epsilon_L^2. \qquad (37)
\end{aligned}$$

Furthermore, we get

$$\alpha_n\,\|s_n^\star\|_{W_2^\sigma(I^K)}^2 \le J_{\bar{\boldsymbol{y}},\alpha,n}(s_n^\star) \le 2\alpha_n\,\|\mathcal{G}_n\|_{W_2^\sigma(I^K)}^2 + 2M\epsilon_L^2. \qquad (38)$$

Now, we can use [8, Theorem 3.5] to deduce that

$$\begin{aligned}
\|s_n^\star - \mathcal{G}_n\|_{L_2(I^K)} &\lesssim h_{C_M,I^K}^{\frac{K}{2}} \left(\sum_{m=1}^M \left(s_n^\star(\boldsymbol{c}_m) - \mathcal{G}_n(\boldsymbol{c}_m)\right)^2\right)^{\frac{1}{2}} + h_{C_M,I^K}^{\tau}\,\|s_n^\star - \mathcal{G}_n\|_{W_2^\tau(I^K)} \\
&\lesssim h_{C_M,I^K}^{\frac{K}{2}} \left(\sqrt{M}\epsilon_L + \sqrt{\alpha_n}\,\|\mathcal{G}_n\|_{W_2^\tau(I^K)}\right) + h_{C_M,I^K}^{\tau} \left(\frac{\sqrt{M}}{\sqrt{\alpha_n}}\epsilon_L + \|\mathcal{G}_n\|_{W_2^\tau(I^K)}\right), \qquad (39)
\end{aligned}$$

where we used (37) and (38) in the last step. If the set $C_M \subset I^K$ is quasi-uniformly distributed in the sense that there is a constant (uniform as $M \to \infty$)

$$q_{C_M} := \frac{1}{2}\min_{\substack{\boldsymbol{c},\boldsymbol{d}\in C_M \\ \boldsymbol{c}\neq\boldsymbol{d}}} \|\boldsymbol{c}-\boldsymbol{d}\|_2 \le h_{C_M,I^K} \le C q_{C_M},$$

we get $h_{C_M,I^K}^{\frac{K}{2}}\sqrt{M} \sim 1$. Hence, for quasi-uniform sets, the error bound (39) reduces to using $s_n^\star = \mathcal{G}_{n;N,M}$

$$\|\mathcal{G}_n - \mathcal{G}_{n;N,M}\|_{L_2(I^K)} \lesssim \left(1 + h_{C_M,I^K}^{\sigma-\frac{K}{2}} \frac{1}{\sqrt{\alpha_n}}\right)\epsilon_L + \left(h_{C_M,I^K}^{\frac{K}{2}}\sqrt{\alpha_n} + h_{C_M,I^K}^{\sigma}\right)\|\mathcal{G}_n\|_{W_2^\sigma(I^K)}.$$

Hence, we obtain

$$\|\mathcal{G}_n - \mathcal{G}_{n;N,M}\|^2_{L_2(I^k;\mathbb{R}^N)} = \sum_{n=1}^{N} \|s_n^\star - \mathcal{G}_n\|_{L_2(I^K)} \leq h^\sigma_{C_M,I^K} \|\mathcal{G}\|_{W_2^\sigma(I^K;\mathbb{R}^N)} + N\epsilon_L.$$

□

These estimates also motivate a choice for the regularization parameter.

Corollary 5.2. *For the choice We choose*

$$\alpha_n = \alpha^\star = h^{2(\sigma-\frac{K}{2})}_{C_M,I^K} = h^{2\sigma-K}_{C_M,I^K}, \tag{40}$$

we obtain

$$\|\mathcal{G}_n - \mathcal{G}_{n;N,M}\|_{L_2(I^K)} \lesssim h^\sigma_{C_M,I^K} \|\mathcal{G}_n\|_{W_2^\sigma(I^K)} + \epsilon_L,$$

where ϵ_L denotes the numerical error as given in (31).

Now we will consider the situation where we construct the set of points C_M in an incremental way.

5.2. **Error estimates for incrementally constructed point sets.** Having the linear system (35) in mind, we can not directly hope to get a low rank approximation of the matrix in that linear system. We will also use the choice for the regularization parameter (40). We consider as fine set $\Upsilon_{N_\infty} \subset \mathbb{X}$ in the general setting of Section 2 a set $C_M \subset I^K$.

Lemma 5.3. *Let $h_{C_M,I^K} \leq h_0$, where h_0 is be defined in (17) with $\mathbb{X} = I^K$. Then, for all$\epsilon \in (0,1)$ there is a constant $c_\epsilon > 0$ such that for all sets Ξ_n created by the discrete power greedy method (12) with the kernel $K^{(\sigma)}$ which satisfies (11), we have that there is a constant $\theta = \sigma - \frac{d}{2}$ such that*

$$h_{\Xi_n,I^K} \leq c_\epsilon n^{-\frac{\theta}{\theta+\epsilon}\frac{1}{K}}. \tag{41}$$

Proof. The proof follows directly from Theorem 2.2. □

We obtain also a result about low rank matrix approximation as

Corollary 5.4. *There is a constant $c_\epsilon > 0$ such that for all sets Ξ_n created by the discrete power greedy method (12), the matrix $\boldsymbol{A}_{\Xi_n}$ as in (9) satisfies the estimate*

$$\left\| \boldsymbol{K}^{(\sigma)}_{C_M,C_M} + \alpha_n \mathbf{Id}_{M\times M} - \boldsymbol{A}_{\Xi_n} \right\|_F \leq C_P M \left(n^{-\frac{\sigma}{K}+\frac{1}{2}} + n^{-\frac{\theta}{\theta+\epsilon}\frac{2\sigma-K}{K}} \right)$$

where α_n is chosen according to (40). If $\sigma \geq 3K$, we get

$$\left\| \boldsymbol{K}^{(\sigma)}_{C_M,C_M} + \alpha_n \mathbf{Id}_{M\times M} - \boldsymbol{A}_{\Xi_n} \right\|_F \leq C_P M n^{-\frac{\sigma}{K}+\frac{1}{2}}.$$

Proof. We can use Theorem 2.1 to obtain

$$\left\| \boldsymbol{K}^{(\sigma)}_{C_M,C_M} - \boldsymbol{A}_{\Xi_n} \right\|_F \leq C_P c_\epsilon M n^{-\frac{\sigma}{K}+\frac{1}{2}}.$$

Now, we observe

$$\|\alpha_n \mathbf{Id}_{M\times M}\|_F \leq \alpha_n M \leq M h^{2\sigma-K}_{C_M,I^K} \leq c_\epsilon M n^{-\frac{\theta}{\theta+\epsilon}\frac{2\sigma-K}{K}},$$

where we used (40) and (41). Hence the claim follows by the triangle inequality if we choose ϵ small enough such that

$$\frac{\theta}{\theta+\epsilon}2\frac{\sigma}{K}-1\geq\frac{\sigma}{K}+\frac{1}{2}\Leftrightarrow\left(\frac{2\theta}{\theta+\epsilon}-1\right)\frac{\sigma}{K}\geq\frac{3}{2}.$$

□

Finally, we get error estimates for the reconstruction of the observation operator using the incrementally constructed set of points.

Theorem 5.5. *The error between the observation operator (22) and the reconstructed observation operator using the data (32) we obtain for $\mathcal{G}_{k;N,\Xi_n}=s_k^\star$ with the solution to the variational problem (34) the error estimate*

$$\|\mathcal{G}_k-\mathcal{G}_{k;N,\Xi_n}\|_{L_2(I^K)}\leq C\epsilon_L+c_\epsilon\left(\sqrt{n}n^{-\frac{\theta}{\theta+\epsilon}\frac{2\sigma-K}{2K}}+n^{-\frac{\theta}{\theta+\epsilon}\frac{\sigma}{K}}\right)\|\mathcal{G}_n\|_{W_2^\sigma(I^K)}, \tag{42}$$

where we use the choice (40) for the regularization parameter and the point set Ξ_n is constructed by the discrete power greedy method (12)

Remark 5.6. *If the point set Ξ_n is quasi-uniform, the estimate (42) looses one order of convergence compared to an interpolation with a quasi-uniform set of points.*

Proof. The proof follows by combing Corollary 5.2 with Lemma 5.3. □

5.3. **Error estimate for the posterior approximation.** The error for the posterior distribution id measured in the *Hellinger distance*. Let ν_i, $i=1,2$ be two measures which are absolutely continuous with respect to the prior measure ν_0. Then the Hellinger distance between ν_1 and ν_2 is given as

$$\mathrm{dist}_{\mathrm{hell}}(\nu_1,\nu_2):=\left(\frac{1}{2}\int_{L_2(I^K)}\left(\sqrt{\frac{d\nu_1}{d\nu_0}}-\sqrt{\frac{d\nu_2}{d\nu_0}}\right)d\nu_0\right)^{\frac{1}{2}}.$$

We define an approximate posterior by (27)

$$\frac{d\nu^{\boldsymbol{y}}}{d\nu_0}(\boldsymbol{c})\approx\frac{1}{\mathbb{E}_{\mu_0}\left[\exp\left(-\|\boldsymbol{y}-\mathcal{G}(\boldsymbol{c})\|_2\right)\right]}\exp\left(-\|\boldsymbol{y}-\mathcal{G}_{N,M}(\boldsymbol{c})\|_2\right).$$

We can use [15, Theorem 4.2] to obtain

$$\mathrm{dist}_{\mathrm{hell}}(\nu,\nu^{\boldsymbol{y}})\leq C\|\mathcal{G}-\mathcal{G}_{:,M,n}\|_{L_2(I^K)}, \tag{43}$$

where the posterior is defined in (25). Hence, we obtain directly the following theorem

Theorem 5.7.

$$\mathrm{dist}_{\mathrm{hell}}(\nu,\nu^{\boldsymbol{y}})\leq C\epsilon_L+c_\epsilon\left(\sqrt{n}n^{-\frac{\theta}{\theta+\epsilon}\frac{2\sigma-K}{2K}}+n^{-\frac{\theta}{\theta+\epsilon}\frac{\sigma}{K}}\right)\|\mathcal{G}_n\|_{W_2^\sigma(I^K)} \tag{44}$$

6. Conclusion

We re-interpreted the covariance of the predictive distribution as reproducing kernel. This interpretation allowed to reformulate certain (simple instances) of techniques from low-rank matrix approximation as simple greedy methods in reproducing kernel Hilbert spaces. In particular, this allowed to use results from the kernel-based literature to derive error estimates for the *adaptive cross approximation* or the *pivoted Cholesky* method as matrix approximation for finitely smooth kernels. We applied the resulting iterative construction of sampling point sets to the inverse problem to reconstruct a parametrized diffusion coefficient by finitely many point evaluations of the solution to the diffusion equation with that coefficient. We explicitly took the error in the solution to the differential equation also into account.

Acknowledgement

The author acknowledges partial support of the Deutsche Forschungsgemeinschaft (DFG) through the Sonderforschungsbereich 1060: The Mathematics of Emergent Effects.

Literatur

[1] M. Bebendorf, *Adaptive Cross Approximation of Multivariate Functions*, Constructive Approximation, 34 (2011), pp. 149–179.
[2] J. Cockayne, C. Oates, T. Sullivan, and M. Girolami, *Probabilistic Meshless Methods for Partial Differential Equations and Bayesian Inverse Problems*, May 2016. available at https://arxiv.org/pdf/1605.07811v1.pdf.
[3] L. Evans, *Partial Differential Equations*, Graduate studies in mathematics, American Mathematical Society, 2010.
[4] P. Giesl and H. Wendland, *Meshless collocation: Error estimates with application to dynamical systems*, SIAM Journal on Numerical Analysis, 45 (2007), pp. 1723–1741.
[5] M. Griebel and C. Rieger, *Reproducing Kernel Hilbert Spaces for Parametric Partial Differential Equations*, SIAM/ASA Journal on Uncertainty Quantification, 5 (2017), pp. 111–137.
[6] H. Harbrecht, M. Peters, and R. Schneider, *On the low-rank approximation by the pivoted Cholesky decomposition*, Appl. Numer. Math., 4 (2012), pp. 428–440.
[7] R. Kempf, H. Wendland, and C. Rieger, *Kernel-based Reconstructions for Parametric PDEs*. Available as INS Preprint No. 1804., 2018.
[8] W. Madych, *An estimate for multivariate interpolation ii*, Journal of Approximation Theory, 142 (2006), pp. 116 – 128.
[9] S. D. Marchi, R. Schaback, and H. Wendland, *Near-optimal data-independent point locations for radial basis function interpolation*, Advances in Computational Mathematics, 23 (2005), pp. 317–330.
[10] M. Mouattamid and R. Schaback, *Recursive Kernels*, Analysis in Theory and Applications, 25 (2009), p. 301.
[11] S. Müller and R. Schaback, *A Newton basis for kernel spaces*, J. Approx. Theory, 161 (2009), pp. 645–655.
[12] M. Pazouki and R. Schaback, *Bases for kernel-based spaces*, Journal of Computational and Applied Mathematics, 236 (2011), pp. 575 – 588. International Workshop on Multivariate Approximation and Interpolation with Applications (MAIA 2010).
[13] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2006.
[14] G. Santin and B. Haasdonk, *Convergence rate of the data-independent P-greedy algorithm in kernel-based approximation*, Dolomites Research Notes on Approximation, 10 (2017), pp. 68–78.
[15] A. M. Stuart and A. L. Teckentrup, *Posterior Consistency for Gaussian Process Approximations of Bayesian Posterior Distributions*, Mathematics of Computation, 87 (2018), pp. 721–753. available at: https://arxiv.org/abs/1603.02004.

[16] H. WENDLAND AND C. RIEGER, *Approximate Interpolation with Applications to Selecting Smoothing Parameters*, Numerische Mathematik, 101 (2005), pp. 729–748.