SIAM J. MATH. DATA SCI.
Vol. 1, No. 1, pp. 208–236

Gaussian Process Landmarking on Manifolds*

Tingran Gao†, Shahar Z. Kovalsky‡, and Ingrid Daubechies§

Abstract. As a means of improving analysis of biological shapes, we propose an algorithm for sampling a Riemannian manifold by sequentially selecting points with maximum uncertainty under a Gaussian process model. This greedy strategy is known to be near-optimal in the experimental design literature, and it appears to outperform the use of user-placed landmarks in representing the geometry of biological objects in our application. In the noiseless regime, we establish an upper bound for the mean squared prediction error (MSPE) in terms of the number of samples and geometric quantities of the manifold, demonstrating that the MSPE for our proposed sequential design decays at a rate comparable to the oracle rate achievable by any sequential or nonsequential optimal design; to the best of our knowledge this is the first result of this type for sequential experimental design. The key is to link the greedy algorithm to reduced basis methods in the context of model reduction for partial differential equations (PDEs). We expect this approach will find additional applications in other fields of research.

Key words. Gaussian process, experimental design, active learning, manifold learning, reduced basis methods, geometric morphometrics

AMS subject classifications. 60G15, 62K05, 65D18

DOI. 10.1137/18M1184035

1. Introduction.

This paper grew out of an attempt to apply principles of the statistical field of *optimal experimental design* to *geometric morphometrics*, a subfield of evolutionary biology that focuses on quantifying the (dis)similarities between pairs of two-dimensional anatomical surfaces based on their spatial configurations. In contrast to methods for statistical estimation and inference, which typically focus on studying the error made by estimators with respect to a deterministically generated or randomly drawn (but fixed) collection of sample observations and on constructing estimators to minimize this error, the paradigm of optimal experimental design is to minimize the empirical risk by an "optimal" choice of sample locations, while the estimator itself and the number of samples are kept fixed [62, 6]. Finding an optimal design amounts to choosing sample points that are most informative for a class of estimators so as to reduce the number of observations; this is most desirable when acquiring even one observation is expensive (e.g., in spatial analysis (geostatistics) [79, 27]

*Received by the editors April 30, 2018; accepted for publication (in revised form) January 8, 2019; published electronically February 12, 2019.

http://www.siam.org/journals/simods/1-1/M118403.html

Funding: This work is supported by Simons Math+X Investigators Award 400837 and NSF CAREER Award BCS-1552848.

†Committee on Computational and Applied Mathematics, Department of Statistics, The University of Chicago, Chicago, IL 60637 (tingrangao@galton.uchicago.edu).

‡Department of Mathematics, Duke University, Durham, NC 27708 (shaharko@math.duke.edu).

§Department of Mathematics and Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 (ingrid@math.duke.edu).

and in computationally demanding computer experiments [72]), but similar ideas have long been exploited in the probabilistic analysis of some classical numerical analysis problems (see, e.g., [78, 92, 65]).

In this paper, we adopt the methodology of optimal experimental design for discretely sampling Riemannian manifolds, and we propose a greedy algorithm that sequentially selects design points based on the uncertainty modeled by a Gaussian process. On anatomical surfaces of interest to geometric morphometrical applications, these design points play the role of *anatomical landmarks*, or simply *landmarks*, which are geometrically or semantically meaningful feature points crafted by evolutionary biologists for quantitatively comparing large collections of biological specimens in the framework of *Procrustes analysis* [39, 33, 40]. The effectiveness of our approach on anatomical surfaces, along with more background information on geometric morphometrics and Procrustes analysis, is demonstrated in a companion paper [37]; though the prototypical application scenario in this paper and [37] is geometric morphometrics, we expect the approach proposed here to be more generally applicable to other scientific domains where compact or sparse data representation is demanded. In contexts different from evolutionary biology, closely related (continuous or discretized) manifold sampling problems are addressed in [5, 53, 42], where smooth manifolds are discretized by optimizing the locations of (a fixed number of) points so as to minimize a Riesz functional, and in [61, 66], which study surface simplification via spectral subsampling or geometric relevance. These approaches, when applied to two-dimensional surfaces, tend to distribute points either empirically with respect to fine geometric details preserved in the discretized point clouds or uniformly over the underlying geometric object, whereas triangular meshes encountered in geometric morphometrics often lack fine geometric details but still demand nonuniform, sparse geometric features that are semantically/biologically meaningful; moreover, it is often not clear whether the desired anatomical landmarks are naturally associated with an energy potential. In contrast, our work is inspired by recent research on active learning with Gaussian processes [25, 63, 45] as well as by related applications in finding landmarks along a manifold [49]. Different from many graph-Laplacian-based manifold landmarking algorithms in semisupervised learning (e.g., [93, 90]), our approach considers a Gaussian process on the manifold whose covariance structure is specified by the heat kernel, with a greedy landmarking strategy that aims to produce a set of geometrically significant samples with adequate coverage for biological traits. Furthermore, in stark contrast with [30, 51] where landmarks are utilized primarily for improving computational efficiency, the landmarks produced by our algorithm explicitly and directly minimize the *mean squared prediction error* (MSPE) and thus bear rich information for machine learning and data mining tasks. The optimality of the proposed greedy procedure is also established (see section 4); this is apparently much less straightforward for nondeterministic, sampling-based manifold landmarking algorithms such as those in [32, 21, 83].

The rest of this paper is organized as follows. The remainder of this introduction motivates our main algorithm and discusses other related work. Section 2 sets notation and provides background materials for Gaussian processes and the construction of heat kernels on Riemannian manifolds (and discretizations thereof) as well as the "reweighted kernel" constructed from these discretized heat kernels; section 3 presents an unsupervised landmarking algorithm for anatomical surfaces inspired by recent work on uncertainty sampling in Gaussian process

active learning [49]; section 4 provides the convergence rate analysis and establishes the MSPE optimality; and section 5 summarizes the current paper with a brief sketch of potential future directions. We defer implementation details of the proposed algorithm for applications in geometric morphometrics to the companion paper [37].

1.1. Motivation. To see the link between landmark identification and active learning with uncertainty sampling [48, 74], let us consider the regression problem of estimating a function $f : V \to \mathbb{R}$ defined over a point cloud $V \subset \mathbb{R}^D$. Rather than construct the estimator from random sample observations, we adopt the point of view of active learning, in which one is allowed to sequentially query the values of f at user-picked vertices $x \in V$. In order to minimize the empirical risk of an estimator $\hat{f}$ within a given number of iterations, the simplest and most commonly used strategy is to first evaluate (under reasonable probabilistic model assumptions) the informativeness of the vertices on the mesh that have not been queried, and then greedily choose to inquire the value of f at the vertex x at which the response value $\hat{f}(x)$—inferred from all previous queries—is most "uncertain" in the sense of attaining highest predictive error (though other uncertainty measures such as the Shannon entropy could be used as well); these sequentially obtained highest-uncertainty points will be treated as morphometrical *landmarks* in our proposed algorithm.

This straightforward application of the active learning strategy summarized above relies on selecting a regression function f of rich biological information. In the absence of a natural candidate regression function f, we seek to reduce in every iteration the maximum "average uncertainty" of a class of regression functions, e.g., specified by a Gaussian process prior [63]. Throughout this paper we will denote by $\mathrm{GP}(m, K)$ the Gaussian process on a smooth, compact Riemannian manifold M with mean function $m : M \to \mathbb{R}$ and covariance function $K : M \times M \to \mathbb{R}$. If we interpret choosing a single most "biologically meaningful" function f as a manual "feature handcrafting" step, the specification of a Gaussian process prior can be viewed as a less restrictive and more stable "ensemble" version; the geometric information can be conveniently encoded into the prior by specifying an appropriate covariance function K. We construct such a covariance function in subsection 2.2 by reweighting the heat kernel of the Riemannian manifold M, adopting (but meanwhile also appending further geometric information to) the methodology of Gaussian process optimal experimental design [70, 72, 35] and *sensitivity analysis* [71, 60] from the statistics literature.

1.2. Our contribution and other related work. The main theoretical contribution of this paper is a convergence rate analysis for the greedy algorithm of uncertainty-based sequential experimental design, which amounts to estimating the uniform rate of decay for the prediction error of a Gaussian process as the number of greedily picked design points increases; on a C^∞-manifold we deduce that the convergence is faster than any inverse polynomial rate, which is also the optimal rate any greedy or nongreedy landmarking algorithm can attain on a generic smooth manifold. This analysis makes use of recent results in the analysis of reduced basis methods by converting the Gaussian process experimental design into a basis selection problem in a reproducing kernel Hilbert space associated with the Gaussian process. To the best of our knowledge, there does not exist in the literature any earlier analysis of this type for greedy algorithms in optimal experimental design; the convergence results obtained from this analysis can also be used to bound the number of iterations in Gaussian process active

learning [25, 45, 49] and maximum entropy design [72, 47, 59]. From a numerical linear algebra perspective, though the rank-1 update procedure detailed in Remark 3.2 coincides with the well-known algorithm of pivoted Cholesky decomposition for symmetric positive definite matrices (cf. subsection 3.2), we are not aware of similar results in that context for the performance of pivoting. We thus expect our theoretical contribution to shed light on a deeper understanding of other manifold landmarking algorithms in active and semisupervised learning [93, 90, 21, 83, 30, 51]. We discuss the implementation details of our algorithm for applications in geometric morphomerics in a companion paper [37].

We point out that, though experimental design is a classical problem in the statistical literature [34, 20, 62], it is only very recently that interest in computationally efficient experimental design algorithms has begun to arise in the computer science community [15, 7, 58, 86, 1, 2]. Most experimental design algorithms based on various types of optimality criteria, including but not limited to A(verage)-, D(eterminant)-, E(igen)-, V(ariance)-, G-optimality, and Bayesian alphabetical optimality, are NP-hard computational in their exact form [23, 19], with the only exception being T(race)-optimality, which is trivial to solve. For computer scientists, the interesting problem is to find polynomial algorithms that efficiently find $(1 + O(\epsilon))$-approximations of the optimal solution to the exact problem, where $\epsilon > 0$ is expected to be as small as possible but depends on the size of the problem and the prefixed budget for the number of design points; often these approximation results also require certain constraints on the dimension of the ambient space, the number of design points, and the total number of candidate points. Different from those approaches, our theoretical contribution assumes no relations between these quantities, and the convergence rate is with respect to the increasing number of landmark points (as opposed to a prefixed budget); nevertheless, similar to results obtained in [15, 7, 58, 86, 1, 2], our proposed algorithm has polynomial complexity and is thus computationally tractable; see subsection 3.2 for more details. We refer interested readers to [62] for more exhaustive discussions of the optimality criteria used in experimental design.

2. Background.

2.1. Heat kernels and Gaussian processes on Riemannian manifolds: A spectral embedding perspective. Let (M, g) be an orientable compact Riemannian manifold of dimension $d \geq 1$ with finite volume, where g is the Riemannian metric on M. Denote by dvol_M the canonical volume form M with coordinate representation

$$\mathrm{dvol}_M(x) = \sqrt{|g(x)|}\, \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^d.$$

The finite volume will be denoted as

$$\mathrm{Vol}(M) = \int_M \mathrm{dvol}_M(x) = \int_M \sqrt{|g(x)|}\, \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^d < \infty,$$

and we will fix the canonical normalized volume form $\mathrm{dvol}_M/\mathrm{Vol}(M)$ as a reference. Throughout this paper, all distributions on M are absolutely continuous with respect to $\mathrm{dvol}_M/\mathrm{Vol}(M)$.

The single-output regression problem on the Riemannian manifold M will be described as follows. Given independent and identically distributed (i.i.d.) observations $\{(X_i, Y_i) \in M \times \mathbb{R} \mid 1 \leq i \leq n\}$ of a random variable (X, Y) on the product probability space $M \times \mathbb{R}$, the goal of

the regression problem is to estimate the conditional expectation

$$f(x) := \mathbb{E}(Y \mid X = x), \tag{2.1}$$

which is often referred to as a regression function of Y on X [81]. For simplicity, the joint distribution of X and Y will always be assumed absolutely continuous with respect to the product measure on $M \times \mathbb{R}$. A Gaussian process (or *Gaussian random field*) on M with mean function $m : M \to \mathbb{R}$ and covariance function $K : M \times M \to \mathbb{R}$ is defined as the stochastic process for which any finite marginal distribution on n fixed points $x_1, \ldots, x_n \in M$ is a multivariate Gaussian distribution with mean vector

$$m_n := (m(x_1), \ldots, m(x_n)) \in \mathbb{R}^n$$

and covariance matrix

$$K_n := \begin{pmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

A Gaussian process with mean function $m : M \to \mathbb{R}$ and covariance function $K : M \times M \to \mathbb{R}$ will be denoted as $\mathrm{GP}(m, K)$. Under model $Y \sim \mathrm{GP}(m, K)$, given observed values $y_1, \ldots, y_n$ at locations $x_1, \ldots, x_n$, the *best linear predictor* (BLP) [79, 72] for the random field at a new point x is given by the conditional expectation

$$Y^*(x) := \mathbb{E}[Y(x) \mid Y(x_1) = y_1, \ldots, Y(x_n) = y_n] = m(x) + k_n(x)^\top K_n^{-1}(Y_n - m_n), \tag{2.2}$$

where $Y_n = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$, $k_n(x) = (K(x, x_1), \ldots, K(x, x_n))^\top \in \mathbb{R}^n$; at any $x \in M$, the expected squared error, or *mean squared prediction error* (MSPE), is defined as

$$\begin{aligned} \mathrm{MSPE}(x; x_1, \ldots, x_n) &:= \mathbb{E}\left[(Y(x) - Y^*(x))^2\right] \\ &= \mathbb{E}\left[(Y(x) - \mathbb{E}[Y(x) \mid Y(x_1) = y_1, \ldots, Y(x_n) = y_n])^2\right] \\ &= K(x, x) - k_n(x)^\top K_n^{-1} k_n(x), \end{aligned} \tag{2.3}$$

which is a function over M. Here the expectation is with respect to all realizations $Y \sim \mathrm{GP}(m, K)$. Squared integral (L^2) or sup (L^∞) norms of the pointwise MSPE are often used as criteria for evaluating the prediction performance over the experimental domain. In geospatial statistics, interpolation with (2.2) and (2.3) is known as *kriging*.

Our analysis in this paper concerns the sup-norm of the prediction error with n design points $x_1, \ldots, x_n$ picked using a greedy algorithm, i.e., the quantity

$$\sigma_n := \sup_{x \in M} \mathrm{MSPE}(x; x_1, \ldots, x_n),$$

where $x_1, \ldots, x_n$ are chosen according to Algorithm 3.1. This quantity is compared with the "oracle" prediction error attainable by any sequential or nonsequential experimental design with n points, i.e.,

$$d_n := \inf_{x_1, \ldots, x_n \in M} \sup_{x \in M} \mathrm{MSPE}(x; x_1, \ldots, x_n).$$

As will be shown in (4.10) in section 4, d_n can be interpreted as the *Kolmogorov width* of approximating a *reproducing kernel Hilbert space* (RKHS) with a reduced basis. The RKHS we consider is a natural one associated with a Gaussian process; see, e.g., [28, 55] for general introductions on RKHS and [82] for RKHS associated with Gaussian processes. In Appendix A, we include a brief sketch of the RKHS theory needed for understanding section 4.

On Riemannian manifolds, there is a natural choice for the kernel function: the heat kernel, i.e., the kernel of the Laplace–Beltrami operator. Denote by $\Delta : C^2(M) \to C^2(M)$ the Laplace–Beltrami operator on M with respect to the metric g, i.e.,

$$\Delta f = \frac{1}{\sqrt{|g|}} \partial_i \left(\sqrt{|g|}\, g^{ij} \partial_j f \right) \quad \forall f \in C^\infty(M),$$

where the sign convention is such that $-\Delta$ is positive semidefinite. If the manifold M has no boundary, the spectrum of $-\Delta$ is well known to be real, nonnegative, discrete, with eigenvalues satisfying $0 = \lambda_0 < \lambda_1 \le \lambda_2 \le \cdots \nearrow \infty$, and with ∞ the only accumulation point of the spectrum; when M has a nonempty boundary we assume the Dirichlet boundary condition, so the same inequalities hold for the eigenvalues. If we denote by ϕ_i the eigenfunction of Δ corresponding to the eigenvalue λ_i, then the set $\{\phi_i \mid i = 0, 1, \dots\}$ constitutes an orthonormal basis for $L^2(M)$ under the standard inner product

$$\langle f_1, f_2 \rangle_M := \int_M f_1(x)\, f_2(x)\, \mathrm{dvol}_M(x).$$

The heat kernel $k_t(x,y) := k(x,y;t) \in C^2(M \times M) \times C^\infty((0,\infty))$ is the fundamental solution of the following heat equation on M:

$$\partial_t u(x,t) = -\Delta u(x,t), \qquad x \in M,\ t \in (0,\infty).$$

That is, if the initial data is specified as

$$u(x, t=0) = v(x),$$

then

$$u(x,t) = \int_M k_t(x,y)\, v(y)\, \mathrm{dvol}_M(y).$$

In terms of the spectral data of Δ (see, e.g., [68, 12]), the heat kernel can be written as

$$k_t(x,y) = \sum_{i=0}^{\infty} e^{-\lambda_i t} \phi_i(x)\, \phi_i(y) \quad \forall t \ge 0,\ x, y \in M. \tag{2.4}$$

For any fixed $t > 0$, the heat kernel defines a Mercer kernel on M by

$$(x,y) \mapsto k_t(x,y) \quad \forall (x,y) \in M \times M,$$

and the feature mapping (see (A.4)) takes the form

$$M \ni x \longmapsto \Phi_t(x) := \left(e^{-\lambda_0 t/2} \phi_0(x), e^{-\lambda_1 t/2} \phi_1(x), \dots, e^{-\lambda_i t/2} \phi_i(x), \dots \right) \in \ell^2, \tag{2.5}$$

where ℓ^2 is the infinite sequence space equipped with a standard inner product; see, e.g., [64, section II.1, Example 3]. Note in particular that

$$k_t(x,y) = \langle \Phi_t(x), \Phi_t(y) \rangle_{\ell^2}. \tag{2.6}$$

In fact, up to a multiplicative constant $c(t) = \sqrt{2}(4\pi)^{\frac{d}{4}} t^{\frac{n+2}{4}}$, the feature mapping $\Phi_t : M \to \ell^2$ has long been studied in spectral geometry [11] and is known to be an embedding of M into ℓ^2; furthermore, with the multiplicative correction by $c(t)$, the pullback of the canonical metric on ℓ^2 is asymptotically equal to the Riemannian metric on M.

In this paper we focus on Gaussian processes on Riemannian manifolds with heat kernels (or "reweighted" counterparts thereof; see subsection 2.2) as covariance functions. There are at least two reasons for heat kernels to be considered as natural candidates for covariance functions of Gaussian processes on manifolds. First, as argued in [18, section 2.5], the abundant geometric information encoded in the Laplace–Beltrami operator makes the heat kernel a canonical choice for Gaussian processes; Gaussian processes defined in this way impose natural geometric priors based on randomly rescaled solutions of the heat equation. Second, by (2.6), a Gaussian process on M with a heat kernel is equivalent to a Gaussian process on the embedded image of M into ℓ^2 under the feature mapping (2.5) with a dot product kernel; this is reminiscent of the methodology of *extrinsic Gaussian process regression* (eGPR) [50] on manifolds—in order to perform Gaussian process regression on a nonlinear manifold, eGPR first embeds the manifold into a Euclidean space using an arbitrary embedding, then performs Gaussian process regression on the embedded image following standard procedures for Gaussian process regression. This spectral embedding interpretation also underlies recent work on constructing Gaussian priors, by means of the graph Laplacian, for uncertainty quantification of graph semisupervised learning [13].

2.2. Discretized and reweighted heat kernels. When the Riemannian manifold M is a submanifold embedded in an ambient Euclidean space $\mathbb{R}^D$ ($D \gg d$) and sampled only at finitely many points $\{x_1, \ldots, x_n\}$, we know from the literature of Laplacian eigenmaps [9, 10] and diffusion maps [26, 76, 77] that the extrinsic squared exponential kernel matrix

$$K = (K_{ij})_{1\le i,j\le n} = \left(\exp\left(-\frac{\|x_i - x_j\|^2}{t} \right) \right)_{1\le i,j\le n} \tag{2.7}$$

is a consistent estimator (up to a multiplicative constant) of the heat kernel of the manifold M if $\{x_i \mid 1 \le i \le n\}$ are sampled uniformly and i.i.d. on M with appropriately adjusted bandwidth parameter $t > 0$ as $n \to \infty$; similar results hold when the squared exponential kernel is replaced with any anisotropic kernel, and additional renormalization techniques can be used to adjust the kernel if the samples are i.i.d. but not uniformly distributed on M; see, e.g., [26] for more details. These theoretical results in manifold learning justify using extrinsic kernel functions in a Gaussian process regression framework when the manifold is an embedded submanifold of an ambient Euclidean space; the kernel (2.7) is also used in [91] for Gaussian process regression on manifolds in a Bayesian setting. Nevertheless, one may use other discrete approximations of the heat kernel in place of (2.7) without affecting our

theoretical results in section 4, as long as the kernel matrix K is positive (semi)definite and defines a valid Gaussian process for our landmarking purposes.

The heat kernel of the Riemannian manifold M defines covariance functions for a family of Gaussian processes on M, but this type of covariance function depends only on the spectral properties of M, whereas in practice we would often like to incorporate prior information addressing relative high/low confidence of the selected landmarks. For example, the response variables might be measured with higher accuracy (or, equivalently, the influence of random observation noise is damped) where the predictor falls on a region on the manifold M with lower curvature. We encode the relative high/low confidence of measurements into a smooth, positive weight function $w : M \to \mathbb{R}_+$, defined on the entire manifold, whereby the higher values of $w(x)$ indicate a relatively higher importance if a predictor variable is sampled near $x \in M$. Since we assume M is closed, w is bounded below away from zero. To "knit" the weight function into the heat kernel, notice that by the reproducing property we have

$$k_t(x,y) = \int_M k_{t/2}(x,z)\, k_{t/2}(z,y)\, \mathrm{dvol}_M(z), \tag{2.8}$$

and we can naturally apply the weight function to deform the volume form, i.e., define

$$k_t^w(x,y) = \int_M k_{t/2}(x,z)\, k_{t/2}(z,y)\, w(z)\, \mathrm{dvol}_M(z). \tag{2.9}$$

Obviously, $k_t^w(\cdot,\cdot) = k_t(\cdot,\cdot)$ on $M \times M$ if we pick $w \equiv 1$ on M, using the expression (2.4) for heat kernel $k_t(\cdot,\cdot)$ and the orthonormality of the eigenfunctions $\{\phi_i \mid i = 0, 1, \dots\}$. Intuitively, (2.9) reweighs the mutual interaction between different regions on M such that the portions with high weights have a more significant influence on the covariance structure of the Gaussian process on M. Results established for GP (m, k_t) can often be directly adapted for GP (m, k_t^w).

In practice, when the manifold is sampled only at finitely many i.i.d. points $\{x_1, \dots, x_n\}$ on M, the reweighted kernel can be calculated from the discrete extrinsic kernel matrix (2.7), with t replaced with $t/2$,

$$K^w = \left(K_{ij}^w\right)_{1\le i,j\le n} = \left(\sum_{k=1}^n e^{-\frac{\|x_i - x_k\|^2}{t/2}} \cdot w(x_k) \cdot e^{-\frac{\|x_k - x_j\|^2}{t/2}}\right)_{1\le i,j\le n} = K^\top W K, \tag{2.10}$$

where W is a diagonal matrix of size $n \times n$ with $w(x_k)$ at its kth diagonal entry, for all $1 \le k \le n$, and K is the discrete squared exponential kernel matrix (2.7). It is worth pointing out that the reweighted kernel K^w no longer equals the kernel K in (2.7) even when we set $w \equiv 1$ at this discrete level. Kernels similar to (2.9) have also appeared in [22] as the symmetrization of an asymmetric anisotropic kernel.

Though the reweighting step appears to be a straightforward implementation trick, it turns out to be crucial in the application of automated geometric morphometrics: when the reweighted kernel is adopted, the landmarking algorithm in section 3 produces biologically much more representative features on anatomical surfaces. We demonstrate this in greater detail in [37].

3. Gaussian process landmarking. We present in this section an algorithm motivated by [49] that automatically places "landmarks" on a compact Riemannian manifold using a Gaussian process active learning strategy. Let us begin with an arbitrary nonparametric regression model in the form of (2.1). Unlike standard supervised learning in which a finite number of sample-label pairs is provided, an active learning algorithm can iteratively decide, based on memory of all previously inquired sample-label pairs, which sample to query for label in the next step. In other words, given sample-label pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ observed up to the nth step, an active learning algorithm can decide which sample X_{n+1} to query for the label information $Y_{n+1} = f(X_{n+1})$ of the regression function f to be estimated; typically, the algorithm assumes full knowledge of the sample domain, has access to the regression function f as a black box, and strives to optimize its query strategy so as to estimate f in as few steps as possible. With a Gaussian process prior $\mathrm{GP}(m, K)$ on the regression function class, the joint distribution of a finite collection of $(n+1)$ response values $(Y_1, \ldots, Y_n, Y_{n+1})$ is assumed to follow a multivariate Gaussian distribution $\mathcal{N}_{n+1}(m(X_1, \ldots, X_{n+1}), K(X_1, \ldots, X_{n+1}))$, where

$$\begin{aligned} &m(X_1, \ldots, X_{n+1}) = \big(m(X_1), \ldots, m(X_{n+1})\big) \in \mathbb{R}^n, \\ &K(X_1, \ldots, X_{n+1}) = \begin{pmatrix} K(X_1, X_1) & \cdots & K(X_1, X_{n+1}) \\ \vdots & & \vdots \\ K(X_{n+1}, X_1) & \cdots & K(X_{n+1}, X_{n+1}) \end{pmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}. \end{aligned} \tag{3.1}$$

For simplicity, the rest of this paper will use the shorthand notation

$$X_n^1 = \big(X_1, \ldots, X_n\big) \in M^n, \quad Y_n^1 = \big(Y_1, \ldots, Y_n\big) \in \mathbb{R}^n \tag{3.2}$$

and

$$\begin{aligned} K_{n,n} &= K(X_1, \ldots, X_n) \in \mathbb{R}^{n\times n}, \\ K\big(X, X_n^1\big) &= \big(K(X, X_1), \ldots, K(X, X_n)\big)^\top \in \mathbb{R}^n. \end{aligned} \tag{3.3}$$

Given n observed samples $(X_1, Y_1), \ldots, (X_n, Y_n)$, at any $X \in M$, the conditional probability of the response value $Y(X) \mid Y_n^1$ follows a normal distribution

$$\mathcal{N}(\xi_n(X), \Sigma_n(X)),$$

where

$$\begin{aligned} \xi_n(X) &= K\big(X, X_n^1\big)^\top K_n^{-1} Y_n^1, \\ \Sigma_n(X) &= K(X, X) - K\big(X, X_n^1\big)^\top K_{n,n}^{-1} K\big(X, X_n^1\big). \end{aligned} \tag{3.4}$$

In our landmarking algorithm, we simply choose X_{n+1} to be the location on the manifold M with the largest variance, i.e.,

$$X_{n+1} := \operatorname*{argmax}_{X \in M} \Sigma_n(X). \tag{3.5}$$

Notice that this successive procedure of "landmarking" $X_1, X_2, \ldots$ on M is independent of the specific choice of regression function in $\mathrm{GP}(m, K)$ since we only need the covariance function $K : M \times M \to \mathbb{R}$.

3.1. Algorithm. The main algorithm of this paper, an unsupervised landmarking procedure for anatomical surfaces, will use a discretized, reweighted kernel constructed from triangular meshes that digitize anatomical surfaces. We now describe this algorithm in full detail. Let M be a two-dimensional compact surface isometrically embedded in $\mathbb{R}^3$, and denote by $\kappa : M \to \mathbb{R}$, $\eta : M \to \mathbb{R}$ the Gaussian curvature and (scalar) mean curvature of M. Define a family of weight function $w_{\lambda,\rho} : M \to \mathbb{R}_{\geq 0}$ parametrized by $\lambda \in [0,1]$ and $\rho > 0$ as

$$w_{\lambda,\rho}(x) = \frac{\lambda |\kappa(x)|^\rho}{\int_M |\kappa(\xi)|^\rho \, \mathrm{dvol}_M(\xi)} + \frac{(1-\lambda)|\eta(x)|^\rho}{\int_M |\eta(\xi)|^\rho \, \mathrm{dvol}_M(\xi)} \quad \forall x \in M. \tag{3.6}$$

This weight function seeks to emphasize the influence of high curvature locations on the surface M on the covariance structure of the Gaussian process prior GP $\left(m, k_t^{w_{\lambda,\rho}}\right)$, where $k_t^{w_{\lambda,\rho}}$ is the reweighted heat kernel defined in (2.9). In this paper we stick with simple kriging (setting $m \equiv 0$ in GP (m, K)), and in our implementation we use default values $\lambda = 1/2$ and $\rho = 1$ (but one may wish to alter these values to fine-tune the landscape of the weight function for a specific application).

For all practical purposes, we only concern ourselves with M being a piecewise linear surface, represented as a discrete triangular mesh $T = (V, E)$ with vertex set $V = \{x_1, \ldots, x_{|V|}\} \subset \mathbb{R}^3$ and edge set E. We calculate the mean and Gaussian curvature functions η, κ on the triangular mesh (V, E) using standard algorithms from computational geometry [24, 3]. The weight function $w_{\lambda,\rho}$ can then be calculated at each vertex x_i by

$$w_{\lambda,\rho}(x_i) = \frac{\lambda |\kappa(x_i)|^\rho}{\sum_{k=1}^{|V|} |\kappa(x_k)|^\rho \nu(x_k)} + \frac{(1-\lambda)|\eta(x_i)|^\rho}{\sum_{k=1}^{|V|} |\eta(x_k)|^\rho \nu(x_k)} \quad \forall x_i \in V, \tag{3.7}$$

where $\nu(x_k)$ is the area of the Voronoi cell of the triangular mesh T centered at x_i. The reweighted heat kernel $k_t^{w_{\lambda,\rho}}$ is then defined on $V \times V$ as

$$k_t^{w_{\lambda,\rho}}(x_i, x_j) = \sum_{k=1}^{|V|} k_{t/2}(x_i, x_k) \, k_{t/2}(x_k, x_j) \, w_{\lambda,\rho}(x_k) \, \nu(x_k), \tag{3.8}$$

where the (unweighted) heat kernel k_t is calculated as in (2.7). Until a fixed total number of landmarks is collected, at step $(n+1)$ the algorithm computes the uncertainty score $\Sigma_{(n+1)}$ on V from the existing n landmarks $\xi_1, \ldots, \xi_n$ by

$$\Sigma_{(n+1)}(x_i) = k_t^{w_{\lambda,\rho}}(x_i, x_i) - k_t^{w_{\lambda,\rho}}\left(x_i, \xi_n^1\right)^\top k_t^{w_{\lambda,\rho}}\left(\xi_n^1, \xi_n^1\right)^{-1} k_t^{w_{\lambda,\rho}}\left(x_i, \xi_n^1\right) \quad \forall x_i \in V, \tag{3.9}$$

where

$$k_t^{w_{\lambda,\rho}}\left(x_i,\xi_n^1\right):=\begin{pmatrix}k_t^{w_{\lambda,\rho}}\left(x_i,\xi_1\right)\\ \vdots\\ k_t^{w_{\lambda,\rho}}\left(x_i,\xi_n\right)\end{pmatrix},$$

$$k_t^{w_{\lambda,\rho}}\left(\xi_n^1,\xi_n^1\right):=\begin{pmatrix}k_t^{w_{\lambda,\rho}}\left(\xi_1,\xi_1\right) & \cdots & k_t^{w_{\lambda,\rho}}\left(\xi_1,\xi_n\right)\\ \vdots & & \vdots\\ k_t^{w_{\lambda,\rho}}\left(\xi_n,\xi_1\right) & \cdots & k_t^{w_{\lambda,\rho}}\left(\xi_n,\xi_n\right)\end{pmatrix},$$

and picks the $(n+1)$th landmark ξ_{n+1} according to the rule

$$\xi_{n+1}=\underset{x_i\in V}{\operatorname{argmax}}\,\Sigma_{(n+1)}\left(x_i\right).$$

If there are more than one maximizer of $\Sigma_{(n+1)}$, we just randomly pick one; at step 1 the algorithm simply picks the vertex maximizing $x\mapsto k_t^{w_{\lambda,\rho}}(x,x)$ on V. See Algorithm 3.1 for a comprehensive description.

Remark 3.1. Algorithm 3.1 can be easily adapted to work with point clouds (where connectivity information is not present) and in higher dimensional spaces, which makes it applicable to a much wider range of input data in geometric morphometrics as well as other applications; see, e.g., [37]. For instance, it suffices to replace step 4 of Algorithm 3.1 with a different discrete curvature (or another type of "importance score") calculation procedure on point clouds (see, e.g., [69, 29]), and replace step 5 with a nearest-neighbor weighted graph adjacency matrix construction. In this paper we require the inputs to be triangular meshes with edge connectivity only for ease of the statement, as computation of discrete curvatures on triangular meshes is much more straightforward.

Remark 3.2. Note that, according to (3.9), each step adds only one new row and one new column to the inverse covariance matrix, which enables us to perform rank-1 updates to the covariance matrix according to the block matrix inversion formula (see, e.g., [63, section A.3])

$$K_n^{-1}=\begin{pmatrix}K_{n-1} & P\\ P^\top & K\left(X_n,X_n\right)\end{pmatrix}^{-1}=\begin{pmatrix}K_{n-1}^{-1}\left(I_{n-1}+\mu PP^\top K_{n-1}^{-1}\right) & -\mu K_{n-1}^{-1}P\\ -\mu P^\top K_{n-1}^{-1} & \mu\end{pmatrix},$$

where

$$P=\left(K\left(X_1,X_n\right),\ldots,K\left(X_{n-1},X_n\right)\right)\in\mathbb{R}^{n-1},$$
$$\mu=\left(K\left(X_n,X_n\right)-P^\top K_{n-1}^{-1}P\right)^{-1}\in\mathbb{R}.$$

This simple trick significantly improves the computational efficiency because it avoids directly inverting the covariance matrix when the number of landmarks becomes large as the iteration progresses.

Before we delve into the theoretical aspects of Algorithm 3.1, let us present a few typical instances of this algorithm in practical use. A more comprehensive evaluation of the applicability of Algorithm 3.1 to geometric morphometrics is deferred to [37]. In a nutshell, the

Algorithm 3.1 GAUSSIAN PROCESS LANDMARKING WITH REWEIGHTED HEAT KERNEL.

procedure GPL(T, L, $\lambda \in [0,1]$, $\rho > 0$, $\epsilon > 0$) ▷ Triangular Mesh $T = (V, E)$, number of landmarks L
  $\kappa, \eta \leftarrow$ DISCRETECURVATURES(T) ▷ calculate discrete Gaussian curvature κ and mean curvature η on T
  $\nu \leftarrow$ VORONOIAREAS(T) ▷ calculate the area of Voronoi cells around each vertex x_i
  $w_{\lambda,\rho} \leftarrow$ CALCULATEWEIGHT($\kappa, \eta, \lambda, \rho, \nu$) ▷ calculate weight function $w_{\lambda,\rho}$ according to (3.7)
  $W \leftarrow \left[\exp\left(-\|x_i - x_j\|^2 \Big/ \epsilon\right)\right]_{1\le i,j\le |V|} \in \mathbb{R}^{|V|\times|V|}$
  $\Lambda \leftarrow \operatorname{diag}\left(w_{\lambda,\rho}(x_1)\,\nu(x_1), \ldots, w_{\lambda,\rho}\left(x_{|V|}\right)\nu\left(x_{|V|}\right)\right) \in \mathbb{R}^{|V|\times|V|}$
  $\xi_1, \ldots, \xi_L \leftarrow \emptyset$ ▷ initialize landmark list
  $\Psi \leftarrow 0$
  $\ell \leftarrow 1$
  $K_{\text{full}} \leftarrow W^\top \Lambda W \in \mathbb{R}^{|V|\times|V|}$
  $K_{\text{trace}} \leftarrow \operatorname{diag}(K_{\text{full}}) \in \mathbb{R}^{|V|}$
  while $\ell < L + 1$ **do**
    if $\ell = 1$ **then**
      $\Sigma \leftarrow K_{\text{trace}}$
    else ▷ calculate uncertainty scores by (3.9)
      $b \leftarrow$ solve linear system $\Psi\left[[\xi_1, \ldots, \xi_\ell], :\right] b = \Psi$
      $\Sigma \leftarrow K_{\text{trace}} - \operatorname{diag}\left(\Psi^\top b\right) \in \mathbb{R}^{|V|}$
    end if
    $\xi_\ell \leftarrow \operatorname{argmax} \Sigma$
    $\Psi \leftarrow K_{\text{full}}[:, [\xi_1, \ldots, \xi_\ell]]$
    $\ell \leftarrow \ell + 1$
  end while
  return $\xi_1, \ldots, \xi_L$
end procedure

Gaussian process landmarking algorithm picks the landmarks on the triangular mesh successively, according to the uncertainty score function Σ at the beginning of each step; at the end of each step the uncertainty score function gets updated, with the information about the newly picked landmark incorporated into the inverse covariance matrix defined as in (3.4). Figure 1 illustrates the first few successive steps on a triangular mesh discretization of a fossil molar of primate *Plesiadapoidea*. Empirically, we observed that the updates on the uncertainty score function are mostly local; i.e., no abrupt changes of the uncertainty score are observed away from a small geodesic neighborhood centered at each new landmark. Guided by uncertainty and a curvature-reweighted covariance function, the Gaussian process landmarking often identifies landmarks of abundant biological information; for instance, the first Gaussian process landmarks are often highly biologically informative and demonstrate a comparable level of coverage with observer landmarks manually picked by human experts. See Figure 2 for a

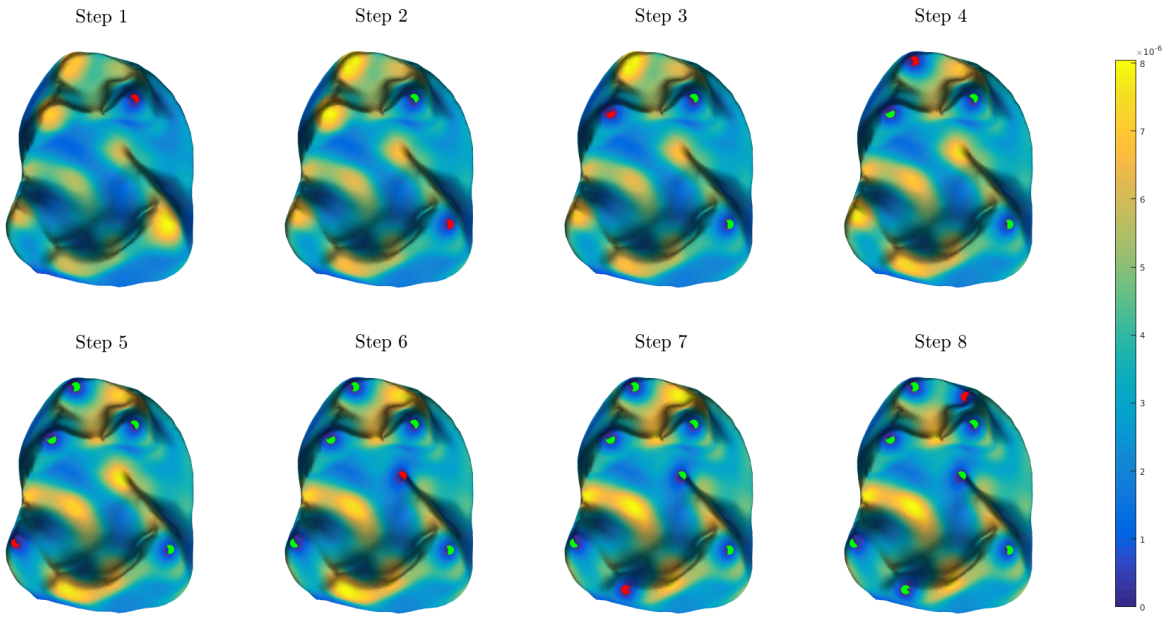
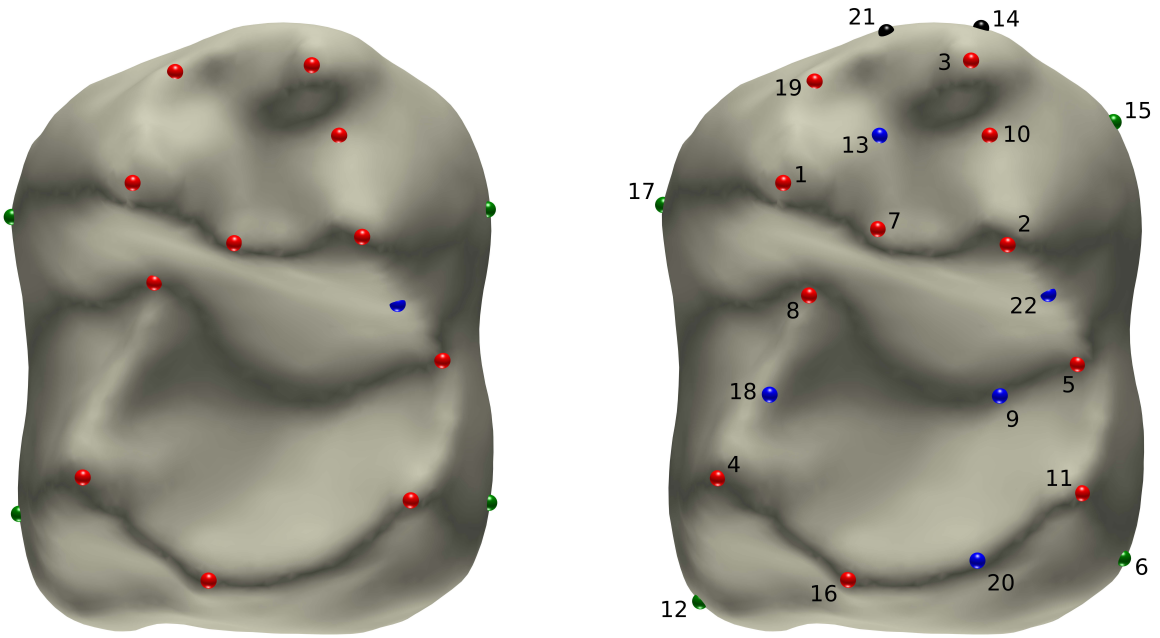


Figure 1. *The first eight landmarks picked successively by Gaussian process landmarking (Algorithm 3.1) on a digitized fossil molar of* Plesiadapoidea *(extinct mammals from the Paleocene and Eocene epochs of North America, Europe, and Asia* [75]*), with the uncertainty scores at the end of each step rendered on the triangular mesh as a heat map. In each subfigure, the preexisting landmarks are colored green and the new landmark is colored red. At each step, the algorithm picks the vertex on the triangular mesh with the highest uncertainty score (computed according to (3.4)) and then updates the score function.*

visual comparison between automatically generated landmarks and observer landmarks manually placed by evolutionary anthropologists on a digitized fossil molar different from the one illustrated in Figure 1.

3.2. Numerical linear algebra perspective. Algorithm 3.1 can be divided into two phases: lines 1–10 focus on constructing the kernel matrix K_{full} from the geometry of the triangular mesh M; from line 11 onward, only numerical linear algebraic manipulations are involved. In fact, the numerical linear algebra part of Algorithm 3.1 is identical to Gaussian elimination (or LU decomposition) with a very particular "diagonal pivoting" strategy, which is different from standard full or partial pivoting in Gaussian elimination. To see this, first note that the variance $\Sigma_n(X)$ in (3.4) is just the diagonal of the Schur complement of the $n \times n$ submatrix of K_{full} corresponding to the n previously chosen landmarks $X_1, \ldots, X_n$, and recall from [80, Ex. 20.3] that this Schur complement arises as the bottom-right $(|V| - n) \times (|V| - n)$ block after the nth elimination step. The greedy criterion (3.5) then amounts to selecting the largest diagonal entry in this Schur complement as the next pivot. Therefore, the second phase of Algorithm 3.1 can be consolidated into the form of a "diagonal-pivoted" LU decomposition, i.e., $K_{\text{full}}P = LU$, in which the first L columns of the permutation matrix P reveal the location of the L chosen landmarks. In fact, since the kernel matrix we choose is symmetric and positive semidefinite, the rank-1 updates in Remark 3.2 most closely resemble the *pivoted Cholesky*



(a) Observer Landmarks **(b) Gaussian Process Landmarks**

Figure 2. ***Left:*** *Sixteen observer landmarks on a digitized fossil molar of a* Teilhardina *(one of the oldest known fossil primates closely related to living tarsiers and anthropoids* [8]*) identified manually by evolutionary anthropologists as ground truth, first published in* [16]*.* ***Right:*** *The first* 22 *landmarks picked by Gaussian process landmarking (Algorithm* 3.1*). The numbers next to each landmark indicate the order of appearance. The Gaussian process landmarks strikingly resemble the observer landmarks: the red landmarks (Number* 1–5*,* 7*,* 8*,* 10*,* 11*,* 16*,* 19*) signal geometric sharp features (cusps or saddle points corresponding to local maximum/minimum Gaussian curvature); the blue landmarks sit either along the curvy cusp ridges and grooves (Number* 13*,* 18*,* 20*,* 22*) or at the basin (Number* 9*), serving the role often played by* semilandmarks *(cf.* [37]*,* section 2.1*]); the four green landmarks (Number* 6*,* 12*,* 15*,* 17*) approximately delimit the "outline" of the tooth in occlusal view.*

decomposition (see, e.g., [43, section 10.3] or [41]). Identical to these classical pivoting-based matrix decomposition algorithms, the time and space computational complexities of the main algorithm in Algorithm 3.1 are thus $O\left(L^3\right)$ and $O\left(n^2\right)$, respectively, where n is the total number of candidate points and L is the desired number of parameters. Note that these complexities are both polynomial and comparable with those in the computer science literature [15, 58, 86, 1, 2]. This numerical linear algebraic perspective motivates us to investigate variants of Algorithm 3.1 based on other numerical linear algebraic algorithms with pivoting in future work.

4. Rate of convergence: Reduced basis methods in reproducing kernel Hilbert spaces. In this subsection we analyze the rate of convergence of our main Gaussian process landmarking algorithm from section 3. While the notion of "convergence rate" in the context of Gaussian process regression (i.e., kriging [56, 79]) or scattered data approximation (see, e.g., [87] and the references therein) refers to how fast the interpolant approaches the true

function, our focus in this paper is the rate of convergence of Algorithm 3.1 itself, i.e., the number of steps the algorithm takes before it terminates. In practice, unless a maximum number of landmarks is predetermined, a natural criterion for terminating the algorithm is to specify a threshold for the sup-norm of the prediction error (2.3) (i.e., the variance (3.5)) over the manifold. We emphasize again that, although this greedy approach is motivated by the probabilistic model of Gaussian processes, the MSPE is completely determined once the kernel function and the design points are specified, as is the greedy algorithmic procedure. Our analysis is centered around bounding the uniform rate at which the pointwise MSPE function (2.3) decays with respect to the number landmarks greedily selected.

To this end, we observe the connection between Algorithm 3.1 and a greedy algorithm studied thoroughly for *reduced basis methods* in [14, 31] in the context of model reduction. While the analyses in [14, 31] assume general Hilbert and Banach spaces, we apply those results to a reproducing kernel Hilbert space (RKHS), denoted as $\mathscr{H}_K$, naturally associated with a Gaussian process $\mathrm{GP}(m, K)$; as will be demonstrated below, the MSPE with respect to n selected landmarks can be interpreted as a measurement of distance from a point to an n-dimensional subspace in $\mathcal{X}_K$, where the subspace is determined by the selected landmarks. We emphasize that, though the connection between the Gaussian process and RKHS is well known (see, e.g., [82] and the references therein), we are not aware of existing literature addressing the resemblance between the two classes of greedy algorithms widely used in Gaussian process experimental design and reduced basis methods.

We begin with a brief summary of the greedy algorithm in reduced basis methods for a general Banach space $(X, \|\cdot\|)$. The algorithm strives to approximate *all* elements of X using a properly constructed linear subspace spanned by (as few as possible) selected elements from a compact subset $\mathscr{F} \subset X$; thus the name "reduced" basis. A popular greedy algorithm for this purpose generates successive approximation spaces by choosing the first basis $f_1 \in \mathscr{F}$ according to

$$f_1 := \operatorname*{argmax}_{f \in \mathscr{F}} \|f\| \tag{4.1}$$

and, successively, when $f_1, \ldots, f_{n-1}$ are picked already, by choosing

$$f_{n+1} := \operatorname*{argmax}_{f \in \mathscr{F}} \operatorname{dist}(f, V_n), \tag{4.2}$$

where

$$V_n = \operatorname{span}\{f_1, f_2, \ldots, f_n\}$$

and

$$\operatorname{dist}(f, V_n) := \inf_{g \in V_n} \|f - g\|.$$

In words, at each step we greedily pick the function that is "farthest away" from the set of already chosen basis elements. Intuitively, this is analogous to the *farthest point sampling* (FPS) algorithm [38, 54] in Banach spaces, with a key difference in the choice of the distance

between a point p and a set of selected points $\{q_1, \ldots, q_n\}$: in FPS such a distance is defined as the maximum over all distances $\{\|p - q_i\| \mid 1 \leq i \leq n\}$, whereas in reduced basis methods the distance is between p and the linear subspace spanned by $\{q_1, \ldots, q_n\}$.

The Gaussian process landmarking algorithm fits naturally into the framework of reduced basis methods as follows. Let us first specialize this construction to the case when X is the RKHS $\mathscr{H}_K \subset L^2(M)$, where M is a compact Riemannian manifold and K is the reproducing kernel. A natural choice for K is the heat kernel $k_t(\cdot, \cdot) : M \times M \to \mathbb{R}$ with a fixed $t > 0$ as in subsection 2.1, but for a submanifold isometrically embedded into an ambient Euclidean space it is common as well to choose the kernel to be the restriction to M of a positive (semi)definite kernel in the ambient Euclidean space such as (2.7) or (2.10), for which Sobolev-type error estimates are known in the literature on scattered data approximation [57, 36]. It follows from standard RKHS theory (see, e.g., (A.5)) that

$$\mathscr{H}_K = \overline{\operatorname{span}\left\{\sum_{i \in I} a_i K(\cdot, x_i) \mid a_i \in \mathbb{R}, x_i \in M, \operatorname{card}(I) < \infty\right\}}, \tag{4.3}$$

and by the compactness of M and the regularity of the kernel function, we have for any $x \in M$

$$\langle K(\cdot, x), K(\cdot, x)\rangle_{\mathscr{H}_K} = K(x, x) \leq \|K\|_{\infty, M \times M} < \infty,$$

which justifies the compactness of

$$\mathscr{F} := \operatorname{span}\{K(\cdot, x) \mid x \in M\} \tag{4.4}$$

as a subset of $\mathscr{H}_K$ since $\mathscr{H}_K$ embeds into $L^2(M)$ compactly [4, 67]. In fact, since we only used the compactness of M and the boundedness of K on $M \times M$, the argument above for the compactness of $\mathscr{F}$ can be extended to any Gaussian process defined on a compact metric space with a bounded kernel. The initialization step (4.1) now amounts to selecting $K(\cdot, x)$ from $\mathscr{F}$ that maximizes

$$\|K(\cdot, x)\|^2_{\mathscr{H}_K} = \langle K(\cdot, x), K(\cdot, x)\rangle_{\mathscr{H}_K} = K(x, x),$$

which is identical to (3.5) when $n = 1$ (or, equivalently, line 14 in Algorithm 3.1); furthermore, given $n \geq 1$ previously selected basis functions $K(\cdot, x_1), \ldots, K(\cdot, x_n)$, the $(n+1)$th basis function will be chosen according to (4.2), i.e., $f_{n+1} = K(\cdot, x_n)$ maximizes the infimum

$$\begin{aligned}
\inf_{g \in \operatorname{span}\{K(\cdot, x_1), \ldots, K(\cdot, x_n)\}} \|K(\cdot, x) - g\|^2_{\mathscr{H}_K} &= \inf_{a_1, \ldots, a_n \in \mathbb{R}} \left\| K(\cdot, x) - \sum_{i=1}^n a_i K(\cdot, x_i) \right\|^2_{\mathscr{H}_K} \\
&= \inf_{a_1, \ldots, a_n \in \mathbb{R}} K(x, x) - 2\sum_{i=1}^n a_i K(x, x_i) + \sum_{i=1}^n \sum_{j=1}^n a_i a_j K(x_i, x_j) \\
&\overset{(*)}{=} K(x, x) - K\left(x, x_n^1\right)^\top K_{n,n}^{-1} K\left(x, x_n^1\right), \qquad (4.5)
\end{aligned}$$

where the notation is as in (3.2) and (3.3), i.e.,

$$K\left(x, x_n^1\right) := \begin{pmatrix} K(x, x_1) \\ \vdots \\ K(x, x_n) \end{pmatrix}, \quad K_{n,n} := \begin{pmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{pmatrix}.$$

The equality $(*)$ follows from the observation that, for any fixed $x \in M$, the minimizing vector $\mathbf{a} := (a_1, \ldots, a_n)^\top \in \mathbb{R}^n$ satisfies

$$K\left(x, x_n^1\right) = K_{n,n}\mathbf{a} \quad \Leftrightarrow \quad \mathbf{a} = K_{n,n}^{-1} K\left(x, x_n^1\right).$$

It is clear at this point that maximizing the rightmost quantity in (4.5) is equivalent to following the greedy landmark selection criterion (3.5) at the $(n+1)$th step. We thus conclude that Algorithm 3.1 is equivalent to the greedy algorithm for reduced basis methods in $\mathscr{H}_K$, an RKHS modeled on the compact manifold M. The following lemma summarizes this observation for future reference.

Lemma 4.1. *Let M be a compact Riemannian manifold, and let $K : M \times M \to \mathbb{R}$ be a positive semidefinite kernel function. Consider the RKHS $\mathscr{H}_K \subset L^2(M)$ as defined in (4.3). For any $x \in M$ and a collection of n points $\mathscr{X}_n = \{x_1, x_2, \ldots, x_n\} \subset M$, the orthogonal projection P_n from $\mathscr{H}_K$ to $V_n = \operatorname{span}\{K(\cdot, x_i) \mid 1 \le i \le n\}$ is*

$$P_n\left(K(\cdot, x)\right) = \sum_{i=1}^{n} a_i^*(x)\, K(\cdot, x_i),$$

where $a_i^* : M \to \mathbb{R}$ is the inner product of vector $\left(K(x, x_1), \ldots, K(x, x_n)\right)$ with the ith row of

$$\begin{pmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{pmatrix}^{-1}.$$

In particular, a_i^* has the same regularity as the kernel Φ for all $1 \le i \le n$. Moreover, the squared distance between $K(\cdot, x)$ and the linear subspace $V_n \subset \mathscr{H}_K$ has the closed-form expression

$$\begin{aligned} P_{K,\mathscr{X}_n}(x) :&= \left\|K(\cdot, x) - P_n\left(K(\cdot, x)\right)\right\|_{\mathscr{H}_K}^2 \\ &= \min_{a_1,\ldots,a_n \in \mathbb{R}} \left\|K(\cdot, x) - \sum_{i=1}^{n} a_i K(\cdot, x_i)\right\|_{\mathscr{H}_K}^2 \\ &= K(x, x) - K\left(x, x_n^1\right)^\top \begin{pmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{pmatrix}^{-1} K\left(x, x_n^1\right), \end{aligned} \tag{4.6}$$

where

$$K\left(x, x_n^1\right) := \left(K(x, x_1), \ldots, K(x, x_n)\right)^\top \in \mathbb{R}^n.$$

Consequently, for any Gaussian process defined on M with covariance structure given by the kernel function K, the MSPE of the Gaussian process conditioned on the observations at $x_1, \ldots, x_n \in M$ equals the distance between $K(\cdot, x)$ and the subspace V_n spanned by $K(\cdot, x_1), \ldots, K(\cdot, x_n)$.

The function $P_{K,\mathscr{X}_n} : M \to \mathbb{R}_{\geq 0}$ defined in (4.6) is in fact the squared *power function* in the literature on scattered data approximation; see, e.g., [87, Definition 11.2].

The convergence rate of greedy algorithms for reduced basis methods has been investigated in a series of works [17, 14, 31]. The general paradigm is to compare the maximum approximation error incurred after the nth greedy step, denoted as

$$\sigma_n := \operatorname{dist}(f_{n+1}, V_n) = \max_{f \in \mathscr{F}} \operatorname{dist}(f, V_n), \tag{4.7}$$

with the *Kolmogorov width* (cf. [52]), a quantity characterizing the theoretical optimal error of approximation using any n-dimensional linear subspace generated from any greedy or nongreedy algorithms, defined as

$$d_n := \inf_{Y} \sup_{f \in \mathscr{F}} \operatorname{dist}(f, Y), \tag{4.8}$$

where the first infimum is taken over all n-dimensional subspaces Y of X. When $n = 1$, both σ_1 and d_1 reduce to the ∞-bound of the kernel function on $M \times M$, i.e., $\|K\|_{\infty, M\times M}$. Note that by definitions (4.7) and (4.8) both sequences $\{\sigma_n \mid n \in \mathbb{N}\}$ and $\{d_n \mid n \in \mathbb{N}\}$ are monotonically nondecreasing since $V_1 \subsetneqq V_2 \subsetneqq \cdots$; see also [14, section 1.3]. In [31] the following comparison between $\{\sigma_n \mid n \in \mathbb{N}\}$ and $\{d_n \mid n \in \mathbb{N}\}$ was established.

Theorem 4.2 ([31, Theorem 3.2] (the $\gamma = 1$ case)). *For any $N \geq 0$, $n \geq 1$, and $1 \leq m < n$, there holds*

$$\prod_{\ell=1}^{n} \sigma_{N+\ell}^2 \leq \left(\frac{n}{m}\right)^m \left(\frac{n}{n-m}\right)^{n-m} \sigma_{N+1}^{2m} d_m^{2n-2m}.$$

This result can be used to establish a direct comparison between the performance of greedy and optimal basis selection procedures. For instance, setting $N = 0$ and taking advantage of the monotonicity of the sequence $\{\sigma_n \mid n \in \mathbb{N}\}$, one has from Theorem 4.2 that

$$\sigma_n \leq \sqrt{2} \min_{1 \leq m < n} \|K\|_{\infty, M\times M}^{\frac{m}{n}} d_m^{\frac{n-m}{n}}$$

for all $n \in \mathbb{N}$. Using the monotonicity of $\{\sigma_n \mid n \in \mathbb{N}\}$, by setting $m = \lfloor n/2 \rfloor$ we have the even more compact inequality

$$\sigma_n \leq \sqrt{2} \|K\|_{\infty, M\times M}^{\frac{1}{2}} d_{\lfloor n/2 \rfloor}^{\frac{1}{2}} \quad \forall n \in \mathbb{N}, n \geq 2. \tag{4.9}$$

If we have a bound for $\{d_n \mid n \in \mathbb{N}\}$, inequality (4.9) can be directly invoked to establish a bound for $\{\sigma_n \mid n \in \mathbb{N}\}$ at the expense of comparing σ_n with d_{2n}; in the regime $n \to \infty$ we may even expect the same rate of convergence at the expense of a larger constant. We emphasize

here that the definition of $\{d_n \mid n \in \mathbb{N}\}$ only involves elements in a compact subset $\mathscr{F}$ of the ambient Hilbert space $\mathscr{H}_K$; in our setting, the compact subset (4.4) consists of only functions of the form $K(\cdot, x)$ for some $x \in M$, and thus

$$\begin{aligned} d_n &= \inf_{x_1,\ldots,x_n \in M} \sup_{x \in M} \operatorname{dist}\left(K(\cdot,x), \operatorname{span}\left\{K(\cdot,x_i) \mid 1 \le i \le n\right\}\right) \\ &= \inf_{x_1,\ldots,x_n \in M} \sup_{x \in M} \left[K(x,x) - K\left(x, x_n^1\right)^\top K_{n,n}^{-1} K\left(x, x_n^1\right)\right]. \end{aligned} \tag{4.10}$$

To ease notation, we will always denote $\mathscr{X}_n := \{x_1, \ldots, x_n\}$ as in Lemma 4.1. Write the maximum value of the function $P_{K,\mathscr{X}_n}$ over M as

$$\Pi_{K,\mathscr{X}_n} := \sup_{x \in M} P_{K,\mathscr{X}_n}(x). \tag{4.11}$$

The Kolmogorov width d_n can be put in this notation as

$$d_n = \inf_{x_1,\ldots,x_n \in M} \Pi_{K,\mathscr{X}_n}. \tag{4.12}$$

The problem of bounding $\{d_n \mid n \in \mathbb{N}\}$ thus reduces to bounding the infimum of $\Pi_{K,\mathscr{X}_n}$ over all n-dimensional linear subspaces of $\mathscr{F}$.

When M is an open, bounded subset of a standard Euclidean space, upper bounds for $\Pi_{K,\mathscr{X}_n}$ are often established—in a kernel-adaptive fashion—using the *fill distance* [87, Chapter 11]

$$h_{\mathscr{X}_n} := \sup_{x \in M} \min_{x_j \in \mathscr{X}_n} \|x - x_j\|, \tag{4.13}$$

where $\|\cdot\|$ is the Euclidean norm of the ambient space. For instance, when K is a squared exponential kernel (2.7) and the domain is a cube (or, more generally, the domain should at least be compact and convex, as pointed out in [85, Theorem 1]) in a Euclidean space, Wendland [87, Theorem 11.22] asserts that

$$\Pi_{K,\mathscr{X}_n} \le \exp\left[c \frac{\log h_{\mathscr{X}_n}}{h_{\mathscr{X}_n}}\right] \quad \forall h_{\mathscr{X}_n} \le h_0 \tag{4.14}$$

for some constants $c > 0$, $h_0 > 0$ depending only on M and the kernel bandwidth $t > 0$ in (2.7). Similar bounds have been established in [89] for Matérn kernels, but the convergence rate is only polynomial. In this case, by the monotonicity of the function $x \mapsto \log x / x$ for $x \in (0, e)$, we have, for all sufficiently small $h_{\mathscr{X}_n}$,

$$d_n = \inf_{x_1,\ldots,x_n \in M} \Pi_{K,\mathscr{X}_n} \le \exp\left[c \frac{\log h_n}{h_n}\right],$$

where

$$h_n := \inf_{\mathscr{X}_n \subset M,\ |\mathscr{X}_n| = n} h_{\mathscr{X}_n} \tag{4.15}$$

is the minimum fill distance attainable for any n sample points on M. We thus have the following theorem for the convergence rate of Algorithm 3.1 for any compact, convex set in a Euclidean space.

Theorem 4.3. *Let $\Omega \subset \mathbb{R}^D$ be a compact and convex subset of the D-dimensional Euclidean space, and consider a Gaussian process $\mathrm{GP}(m, K)$ defined on Ω, with the covariance function K being of the squared exponential form (2.7) with respect to the ambient D-dimensional Euclidean distance. Let $X_1, X_2, \ldots$ denote the sequence of landmarks greedily picked on Ω according to Algorithm 3.1, and define for any $n \in \mathbb{N}$ the maximum MSPE on Ω with respect to the first n landmarks $X_1, \ldots, X_n$ as*

$$\sigma_n = \max_{x \in \Omega} \left[K(x,x) - K\left(x, X_n^1\right)^\top K_n^{-1} K\left(x, X_n^1\right) \right],$$

where the notation $K\left(x, X_n^1\right)$ and K_n are defined in section 3. Then

$$\sigma_n = O\left(\beta^{\frac{\log h_{\lfloor n/2 \rfloor}}{h_{\lfloor n/2 \rfloor}}}\right) \quad \text{as } n \to \infty \tag{4.16}$$

for some positive constant $\beta > 1$ depending only on the geometry of the domain Ω and the bandwidth of the squared exponential kernel K; h_n is the minimum fill distance of n arbitrary points on Ω (cf. (4.15)).

Proof. By the monotonicity of the sequence $\{\sigma_n \mid n \in \mathbb{N}\}$, it suffices to establish the convergence rate for a subsequence. Using directly (4.9), (4.12), (4.14), and the definition of h_n in (4.15), we have the inequality for all $\mathbb{N} \ni n \geq N$:

$$\sigma_{2n} \leq \sqrt{2}\, \|K\|_{\infty,\Omega\times\Omega}^{\frac{1}{2}} \exp\left[\frac{c}{2}\frac{\log h_n}{h_n}\right] = \sqrt{2}\, \|K\|_{\infty,\Omega\times\Omega}^{\frac{1}{2}} \beta^{\frac{\log h_n}{h_n}},$$

where $\beta := \exp(c/2) > 1$. Here the positive constants $N = N(\Omega, t) > 0$ and $c = c(\Omega, t) > 0$ depend only on the geometry of Ω and the bandwidth of the squared exponential kernel. This completes the proof. ∎

Convex bodies in $\mathbb{R}^D$ are far too restricted as a class of geometric objects for modeling anatomical surfaces in our main application [37]. The rest of this section will be devoted to generalizing the convergence rate for squared exponential kernels (2.7) to their reweighted counterparts (2.10) and, more importantly, for submanifolds of the Euclidean space. The crucial ingredient is an estimate of the type (4.14) bounding the sup-norm of the squared power function using fill distances, tailored for restrictions of the squared exponential kernel

$$K_\epsilon(x,y) = \exp\left(-\frac{1}{2\epsilon}\|x-y\|^2\right), \quad x, y \in M, \tag{4.17}$$

as well as the reweighted version

$$K_\epsilon^w(x,y) = \int_M w(z) \exp\left[-\frac{1}{2\epsilon}\left(\|x-z\|^2 + \|z-y\|^2\right)\right] \mathrm{dvol}_M(z), \quad x, y \in M, \tag{4.18}$$

where $w : M \to \mathbb{R}_{\geq 0}$ is a nonnegative weight function. Note that when $w(x) \equiv 1$, for all $x \in M$ the reweighted kernel (4.18) does not coincide with the squared exponential kernel (4.17)—not even up to normalization—since the domain of integration is M instead of the entire $\mathbb{R}^D$; nor

does it seem to work to naïvely enclose the compact manifold M with a compact, convex subset Ω of the ambient space and reuse Theorem 4.3 by extending/restricting functions to/from M to Ω , since the samples are constrained to lie on M, but the convergence will be in terms of fill distances in Ω. Nevertheless, the desired bound can be established using local parametrizations of the manifold, i.e., working within each local Lipschitz coordinate chart and taking advantage of the compactness of M.

We will henceforth impose no additional assumptions, other than compactness and smoothness, on the geometry of the Riemannian manifold M. In the first step we refer to a known uniform estimate from [87, Theorem 17.21] for power functions on a compact Riemannian manifold.

Lemma 4.4. *Let M be a d-dimensional C^ℓ compact manifold isometrically embedded in $\mathbb{R}^D$ (where $D > d$), and let $\Phi \in C^{2k}(M \times M)$ be any positive definite kernel function on $M \times M$ with $2k \leq \ell$. There exists a positive constant $h_0 = h_0(M) > 0$ depending only on the geometry of the manifold M such that, for any collection of n distinct points $\mathscr{X}_n = \{x_1, \ldots, x_n\}$ on M with $h_{\mathscr{X}_n} \leq h_0$, the following inequality holds:*

$$\Pi_{\Phi,\mathscr{X}_n} = \sup_{x \in M} P_{\Phi,\mathscr{X}_n}(x) \leq C h_{\mathscr{X}_n}^{2k},$$

where $C = C(k, M, \Phi) > 0$ is a positive constant depending only on the manifold M and the kernel function Φ. This of course further implies for all $h_n \leq h_0$ that

$$\inf_{\mathscr{X}_n \subset M, |\mathscr{X}_n| = n} \Pi_{\Phi,\mathscr{X}_n} \leq C h_n^{2k},$$

where h_n is the minimum fill distance of n arbitrary points on Ω (cf. (4.15)).

Proof. This is essentially [87, Theorem 17.21], with the only adaptation being that the definition of the power function throughout [87] is the square root of the $P_{\Phi,\mathscr{X}_n}$ in our definition (4.11). ∎

Lemma 4.4 suggests that the convergence of Algorithm 3.1 is faster than any polynomial of h_n. The dependence on h_n can be made more direct in terms of the number of samples n by the following geometric lemma.

Lemma 4.5. *Let M be a d-dimensional C^ℓ compact Riemannian manifold isometrically embedded in $\mathbb{R}^D$ (where $D > d$). Denote by ω_{d-1} the surface measure of the unit sphere in $\mathbb{R}^d$, and by $\mathrm{Vol}(M)$ the volume of M induced by the Riemannian metric. There exists a positive constant $N = N(M) > 0$ depending only on the manifold M such that*

$$h_n \leq \left(\frac{2^{d+1} d}{\omega_{d-1}} \mathrm{Vol}(M)\right)^{\frac{1}{d}} \cdot n^{-\frac{1}{d}} \quad \text{for any } \mathbb{N} \ni n \geq N.$$

Proof. For any $r > 0$ and $x \in M$, we denote by $B_r^D(x)$ the (extrinsic) D-dimensional Euclidean ball centered at $x \in M$ and set $B_r(x) := B_r^D(x) \cap M$. In other words, $B_r(x)$ is a ball of radius r centered at $x \in M$ with respect to the "chordal" metric on M induced from the ambient Euclidean space $\mathbb{R}^D$. Define the *covering number* and the *packing number* for M

with respect to the chordal metric balls by

$$\begin{aligned}\mathscr{N}(r) &:= \mathscr{N}(M, \|\cdot\|_D, r)\\ &:= \min_{n\in\mathbb{N}}\left\{M \subset \bigcup_{i=1}^{n} B_r(x_i) \mid x_i \in M, 1 \le i \le n\right\},\\ \mathscr{P}(r) &:= \mathscr{P}(M, \|\cdot\|_D, r)\\ &:= \max_{n\in\mathbb{N}}\left\{\bigcup_{i=1}^{n} B_{r/2}(x_i) \subset M, B_{r/2}(x_i)\cap B_{r/2}(x_j) = \emptyset\right.\\ &\qquad\qquad \left.\forall\ 1 \le i \ne j \le n \,\middle|\, x_i \in M, 1 \le i \le n\right\}.\end{aligned}$$

By the definition of fill distance and h_n (cf. (4.15)), the covering number $\mathscr{N}(h_n)$ is lower bounded by n; furthermore, by the straightforward inequality $\mathscr{P}(r) \ge \mathscr{N}(r)$ for all $r > 0$, we have

$$n < \mathscr{N}(h_n) \le \mathscr{P}(h_n);$$

i.e., there exists a collection of n points $x_1, \ldots, x_n \in M$ such that the n chordal metric balls $\{B_{h_n/2}(x_i) \mid 1 \le i \le n\}$ form a packing of M. Thus

$$\sum_{i=1}^{n} \mathrm{Vol}\left(B_{h_n/2}(x_i)\right) \le \mathrm{Vol}(M) < \infty,$$

where the last inequality follows from the compactness of M. The volume of each $B_{h_n/2}(x_i)$ can be expanded asymptotically for small h_n as (cf. [46])

$$\mathrm{Vol}\left(B_{h_n/2}(x)\right) = \frac{\omega_{d-1}}{d}\left(\frac{h_n}{2}\right)^d\left[1 + \frac{2\|B\|_x^2 - \|H\|_x^2}{8(d+2)}\left(\frac{h_n}{2}\right)^2\right] + O\left(h_n^{d+3}\right) \text{ as } h_n \to 0, \tag{4.19}$$

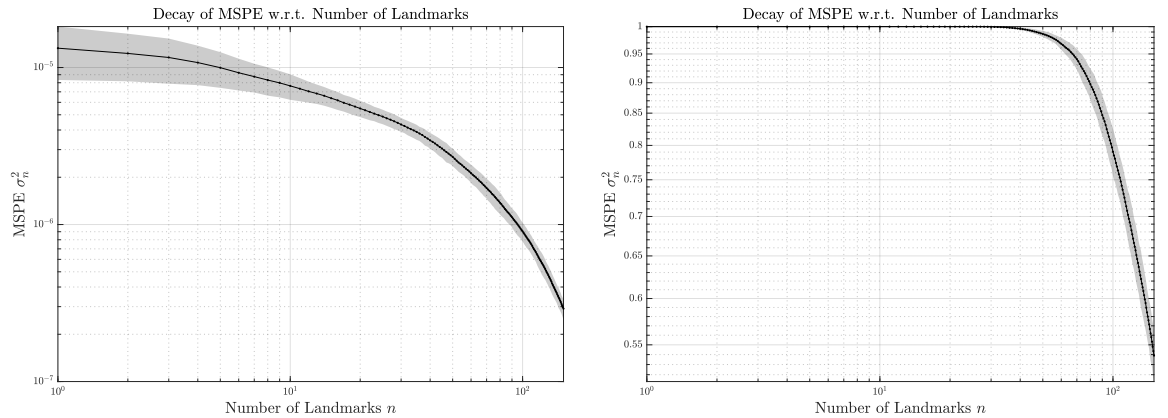
where ω_{d-1} is the surface measure of the unit sphere in $\mathbb{R}^d$, B is the second fundamental form of M, and H is the mean curvature normal. The compactness of M ensures the boundedness of all these extrinsic curvature terms. Pick n sufficiently large so that h_n is sufficiently small (again by the compactness of M) to ensure

$$\mathrm{Vol}\left(B_{h_n/2}(x)\right) \ge \frac{\omega_{d-1}}{2d}\left(\frac{h_n}{2}\right)^d.$$

It then follows from (4.19) that

$$\frac{n\omega_{d-1}}{2d}\left(\frac{h_n}{2}\right)^d \le \mathrm{Vol}(M) \quad \Rightarrow \quad h_n \le \left(\frac{2^{d+1}d}{\omega_{d-1}}\mathrm{Vol}(M)\right)^{\frac{1}{d}} \cdot n^{-\frac{1}{d}}.$$ ∎

We are now ready to conclude that Algorithm 3.1 converges faster than any inverse polynomial in the number of samples with our specific choice of kernel functions, regardless of the presence of reweighting.



(a) Reweighted kernel (4.18)

(b) Euclidean kernel (4.17)

Figure 3. *Log-log plots illustrating the convergence of MSPEs with respect to the number of Gaussian process landmarks produced using the reweighted kernel* (4.18) *or the Euclidean kernel* (4.17)*, for a collection of* 116 *second mandibular molars of prosimian primates and closely related nonprimates* [16, 37]*. Each point on either curve is obtained by averaging the* 116 *MSPEs over the entire dataset, and the transparent bands represent pointwise confidence bands of two standard deviations. For both plots we vary the number of landmarks from* 1 *to* 150*; the total number of vertices on each of the* 116 *triangular meshes vary around* 5000*. For both plots, the MSPE decays linearly for sufficiently large* n *on a log-log scale, suggesting exponential convergence with respect to the number of Gaussian process landmarks.*

Theorem 4.6. *Let* M *be a* d*-dimensional* C^∞ *compact manifold isometrically embedded in* $\mathbb{R}^D$ *(where* $D > d$*), and let* $\Phi \in C^\infty(M \times M)$ *be any positive definite kernel function on* M*. For any* $k \in \mathbb{N}$*, there exist positive constants* $N = N(M) > 0$ *and* $C_k = C_k(M, \Phi) > 0$ *such that*

$$\sigma_n \leq C_k n^{-\frac{k}{d}} \quad \forall n \geq N.$$

Equivalently speaking, Algorithm 3.1 *converges at rate* $O(n^{-\frac{k}{d}})$ *for all* $k \in \mathbb{N}$ *but with constants possibly depending on* k*.*

Proof. Use Lemmas 4.4 and 4.5 and the regularity of the kernel function Φ. ∎

It is natural to conjecture that a rate of convergence faster than the conclusion of Theorem 4.6, for instance, an exponential rate of convergence, should hold for the reweighted kernel (4.18), or at least for the Euclidean radial basis kernel (4.17); this can be empirically validated with numerical experiments; see, e.g., the log-log plots in Figure 3 depicting the decay of MSPE (i.e., σ_n^2) with respect to the increasing number of landmarks (i.e., n). Unfortunately, Theorem 4.6 is about as far as we can get with our current techniques, unless we impose additional assumptions on the regularity of the manifolds of interest. It is tempting to proceed directly as in [87, Theorem 17.21] by working locally on coordinate charts and citing the exponential convergence result for radial basis kernels in [87, Theorem 11.22]; unfortunately, even though kernel K_ϵ is of radial basis type in the ambient space $\mathbb{R}^D$, it is generally no longer of radial basis type in local coordinate charts, unless one imposes additional restrictive assumptions on the growth of the derivatives of local parametrization maps (e.g., all

coordinate maps are affine). We will not pursue the theoretical aspects of these additional assumptions in this paper.

Remark 4.7. The asymptotic optimality of the rate established in Theorem 4.6 for Gaussian process landmarking follows from Theorem 4.2. In other words, the Gaussian process landmarking algorithm leads to a rate of decay of the ∞-norm of the pointwise MSPE that is at least as fast as any other landmarking algorithm, including random or uniform sampling on the manifold. In our application of comparative biology that motivated this paper, it is more important that Gaussian process landmarking is capable of identifying biologically meaningful and operationally homologous points across the anatomical surfaces even when the number of landmarks is not large ($n \ll \infty$); see [37] for more details. A more thorough theory explaining this advantageous aspect of Gaussian process landmarking will be left for future work.

5. Discussion and future work.

This paper discusses a greedy algorithm for automatically selecting representative points on compact manifolds, motivated by the methodology of experimental design with Gaussian process prior in statistics. With a carefully modified heat kernel specified as the covariance function in the Gaussian process prior, our algorithm is capable of producing biologically highly meaningful feature points on some anatomical surfaces. Application of this landmarking scheme for real anatomical datasets is detailed in the companion paper [37].

A future direction of interest is to build theoretical analysis for the optimal experimental design aspects of manifold learning: whereas existing manifold learning algorithms estimate the underlying manifold *from* discrete samples, our algorithm concerns economical strategies for encoding geometric information *into* discrete samples. The landmarking procedure can also be interpreted as a compression scheme for manifolds; correspondingly, standard manifold learning algorithms may be understood as a decoding mechanism. Our theory is also of potential interest in adaptive matrix sensing and image completion problems, in which sensing procedures and subsampling schemes can be designed to collect more information for ease of reconstruction. Some related works of this type include [7, 84, 86] and the references therein.

The current paper stems from an attempt to impose Gaussian process priors on diffeomorphisms between distinct but comparable biological structures, with which a rigorous Bayesian statistical framework for biological surface registration may be developed. The motivation is to measure the uncertainty of pairwise bijective correspondences automatically computed from geometry processing and computer vision techniques. We hope this MSPE-based sequential landmarking algorithm will shed light on generalizing covariance structures from a single shape to pairs or even collections of shapes for collection shape analysis.

Appendix A. Reproducing kernel Hilbert spaces.

For any positive semidefinite symmetric kernel function $K : M \times M \to \mathbb{R}$ defined on a complete metric measure space M, Mercer's theorem [28, Theorem 3.6] states that K admits a uniformly convergent expansion of the form

$$K(x,y) = \sum_{i=0}^{\infty} e^{-\lambda_i} \phi_i(x)\, \phi_i(y) \quad \forall x, y \in M,$$

where $\{\phi_i\}_{i=0}^{\infty} \subset L^2(M)$ are the eigenfunctions of the integral operator

$$T_K : L^2(M) \to L^2(M),$$
$$T_K f(x) := \int_M K(x,y) f(y) \, \mathrm{dvol}_M(y) \quad \forall f \in L^2(M),$$

and $e^{-\lambda_i}, i = 0, 1, \ldots$, ordered so that $e^{-\lambda_0} \geq e^{-\lambda_1} \geq e^{-\lambda_2} \geq \cdots$, are the eigenvalues of this integral operator corresponding to the eigenfunctions $\phi_i, i = 0, 1, \ldots$, respectively. Regression under this framework amounts to restricting the regression function to lie in the Hilbert space

$$\mathscr{H}_K := \left\{ f = \sum_{i=0}^{\infty} \alpha_i \phi_i \,\middle|\, \alpha_i \in \mathbb{R}, \sum_{i=0}^{\infty} e^{\lambda_i} \alpha_i^2 < \infty \right\} \tag{A.1}$$

on which the inner product is defined as

$$\langle f, g \rangle_{\mathscr{H}_K} = \sum_{i=0}^{\infty} e^{\lambda_i} \langle f, \phi_i \rangle_{L^2(M)} \langle g, \phi_i \rangle_{L^2(M)}. \tag{A.2}$$

The reproducing property is reflected in the identity

$$\langle K(\cdot, x), K(\cdot, y) \rangle_{\mathscr{H}_K} = K(x,y) \quad \forall x, y \in M. \tag{A.3}$$

Borrowing terminologies from kernel-based learning methods (see, e.g., [28, 73]), the eigenfunctions and eigenvalues of T_K define a *feature mapping*

$$M \ni x \longmapsto \Phi(x) := \left(e^{-\lambda_0/2} \phi_0(x), e^{-\lambda_1/2} \phi_1(x), \ldots, e^{-\lambda_i/2} \phi_i(x), \ldots \right) \in \ell^2 \tag{A.4}$$

such that the kernel value $K(x,y)$ at an arbitrary pair $x, y \in M$ is given exactly by the inner product of $\Phi(x)$ and $\Phi(y)$ in the feature space ℓ^2, i.e.,

$$K(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\ell^2} \quad \forall x, y \in M.$$

This interpretation leads to the following equivalent form of the RKHS (A.1):

$$\begin{aligned} \mathscr{H}_K &= \left\{ f = \sum_{i=0}^{\infty} \beta_i \cdot e^{-\lambda_i/2} \phi_i = \langle \beta, \Phi \rangle_{\ell^2} \,\middle|\, \beta = (\beta_0, \beta_1, \ldots, \beta_i, \ldots) \in \ell^2 \right\} \\ &= \mathrm{span} \left\{ \sum_{i \in I} a_i K(\cdot, x_i) \mid a_i \in \mathbb{R}, x_i \in M, \mathrm{card}(I) < \infty \right\}. \end{aligned} \tag{A.5}$$

In other words, the RKHS framework embeds the Riemannian manifold M into an infinite dimensional Hilbert space ℓ^2 and converts the (generically) nonlinear regression problem on M into a linear regression problem on a subset of ℓ^2. We refer interested readers to [11, 44, 88] for more discussions of this type of embedding in the nonlinear dimension reduction literature.

Acknowledgments. The first author would like to thank Peng Chen (UT Austin) for pointers to the reduced basis method literature, Chen-Yun Lin (Duke) for many useful discussions on heat kernel estimates, Daniel Sanz-Alonso (University of Chicago) for general discussions on Gaussian processes and RKHS, and Yingzhou Li (Duke) and Jianfeng Lu (Duke) for discussions on the numerical linear algebra perspective. The authors would also like to thank Shaobo Han, Rui Tuo, Sayan Mukherjee, Robert Ravier, Shan Shan, and Albert Cohen for many inspirational discussions, as well as Ethan Fulwood, Bernadette Perchalski, Julia Winchester, and Arianna Harrington for assistance with collecting observer landmarks. Last but not least, the authors sincerely appreciate the constructive feedback from the anonymous reviewers.

REFERENCES

[1] Z. Allen-Zhu, Y. Li, A. Singh, and Y. Wang, *Near-optimal design of experiments via regret minimization*, in Proceedings of the International Conference on Machine Learning, 2017, pp. 126–135.

[2] Z. Allen-Zhu, Y. Li, A. Singh, and Y. Wang, *Near-Optimal Discrete Optimization for Experimental Design: A Regret Minimization Approach*, preprint, https://arxiv.org/abs/1711.05174, 2017.

[3] P. Alliez, D. Cohen-Steiner, O. Devillers, B. Lévy, and M. Desbrun, *Anisotropic polygonal remeshing*, ACM Trans. Graph., 22 (2003), pp. 485–493, https://doi.org/10.1145/882262.882296.

[4] n. aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[5] M. Atiyah and P. Sutcliffe, *The geometry of point particles*, Proc. R. Soc. Lond. A Math. Phys. Eng. Sci., 458 (2002), pp. 1089–1115.

[6] A. Atkinson, A. Donev, and R. Tobias, *Optimum Experimental Designs, with SAS*, Oxford Statist. Sci. Ser. 34, Oxford University Press, 2007.

[7] H. Avron and C. Boutsidis, *Faster subset selection for matrices and applications*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 1464–1499, https://doi.org/10.1137/120867287.

[8] C. Beard, *Teilhardina,* in The International Encyclopedia of Primatology, John Wiley & Sons, 2017, pp. 1369–1370, https://doi.org/10.1002/9781119179313.wbprim0444.

[9] M. Belkin and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Comput., 15 (2003), pp. 1373–1396, https://doi.org/10.1162/089976603321780317.

[10] M. Belkin and P. Niyogi, *Towards a Theoretical Foundation for Laplacian-Based Manifold Methods*, in Learning Theory, Springer, 2005, pp. 486–500.

[11] P. Bérard, G. Besson, and S. Gallot, *Embedding Riemannian manifolds by their heat kernel*, Geom. Funct. Anal., 4 (1994), pp. 373–398, https://doi.org/10.1007/BF01896401.

[12] N. Berline, E. Getzler, and M. Vergne, *Heat Kernels and Dirac Operators*, Springer, 2004.

[13] A. L. Bertozzi, X. Luo, A. M. Stuart, and K. C. Zygalakis, *Uncertainty quantification in graph-based classification of high dimensional data,* SIAM/ASA J. Uncertainty Quantification, 6 (2018), pp. 568–595, https://doi.org/10.1137/17M1134214.

[14] 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 (2011), pp. 1457–1472, https://doi.org/10.1137/100795772.

[15] M. Bouhtou, S. Gaubert, and G. Sagnol, *Submodularity and randomized rounding techniques for optimal experimental design*, Electron. Notes Discrete Math., 36 (2010), pp. 679–686.

[16] D. M. Boyer, Y. Lipman, E. St. Clair, J. Puente, B. A. Patel, T. Funkhouser, J. Jernvall, and I. Daubechies, *Algorithms to automatically quantify the geometric similarity of anatomical surfaces*, Proc. Natl. Acad. Sci. USA, 108 (2011), pp. 18221–18226, https://doi.org/10.1073/pnas.1112822108.

[17] A. Buffa, Y. Maday, A. T. Patera, C. Prud'homme, and G. Turinici, *A priori convergence of the greedy algorithm for the parametrized reduced basis method*, ESAIM Math. Model. Numer. Anal., 46 (2012), pp. 595–603.

[18] I. Castillo, G. Kerkyacharian, and D. Picard, *Thomas Bayes' walk on manifolds*, Probab. Theory Related Fields, 158 (2014), pp. 665–710, https://doi.org/10.1007/s00440-013-0493-0.

[19] M. Černỳ and M. Hladík, *Two complexity results on C-optimality in experimental design*, Comput. Optim. Appl., 51 (2012), pp. 1397–1408.
[20] K. Chaloner and I. Verdinelli, *Bayesian Experimental design: A review*, Statist. Sci., 10 (1995), pp. 273–304.
[21] K. Chaudhuri, S. M. Kakade, P. Netrapalli, and S. Sanghavi, *Convergence rates of active learning for maximum likelihood estimation*, in Advances in Neural Information Processing Systems, 2015, pp. 1090–1098.
[22] X. Cheng, A. Cloninger, and R. R. Coifman, *Two-Sample Statistics Based on Anisotropic Kernels*, preprint, https://arxiv.org/abs/1709.05006, 2017.
[23] A. Çivril and M. Magdon-Ismail, *On selecting a maximum volume sub-matrix of a matrix and related problems*, Theoret. Comput. Sci., 410 (2009), pp. 4801–4811.
[24] D. Cohen-Steiner and J.-M. Morvan, *Restricted Delaunay triangulations and normal cycle*, in Proceedings of the 19th Annual Symposium on Computational Geometry, ACM, 2003, pp. 312–321.
[25] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, *Active learning with statistical models*, J. Artificial Intelligence Res., 4 (1996), pp. 129–145.
[26] R. R. Coifman and S. Lafon, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5–30, https://doi.org/10.1016/j.acha.2006.04.006.
[27] N. Cressie, *Statistics for Spatial Data*, Wiley Ser. Probab. Statist., John Wiley & Sons, 2015.
[28] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.
[29] L. Cuel, J.-O. Lachaud, Q. Mérigot, and B. Thibert, *Robust geometry estimation using the generalized Voronoi covariance measure*, SIAM J. Imaging Sci., 8 (2015), pp. 1293–1314, https://doi.org/10.1137/140977552.
[30] V. De Silva and J. B. Tenenbaum, *Sparse Multidimensional Scaling Using Landmark Points*, Technical report, Stanford University, 2004.
[31] R. DeVore, G. Petrova, and P. Wojtaszczyk, *Greedy algorithms for reduced bases in Banach spaces*, Construct. Approx., 37 (2013), pp. 455–466.
[32] P. Dhillon, Y. Lu, D. P. Foster, and L. Ungar, *New subsampling algorithms for fast least squares regression*, in Advances in Neural Information Processing Systems, 2013, pp. 360–368.
[33] I. L. Dryden and K. V. Mardia, *Statistical Shape Analysis*, John Wiley & Sons, 1998.
[34] V. Fedorov, *Theory of Optimal Experiments*, translated from the Russian and edited by W. J. Studden and E. M. Klimko, Probab. Math. Statist. 12, Academic Press, 1972.
[35] T. E. Fricker, J. E. Oakley, and N. M. Urban, *Multivariate Gaussian process emulators with nonseparable covariance structures*, Technometrics, 55 (2013), pp. 47–56.
[36] E. Fuselier and G. B. Wright, *Scattered data interpolation on embedded submanifolds with restricted positive definite kernels: Sobolev error estimates*, SIAM J. Numer. Anal., 50 (2012), pp. 1753–1776, https://doi.org/10.1137/110821846.
[37] T. Gao, S. Z. Kovalsky, D. M. Boyer, and I. Daubechies, *Gaussian process landmarking for three-dimensional geometric morphometrics*, SIAM J. Math. Data Sci., 1 (2019), pp. 237–267, https://doi.org/10.1137/18M1203481.
[38] T. F. Gonzalez, *Clustering to minimize the maximum intercluster distance*, Theoret. Comput. Sci., 38 (1985), pp. 293–306.
[39] J. C. Gower, *Generalized procrustes analysis*, Psychometrika, 40 (1975), pp. 33–51, https://doi.org/10.1007/BF02291478.
[40] J. C. Gower and G. B. Dijksterhuis, *Procrustes Problems*, Oxford Statist. Sci. Ser. 3, Oxford University Press, 2004.
[41] H. Harbrecht, M. Peters, and R. Schneider, *On the low-rank approximation by the pivoted Cholesky decomposition*, Appl. Numer. Math., 62 (2012), pp. 428–440, https://doi.org/10.1016/j.apnum.2011.10.001.
[42] D. Hardin and E. Saff, *Minimal Riesz energy point configurations for rectifiable D-dimensional manifolds*, Adv. in Math., 193 (2005), pp. 174–204.
[43] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, 2002, https://doi.org/10.1137/1.9780898718027.
[44] P. W. Jones, M. Maggioni, and R. Schul, *Manifold parametrizations by eigenfunctions of the Lapla-*

cian and heat kernels, Proc. Natl. Acad. Sci. USA, 105 (2008), pp. 1803–1808, https://doi.org/10.1073/pnas.0710175104.

[45] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell, *Active learning with Gaussian processes for object categorization*, in Proceedings of the IEEE 11th International Conference on Computer Vision, IEEE, 2007, pp. 1–8.

[46] L. Karp and M. Pinsky, *Volume of a small extrinsic ball in a submanifold*, Bull. London Math. Soc., 21 (1989), pp. 87–92.

[47] A. Krause, A. Singh, and C. Guestrin, *Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies*, J. Mach. Learn. Res., 9 (2008), pp. 235–284.

[48] D. D. Lewis and W. A. Gale, *A sequential algorithm for training text classifiers*, in Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Springer-Verlag, 1994, pp. 3–12.

[49] D. Liang and J. Paisley, *Landmarking manifolds with Gaussian processes*, in Proceedings of the 32nd International Conference on Machine Learning, Vol. 37, JMLR.org, 2015, pp. 466–474.

[50] L. Lin, M. Niu, P. Cheung, and D. Dunson, *Extrinsic Gaussian Processes for Regression and Classification on Manifolds*, preprint, https://arxiv.org/abs/1706.08757, 2017.

[51] A. W. Long and A. L. Ferguson, *Landmark diffusion maps (L-dMaps): Accelerated manifold learning out-of-sample extension*, Appl. Comput. Harmon. Anal., published online August 31, 2017, https://doi.org/10.1016/j.acha.2017.08.004

[52] G. G. Lorentz, M. von Golitschek, and Y. Makovoz, *Constructive Approximation: Advanced Problems*, Grundlehren Math. Wiss. 304, Springer, Berlin, 1996.

[53] A. Martinez-Finkelshtein, V. Maymeskul, E. Rakhmanov, and E. Saff, *Asymptotics for minimal discrete Riesz energy on curves in R^d*, Canad. J. Math, 56 (2004), pp. 529–552.

[54] C. Moenning and N. A. Dodgson, *Fast Marching Farthest Point Sampling*, Tech. report, Computer Laboratory, University of Cambridge, 2003.

[55] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, MIT Press, 2012.

[56] S. Molnar, *On the convergence of the kriging method*, Ann. Univ. Sci. Budapest. Sect. Comput., 6 (1985), pp. 81–90.

[57] F. J. Narcowich, J. D. Ward, and H. Wendland, *Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions*, Construct. Approx., 24 (2006), pp. 175–186.

[58] A. Nikolov, *Randomized rounding for the largest simplex problem*, in Proceedings of the 47th Annual ACM Symposium on Theory of Computing, ACM, 2015, pp. 861–870.

[59] S. Niranjan, A. Krause, S. M. Kakade, and M. Seeger, *Gaussian process optimization in the bandit setting: No regret and experimental design*, in Proceedings of the 27th International Conference on Machine Learning, Omnipress, 2010, pp. 1015–1022.

[60] J. E. Oakley and A. O'Hagan, *Probabilistic sensitivity analysis of complex models: A Bayesian approach*, J. R. Statist. Soc. Ser. B Statist. Methodol., 66 (2004), pp. 751–769.

[61] A. C. Öztireli, M. Alexa, and M. Gross, *Spectral sampling of manifolds*, ACM Trans. Graphics (TOG), 29 (2010), 168.

[62] F. Pukelsheim, *Optimal Design of Experiments*, Classics Appl. Math. 50, SIAM, 2006, https://doi.org/10.1137/1.9780898719109.

[63] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Adapt. Comput. Mach. Learn., MIT Press, 2006.

[64] M. Reed and B. Simon, *Methods of Modern Mathematical Physics. Vol.* 1, *Functional Analysis*, Academic Press, 1980.

[65] K. Ritter, *Average-Case Analysis of Numerical Problems*, Springer, 2007.

[66] E. Rodolá, A. Albarelli, D. Cremers, and A. Torsello, *A simple and effective relevance-based point sampling for 3D shapes*, Pattern Recognition Lett., 59 (2015), pp. 41–47, https://doi.org/10.1016/j.patrec.2015.03.009.

[67] L. Rosasco, M. Belkin, and E. D. Vito, *On learning with integral operators*, J. Mach. Learn. Res., 11 (2010), pp. 905–934.

[68] S. Rosenberg, *The Laplacian on a Riemannian Manifold: An Introduction to Analysis on Manifolds*, London Math. Soc. Stud. Texts 31, Cambridge University Press, 1997.

[69] R. B. Rusu and S. Cousins, *3D is here: Point Cloud Library (PCL)*, in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2011.
[70] J. Sacks, S. B. Schiller, and W. J. Welch, *Designs for computer experiments*, Technometrics, 31 (1989), pp. 41–47.
[71] A. Saltelli and S. Tarantola, *On the relative importance of input factors in mathematical models: Safety assessment for nuclear waste disposal*, J. Amer. Statist. Assoc., 97 (2002), pp. 702–709.
[72] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*, Springer Ser. Statist., Springer Science+Business Media, 2013.
[73] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, 2001.
[74] B. Settles, *Active Learning Literature Survey*, Tech. report 1648, Computer Sciences, University of Wisconsin–Madison, 2010, http://burrsettles.com/pub/settles.activelearning.pdf.
[75] M. T. Silcox, *Plesiadapiform,* in The International Encyclopedia of Primatology, John Wiley & Sons, 2017, pp. 999–1000, https://doi.org/10.1002/9781119179313.wbprim0038.
[76] A. Singer, *From Graph to manifold Laplacian: The convergence rate*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 128–134.
[77] A. Singer and H.-T. Wu, *Vector diffusion maps and the connection Laplacian*, Comm. Pure Appl. Math., 65 (2012), pp. 1067–1144, https://doi.org/10.1002/cpa.21395.
[78] K. Smith, *On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations*, Biometrika, 12 (1918), pp. 1–85.
[79] M. L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*, Springer Science+Business Media, 2012.
[80] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.
[81] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, Springer, 2008.
[82] A. W. van der Vaart and J. H. van Zanten, *Reproducing kernel Hilbert spaces of Gaussian priors*, in Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh, Institute of Mathematical Statistics, 2008, pp. 200–222.
[83] C. Wachinger and P. Golland, *Diverse Landmark Sampling from Determinantal Point Processes for Scalable Manifold Learning*, preprint, https://arxiv.org/abs/1503.03506, 2015.
[84] S. Wang and Z. Zhang, *Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling*, J. Mach. Learn. Res., 14 (2013), pp. 2729–2769.
[85] W. Wang, R. Tuo, and C. F. J. Wu, *Universal Convergence of Kriging*, preprint, https://arxiv.org/abs/1710.06959v1, 2017.
[86] Y. Wang, A. W. Yu, and A. Singh, *On computationally tractable selection of experiments in measurement-constrained regression models*, J. Mach. Learn. Res., 18 (2017), pp. 5238–5278.
[87] H. Wendland, *Scattered Data Approximation*, Cambridge Monogr. Appl. Comput. Math. 17, Cambridge University Press, 2004.
[88] H.-T. Wu, *Embedding Riemannian manifolds by the heat kernel of the connection Laplacian*, Adv. in Math., 304 (2017), pp. 1055–1079.
[89] Z.-M. Wu and R. Schaback, *Local error estimates for radial basis function interpolation of scattered data*, IMA J. Numer. Anal., 13 (1993), pp. 13–27.
[90] H. Xu, H. Zha, R.-C. Li, and M. A. Davenport, *Active manifold learning via Gershgorin circle guided sample selection*, in Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, 2015, pp. 3108–3114.
[91] Y. Yang and D. B. Dunson, *Bayesian manifold regression*, Ann. Statist., 44 (2016), pp. 876–905.
[92] D. Ylvisaker, *Designs on random fields*, in A Survey of Statistical Design and Linear Models (Proc. Internat. Sympos., Colorado State Univ., Ft. Collins, CO, 1973), J. N. Srivastava, ed., North-Holland, 1975, pp. 593–607.
[93] X. Zhu, J. Lafferty, and Z. Ghahramani, *Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions*, in Proceedings of the ICML 2003 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, Washington, DC, 2003, pp. 58–65, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.5.5447.