The Annals of Statistics
2016, Vol. 44, No. 2, 876–905
DOI: 10.1214/15-AOS1390

BAYESIAN MANIFOLD REGRESSION

BY YUN YANG[1] AND DAVID B. DUNSON

University of California, Berkeley and Duke University

There is increasing interest in the problem of nonparametric regression with high-dimensional predictors. When the number of predictors D is large, one encounters a daunting problem in attempting to estimate a D-dimensional surface based on limited data. Fortunately, in many applications, the support of the data is concentrated on a d-dimensional subspace with $d \ll D$. Manifold learning attempts to estimate this subspace. Our focus is on developing computationally tractable and theoretically supported Bayesian nonparametric regression methods in this context. When the subspace corresponds to a locally-Euclidean compact Riemannian manifold, we show that a Gaussian process regression approach can be applied that leads to the minimax optimal adaptive rate in estimating the regression function under some conditions. The proposed model bypasses the need to estimate the manifold, and can be implemented using standard algorithms for posterior computation in Gaussian processes. Finite sample performance is illustrated in a data analysis example.

1. Introduction. Dimensionality reduction in nonparametric regression is of increasing interest given the routine collection of high-dimensional predictors. Our focus is on the regression model

$$Y_i = f(X_i) + w_i, w_i \sim N(0, \sigma^2), \qquad i = 1, \ldots, n, \tag{1.1}$$

where $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^D$, f is an unknown regression function, and w_i is a residual having variance σ^2. We face problems in estimating f accurately due to the moderate to large number of predictors D. Fortunately, in many applications, the predictors have support that is concentrated near a d-dimensional subspace $\mathcal{M}$. If one can learn the mapping from the ambient space to this subspace, the dimensionality of the regression function can be reduced massively from D to d, so that f can be much more accurately estimated.

There is an increasingly vast literature on subspace learning, but there remains a lack of approaches that allow flexible nonlinear dimensionality reduction, are scalable computationally to moderate to large D, have theoretical guarantees and provide a characterization of uncertainty. Castillo et al. [10] directly constructed

Received December 2014; revised September 2015.

[1]Supported by Grant ES017436 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH).

MSC2010 subject classifications. Primary 62H30, 62-07; secondary 65U05, 68T05.

Key words and phrases. Asymptotics, contraction rates, dimensionality reduction, Gaussian process, manifold learning, nonparametric Bayes, subspace learning.

a nonstationary Gaussian process prior on a known manifold though rescaling the solutions of the heat equation. However, in many cases, the manifold is not known in advance.

With this motivation, we focus on Bayesian nonparametric regression methods that allow $\mathcal{M}$ to be an unknown Riemannian manifold. One natural direction is to choose a prior to allow uncertainty in $\mathcal{M}$, while also placing priors on the mapping from x_i to $\mathcal{M}$, the regression function relating the lower-dimensional features to the response, and the residual variance. Some related attempts have been made in the literature. Tokdar et al. [31] propose a logistic Gaussian process model, which allows the conditional response density $f(y|x)$ to be unknown and changing flexibly with x, while reducing dimension through projection to a linear subspace. Their approach is elegant and theoretically grounded, but does not scale efficiently as D increases and is limited by the linear subspace assumption. Also making the linear subspace assumption, [25] proposed a Bayesian finite mixture model for sufficient dimension reduction. Page et al. [24] instead propose a method for Bayesian nonparametric learning of an affine subspace motivated by classification problems.

There is also a limited literature on Bayesian nonlinear dimensionality reduction. Gaussian process latent variable models (GP-LVMs) (Lawrence [19]) were introduced as a nonlinear alternative to PCA for visualization of high-dimensional data. Kundu and Dunson [18] proposed a related approach that defines separate Gaussian process regression models for the response and each predictor, with these models incorporating shared latent variables to induce dependence. The latent variables can be viewed as coordinates on a lower dimensional manifold, but daunting problems arise in attempting to learn the number of latent variables, the distribution of the latent variables, and the individual mapping functions while maintaining identifiability restrictions. Chen et al. [11] instead approximate the manifold through patching together hyperplanes. Such mixtures of linear subspace-based methods may require a large number of subspaces to obtain an accurate approximation even when d is small.

It is clear that probabilistic models for learning the manifold face daunting statistical and computational hurdles. In this article, we take a very different approach in attempting to define a simple and computationally tractable model, which bypasses the need to estimate $\mathcal{M}$ but can exploit the lower-dimensional manifold structure when it exists. In particular, our goal is to define an approach that obtains a minimax-optimal adaptive rate in estimating f, with the rate adaptive to the manifold and smoothness of the regression function. Surprisingly, we show that this can be achieved with a simple Gaussian process prior.

Section 2 provides background and our main results. Section 3 discusses two approaches to construct intrinsic dimension adaptive estimators. Section 4 contains a toy example and a simulation study of finite sample performance relative to competitors. Section 5 provides auxiliary results that are crucial for proving

the main results. Technical proofs are deferred to Section 6. A review of necessary geometric properties and selected proofs are included in the supplementary material [36].

2. Gaussian processes on manifolds.

2.1. *Background.* Gaussian processes (GP) are widely used as prior distributions for unknown functions. For example, in the nonparametric regression (1.1), a GP can be specified as a prior for the unknown function f. In classification, the conditional distribution of the binary response Y_i is related to the predictor X_i through a known link function h and a regression function f as $Y_i|X_i \sim \mathrm{Ber}[h\{f(X_i)\}]$, where f is again given a GP prior. The following developments will mainly focus on the regression case. The GP with squared exponential covariance is a commonly used prior in the literature. The law of the centered squared exponential GP $\{W_x : x \in \mathcal{X}\}$ is entirely determined by its covariance function,

$$K^a(x, y) = EW_xW_y = \exp\bigl(-a^2\|x - y\|^2/2\bigr), \tag{2.1}$$

where the predictor domain $\mathcal{X}$ is a subset of $\mathbb{R}^D$, $\|\cdot\|$ is the usual Euclidean norm and a is a length scale parameter. Although we focus on the squared exponential case, our results can be extended to a broader class of covariance functions with exponentially decaying spectral density, including standard choices such as Matérn, with some elaboration. We use $\mathrm{GP}(m, K)$ to denote a GP with mean function $m : \mathcal{X} \to \mathbb{R}$ and covariance function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

Given n independent observations, the minimax rate of estimating a D-variate function that is only known to be Hölder s-smooth is $n^{-s/(2s+D)}$ [28]. [35] proved that, for Hölder s-smooth functions, a prior specified as

$$W^A|A \sim \mathrm{GP}\bigl(0, K^A\bigr), \qquad A^D \sim \mathrm{Ga}(a_0, b_0), \tag{2.2}$$

for $\mathrm{Ga}(a_0, b_0)$ the Gamma distribution with p.d.f. $p(t) \propto t^{a_0-1}e^{-b_0t}$ leads to the minimax rate $n^{-s/(2s+D)}$ up to a logarithmic factor $(\log n)^\beta$ with $\beta \sim D$ adaptively over all $s > 0$ without knowing s in advance. The superscript in W^A indicates the dependence on the random scaling or inverse bandwidth parameter A.

In many real problems, the predictor X can be represented as a vector in high dimensional Euclidean space $\mathbb{R}^D$, where D is called the ambient dimensionality. When D is large, assumptions are required to conquer the notorious curse of dimensionality. One common assumption requires that f only depends on a small number $d \ll n$ of components of the vector X that are identified as important. In the GP prior framework, [27] proposed to use "spike and slab" type point mass mixture priors for different scaling parameters for each component of X to do Bayesian variable selection. Assuming the function is flat in all but d directions, [3] showed that a dimension-specific scaling prior for inverse bandwidth parameters can lead to a near minimax rate for anisotropic smooth functions. We instead

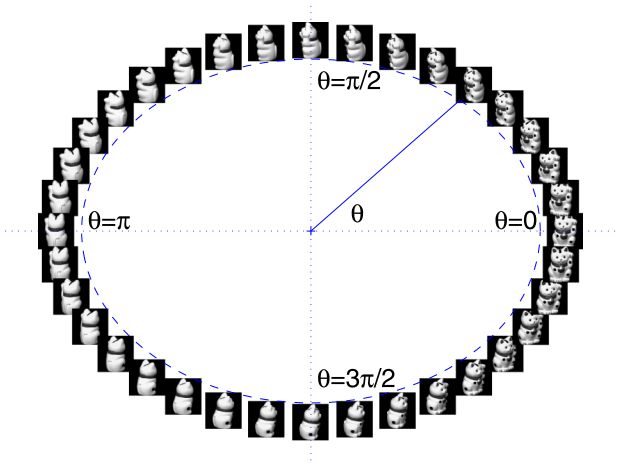

FIG. 1. *In these data*, 72 *size* 128×128 *images were taken for a "lucky cat" from different angles*: *one at every* 5 *degrees of rotation*. 36 *images are displayed in this figure*.

assume that the predictor lies on a manifold $\mathcal{M}$ of intrinsic dimension d with $d \ll D$. An example is shown in Figure 1. These data [23] consist of 72 images of a "lucky cat" taken from different angles $5°, 10°, \ldots$. The predictor $X \in \mathbb{R}^{128^2}$ is obtained by vectorizing the 128×128 image. The response Y is a continuous function f of the rotation angle $\theta \in [0, 2\pi]$ satisfying $f(0) = f(2\pi)$, such as sin or cos functions. Intuitively, the predictor X concentrates on a circle in $D = 128^2$-dim ambient space, and thus the intrinsic dimension d of X is equal to one, the dimension of the rotation angle θ.

2.2. *Bayesian regression on manifold*. When $X \in \mathcal{M}$ with the manifold $\mathcal{M}$ d-dimensional, a natural question is whether we can achieve the intrinsic rate $n^{-s/(2s+d)}$ for f Hölder s-smooth without estimating $\mathcal{M}$.

Kpotufe [16] and Kpotufe and Dasgupta [17] used random projection trees to partition the ambient space and constructed a piecewise constant estimator based on the partition. The authors showed that their estimator has a convergence rate at least $n^{-1/(2+k)}$ for Lipschitz continuous functions that is adaptive to the intrinsic dimension d, where k is guaranteed to be of order $O(d \log d)$. A more general framework is considered in [7] and [6], which covers the case where covariates lie on a low dimensional manifold in $\mathbb{R}^D$. They studied partition-based estimators and proved an $n^{-r/(2r+1)}$ rate, where r depends on how well the truth f can be approximated by their class. However, it is not clear whether their class of piecewise polynomial functions in $\mathbb{R}^D$ can adapt to manifold structures.

Ye and Zhou [37] showed that a least squares regularized algorithm with appropriate d-dependent regularization parameter can achieve a convergence rate at least $n^{-s/(8s+4d)}(\log n)^{2s/(8s+4d)}$ for functions with Hölder smoothness $s \leq 1$. Bickel

and Li [5] proved that local polynomial regression with d-dependent bandwidth can attain the minimax rate $n^{-s/(2s+d)}$ for functions with Hölder smoothness $s \leq 2$. However, similar adaptive properties have not been established for a Bayesian procedure. In this paper, we prove that a GP prior on the regression function with a proper prior for the scaling parameter leads to the minimax rate for functions with Hölder smoothness $s \leq \{2, \gamma - 1\}$, where γ is the smoothness of the manifold $\mathcal{M}$. Moreover, we describe two approaches to construct an intrinsic dimension adaptive estimator in Section 3. The first estimator of d is independent of the GP prior and only based on the covariates $\{X_i\}$, since most information about the manifold is contained in $\{X_i\}$. The second estimator of d is based on cross validation and uses the posterior mean of the GP prior. Unlike the first estimator, the second estimator cannot guarantee consistently estimating d, but still yields an optimal convergence rate for estimating the regression function f. However, the second estimator does not need any regularity assumption on the distribution of X_i other than i.i.d. In the remainder of this section, we first propose the model, and then provide a heuristic argument explaining the possibility of manifold adaptation.

Analogous to (2.2), we propose the prior for the regression function f as

$$W^A | A \sim \mathrm{GP}\big(0, K^A\big), \qquad A^d \sim \mathrm{Ga}(a_0, b_0), \tag{2.3}$$

where d is the intrinsic dimension of the manifold $\mathcal{M}$ and K^a is defined as in (2.1) with $\|\cdot\|$ the Euclidean norm of the ambient space $\mathbb{R}^D$. Adaptation to unknown intrinsic dimensionality is considered in Section 3. Although the GP in (2.3) is specified through embedding in the $\mathbb{R}^D$ ambient space, we essentially obtain a GP on $\mathcal{M}$ if we view the covariance function K^a as a bivariate function defined on $\mathcal{M} \times \mathcal{M}$. Moreover, this prior has two major differences with usual GPs or GP with Bayesian variable selection:

1. Unlike GP with Bayesian variable selection, all predictors are used in the calculation of the covariance function K^a;
2. The dimension D in the prior for inverse bandwidth A is replaced with the intrinsic dimension d.

Intuitively, one would expect that a geodesic distance should be used in the squared exponential covariance function (2.1). However, there are two main advantages of using the Euclidean distance instead of a geodesic distance. First, when a geodesic distance is used, the covariance function may fail to satisfy the positive definiteness requirement. In contrast, with the Euclidean distance in (2.1), K^a is ensured to be positive definite. Second, for a given manifold $\mathcal{M}$, a geodesic distance can be specified in many ways through different Riemannian metrics on $\mathcal{M}$. However, different geodesic distances are equivalent to each other and to the Euclidean distance on $\mathbb{R}^D$. Therefore, by using the Euclidean distance, we bypass the need to estimate a geodesic distance, but still reflect the geometric structure of the observed predictors in terms of pairwise distances. In addition, although

we use the full data in the calculation of the covariance function, computation is fast for moderate sample sizes n regardless of the size of D since only pairwise Euclidean distances among D-dimensional predictors are involved, whose computational complexity scales linearly in D.

In this work, we primarily focus on compact manifolds without boundary. The study of manifolds with boundaries is beyond the scope of this paper, since boundaries usually have smaller dimensions than the intrinsic dimension of the manifold. As a consequence, in order to achieve optimal rate on boundaries, we may need to consider nonstationary Gaussian process priors, whose length scale parameter A varies on the manifold. However, if we stick with the prior (2.3), then we conjecture that the rate is still optimal in the interior, but suboptimal on the boundaries.

Now we provide some heuristic explanations on why the rate is adaptive to the predictor manifold. Although the ambient space is $\mathbb{R}^D$, the support $\mathcal{M}$ of the predictor X is a d dimension submanifold of $\mathbb{R}^D$. As a result, the GP prior specified in Section 2.1 has all probability mass on the functions supported on this manifold, leading the posterior contraction rate to entirely depend on the evaluations of f on $\mathcal{M}$. Following [13, 14], the posterior contraction rate of the GP prior is said to be at least ε_n under a semimetric d_n if

$$\Pi\big(d_n(f, f_0) > \varepsilon_n | S_n\big) \to 0 \qquad \text{in probability as } n \to \infty,$$

where $S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ denotes the dataset and $\Pi(A|S_n)$ is the posterior of A. Here, the semimetric d_n measures the discrepancy between f and the truth f_0. For example, d_n can be chosen as the empirical L_2 metric $\|\cdot\|_n$ defined through $\|f - f_0\|_n^2 = (1/n)\sum_{i=1}^n (f(x_i) - f_0(x_i))^2$ in fixed design and the L_2 metric $\|\cdot\|_2$ defined through $\|f - f_0\|_2^2 \triangleq \int_{\mathcal{M}} (f(x) - f_0(x))^2 Q(dx)$, with Q the marginal distribution for predictor X, in random design. Since the semimetric d_n solely depends on evaluations of f on the manifold $\mathcal{M}$, we only need to fit and infer f on $\mathcal{M}$ for the purpose of making predictions. Consider a special case when the points on manifold $\mathcal{M}$ have a global smooth representation $x = \phi(t)$, where $t \in \mathbb{R}^d$ is the global latent coordinate of x. Then the regression function

$$f(x) = f\big[\phi(t)\big] \triangleq h(t), \qquad t \in \mathbb{R}^d, \tag{2.4}$$

is essentially a d-variate s-smooth function when ϕ is sufficiently smooth. Consequently, estimating the function f on $\mathbb{R}^D$ boils down to the estimation of a function h on $\mathbb{R}^d$, which makes the intrinsic rate attainable.

2.3. *Convergence rate under fixed design.* Let $C^s(\mathcal{M})$ be the Banach space of functions on $\mathcal{M}$ with Hölder smoothness s. The following theorem is our main result which provides posterior convergence rate under fixed design.

THEOREM 2.1. *Assume that $\mathcal{M}$ is a d-dimensional compact C^γ submanifold of R^D. For any $f_0 \in C^s(\mathcal{M})$ with $s \le \min\{2, \gamma - 1\}$, if we specify the prior as* (2.2),

then (5.1) *below will be satisfied for* ε_n *a multiple of* $n^{-s/(2s+d)}(\log n)^{\kappa_1}$ *and* $\bar{\varepsilon}_n$ *a multiple of* $\varepsilon_n(\log n)^{\kappa_2}$ *with* $\kappa_1 = (1+d)/(2+d/s)$ *and* $\kappa_2 = (1+d)/2$. *This implies that the posterior contraction rate with respect to* $\|\cdot\|_n$ *will be at least a multiple of* $n^{-s/(2s+d)}(\log n)^{d+1}$.

The ambient space dimensionality D implicitly influences the rate through a multiplicative constant. This theorem suggests that the Bayesian model (2.3) can adapt to both the low dimensional manifold structure of X and the smoothness $s \leq 2$ of the regression function. The condition $s \leq 2$ on the smoothness is due to the order of error in approximating the intrinsic distance $d_{\mathcal{M}}$ by the Euclidean distance d (Proposition 7.5 in the supplementary material [36]).

Generally, the intrinsic dimension d is unknown and needs to be estimated. In the case when the intrinsic dimensionality d is misspecified as d', the following result still ensures the rate to be much better than $n^{-O(1/D)}$ when d' is not too small, although the rate may become suboptimal.

THEOREM 2.2. *Assume the same conditions as in Theorem* 2.1, *but with the prior specified as* (2.2) *with* $d' \neq d$ *and* $d' > d^2/(2s+d)$.

1. *If* $d' > d$, *then the posterior contraction rate with respect to* $\|\cdot\|_n$ *will be at least a multiple of* $n^{-s/(2s+d')}(\log n)^{\kappa}$, *where* $\kappa = (1+d)/(2+d'/s)$;

2. *If* $\frac{d^2}{2s+d} < d' < d$, *then the posterior contraction rate with respect to* $\|\cdot\|_n$ *will be at least a multiple of* $n^{-((2s+d)d'-d^2)/(2(2s+d)d')}(\log n)^{\kappa}$, *where* $\kappa = (d + d^2)/(2d' + dd'/s) + (1+d)/2$.

2.4. *Convergence rate under random design.* Theorem 2.1 characterizes the posterior contraction rate in fixed design. In general, convergence rate in random design is more challenging. Van Der Vaart and Van Zanten [33] obtains posterior convergence rates for regression on Euclidean space $\mathbb{R}^d$ using GP priors. However, they require $s \geq d/2$ for estimating an s-smooth function. This assumption restricts the applicability of Theorem 2.1 as it assumes $s \leq 2$. Van Der Vaart and Van Zanten [33] also makes a crucial assumption that the prior puts all its mass over s-smooth function spaces, which excludes the interesting case of estimating nonanalytic functions with the squared exponential covariance function.

Instead of directly proving that Theorem 2.1 works in random design, we take a different approach by post-processing the posterior. We show that the post-processed posterior can achieve the same rate $\bar{\varepsilon}_n$ with respect to $\|\cdot\|_2$, and the corresponding Bayes estimator $\hat{f}$ satisfies $\|\hat{f} - f_0\|_2 \lesssim \bar{\varepsilon}_n$ with high probability. Here, we use the notation $a_n \lesssim b_n$ to indicate $a_n = O(b_n)$ as $n \to \infty$. Usual empirical process theory (Lemma 6.2, or [32]) requires the function space to be uniformly bounded in order for the empirical norm $\|\cdot\|_n$ to be comparable to the L_2 norm $\|\cdot\|_2$ uniformly over the space. This motivates us to truncate the functions sampled from the posterior distribution. The idea of post-processing a posterior

has also been considered in [21] for Bayesian monotone regression, where posterior samples are projected into the monotone function space.

Let A be any upper bound for $\|f_0\|_\infty$. In practice, one may choose A based on prior knowledge or set A to be a large multiple of $\max_{i=1,\dots,m}|\tilde{Y}_i|$ based on another dataset $\{(\tilde{X}_i, \tilde{Y}_i)\}_{i=1}^m$ independent of S_n. In the rest of the paper, we assume that an upper bound A for $\|f_0\|_\infty$ is known. For any function f, denote its truncation by A as $f_A = (f \vee (-A)) \wedge A$. Then our post-processed posterior is the posterior of f_A and the corresponding estimator $\hat{f}_A$ is given by $\hat{f}_A(x) = \int f_A(x)\, d\Pi(f|S_n)$, which is the posterior expectation of f_A. In practice, $\hat{f}_A$ can be easily obtained by taking the average of $\{f_A^{(j)} : j = 1, \dots, N\}$ where $\{f^{(j)} : j = 1, \dots, N\}$ are sampled from the posterior distribution of f. The reasons for truncating f in the posterior are two-fold. For practical purposes, truncation will never deteriorate an estimator, that is, $|f_A(x) - f_0(x)| \le |f(x) - f_0(x)|$ for all x as long as $A \ge \|f_0\|_\infty$. For theoretical purposes, we require the estimator to be bounded in order to compare $\|\cdot\|_n$ and $\|\cdot\|_2$ by applying results in empirical process theory.

The following theorem shows that the truncated GP posterior contracts to f_0 at a near minimax-optimal rate with respect to both $\|\cdot\|_n$ and $\|\cdot\|_2$. Moreover, the corresponding estimator $\hat{f}_A$ is near minimax-optimal in both fixed design and random design. A proof is provided in Section 6.

THEOREM 2.3. *Assume the same conditions as in Theorem 2.1. In addition, suppose that $X_1, \dots, X_n$ are i.i.d. samples coming from a distribution Q supported on the manifold $\mathcal{M}$ and the true function f_0 satisfies $\|f_0\|_\infty \le A$. Then*

$$\Pi\big(\max\{\|f_A - f_0\|_n, \|f_A - f_0\|_2\} > \varepsilon_n | S_n\big) \to 0 \qquad \text{in probability as } n \to \infty.$$

Moreover, with probability tending to one, the following holds:

$$\max\{\|\hat{f}_A - f_0\|_n, \|\hat{f}_A - f_0\|_2\} \le C\bar{\varepsilon}_n \lesssim n^{-s/(2s+d)}(\log n)^{d+1},$$

with $\bar{\varepsilon}_n$ given in Theorem 2.1 and C a positive constant.

REMARK. The assumption that an upper bound A on $\|f_0\|_\infty$ is known is restrictive. As a reviewer pointed out, there is still room for improvement on the convergence rate under random design, for example, our boundedness assumption $\|f_0\|_\infty \le A$ takes away part of the difficulty of the problem, and proving boundedness in L^∞ for posterior distributions can be highly nontrivial. We admit that this boundedness assumption may not be theoretically appealing, and needs further investigation; nevertheless, it is a mild assumption in practice—under this assumption, our truncation procedure provides a practically useful way to circumvent the difficult problem of proving boundedness in L^∞. Moreover, by using this "truncated" posterior distribution we no longer need the strong $s \ge d/2$ condition made in [33]. An alternative way to restrict the support of the posterior measure into a given function class is to replace the prior $\Pi(\cdot)$ with the conditional measure

$\Pi(\cdot|B)$, where the set B encodes the constraint. For example, [15] (Section 2.4) sets $B = \{\|f\|_\infty \le A\}$ to obtain a posterior measure supported on bounded functions, and [4] (Corollary 7.2) uses $B = \{\|f\|_\gamma \le A\}$ to ensure that the posterior measure is supported on C^γ, the Banach space of Hölder γ-smooth functions equipped with norm $\|\cdot\|_\gamma$. This alternative "conditioning" procedure appears to be more difficult to implement in practice than our "truncating" procedure.

2.5. *Dimensionality reduction.* Tenenbaum et al. [29] and Roweis and Saul [26] initiated the area of manifold learning, which aims to design nonlinear dimensionality reduction algorithms to map high dimensional data into a low dimensional feature space under the assumption that data fall on an embedded nonlinear manifold within the high dimensional ambient space. A combination of manifold learning and usual nonparametric regression leads to a two-stage approach, in which a dimensionality reduction map from the original ambient space $\mathbb{R}^D$ to a feature space $\mathbb{R}^{\tilde{d}}$ is estimated in the first stage and a nonparametric regression analysis with low dimensional features as predictors is conducted in the second stage. As a byproduct of Theorem 2.1, we provide a theoretical justification for this two stage approach under some mild conditions.

Let $\Psi : \mathbb{R}^D \to \mathbb{R}^{\tilde{d}}$ be a dimensionality reduction map, which may be constructed based on the design points $\{X_i\}$. For identifiability, we require the restriction $\Psi_{\mathcal{M}}$ of Ψ on the manifold $\mathcal{M}$ to be a diffeomorphism, that is, $\Psi_{\mathcal{M}}$ is injective and both $\Psi_{\mathcal{M}}$ and its inverse are smooth. This requires the projection dimension $\tilde{d}$ to satisfy $\tilde{d} \ge d$. Diffeomorphism is the least and only requirement so that both the intrinsic dimension d of the predictor and the smoothness s of regression function f are invariant. If we view $\Psi(\mathbb{R}^D)$ as the new ambient space, then the corresponding new regression function $\tilde{f}$ is induced by f via

$$\tilde{f}(\tilde{x}) = f\big[\Psi_{\mathcal{M}}^{-1}(\tilde{x})\big] \qquad \text{for all } \tilde{x} \in \Psi_{\mathcal{M}}(\mathcal{M}).$$

Accordingly, the empirical norm of $\tilde{f}$ under fixed design becomes $\|\tilde{f}\|_n^2 = \sum_{i=1}^n |\tilde{f}(\Psi(X_i))|^2$. By the identifiability condition on Ψ, $\tilde{f}$ is a well-defined function on the manifold $\mathcal{M}$, represented as a submanifold in the ambient space $\mathbb{R}^{\tilde{d}}$, and has the same smoothness as f. Therefore, by specifying a GP prior (2.2) directly on $\mathbb{R}^{\tilde{d}}$, we are able to achieve a posterior contraction rate at least $n^{-s/(2s+d)}(\log n)^{d+1}$, as indicated by the following theorem.

THEOREM 2.4. *Assume that $\mathcal{M}$ is a d-dimensional compact C^γ submanifold of R^D. Suppose that $\Psi : \mathbb{R}^D \to \mathbb{R}^{\tilde{d}}$ is an ambient space mapping (dimension reduction) such that when restricted on $\mathcal{M}$, Ψ is a $C^{\gamma'}$-diffeomorphism. Then by specifying the prior (2.2) with $\{\Psi(X_i)\}_{i=1}^n$ as observed predictors and the Euclidean norm of $\mathbb{R}^{\tilde{d}}$ as the norm $\|\cdot\|$ in (2.1), we have that for any $f_0 \in C^s(\mathcal{M})$ with $s \le \min\{2, \gamma - 1, \gamma' - 1\}$, (5.1) will be satisfied for $\varepsilon_n = n^{-s/(2s+d)}(\log n)^{\kappa_1}$*

and $\bar{\varepsilon}_n = \varepsilon_n (\log n)^{\kappa_2}$ *with* $\kappa_1 = (1+d)/(2+d/s)$ *and* $\kappa_2 = (1+d)/2$. *This implies that the posterior contraction rate with respect to* $\|\cdot\|_n$ *will be at least* $\varepsilon_n = n^{-s/(2s+d)} (\log n)^{d+1}$.

3. Adaptation to intrinsic dimension. To make our approach adaptive to the intrinsic dimension d, we follow an empirical Bayes approach by plugging in an estimator of the intrinsic dimension. Such an estimator can be chosen by focusing either on the consistent estimation of d or on the optimal prediction in terms of f. In the latter approach, the estimator for d may not be consistent but one can still achieve a near minimax-optimal rate for estimating f. Focusing on our truncated estimator in random design, we describe two approaches in the following subsections.

3.1. *Intrinsic dimension estimation.* As d is a hyper-parameter in prior (2.3), in principle one can specify a prior for d over a finite grid $d_1 \le \cdots \le d_p$ and conditioning on $d = d_j$, use (2.3) as a prior for f. Since W^A is conditionally independent of d given A, one can marginalize out d and obtain an equivalent prior for A as a mixture distribution. In the proof of Theorem 2.1, a critical property of the prior of A employed is its tail behavior as $P(A > a) \sim \exp(-Ca^d)$. However, with an extra level of prior for d, the tail $P(A > a)$ of the marginal prior is dominated by $\exp(-Ca^{d_1})$, which has a similar decay rate as the prior $A^{d_1} \sim \mathrm{Ga}(a_0, b_0)$. This illustrates that specifying a prior for d may still lead to a suboptimal rate as suggested by Theorem 2.2.

Intuitively, in random design information about the intrinsic dimension d is contained in the marginal distribution of X, and this information cannot be fully revealed by estimating the conditional distribution $P(Y|X)$. This motivates our first approach of estimating d directly based on the covariates $\{X_i\}$.

Many estimation methods have been proposed for determining the intrinsic dimension of a dataset lying on a manifold [8, 9, 12, 20, 22]. For example, [20] considers a likelihood based approach and [22] relies on singular value decomposition of the local sample covariance matrix. [12] proposes a nearest-neighbor method and analyzes its finite-sample properties. Their estimator $\hat{d}$ takes the form as

$$\hat{d} = \frac{\log 2}{\log \hat{r}^{(k)}(X_1) - \log \hat{r}^{(\lceil k/2 \rceil)}(X_1)},$$

where $\hat{r}^{(k)}(X_1)$ is the distance from X_1 to its kth nearest neighbor in $\{X_i\}$. Let $B(x, r) \subset \mathbb{R}^D$ denote a ball centering at point $x \in \mathcal{M}$ with radius r in the ambient space $\mathbb{R}^D$. Consider the following condition:

ASSUMPTION A. X_i are i.i.d. samples coming from a distribution Q supported on the manifold $\mathcal{M}$ with $Q(X_i \in B(x, r)) = \eta(x, r) r^d$, where the function $\eta : \mathcal{M} \times [0, \infty) \to [0, \infty)$ satisfies: (1) $\inf_{x \in \mathcal{M}} \eta(x, 0) > 0$; (2) For any point $x \in$

$\mathcal{M}$, $\eta(x,\cdot)$ is differentiable in $(0,\infty)$ and right-differentiable at 0; (3) There exist positive constants (B, B', r_0), such that for any $(x,r) \in \mathcal{M} \times [0, r_0)$, $\frac{\partial}{\partial r}\eta(x,r) \le B'\eta(x,r)$ and $|\eta(x,r) - \eta(x,0)| \le B\eta(x,0)r$.

Assumption A requires that the distribution Q of X_i is well spreaded over the manifold and locally resembles a smooth and regular distribution over $\mathbb{R}^d$, which appears natural. Farahmand et al. [12] proved that under Assumption A, if $n \ge k2^d$, then it holds with probability at least $1-\delta$ that

$$|\hat{d} - d| \le C\left\{\left(\frac{k}{n}\right)^{1/d} + \sqrt{\frac{\log(4/\delta)}{k}}\right\},$$

where C is some constant independent of k and n. As a consequence, if we choose $k = n^{1/2}$ and let $\hat{d}_R$ be the closest integer to $\hat{d}$, then $P_0(\hat{d}_R \ne d) \to 0$ as $n \to \infty$, that is, $\hat{d}_R$ is a consistent estimator of d.

We use $\hat{d}_R$ as an estimator of d and plug in $\hat{d}_R$ into our prior (2.3) to obtain an empirical Bayes estimator $\hat{f}_{\mathrm{EB}}$, based on the truncation procedure in Section 2.4. The following corollary summarizes its asymptotic performance.

COROLLARY 3.1. *Assume Assumption A and the same conditions as in Theorem 2.3, then with probability tending to one,*

$$\max\{\|\hat{f}_{\mathrm{EB}} - f_0\|_n, \|\hat{f}_{\mathrm{EB}} - f_0\|_2\} \lesssim n^{-s/(2s+d)}(\log n)^{d+1}.$$

REMARK. Although the estimator $\hat{d}$ is specifically built for random design, it also works in fixed design under some deterministic conditions on the design points $\{X_i\}$. For example, the key step of the proofs in [12] is to show that for any fixed $x \in \mathcal{M}$, the random quantity $\hat{r}^{(k)}(x)$ converges to $r_p(x)$, where $p = k/n$ and $r_p(x)$ is the solution of the equation $p = \eta(x, r_p)r_p^d$, by applying Bernstein's inequality (Lemma 4, [12]). Therefore, if we assume that the fixed-design points $\{X_i\}$ are well distributed so that $n^{-1}|\{i : X_i \in B(x,r)\}| \in [(1-\alpha)\eta(x,r)r^d, (1+\alpha)\eta(x,r)r^d]$ holds for each $(x,r) \in \mathcal{M} \times (0, r_0)$, where $|E|$ denotes the cardinality of a set E, $\alpha \in (0, 1/(4(d+1)))$ is some constant and the function η satisfies Assumption A, then $|\hat{d} - d| \le C(\frac{k}{n})^{1/d}$ holds for some constant C depending on α.

3.2. *Cross validation.* In this subsection, we select a best dimension and its associated estimator as constructed in Section 2.4 based on prediction accuracy on a testing set. This selection rule cannot consistently estimate d but still yields an optimal convergence rate for estimating f (see Theorem 3.2 below). In principle, the selection procedure described in this subsection may be applicable to other hyperparameter selection problems.

We focus on random design. Let $d_{\max}$ be a pre-specified upper bound for d. For example, we can choose $d_{\max} = 20$. Let Π_k denote the prior (2.3) with $d = k$ for $k = 1, \ldots, d_{\max}$. The selection procedure proceeds as follows:

1. Randomly split the whole data set with size $n+m$ into a training set $S_n = \{(X_i, Y_i) : i = 1, \ldots, n\}$ and a testing set $\tilde{S}_m = \{(\tilde{X}_i, \tilde{Y}_i) : i = 1, \ldots, m\}$.

2. For $k = 1, \ldots, d_{\max}$, obtain a truncated Bayes estimator $\hat{f}^{(k)}$ under Π_k defined as the posterior mean $\int f_A \, d\Pi_k(f|S_n)$ as in Section 2.4. Compute its mean squared prediction error (MSPE) $E_m^{(k)} = m^{-1} \sum_{i=1}^m (\hat{f}^{(k)}(\tilde{X}_i) - \tilde{Y}_i)^2$ on the testing set.

3. Let $\hat{d}_{\mathrm{CV}} = \arg\min_k E_m^{(k)}$. The final estimator is defined by $\hat{f}_{\mathrm{CV}} = \hat{f}^{(\hat{d}_{\mathrm{CV}})}$.

The intuition is simple: an estimator with minimal MSPE, which approximately minimizes $\|\hat{f}^{(k)} - f_0\|_2^2$ over k, should be at least better than $\hat{f}^{(d)}$, the estimator under the true intrinsic dimension d. In practice, one can repeat steps 1 and 2 for a number of times and use an averaged MSPE instead of $E_m^{(k)}$ to improve stability. However, the following theorem suggests that one splitting suffices for the adaptation to the dimensionality. Due to space constraints, its proof is provided in the supplementary material [36].

THEOREM 3.2. *Suppose $d \le d_{\max}$ and $A \ge \|f_0\|_\infty$ in the cross validation procedure. If $m\bar{\varepsilon}_n^2 \to \infty$ with $\bar{\varepsilon}_n$ any upper bound to the posterior convergence rate under the true intrinsic dimension d, then under the conditions in Theorem 2.3,*

$$\|\hat{f}_{\mathrm{CV}} - f_0\|_2 \le 2\sqrt{2}\bar{\varepsilon}_n \lesssim n^{-s/(2s+d)} (\log n)^{d+1}$$

holds with probability tending to 1. *Here, the probability is the joint probability of the training set and the testing set.*

REMARK. The condition on m suggests that the size of the testing set is allowed to be of order $O(n^\gamma)$ for any $\gamma > 2s/(2s+d)$, which depends on the unknown smoothness s and intrinsic dimension d, and can be substantially smaller than n if s is small and d is large. This condition provides the right intuition that if we are not expected to accurately estimate the unknown parameter, then only a small number of testing samples are needed in the cross validation step. In practice, one may simply choose m to be a small fraction of n. A reviewer pointed out that again the boundedness assumption $\|f_0\|_\infty \le A$ plays an important role in the application of Bernstein's inequality in the proof of Theorem 3.2. We agree that how to avoid this boundedness assumption is still an open problem worth further theoretical investigation.

4. Numerical examples.

4.1. *Regression on the Swiss roll.* We start with a toy example where X lies on a two-dimensional Swiss roll in the 100-dimensional Euclidean space (Figure 2 plots a typical Swiss roll in the three-dimensional Euclidean space). X is generated as follows. We first sample $T = (T_1, T_2, T_3)^T$ from a two-dimensional Swiss roll in three-dimensional ambient space as

$$T_1 = U\cos(U), \qquad T_2 = V, \qquad T_3 = U\sin(U),$$

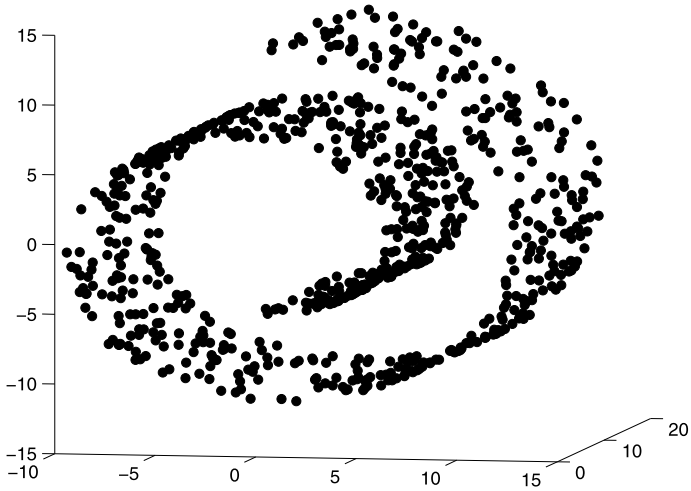

FIG. 2. *A typical Swiss roll in three-dimensional Euclidean space.*

with $U \sim \text{Unif}(\frac{3\pi}{2}, \frac{9\pi}{2})$ and $V \sim \text{Unif}(0, 20)$. Then we transform T into a 100-dimensional vector via $X = \Omega T$, where Ω is a random matrix with size 100-by-3, whose components follow i.i.d. $N(0, 1)$. Ω will be fixed in each synthetic dataset. The response Y depends on X through

$$Y = 4\left(\frac{1}{3\pi}U - \frac{1+3\pi}{2}\right)^2 + \frac{\pi}{20}V + N(0, 0.1^2).$$

To assess the fitting performance, we use the empirical error $\|\hat{f} - f_0\|_n$ of our estimator $\hat{f}$ on the design points as a criterion. In the GP approach, we apply the empirical Bayes approach described in Section 3.1 and run 10,000 iterations with the first 5000 as burn-in in each replicate. We report an average empirical error (AEE) over 100 replicates in Table 1. In this example, the GP estimator has a relatively fast convergence rate even though the dimensionality of the ambient space is large, which is consistent with our theory.

4.2. *Application to the lucky cat data.* The lucky cat data (Figure 1) has intrinsic dimensionality one, which is the dimension of the rotation angle θ. Since we

TABLE 1
Simulation results for the Swiss roll example over 100 *replicates. The numbers in the parentheses indicate standard deviations*

	$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
AEE	0.164 (0.090)	0.143 (0.026)	0.121 (0.012)	0.106 (0.005)	0.095 (0.003)

TABLE 2
Square root of MSPE for the lucky cat data by using two different approaches over 100 *random splitting are displayed. The numbers in the parentheses indicate standard deviations*

	$n = 18$	$n = 36$	$n = 54$
EN	0.416 (0.152)	0.198 (0.042)	0.149 (0.031)
LASSO	0.431 (0.128)	0.232 (0.061)	0.163 (0.038)
GP	0.332 (0.068)	0.128 (0.036)	**0.077 (0.014)**
2GP	**0.181 (0.051)**	**0.124 (0.038)**	0.092 (0.021)

know the true value of θ, we create the truth $f_0(\theta) = \cos\theta$ as a continuous function on the unit circle. The responses are simulated from $Y_i = f_0(\theta_i) + w_i$ by adding independent Gaussian noises $w_i \sim N(0, 0.1^2)$ to the true values. In this model, the total sample size $N = 72$ and the predictors X_i lie in $\mathbb{R}^D$, with $D = 16{,}384$. To assess the impact of the sample size n on the fitting performance, we randomly divide $n = 18$, 36 and 64 samples into a training set and treat the rest as a testing set. The training set is used for fitting a model and the testing set is used for quantifying the estimation accuracy. For each training size n, we repeat this procedure for $m = 100$ times and calculate the square root of the following mean squared prediction error (MSPE) on the testing set,

$$\sum_{l=1}^{m} \frac{1}{N-n} \sum_{i \in T_l} \left\| \hat{Y}_i - f_0(\theta_i) \right\|^2,$$

where T_l is the lth testing set and $\hat{Y}_i$ is an estimation of $E[Y|X_i] = f_0(\theta_i)$. We apply two GP based algorithms on this data set: 1. vanilla GP specified by (2.3); 2. Two stage GP (2GP) where the D-dimensional predictors were projected into $\mathbb{R}^2$ by using Laplacian eigenmap [2] in the first stage and then a GP with projected features as predictors was fitted in the second stage. To assess the prediction performance, we also compare our GP prior based models (2.3) with lasso [30] and elastic net (EN) [38] under the same settings. We choose these two competing models because they are among the most widely used methods in high dimensional regression settings and perform especially good when the true model is sparse. In the GP models, we set $d = 1$ since the sample size for this dataset is too small for most dimension estimation algorithms to reliably estimate d. In addition, for each simulation, we run 10,000 iterations with the first 5000 as burn-in.

The results are shown in Table 2. As we can see, under each training size n, GP performs the best. Moreover, as n increases, the prediction error of GP decays much faster than EN and Lasso: when $n = 18$, the square root of MSPEs by using EN and lasso are about 125% of that by using GP; however, as n increases to 54, this ratio becomes about 200%. Moreover, the standard deviations of the square roots of MSPEs by using GP are also significantly lower than those by using lasso

and EN. It is not surprising that 2GP has better performance than GP when n is small since the dimensionality reduction map Ψ is constructed using the whole dataset (the available Laplacian eigenmap code we used cannot do interpolations). Therefore, as the training size n becomes closer to the total sample size 72, GP becomes better. In addition, GP is computationally faster than 2GP due to the use of the manifold learning algorithm in the first stage of 2GP.

5. Auxiliary results. In the GP prior (2.3), the covariance function $K^a : \mathcal{M} \times \mathcal{M} \to R$ is defined on the submanifold $\mathcal{M}$. Therefore, (2.3) essentially defines a GP on $\mathcal{M}$ and we can study its posterior contraction rate as a prior for functions on the manifold. In this section, we combine geometry properties and Bayesian nonparametric asymptotic theory to prove the theorems in Section 2.

5.1. *Reproducing kernel Hilbert space on the manifold.* Being viewed as a covariance function defined on $[0,1]^D \times [0,1]^D$, $K^a(\cdot,\cdot)$ corresponds to a reproducing kernel Hilbert space (RKHS) $\mathbb{H}^a$, which is defined as the completion of $\mathcal{H}$, the linear space of all functions $x \mapsto \sum_{i=1}^m a_i K^a(x_i, x), x \in [0,1]^D$, indexed by $a_1, \ldots, a_m \in \mathbb{R}$ and $x_1, \ldots, x_m \in [0,1]^D, m \in \mathbb{N}$, relative to the norm induced by the inner product $\langle K^a(x,\cdot), K^a(y,\cdot)\rangle_{\mathbb{H}^a} = K^a(x,y)$. Similarly, $K^a(\cdot,\cdot)$ can also be viewed as a covariance function defined on $\mathcal{M} \times \mathcal{M}$, with the associated RKHS denoted by $\tilde{\mathbb{H}}^a$. Here, $\tilde{\mathbb{H}}^a$ is the completion of $\tilde{\mathcal{H}}$, which is the linear space of all functions $x \mapsto \sum_{i=1}^m a_i K^a(x_i, x), x \in \mathcal{M}$, indexed by $a_1, \ldots, a_m \in \mathbb{R}$ and $x_1, \ldots, x_m \in \mathcal{M}, m \in \mathbb{N}$.

Many probabilistic properties of GPs are closely related to the RKHS associated with its covariance function. Readers can refer to [1] and [34] for an introduction on the RKHS theory for GPs on Euclidean spaces. In order to generalize properties of the RKHS in Euclidean spaces to submanifolds, we need a link to transfer the theory. The next lemma achieves this by characterizing the relationship between $\mathbb{H}^a$ and $\tilde{\mathbb{H}}^a$.

LEMMA 5.1. *For any $f \in \tilde{\mathbb{H}}^a$, there exists $g \in \mathbb{H}^a$ such that $g|_{\mathcal{M}} = f$ and $\|g\|_{\mathbb{H}^a} = \|f\|_{\tilde{\mathbb{H}}^a}$, where $g|_{\mathcal{M}}$ is the restriction of g on $\mathcal{M}$. Moreover, for any other $g' \in \mathbb{H}^a$ with $g'|_{\mathcal{M}} = f$, it holds that $\|g'\|_{\mathbb{H}^a} \geq \|f\|_{\tilde{\mathbb{H}}^a}$. This implies $\|f\|_{\tilde{\mathbb{H}}^a} = \inf_{g \in \mathbb{H}^a, g|_{\mathcal{M}} = f} \|g\|_{\mathbb{H}^a}$.*

This lemma implies that any element f in the RKHS $\tilde{\mathbb{H}}^a$ can be considered as the restriction of some element g in the RKHS $\mathbb{H}^a$. Particularly, there exists a unique such element g in $\mathbb{H}^a$ such that the norm is preserved, that is, $\|g\|_{\mathbb{H}^a} = \|f\|_{\tilde{\mathbb{H}}^a}$.

5.2. *Background on posterior convergence rate for GP.* As shown in [13, 14], in order to characterize the posterior contraction rate in a Bayesian nonparametric problem, such as density estimation, fixed design regression or classification, we

need to verify some conditions on the prior measure Π. Specifically, we describe a set of sufficient conditions on the randomly rescaled GP prior (2.2) given in [35]. Let $\mathcal{X}$ be the predictor space and $f_0 : \mathcal{X} \to \mathbb{R}$ be the true function that is of interest, for example, the log density $\log p(x)$ in density estimation, the conditional expection $E[Y|X]$ in regression and the logistic-transformed conditional probability $\operatorname{logit} P(Y = 1|X)$ in classification. We will not consider density estimation since the knowledge of the support $\mathcal{M}$ is needed so that e^{f_0} can be properly normalized to produce a valid density function. Let ε_n and $\bar{\varepsilon}_n$ be two sequences. If there exist Borel measurable subsets B_n of $C(\mathcal{X})$ such that for n sufficiently large,

$$\begin{aligned} P\big(\|W^A - f_0\|_\infty \le \varepsilon_n\big) &\ge e^{-n\varepsilon_n^2}, \\ P\big(W^A \notin B_n\big) &\le e^{-4n\varepsilon_n^2}, \\ \log N\big(\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty\big) &\le n\bar{\varepsilon}_n^2, \end{aligned} \tag{5.1}$$

where $W^A \sim \Pi$ and $\|\cdot\|_\infty$ is the sup-norm on $C(\mathcal{X})$, then the posterior contraction rate would be at least $\varepsilon_n \vee \bar{\varepsilon}_n$ under $\|\cdot\|_n$. In our case, $\mathcal{X}$ is the d-dimensional submanifold $\mathcal{M}$ in the ambient space $\mathbb{R}^D$. We require $\mathcal{M}$ to be compact because the space of continuous functions on a compact metric space is a separable Banach space, which is fundamental to apply the theory from [35]. To verify the first concentration condition, we need to give an upper bound to the so-called concentration function [35] $\phi_{f_0}^a(\varepsilon)$ of the GP W^a around the truth f_0 for any given a and ε. $\phi_{f_0}^a(\varepsilon)$ is composed of two terms. Both terms depend on the RKHS $\tilde{\mathbb{H}}^a$ associated with the covariance function of the GP W^a. The first term is the decentering function $\inf\{\|h\|_{\tilde{\mathbb{H}}^a}^2 : \|h - f_0\|_\infty < \varepsilon\}$, where $\|\cdot\|_{\tilde{\mathbb{H}}^a}$ denotes the RKHS norm. This quantity measures how well the truth f_0 is approximated by the elements in the RKHS. The second term is the negative log small ball probability $-\log P(\|W^a\|_\infty < \varepsilon)$, which intimately depends on the covering entropy $\log N(\varepsilon_n, \tilde{\mathbb{H}}_1^a, \|\cdot\|_\infty)$ of the unit ball in the RKHS $\tilde{\mathbb{H}}^a$. As a result of this dependence, the second and third conditions in (5.1) are often satisfied as byproducts of the first condition by applying Borell's inequality [34].

As pointed out by [35], the key to ensure the smoothness adaptability of the GP prior on the Euclidean space is the sub-exponentially decaying tail of the spectral density of its stationary covariance function. This property on the spectral density is satisfied for the squared exponential and the Matérn class covariance functions. More specifically, by Bochner's theorem the squared exponential covariance function $K^1(x, y) = \exp\{-\|x - y\|^2/2\}$ (with unit inverse bandwidth) on $\mathbb{R}^D$ has a spectral representation as

$$K^1(x, y) = \int_{\mathbb{R}^D} e^{-i(\lambda, x-y)} \mu(d\lambda),$$

where μ is its spectral measure with sub-Gaussian tail. Sub-Gaussian tail satisfies the sub-exponential tail requirement, that is, for some (in fact, any) $\delta > 0$,

$$\int e^{\delta\|\lambda\|}\mu(d\lambda) < \infty. \tag{5.2}$$

For convenience, we will focus on the squared exponential covariance function. Extending the results to other types of covariance functions with sub-exponential decaying spectral densities can also be done with more efforts.

5.3. *Decentering function.* To estimate the decentering function, we need construct a function $I_a(f)$ on the manifold $\mathcal{M}$ for approximating any differentiable function f on $\mathcal{M}$, so that the magnitude of the RKHS norm $\|I_a(f)\|_{\tilde{\mathbb{H}}^a}$ can be controlled. Unlike in the Euclidean space where functions in the RKHS $\mathbb{H}^a$ can be represented via Fourier transformation [35], there is no general way to represent and calculate RKHS norms of functions in the RKHS $\tilde{\mathbb{H}}^a$ on a manifold. In the next lemma, we provide a specific way to construct an approximation $I_a(f)$ with a controllable RKHS norm via convolving f with K^a on the manifold $\mathcal{M}$:

$$\begin{aligned} I_a(f)(x) &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathcal{M}} K^a(x,y) f(y)\, dV(y) \\ &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathcal{M}} \exp\left\{-\frac{a^2\|x-y\|^2}{2}\right\} f(y)\, dV(y), \qquad x \in \mathcal{M}, \end{aligned} \tag{5.3}$$

where V is the Riemannian volume form of $\mathcal{M}$. Heuristically, for large a, the above integrand only has a nonnegligible value in a small neighborhood around x. Therefore, we can conduct a change of variable in the above integral with transformation $\phi^x : B_\delta \to W$ defined by (7.2) in the supplementary material [36] in a small neighborhood W of x:

$$\begin{aligned} I_a(f)(x) &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d} \exp\left\{-\frac{a^2\|\phi^x(u)-\phi^x(0)\|^2}{2}\right\} \\ &\quad \times f\big(\phi^x(u)\big)\sqrt{\det\big(g_{ij}^{\phi}(u)\big)}\, du \\ &\approx \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d} \exp\left\{-\frac{a^2\|u\|^2}{2}\right\} f\big(\phi^x(u)\big)\, du \\ &\approx f\big(\phi^x(0)\big) = f(x), \qquad x \in \mathcal{M}, \end{aligned}$$

where the above approximation holds since: 1. $\phi^x(0) = x$; 2. ϕ^x preserves local distances [supplementary material [36], Proposition 7.5(3)]; 3. the Jacobian $\sqrt{\det(g_{ij}^{\phi}(u))}$ is close to one [supplementary material [36], Proposition 7.5(2)]. From this heuristic argument, we can see that the approximation error $\|I_a(w) - f_0\|_\infty$ is determined by two factors: the convolution error

$|(\frac{a}{\sqrt{2\pi}})^d \int_{\mathbb{R}^d} \exp\{-\frac{a^2\|u\|^2}{2}\} f(\phi^x(u))\,du - f(x)|$ and the nonflat error caused by the nonzero curvature of $\mathcal{M}$. Moreover, we can expand each of these errors as a polynomial of $1/a$ and call the expansion term related to $1/a^k$ as the kth order error.

When $\mathcal{M}$ is Euclidean space $\mathbb{R}^d$, the nonflat error is zero, and by Taylor expansion the convolution error has order s if $f_0 \in C^s(\mathbb{R}^d)$ and $s \le 2$, where $C^s(\mathbb{R}^d)$ is the Holder class of s-smooth functions on $\mathbb{R}^d$. The constraint $s \le 2$ occurs because the Gaussian kernel $\exp\{-\|(x-y)\|^2/2\}$ only has a vanishing moment up to first order: $\int x \exp(-\|(x-y)\|^2/2)\,dx = 0$. More generally, the convolution error has order up to $s+1$ if the convolution kernel K has vanishing moments up to order s, that is, $\int x^t K(x)\,dx = 0, t = 1, \ldots, s$. On the other hand, for general manifold $\mathcal{M}$ with nonvanishing curvature tensor, the nonflat error always has order two (see the proof of Lemma 5.2). This implies that even with a carefully chosen covariance function that can improve the convolution error to higher order, the overall approximation still tends to exhibit second-order error due to the deterioration caused by the nonzero curvature of the manifold. The following lemma formalizes the above heuristic argument on the order of the approximation error caused by using (5.3) as an approximation to f and provides an upper bound on the decentering function.

LEMMA 5.2. *Assume that $\mathcal{M}$ is a d-dimensional compact C^γ submanifold of R^D. Then for any $f \in C^s(\mathcal{M})$ with $s \le \min\{2, \gamma\}$, there exist constants $a_0 \ge 1$, $C > 0$ and $B > 0$ depending only on μ, $\mathcal{M}$ and f such that for all $a \ge a_0$,*

$$\inf\Big\{\|h\|^2_{\tilde{\mathbb{H}}^a} : \sup_{x\in\mathcal{M}} |h(x) - f(x)| \le Ca^{-s}\Big\} \le Ba^d.$$

5.4. *Centered small ball probability.* As indicated by the proof of Lemma 4.6 in [35], in order to prove an upper bound on $-\log P(\|W^a\|_\infty < \varepsilon)$, we only need to provide an upper bound on the covering entropy $\log N(\varepsilon, \tilde{\mathbb{H}}^a_1, \|\cdot\|_\infty)$ of the unit ball in the RKHS $\tilde{\mathbb{H}}^a$. Following the discussion in Section 4.1, we want to link $\tilde{\mathbb{H}}^a$ to $\mathbb{H}^a$, the associated RKHS defined on the ambient space $\mathbb{R}^D$. Therefore, we need a lemma to characterize this space $\mathbb{H}^a$ ([35], Lemma 4.1).

LEMMA 5.3. *$\mathbb{H}^a$ is the set of real parts of the functions*

$$x \mapsto \int e^{i(\lambda,x)} \psi(\lambda) \mu_a(d\lambda),$$

when ψ runs through the complex Hilbert space $L_2(\mu_a)$. Moreover, the RKHS norm of the above function is $\|\psi\|_{L_2(\mu_a)}$, where μ_a is the spectral measure of the covariance function K^a.

Based on this representation of $\mathbb{H}^a$ on $\mathbb{R}^D$, [35] proved an upper bound $Ka^D(\log\frac{1}{\varepsilon})^{D+1}$ for $\log N(\varepsilon, \tilde{\mathbb{H}}^a_1, \|\cdot\|_\infty)$ through constructing an ε-covering set

composed of piecewise polynomials. However, there is no straightforward generalization of their scheme from Euclidean spaces to manifolds. The following lemma provides an upper bound on the covering entropy of $\tilde{\mathbb{H}}_1^a$, where the exponents only depend on the intrinsic dimension d. The main novelty in our proof is the construction of an ε-covering set composed of piecewise transformed polynomials (6.9) via analytically extending the truncated Taylor polynomial approximations (6.6) to the elements in $\tilde{\mathbb{H}}_1^a$. As the proof indicates, the d in a^d relates to the covering dimension d of $\mathcal{M}$, i.e. the ε-covering number $N(\varepsilon, \mathcal{M}, \varepsilon)$ of $\mathcal{M}$ is proportional to $1/\varepsilon^d$. The d in $(\log \frac{1}{\varepsilon})^{d+1}$ relates to the order of the number k^d of coefficients in piecewise transformed polynomials of degree k in d variables.

LEMMA 5.4. *Assume that $\mathcal{M}$ is a d-dimensional C^γ compact submanifold of R^D with $\gamma \geq 2$. Then for the squared exponential covariance function K^a, there exists a constant K depending only on d, D and $\mathcal{M}$, such that for all $\varepsilon < 1/2$ and $a > \max\{a_1, \varepsilon^{-1/(\gamma-1)}\}$, where constant $a_1 = \delta/(2\delta_0\sqrt{d})$, δ_0 is defined in Lemma 7.7 in the supplementary material* [36] *and δ is the constant in* (5.2), *it holds that*

$$\log N\big(\varepsilon, \tilde{\mathbb{H}}_1^a, \|\cdot\|_\infty\big) \leq K a^d \bigg(\log \frac{1}{\varepsilon}\bigg)^{d+1}.$$

Similar to Lemma 4.6 in [35], Lemma 5.4 implies an upper bound on $-\log P(\|W^a\|_\infty < \varepsilon)$.

LEMMA 5.5. *Assume that $\mathcal{M}$ is a d-dimensional compact C^γ submanifold of $\mathbb{R}^D$ with $\gamma \geq 2$. If K^a is the squared exponential covariance function with inverse bandwidth a, then for some $a_1 > 0$, there exist constants C and ε_0 that only depend on a_1, μ, d, D and $\mathcal{M}$, such that for all $a \geq \max\{a_1, \varepsilon^{-1/(\gamma-1)}\}$ and $\varepsilon < \varepsilon_0$,*

$$-\log P\Big(\sup_{x\in\mathcal{M}} |W_x^a| \leq \varepsilon\Big) \leq C a^d \bigg(\log \frac{a}{\varepsilon}\bigg)^{d+1}.$$

5.5. *Posterior contraction rate of GP on manifolds.* By using the manifold adapted lemmas in Section 5.3 to 5.4, the proofs of Theorems 2.1 and 2.2 follow similar ideas as the proof of Theorem 3.1 in [35] and are provided in the supplementary material [36].

6. Proofs. In this section, we provide proofs for the key results in the paper.

6.1. *Proof of Lemma 5.1.* Consider the map $\Phi : \tilde{\mathcal{H}} \to \mathcal{H}$ that maps the function

$$\sum_{i=1}^m a_i K^a(x_i, \cdot) \in \tilde{\mathcal{H}}, \qquad a_1, \ldots, a_m \in \mathbb{R}, x_1, \ldots, x_m \in \mathcal{M}, m \in \mathbb{N}$$

on $\mathcal{M}$ to the function of the same form $\sum_{i=1}^{m} a_i K^a(x_i, \cdot) \in \mathcal{H}$, but viewed as a function on $[0,1]^D$. By definitions of RKHS norms, Φ is an isometry between $\tilde{\mathcal{H}}$ and a linear subspace of $\mathcal{H}$. As a result, Φ can be extended to an isometry between $\tilde{\mathbb{H}}^a$ and a complete subspace of $\mathbb{H}^a$. To prove the first part of this lemma, it suffices to verify that for any $f \in \tilde{\mathbb{H}}^a$, $g = \Phi(f)|_{\mathcal{M}} = f$. Assume that the sequence $\{f_n\} \in \tilde{\mathcal{H}}$ satisfies $\|f_n - f\|_{\tilde{\mathbb{H}}^a} \to 0$, as $n \to \infty$, then by the definition of Φ on $\tilde{\mathcal{H}}$, $\Phi(f_n)|_{\mathcal{M}} = f_n$. For any $x \in [0,1]^D$, by the reproducing property and Cauchy–Schwarz inequality,

$$\begin{aligned} \big|\Phi(f_n)(x) - g(x)\big| &= \big|\langle K^a(x, \cdot), \Phi(f_n) - g\rangle_{\mathbb{H}^a}\big| \\ &\le \sqrt{K^a(x,x)}\big\|\Phi(f_n) - \Phi(f)\big\|_{\mathbb{H}^a} \\ &= \|f_n - f\|_{\tilde{\mathbb{H}}^a} \to 0 \qquad \text{as } n \to \infty, \end{aligned}$$

where the last step follows since Φ is an isometry. This indicates that g can be obtained as the pointwise limit of $\Phi(f_n)$ as a function on $[0,1]^D$ and in the special case when $x \in \mathcal{M}$,

$$g(x) = \lim_{n\to\infty} \Phi(f_n)(x) = \lim_{n\to\infty} f_n(x) = f(x).$$

Denote the orthogonal complement of $\Phi(\tilde{\mathbb{H}}^a)$ in $\mathbb{H}^a$ as $\Phi(\tilde{\mathbb{H}}^a)^{\perp}$. Since $(g' - g)|_{\mathcal{M}} = 0$, we have $\langle K^a(x, \cdot), g - g'\rangle_{\mathbb{H}^a} = 0$ for all $x \in \mathcal{M}$, which implies $g - g' \perp \Phi(\tilde{\mathbb{H}}^a)$, that is, $g' - g \in \Phi(\tilde{\mathbb{H}}^a)^{\perp}$. Then by applying the Pythagorean theorem, we obtain

$$\|g'\|_{\mathbb{H}^a}^2 = \|g\|_{\mathbb{H}^a}^2 + \|g - g'\|_{\mathbb{H}^a}^2 \ge \|g\|_{\mathbb{H}^a}^2.$$

6.2. *Proof of Lemma 5.2.* The proof consists of two parts. In the first part, we prove that the approximation error of $I_a(f)$ can be decomposed into four terms. The first term T_1 is the convolution error defined in our previous heuristic argument. The second term T_2 is caused by localization of the integration, which is negligible due to the exponential decay of the squared exponential covariance function. The third and fourth terms T_3, T_4 correspond to the nonflat error. T_3 is caused by the error $|\|\phi^q(u) - q\|^2 - \|u\|^2|$ of approximating the geodesic distance with Euclidean distance, and T_4 the error $|\sqrt{\det(g_{ij}^{\phi}(u))} - 1|$ of approximating the local Jacobian by 1. Therefore, the overall approximation error $|I_a(f)(x) - f(x)|$ has order s in the sense that for some constant $C > 0$ dependent on $\mathcal{M}$ and f:

$$\sup_{x\in\mathcal{M}} \big|I_a(f)(x) - f(x)\big| \le Ca^{-s}, \qquad s \le \min\{2, \gamma\}. \tag{6.1}$$

In the second part, we prove that $I_a(f)$ belongs to $\tilde{\mathbb{H}}^a$ and has a squared RKHS norm $\|I_a(f)\|_{\tilde{\mathbb{H}}^a}^2 \le Ba^d$, where B is a positive constant independent of a.

Step 1 (*Estimation of the approximation error*): This part follows similar ideas as the proof of Theorem 1 in [37], where they have shown that (6.1) holds for

$s \leq 1$. Our proof generalizes their results to $s \leq 2$ and, therefore, involves a more careful treatment.

By Proposition 7.5 in the supplementary material [36], for each $p \in \mathcal{M}$, there exists a neighborhood W_p and an associated δ_p satisfying the two conditions in Proposition 7.4 and equations (7.4)–(7.6) in the supplementary material [36]. By compactness, $\mathcal{M}$ can be covered by $\bigcup_{p\in\mathcal{P}} W_p$ for a finite subset $\mathcal{P}$ of $\mathcal{M}$. Then $\sup_{x\in\mathcal{M}} |I_a(f)(x) - f(x)| = \sup_{p\in\mathcal{P}}\{\sup_{x\in W_p} |I_a(f)(x) - f(x)|\}$. Let $\delta^* = \min_{p\in\mathcal{P}}\{\min\{\delta_p, 1/\sqrt{2C_p}\}\} > 0$, where C_p is defined in equation (7.6) in the supplementary material [36]. Choose $a_0 \geq 1$ sufficiently large such that $C_0\sqrt{(2d+8)\log a_0}/a_0 < \delta^*$, where C_0 is the C_2 in Lemma 7.6 in the supplementary material [36].

Let $q \in W_p$ and $a > a_0$. Define $B_a^q = \{x \in \mathcal{M} : d_{\mathcal{M}}(q,x) < C_0\sqrt{(2d+8)\log a}/a\}$. Combining equation (7.3) in the supplementary material [36] and the fact that $\mathcal{E}_q$ is a diffeomorphism on $B_{\delta^*}(0)$,

$$B_a^q = \left\{\mathcal{E}_q\left(\sum_{i=1}^d u_i e_i^q\right) : u \in \tilde{B}_a\right\} \subset \mathcal{E}_q\big(B_{\delta^*}(0)\big),$$

where $\tilde{B}_a = \{u : \|u\| < C_0\sqrt{(2d+8)\log a}/a\} \subset B_{\delta^*}(0)$.

Denote $\phi^q(u) = \mathcal{E}_q(\sum_{i=1}^d u_i e_i^q)$. Then $B_a^q = \phi^q(\tilde{B}_a)$. By Definition (7.1) in the supplementary material [36],

$$\begin{aligned}&\left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{B_a^q} K^a(q,y) f(y)\, dV(y)\\ &\quad= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\tilde{B}_a} \exp\left\{-\frac{a^2\|q-\phi^q(u)\|^2}{2}\right\} f\big(\phi^q(u)\big)\sqrt{\det\big(g_{ij}^q\big)}(u)\, du.\end{aligned}$$

Therefore, by (5.3) we have the following decomposition:

$$I_a(f)(q) - f(q) = T_1 + T_2 + T_3 + T_4,$$

where

$$\begin{aligned}T_1 &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\tilde{B}_a} \exp\left\{-\frac{a^2\|u\|^2}{2}\right\}\big[f\big(\phi^q(u)\big) - f\big(\phi^q(0)\big)\big]\, du,\\
T_2 &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathcal{M}\setminus B_a^q} K^a(q,y) f(y)\, dV(y)\\
&\quad - \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d\setminus\tilde{B}_a} \exp\left\{-\frac{a^2\|u\|^2}{2}\right\} f(q)\, du,\\
T_3 &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\tilde{B}_a} \left\{\exp\left\{-\frac{a^2\|q-\phi^q(u)\|^2}{2}\right\}\right.\end{aligned}$$

$$- \exp\biggl\{-\frac{a^2\|u\|^2}{2}\biggr\}\biggr\} f\bigl(\phi^q(u)\bigr)\,du,$$

$$T_4 = \biggl(\frac{a}{\sqrt{2\pi}}\biggr)^d \int_{\tilde{B}_a} \exp\biggl\{-\frac{a^2\|q-\phi^q(u)\|^2}{2}\biggr\} f\bigl(\phi^q(u)\bigr)\Bigl(\sqrt{\det\bigl(g_{ij}^q\bigr)}(u)-1\Bigr)\,du.$$

Step 1.1 (*Estimation of* T_1): Let $g = f\circ\phi^q$. Since $f \in C^s(\mathcal{M})$ and $(\phi^q, B_{\delta^*}(0))$ is a C^γ coordinate chart, we have $g \in C^s(\mathbb{R}^d)$ and, therefore,

$$g(u)-g(0) = \begin{cases} R(u,s), & \text{if } 0 < s \le \min\{1,\gamma\}, \\ \displaystyle\sum_{i=1}^d \frac{\partial g}{\partial u_i}(0)u_i + R(u,s), & \text{if } 1 < s \le \min\{2,\gamma\}, \end{cases}$$

where the remainder term $|R(u,s)| \le C_1\|u\|^s$ for all $0 < s \le \min\{2,\gamma\}$. Since $\tilde{B}_a$ is symmetric,

$$\int_{\tilde{B}_a} \exp\biggl\{-\frac{a^2\|u\|^2}{2}\biggr\} u_i\,du = 0, \qquad i = 1,\dots,d,$$

and, therefore,

$$|T_1| \le C_1\biggl(\frac{a}{\sqrt{2\pi}}\biggr)^d \int_{\tilde{B}_a} \exp\biggl\{-\frac{a^2\|u\|^2}{2}\biggr\}\|u\|^s\,du = C_2 a^{-s}.$$

Step 1.2 (*Estimation of* T_2): Denote $T_2 = S_1 + S_2$ where S_1 and S_2 are the first term and second term of T_2, respectively. By Lemma 7.6 in the supplementary material [36], for $y \in \mathcal{M}\setminus B_a^q$, $\|q-y\| \ge d_{\mathcal{M}}(q,y)/C_0 \ge \sqrt{(2d+8)\log a}/a$. Therefore,

$$\begin{aligned} |S_1| &= \biggl|\biggl(\frac{a}{\sqrt{2\pi}}\biggr)^d \int_{\mathcal{M}\setminus B_a^q} \exp\biggl\{-\frac{a^2\|q-y\|^2}{2}\biggr\} f(y)\,dV(y)\biggr| \\ &\le \|f\|_\infty \operatorname{Vol}(\mathcal{M})\biggl(\frac{a}{\sqrt{2\pi}}\biggr)^d \exp\biggl\{-\frac{(2d+8)\log a}{2}\biggr\} \\ &= C_3 a^{-4} \le C_3 a^{-s}. \end{aligned}$$

As for S_2, we have

$$\begin{aligned} |S_2| &\le \|f\|_\infty\biggl(\frac{a}{\sqrt{2\pi}}\biggr)^d \int_{\|u\|\ge C_0\sqrt{(2d+8)\log a}/a} \exp\biggl\{-\frac{a^2\|u\|^2}{2}\biggr\}\,du \\ &\le \|f\|_\infty\biggl(\frac{a}{\sqrt{2\pi}}\biggr)^d \int_{\mathbb{R}^d} \exp\biggl\{-\frac{C_0^2(2d+8)\log a}{4}\biggr\}\exp\biggl\{-\frac{a^2\|u\|^2}{4}\biggr\}\,du \\ &= C_4 a^{-C_0^2(d/2+2)} \le C_4 a^{-s}, \end{aligned}$$

since $d \ge 1$, $C_0 \ge 1$ and $a \ge a_0 \ge 1$.

Combining the above inequalities for S_1 and S_2, we obtain

$$|T_2| \le (C_3 + C_4)a^{-s} = C_5 a^{-s}.$$

Step 1.3 (*Estimation of* T_3): By equation (7.6) in Proposition 7.5 and equation (7.3) in the supplementary material, we have

$$\begin{aligned} \big|\|u\|^2 - \big\|q - \phi^q(u)\big\|^2\big| &= \big|d^2_{\mathcal{M}}\big(q, \phi^q(u)\big) - \big\|q - \phi^q(u)\big\|^2\big| \\ &\le C_p d^4_{\mathcal{M}}\big(q, \phi^q(u)\big) = C_p\|u\|^4. \end{aligned} \tag{6.2}$$

Therefore, by using the inequality $|e^{-a} - e^{-b}| \le |a-b|\max\{e^{-a}, e^{-b}\}$ for $a, b > 0$, we obtain

$$\begin{aligned} |T_3| \le \|f\|_\infty \Big(\frac{a}{\sqrt{2\pi}}\Big)^d \int_{\tilde{B}_a} \max\Big\{\exp\Big\{-\frac{a^2\|q-\phi^q(u)\|^2}{2}\Big\} \\ \times \exp\Big\{-\frac{a^2\|u\|^2}{2}\Big\}\Big\}\frac{a^2\|u\|^4}{2}\,du. \end{aligned}$$

By equation (6.2) and the fact that $u \in \tilde{B}_a$, we obtain $\|u\|^2 \le (\delta^*)^2 \le 1/(2C_p)$. Consequently,

$$\big|\|u\|^2 - \big\|q - \phi^q(u)\big\|^2\big| \le \tfrac{1}{2}\|u\|^2, \qquad \big\|q - \phi^q(u)\big\|^2 \ge \tfrac{1}{2}\|u\|^2. \tag{6.3}$$

Therefore,

$$|T_3| \le \|f\|_\infty \Big(\frac{a}{\sqrt{2\pi}}\Big)^d \int_{\tilde{B}_a} \exp\Big\{-\frac{a^2\|u\|^2}{4}\Big\}\frac{a^2\|u\|^4}{2}\,du = C_6 a^{-2} \le C_6 a^{-s},$$

since $a \ge a_0 \ge 1$.

Step 1.4 (*Estimation of* T_4): By equation (7.5) in Proposition 7.5 in the supplementary material [36], there exists a constant C_7 depending on the Ricci tensor of the manifold $\mathcal{M}$, such that

$$\Big|\sqrt{\det\big(g^q_{ij}\big)}(u) - 1\Big| \le C_7\|u\|^2.$$

Therefore, by applying equation (6.3) again, we obtain

$$|T_4| \le C_4\|f\|_\infty \Big(\frac{a}{\sqrt{2\pi}}\Big)^d \int_{\tilde{B}_a} \exp\Big\{-\frac{a^2\|u\|^2}{4}\Big\}\|u\|^2\,du = C_8 a^{-2} \le C_8 a^{-s}.$$

Combining the above estimates for T_1, T_2, T_3 and T_4, we have

$$\sup_{x\in\mathcal{M}} \big|I_a(f)(q)(x) - f(q)(x)\big| \le (C_2 + C_3 + C_6 + C_8)a^{-s} = Ca^{-s}.$$

Step 2 (*Estimation of the RKHS norm*): Since $\langle K^a(x,\cdot), K^a(y,\cdot)\rangle_{\tilde{\mathbb{H}}^a} = K^a(x, y)$, we have

$$\begin{aligned} \big\|I_a(f)\big\|_{\tilde{\mathbb{H}}^a} &= \Big(\frac{a}{\sqrt{2\pi}}\Big)^{2d} \int_{\mathcal{M}}\int_{\mathcal{M}} K^a(x,y) f(x) f(y)\,dV(x)\,dV(y) \\ &\le \|f\|^2_\infty \Big(\frac{a}{\sqrt{2\pi}}\Big)^d \int_{\mathcal{M}} dV(x) \Big(\frac{a}{\sqrt{2\pi}}\Big)^d \int_{\mathcal{M}} K^a(x,y)\,dV(y). \end{aligned}$$

Applying the results of the first part to function $f \equiv 1$, we have

$$\left|\left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathcal{M}} K^a(x,y)\,dV(y) - 1\right| \le Ca^{-2} \le C,$$

since $a \ge a_0 \ge 1$. Therefore,

$$\|I_a(f)\|_{\tilde{\mathbb{H}}^a} \le (1+C)\|f\|_\infty^2 \left(\frac{a}{\sqrt{2\pi}}\right)^d \mathrm{Vol}(\mathcal{M}) = Ba^d.$$

6.3. *Proof of Lemma 5.4.* By Lemmas 5.1 and 5.3, a typical element of $\tilde{\mathbb{H}}^a$ can be written as the real part of the function

$$h_\psi(x) = \int e^{i(\lambda,x)}\psi(\lambda)\mu_a(d\lambda) \qquad \text{for } x \in \mathcal{M}$$

for $\psi : \mathbb{R}^D \to \mathbb{C}$ a function with $\int |\psi|^2 \mu_a(d\lambda) \le 1$. This function can be extended to $\mathbb{R}^D$ by allowing $x \in \mathbb{R}^D$. For any given point $p \in \mathcal{M}$, by (7.2) in the supplementary material [36], we have a local coordinate $\phi^p : B_{\delta_0}(0) \subset \mathbb{R}^d \to \mathbb{R}^D$ induced by the exponential map $\mathcal{E}_p$. Therefore, for $x \in \phi_p(B_{\delta_0}(0))$, $h_\psi(x)$ can be written in local q-normal coordinates as

$$h_{\psi,p}(u) = h_\psi\big(\phi^p(u)\big) = \int e^{i(\lambda,\phi^p(u))}\psi(\lambda)\mu_a(d\lambda), \qquad u \in B_{\delta_0}(0). \tag{6.4}$$

Similar to the idea in the proof of Lemma 4.5 in [35], we want to extend the function $h_{\psi,p}$ to an analytical function $z \mapsto \int e^{i(\lambda,\phi^p(z))}\psi(\lambda)\mu_a(d\lambda)$ on the set $\Omega = \{z \in \mathbb{C}^d : \|\mathrm{Re}\, z\| < \delta_0, \|\mathrm{Im}\, z\| < \rho/a\}$ for some $\rho > 0$. Then we can obtain upper bounds on the mixed partial derivatives of the analytic extension $h_{\psi,p}$ via the Cauchy formula, and finally construct an ε-covering set of $\tilde{\mathbb{H}}_1^a$ by piecewise polynomials defined on $\mathcal{M}$. Unfortunately, this analytical extension is impossible unless $\phi^p(u)$ is a polynomial. This motivates us to approximate $\phi^p(u)$ by its γth order Taylor polynomial $P_{p,\gamma}(u)$. More specifically, by Lemma 8.2 and the discussion after Lemma 7.7 in the supplementary material [36], the error caused by approximating $\phi^p(u)$ by $P_{p,\gamma}(u)$ is

$$\big|h_\psi\big(\phi^p(u)\big) - h_\psi\big(P_{p,\gamma}(u)\big)\big| \le a\big\|\phi^p(u) - P_{p,\gamma}(u)\big\| \le Ca\|u\|^\gamma. \tag{6.5}$$

For notation simplicity, fix p as a center and denote the function $h_\psi(P_{p,\gamma}(u))$ by $r(u)$ for $u \in B_{\delta_0}$. Since $P_{p,\gamma}(u)$ is a polynomial of degree γ, by viewing the function r as a function of argument u ranging over the product of the imaginary axes in $\mathbb{C}^d$, we can extend

$$r(u) = \int e^{i(\lambda,P_{p,\gamma}(u))}\psi(\lambda)\mu_a(d\lambda), \qquad u \in B_{\delta_0}(0) \tag{6.6}$$

to an analytical function $z \mapsto \int e^{i(\lambda,P_{p,\gamma}(z))}\psi(\lambda)\mu_a(d\lambda)$ on the set $\Omega = \{z \in \mathbb{C}^d : \|\mathrm{Re}\, z\| < \delta_0, \|\mathrm{Im}\, z\| < \rho/a\}$ for $\rho = \delta$, where $\delta < 1/2$ is defined in (5.2).

Moreover, by Cauchy–Schwarz inequality, $|r(z)| \le C$ for $z \in \Omega$ and $C^2 = \int e^{\delta\|\lambda\|}\mu(d\lambda) < \infty$. Therefore, by the Cauchy formula, with D^n denoting the partial derivative of orders $n = (n_1, \ldots, n_d)$ and $n! = n_1! \cdots n_d!$, we have the following bound for partial derivatives of r at any $u \in B_{\delta_0}(0)$,

$$\left|\frac{D^n r(u)}{n!}\right| \le \frac{C}{R^n}, \tag{6.7}$$

where $R = \rho/(a\sqrt{d})$. Based on the inequalities (6.5) and (6.7), we can construct an ε-covering set of $\tilde{\mathbb{H}}_1^a$ as follows.

Set $a_1 = \delta/(2\delta_0\sqrt{d})$, then for any $a > a_1$, $R < 2\delta_0$. Since $\mathcal{M} \subset [0,1]^D$, with C_2 defined in Lemma 7.6 in the supplementary material [36], let $\{p_1, \ldots, p_m\}$ be an $R/(2C_2)$-net in $\mathcal{M}$ for the Euclidean distance, and let $\mathcal{M} = \bigcup_i B_i$ be a partition of $\mathcal{M}$ into sets $B_1, \ldots, B_m$ obtained by assigning every $x \in \mathcal{M}$ to the closest $p_i \in \{p_1, \ldots, p_m\}$. By (7.3) and Lemma 7.6 in the supplementary material [36]

$$\left|(\phi^{p_i})^{-1}(x)\right| < C_2\frac{R}{2C_2} = \frac{R}{2} < \delta_0, \tag{6.8}$$

where ϕ_{p_i} is the local normal coordinate chart at p_i. Therefore, we can consider the piecewise transformed polynomials $P = \sum_{i=1}^m P_{i,a_i} 1_{B_i}$, with

$$P_{i,a_i}(x) = \sum_{n_\cdot \le k} a_{i,n}\left[(\phi^{p_i})^{-1}(x)\right]^n, \qquad x \in \phi^{p_i}\big(B_{\delta_0}(0)\big). \tag{6.9}$$

Here, the sum ranges over all multi-index vectors $n = (n_1, \ldots, n_d) \in (\mathbb{N} \cup \{0\})^d$ with $n_\cdot = n_1 + \cdots + n_d \le k$. Moreover, for $y = (y_1, \ldots, y_d) \in \mathbb{R}^d$, the notation y^n used above is short for $y_1^{n_1} y_2^{n_2} \cdots y_d^{n_d}$. We obtain a finite set of functions by discretizing the coefficients $a_{i,n}$ for each i and n over a grid of mesh width ε/R^n-net in the interval $[-C/R^n, C/R^n]$ [by (6.7)]. The log cardinality of this set is bounded by

$$\log\Bigg(\prod_i \prod_{n:n_\cdot \le k} \#a_{i,n}\Bigg) \le m \log\Bigg(\prod_{n:n_\cdot \le k} \frac{2C/R^n}{\varepsilon/R^n}\Bigg) \le mk^d \log\Big(\frac{2C}{\varepsilon}\Big).$$

Since $R = \rho/(a\sqrt{d})$, we can choose $m = N(\mathcal{M}, \|\cdot\|, \rho/(2C_0 a d^{1/2})) \simeq a^d$. To complete the proof, it suffices to show that for k of order $\log(1/\varepsilon)$, the resulting set of functions is a $K\varepsilon$-net for constant K depending only on μ.

For any function $f \in \tilde{\mathbb{H}}_1^a$, by Lemma 5.1, we can find a $g \in \tilde{\mathbb{H}}_1^a$ such that $g|_{\mathcal{M}} = f$. Assume that r_g (subscript g indicates the dependence on g) is the local polynomial approximation for g defined as (6.6). Then we have a partial derivative bound on r_g as

$$\left|\frac{D^n r_g(p_i)}{n!}\right| \le \frac{C}{R^n}.$$

Therefore, there exists a universal constant K and appropriately chosen a_i in (6.9), such that for any $z \in B_i \subset \mathcal{M}$,

$$\left|\sum_{n.>k} \frac{D^n r_g(p_i)}{n!}(z-p_i)^n\right| \le \sum_{n.>k} \frac{C}{R^n}(R/2)^n \le C \sum_{l=k+1}^{\infty} \frac{l^{d-1}}{2^l} \le KC\left(\frac{2}{3}\right)^k,$$

$$\left|\sum_{n.\le k} \frac{D^n r_g(p_i)}{n!}(z-p_i)^n - P_{i,n_i}(z)\right| \le \sum_{n.\le k} \frac{\varepsilon}{R^n}(R/2)^n \le \sum_{l=1}^{k} \frac{l^{d-1}}{2^l}\varepsilon \le K\varepsilon.$$

Moreover, by (6.5) and (6.8),

$$|g(z) - r_g(z)| \le Ca\left\|\left(\phi^{p_i}\right)^{-1}(z)\right\|^{\gamma} \le aR^{\gamma} \le Ka^{-(\gamma-1)} < K\varepsilon,$$

where the last step follows by the condition on a.

Consequently, we obtain

$$|f(z) - P_{i,n_i}(z)| = |g(z) - P_{i,n_i}(z)| \le |g(z) - r_g(z)| + |r_g(z) - P_{i,n_i}(z)| \le KC\left(\frac{2}{3}\right)^k + 2K\varepsilon.$$

This suggests that the piecewise polynomials form a $3K\varepsilon$-net for k sufficiently large so that $(2/3)^k$ is smaller than $K\varepsilon$.

6.4. *Proof of Theorem 2.3.* Let $P_0^{(n)}$ denote the joint distribution of $\{Y_i\}_{i=1}^n$ in fixed design or the joint distribution of $\{(X_i, Y_i)\}_{i=1}^n$ in random design. First, we consider fixed design. We will apply the following lemma which strengthens the result in Theorem 2.1. A proof of this lemma is provided in Section 9 of the supplementary material [36].

LEMMA 6.1. *Under the conditions of Theorem 2.1, there exists a sequence of measurable sets A_n satisfying $P_0^{(n)}(A_n^c) \to 0$, such that for some positive constant c and any $t \ge 1$,*

$$P_0^{(n)} I(A_n)\Pi\left(\|f - f_0\|_n \ge t\bar{\varepsilon}_n | S_n\right) \le e^{-cn\varepsilon_n^2 t^2}.$$

By plugging in $t = 1, 2, \ldots$ into the above display, dividing both sides by $\exp\{-cn\varepsilon_n^2 t^2/2\}$ and taking a summation, we can obtain

$$P_0^{(n)} I(A_n) \sum_{k=1}^{\infty} \Pi\left(\|f - f_0\|_n \ge k\bar{\varepsilon}_n | S_n\right) e^{cn\varepsilon_n^2 k^2/2} \le \sum_{k=1}^{\infty} e^{-cn\varepsilon_n^2 k^2/2} \to 0,$$

as $n \to \infty$. As a consequence, there exists a sequence of events $\{C_n\}$ satisfying $C_n \subset A_n$ and $P_0^{(n)}(C_n) \to 1$ as $n \to \infty$, such that under C_n the following inequality holds uniformly for all $k = 1, 2, \ldots$:

$$\Pi\big(\|f - f_0\|_n \geq k\bar{\varepsilon}_n | S_n\big) \leq e^{-cn\varepsilon_n^2 k^2/2}.$$

For any $t \geq 1$, there always exists an integer $k_t \geq 1$ such that $k_t \leq t < k_t + 1$. Then the preceding display implies

$$\begin{aligned}\Pi\big(\|f - f_0\|_n \geq t\bar{\varepsilon}_n | S_n\big) &\leq \Pi\big(\|f - f_0\|_n \geq (k_t + 1)\bar{\varepsilon}_n | S_n\big) \\ &\leq e^{-cn\varepsilon_n^2 (k_t+1)^2/2} \leq e^{-cn\varepsilon_n^2 k_t^2/8} \leq e^{-cn\varepsilon_n^2 t^2/8}.\end{aligned} \tag{6.10}$$

Therefore, by Fubini's theorem we have that under the event C_n

$$\begin{aligned}\int \|f - f_0\|_n^2 \, d\Pi(f|S_n) &= \int_0^\infty \Pi\big(\|f - f_0\|_n^2 \geq s | S_n\big)\, ds \\ &\leq \bar{\varepsilon}_n^2 + \int_{\bar{\varepsilon}_n^2}^\infty \Pi\big(\|f - f_0\|_n^2 \geq s | S_n\big)\, ds \\ &\leq \bar{\varepsilon}_n^2 + \int_{\varepsilon_n^2}^\infty e^{-cns/8}\, ds = \bar{\varepsilon}_n^2 + \frac{8}{cn} e^{-cn\varepsilon_n^2/8} \leq 2\bar{\varepsilon}_n^2\end{aligned} \tag{6.11}$$

for n sufficiently large so that $\frac{8}{cn} e^{-cn\varepsilon_n^2/8} \leq \bar{\varepsilon}_n^2$.

Since $\hat{f}_A = \int f_A \, d\Pi(f|S_n)$, we have decomposition $\int \|f_A - f_0\|_n^2 \, d\Pi(f|S_n) = \int \|f_A - \hat{f}_A\|_n^2 \, d\Pi(f|S_n) + \|\hat{f}_A - f_0\|_n^2$. Combining this decomposition with the fact that $|f_A(x) - f_0(x)| \leq |f(x) - f_0(x)|$ for any x, we obtain

$$\|\hat{f}_A - f_0\|_n^2 \leq \int \|f - f_0\|_n^2 \, d\Pi(f|S_n).$$

By combining the preceding display with (6.11) and noticing $P_0^{(n)}(C_n) \to 1$, we can conclude that $\|\hat{f}_A - f_0\|_n^2 \leq 2\bar{\varepsilon}_n^2$ holds with probability tending to 1 as $n \to \infty$.

The proof for random design is more involved. We will utilize the following result for comparing $\|\cdot\|_n$ and $\|\cdot\|_2$ based on empirical process theory ([32], Lemma 5.16). Let $H_B(\varepsilon, \mathcal{F}, \|\cdot\|)$ denote the ε-bracketing entropy of a function space $\mathcal{F}$ with respect to a norm $\|\cdot\|$.

LEMMA 6.2. *Suppose* $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq A$ *and* X_i *are i.i.d. with distribution* Q. *For* δ *satisfying* $nA^{-2}\delta^2 \geq H_B(A^{-1}\delta, \mathcal{F}, \|\cdot\|_2)$ *and* $\eta \in (0, 1)$, *we have*

$$Q^{(n)}\Bigg(\sup_{f \in \mathcal{F}, \|f\|_2 \geq 2^5\delta/\eta} \left| \frac{\|f\|_n}{\|f\|_2} - 1 \right| \geq \eta \Bigg) \leq 8 \exp\big(-CnA^{-2}\delta^2\eta^2\big),$$

where the constant $C > 0$ *does not depend on* A *and* $Q^{(n)}$ *denotes the joint distribution of* $\{X_i\}_{i=1}^n$.

According to the proof of Theorem 2.1 in the supplementary material [36], there exists a sequence of sieves $\{B_n\}$ that satisfies

$$N(3\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) \le n\bar{\varepsilon}_n^2 \quad \text{and} \quad \Pi(f \notin B_n) \le e^{-4n\varepsilon_n^2}.$$

By the second inequality above and Lemma 1 in [14], we obtain that the posterior probability $\Pi(f \notin B_n|S_n) \le e^{-c_1 n\varepsilon_n^2}$ under the event A_n with some constant $c_1 > 0$. This inequality together with (6.10) implies

$$\Pi(\|f - f_0\|_n \le \bar{\varepsilon}_n, f \in B_n|S_n) \ge 1 - 2e^{-c_2 n\varepsilon_n^2}$$

under the event $A_n \cap C_n = C_n$, where $c_2 = \min\{c/8, c_1\}$.

Let $B_{n,A} = \{f_A : f \in B_n\}$. If $\{f^{(j)}\}$ forms an ε-net of B_n, then $\{f_A^{(j)}\}$ forms an ε-net of $B_{n,A}$. As a result, the covering entropy of $B_{n,A}$ is bounded by that of B_n. Combining this, the two preceding displays, inequality (6.10) and the fact that an ε-bracket entropy is always bounded by an ε-covering entropy with respect to $\|\cdot\|_\infty$, we obtain that under the event $A_n \cap C_n = C_n$,

$$\Pi(\|f_A - f_0\|_n \le \bar{\varepsilon}_n, f_A \in B_{n,A}|S_n) \ge 1 - 2e^{-c_2 n\varepsilon_n^2}, \tag{6.12}$$

$$H_B(3\bar{\varepsilon}_n, B_{n,A}, \|\cdot\|_2) \le n\bar{\varepsilon}_n^2. \tag{6.13}$$

Given inequality (6.13), we can apply Lemma 6.2 with $\mathcal{F} = B_{n,A} - f_0$, $\delta = \bar{\varepsilon}_n$ and $\eta = 1/2$ to obtain that there exists a sequence of events E_n with $P_0^{(n)}(E_n) = Q^{(n)}(E_n) \to 1$ as $n \to \infty$ such that under E_n,

$$\frac{1}{2} \le \sup_{f_A \in B_{n,A}, \|f_A - f_0\|_2 \ge 64\bar{\varepsilon}_n} \frac{\|f_A - f_0\|_n}{\|f_A - f_0\|_2} \le \frac{3}{2}.$$

By combining the above with inequality (6.12), we obtain $\Pi(\|f_A - f_0\|_2 \le 64\bar{\varepsilon}_n|S_n) \ge 1 - 2e^{-c_2 n\varepsilon_n^2}$ and then

$$\int \|f_A - f_0\|_2^2 \, d\Pi(f|S_n) \le 64^2\bar{\varepsilon}_n^2 + 4A^2\Pi(\|f_A - f_0\|_2 \ge 64\bar{\varepsilon}_n|S_n) \le 4097\bar{\varepsilon}_n^2$$

under the event $C_n \cap E_n$ for n sufficiently large so that $8A^2e^{-c_2 n\varepsilon_n^2} \le \bar{\varepsilon}_n^2$. Therefore, we have $\|\hat{f}_A - f_0\|_2^2 \le \int \|f_A - f_0\|_2^2 \, d\Pi(f|S_n) \le 4097\bar{\varepsilon}_n^2$ with probability at least $P_0^{(n)}(C_n \cap E_n) \to 1$ as $n \to \infty$.

SUPPLEMENTARY MATERIAL

Reviews of geometric properties and proofs of Theorems 2.1, 2.2, 2.4 and 3.2 (DOI: 10.1214/15-AOS1390SUPP; .pdf). Concepts and results in differential and Riemannian geometry were reviewed in Section 7, where new results are included with proofs. Then proofs of Theorems 2.1, 2.2, 2.4 and 3.2 are provided in Section 8.

REFERENCES

[1] ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404. MR0051437

[2] BELKIN, M. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15** 1373–1396.

[3] BHATTACHARYA, A., PATI, D. and DUNSON, D. (2014). Anisotropic function estimation using multi-bandwidth Gaussian processes. *Ann. Statist.* **42** 352–381. MR3189489

[4] BICKEL, P. J. and KLEIJN, B. J. K. (2012). The semiparametric Bernstein–von Mises theorem. *Ann. Statist.* **40** 206–237. MR3013185

[5] BICKEL, P. J. and LI, B. (2007). Local polynomial regression on unknown manifolds. In *Complex Datasets and Inverse Problems. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **54** 177–186. IMS, Beachwood, OH. MR2459188

[6] BINEV, P., COHEN, A., DAHMEN, W. and DEVORE, R. (2007). Universal algorithms for learning theory. II. Piecewise polynomial functions. *Constr. Approx.* **26** 127–152. MR2327596

[7] BINEV, P., COHEN, A., DAHMEN, W., DEVORE, R. and TEMLYAKOV, V. (2005). Universal algorithms for learning theory. I. Piecewise constant functions. *J. Mach. Learn. Res.* **6** 1297–1321. MR2249856

[8] CAMASTRA, F. and VINVIARELLI, A. (2002). Estimating the intrinsic dimension of data with a fractal-based method. *IEEE P.A.M.I.* **24** 1404–1407.

[9] CARTER, K. M., RAICH, R. and HERO, A. O. III (2010). On local intrinsic dimension estimation and its applications. *IEEE Trans. Signal Process.* **58** 650–663. MR2750463

[10] CASTILLO, I., KERKYACHARIAN, G. and PICARD, D. (2013). Thomas Bayes' walk on manifolds. *Probab. Theory Related Fields* **158** 665–710.

[11] CHEN, M., SILVA, J., PAISLEY, J., WANG, C., DUNSON, D. and CARIN, L. (2010). Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Trans. Signal Process.* **58** 6140–6155. MR2790088

[12] FARAHMAND, A. M., SZEPESVÁI, C. and AUDIBERT, J. (2007). Manifold-adaptive dimension estimation. In *ICML* 2007 265–272. ACM Press, New York.

[13] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

[14] GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274

[15] GINÉ, E. and NICKL, R. (2011). Rates on contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist.* **39** 2883–2911. MR3012395

[16] KPOTUFE, S. (2009). Escaping the curse of dimensionality with a tree-based regressor. In *COLT* 2009—*The* 22*nd Conference on Learning Theory, June* 18–21. Montreal, QC.

[17] KPOTUFE, S. and DASGUPTA, S. (2012). A tree-based regressor that adapts to intrinsic dimension. *J. Comput. System Sci.* **78** 1496–1515. MR2926146

[18] KUNDU, S. and DUNSON, D. B. (2011). Latent factor models for density estimation. Available at arXiv:1108.2720v2.

[19] LAWRENCE, N. D. (2003). Gaussian process latent variable models for visualisation of high dimensional data. *Neural Information Processing Systems* **16** 329–336.

[20] LEVINA, E. and BICKEL, P. (2004). Maximun likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems* **17**. MIT Press, Cambridge, MA.

[21] LIN, L. and DUNSON, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika* **101** 303–317. MR3215349

[22] LITTLE, A. V., LEE, J., JUNG, Y. M. and MAGGIONI, M. (2009). Estimation of intrinsic dimensionality of samples from noisy low-dimensional manifolds in high dimensions with multiscale SVD. In 2009 *IEEE/SP* 15*th Workshop on Statistical Signal Processing* 85–88. IEEE, Cardiff.

[23] NENE, S. A., NAYAR, S. K. and MURASE, H. (1996). Columbia object image library (COIL-100). Technical report, Columbia Univ., New York.

[24] PAGE, G., BHATTACHARYA, A. and DUNSON, D. (2013). Classification via Bayesian nonparametric learning of affine subspaces. *J. Amer. Statist. Assoc.* **108** 187–201. MR3174612

[25] REICH, B. J., BONDELL, H. D. and LI, L. (2011). Sufficient dimension reduction via Bayesian mixture modeling. *Biometrics* **67** 886–895. MR2829263

[26] ROWEIS, S. T. and SAUL, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323–2326.

[27] SAVITSKY, T., VANNUCCI, M. and SHA, N. (2011). Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statist. Sci.* **26** 130–149. MR2849913

[28] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642

[29] TENENBAUM, J. B., DE SILVA, V. and LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.

[30] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 273–282.

[31] TOKDAR, S. T., ZHU, Y. M. and GHOSH, J. K. (2010). Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Anal.* **5** 319–344. MR2719655

[32] VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

[33] VAN DER VAART, A. and VAN ZANTEN, H. (2011). Information rates of nonparametric Gaussian process methods. *J. Mach. Learn. Res.* **12** 2095–2119. MR2819028

[34] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics*: *Contributions in Honor of Jayanta K. Ghosh. Inst. Math. Stat. Collect.* **3** 200–222. IMS, Beachwood, OH. MR2459226

[35] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442

[36] YANG, Y. and DUNSON, D. B. (2015). Supplement to "Bayesian manifold regression." DOI:10.1214/15-AOS1390SUPP.

[37] YE, G.-B. and ZHOU, D.-X. (2008). Learning and approximation by Gaussians on Riemannian manifolds. *Adv. Comput. Math.* **29** 291–310. MR2438346

[38] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

DEPARTMENT OF EECS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: yy84@berkeley.edu

DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
BOX 90251
DURHAM, NORTH CAROLINA 27708-0251
USA
E-MAIL: dunson@duke.edu