Adv Comput Math
DOI 10.1007/s10444-007-9049-0

Learning and approximation by Gaussians on Riemannian manifolds

Gui-Bo Ye · Ding-Xuan Zhou

Received: 3 September 2006 / Accepted: 6 February 2007

Abstract Learning function relations or understanding structures of data lying in manifolds embedded in huge dimensional Euclidean spaces is an important topic in learning theory. In this paper we study the approximation and learning by Gaussians of functions defined on a d-dimensional connected compact C^∞ Riemannian submanifold of $\mathbb{R}^n$ which is isometrically embedded. We show that the convolution with the Gaussian kernel with variance σ provides the uniform approximation order of $O(\sigma^s)$ when the approximated function is Lipschitz $s \in (0, 1]$. The uniform normal neighborhoods of a compact Riemannian manifold play a central role in deriving the approximation order. This approximation result is used to investigate the regression learning algorithm generated by the multi-kernel least square regularization scheme associated with Gaussian kernels with flexible variances. When the regression function is Lipschitz s, our learning rate is $(\log^2 m)/m)^{s/(8s+4d)}$ where m is the sample size. When the manifold dimension d is smaller than the dimension n of the underlying Euclidean space, this rate is much faster compared with those in the literature. By comparing approximation orders, we also show the essential difference between approximation schemes with flexible variances and those with a single variance.

Communicated by Charles A. Micchelli.

Supported partially by the Research Grants Council of Hong Kong [Project No. CityU 103405], City University of Hong Kong [Project No. 7001983], National Science Fund for Distinguished Young Scholars of China [Project No. 10529101], and National Basic Research Program of China [Project No. 973-2006CB303102].

G.-B. Ye
School of Mathematical Sciences, Fudan University,
Shanghai 200433, People's Republic of China
e-mail: yeguibo@hotmail.com

D.-X. Zhou (✉)
Department of Mathematics, City University of Hong Kong,
Kowloon, Hong Kong, China
e-mail: mazhou@cityu.edu.hk

Keywords Learning theory · Reproducing kernel Hilbert spaces · Gaussian kernels · Approximation · Riemannian manifolds · Multi-kernel least square regularization scheme

Mathematics Subject Classifications (2000) 68T05 · 62J02

1 Introduction and main results

Learning theory studies learning of function relations or data structures from samples. The desired function or data arise from physical or biological systems, engineering problems, financial studies, and many other fields. They can be effectively modelled or analyzed on an input space X which is often a low-dimensional manifold embedded in a Euclidean space $\mathbb{R}^n$. In many applications such as gene expression analysis, it is observed that the dimension n of the underlying Euclidean space is huge while the intrinsic dimension d of the manifold X is much smaller $d << n$. This has led to the hot topic of manifold learning and several important learning strategies including dimensionality reduction [4, 12], feature selection [6, 18, 21], semi-supervised learning [3, 21], and learning topological statistics [14].

To model such situations, as in the literature of manifold learning [2, 14, 18], we assume throughout the paper that X is a d-dimensional connected compact C^∞ submanifold of $\mathbb{R}^n$ which is isometrically embedded. For detailed definition and properties, see [5, 7] and the description in Section 2. In particular, we know that X is a metric space with the metric d_X and the inclusion map $\Phi : (X, d_X) \hookrightarrow (\mathbb{R}^n, \|\cdot\|)$ is well defined and continuous (actually it is C^∞). Here $\|\cdot\|$ is the norm in $\mathbb{R}^n$.

The purpose of this paper is to investigate the approximation of functions on Riemannian manifolds by Gaussians and its applications in quantitative analysis of learning algorithms. The Gaussians form a family of functions with an index $\sigma \in (0, \infty)$ defined for $x, y \in X$ or on the whole underlying space $\mathbb{R}^n$ by

$$K_\sigma(x, y) = \exp\left\{-\frac{\|x - y\|^2}{2\sigma^2}\right\}. \tag{1}$$

When X has nonempty interior as a subset of $\mathbb{R}^n (d = n)$, the approximation of functions from various function spaces by Gaussians is a classical topic in approximation theory [11] and its applications in error analysis of learning algorithms are well understood [17]. When X is a Riemannian submanifold of $\mathbb{R}^n$ with dimension $d < n$, things are totally different and little is known. In fact, our assumption that the embedding map Φ is the inclusion plays an essential role. For a general embedding map (which always exists according to the Nash Embedding Theorem), we do not know how to establish similar analysis for the approximation.

We consider the approximation in the space $C(X)$ of continuous functions on X with the norm $\|f\|_{C(X)} = \max_{x\in X} |f(x)|$. The approximation scheme is given by a family of linear operators $\{I_\sigma : C(X) \to C(X)\}_{\sigma>0}$ as

$$\begin{aligned} I_\sigma(f)(x) &= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^d} \int_X K_\sigma(x, y) f(y) dV(y) \\ &= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^d} \int_X \exp\left\{-\frac{\|x - y\|^2}{2\sigma^2}\right\} f(y) dV(y), \qquad x \in X, \end{aligned} \tag{2}$$

where V is the Riemannian volume measure of X. For details on the Riemannian volume measure of a Riemannian manifold, see Section 2. It is a generalization of the Lebesgue measure in a Euclidean space. A d-dimensional manifold is, roughly speaking, a topological space which is locally Euclidean of dimension d. This verifies the use of the power d in the scaling factor $\frac{1}{(\sqrt{2\pi}\sigma)^d}$.

To get explicit approximation orders, we need some smoothness of the approximated function f. Here we use the Lipschitz smoothness.

Definition 1 Let X be a Riemannian manifold with the metric d_X and $0 < s \leq 1$. We say that a bounded continuous function f on X is in the Lipschitz s space $\mathrm{Lip}(s) = \mathrm{Lip}(s, X)$ if there exists a constant $C > 0$ such that for all $x, y \in X$,

$$|f(x) - f(y)| \leq C(d_X(x, y))^s.$$

The norm in the space $\mathrm{Lip}(s)$ is defined as

$$\|f\|_{Lip(s)} := |f|_{Lip(s)} + \|f\|_{C(X)}$$

where $|f|_{Lip(s)}$ is the seminorm

$$|f|_{Lip(s)} := \sup_{x \neq y \in X} \frac{|f(x) - f(y)|}{(d_X(x, y))^s}.$$

The space $\mathrm{Lip}(s)$ is a Banach space. The smoothness of a function $f \in \mathrm{Lip}(s)$ is measured by the index s. The bigger the index s is, the smoother the function f is.

Our first main result is the following theorem which will be proved in Section 3.

Theorem 1 *Let X be a connected compact C^∞ submanifold of $\mathbb{R}^n$ which is isometrically embedded and of dimension d. Define $I_\sigma : C(X) \to C(X)$ for $\sigma > 0$ by (2). If $f \in Lip(s)$ with $0 < s \leq 1$, then there holds*

$$\|I_\sigma(f) - f\|_{C(X)} \leq C_X \|f\|_{Lip(s)} \sigma^s \qquad \forall \sigma > 0, \tag{3}$$

where C_X is a positive constant independent of f or σ.

Due to a phenomenon of saturation, the order of approximation in (3) cannot be increased to $s > 2$ on function spaces of higher order Lipschitz smoothness (see [11]), though extensions to orders with $1 \leq s < 2$ are possible. Convergence rates like (3) may be established for other function spaces such as $L^p(X)$ and for non-compact manifolds (see the special example in Proposition 1 below), which is out of the scope of the paper.

Our second main result is the error analysis for the regression algorithm generated by the multi-kernel least-square regularization scheme associated with Gaussians with flexible variance [22] which is an application of Theorem 1 in learning theory.

In regression problems, we assume that the sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ is independently drawn according to a probability measure ρ on $Z := X \times Y$ with $Y = \mathbb{R}$. The measure ρ can be decomposed into the marginal distribution ρ_X on X and the

conditional distribution $\rho(\cdot|x)$ at $x \in X$. The goal of the regression problem is to learn the regression function

$$f_\rho(x) = \int_Y y d\rho(y|x), \qquad x \in X \tag{4}$$

from the sample $\mathbf{z}$ by some learning algorithms. The algorithm considered here is a kernel method.

A *Mercer kernel* $K : X \times X \to \mathbb{R}$ is a continuous and symmetric function which is positive semidefinite in the sense that for any finite set of points $\{x_1, \cdots, x_\ell\} \subseteq X$, the matrix $(K(x_i, x_j))_{i,j=1}^\ell$ is positive semidefinite. The *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}_K$ associated with the kernel K is defined [1] to be the completion of the linear span of the set of functions $\{K_x = K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ given by $\langle K_x, K_y \rangle_K = K(x, y)$. Its reproducing property plays a special role in learning theory:

$$\langle K_x, f \rangle_K = f(x), \qquad x \in X, \quad f \in \mathcal{H}_K. \tag{5}$$

For each $\sigma \in (0, \infty)$, the Gaussian (1) on X is a Mercer kernel [9] and the corresponding RKHS is denoted by $\mathcal{H}_{K_\sigma}$ with associated norm $\| \cdot \|_{K_\sigma}$.

We consider a multi-kernel regularization scheme. It is a least-square regularized algorithm for the regression problem associated with the Gaussians with flexible variance $(0 < \sigma < \infty)$ defined by

$$f_{\mathbf{z},\lambda} := \arg \min_{\sigma \in (0,+\infty)} \min_{f \in \mathcal{H}_{K_\sigma}} \left\{ \frac{1}{m} \sum_{i=1}^m \big(f(x_i) - y_i \big)^2 + \lambda \| f \|_{K_\sigma}^2 \right\}. \tag{6}$$

Here $\lambda > 0$ is called the *regularization parameter*.

Theorem 2 *Let X be a connected compact C^∞ submanifold of $\mathbb{R}^n$ which is isometrically embedded and of dimension d. Let $f_{\mathbf{z},\lambda}$ be defined by (6) with a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ independently drawn according to ρ. If $f_\rho \in Lip(s)$ for some $0 < s \le 1$, then by taking $\lambda = \left(\frac{\log^2 m}{m}\right)^{\frac{s+d}{8s+4d}}$, we have*

$$\mathbb{E}_{\mathbf{z} \in Z^m} \left\{ \| f_{\mathbf{z},\lambda} - f_\rho \|_{L^2_{\rho_X}}^2 \right\} = O\left(\left(\frac{\log^2 m}{m} \right)^{\frac{s}{8s+4d}} \right). \tag{7}$$

The least-square regularization scheme with one Mercer kernel has been well understood in learning theory for various purposes including regression and classification [10, 13, 20]. But the approximation ability of the regularization scheme with one fixed Gaussian is weak, as shown in [19] and in Section 5 of this paper.

When all the Gaussians with $\sigma \in (0, \infty)$ are included in the multi-kernel regularization scheme (6), the approximation ability of the scheme and hence the learning rates are greatly improved. This has been studied recently in [17, 25] and [22] for classification. The sample error for the algorithm (6) was estimated in [25] by means of bounding empirical covering numbers for the union of unit balls of the corresponding infinitely many reproducing kernel Hilbert spaces. When ρ_X is the Lebesgue measure, X is a domain with Lipschitz boundary of $\mathbb{R}^n$ meaning that $d = n$, and f_ρ lies in the Sobolev space $H^s(X)$, some approximation error estimates by means of

the Fourier transform on $\mathbb{R}^n$ led to $\mathbb{E}_{\mathbf{z}\in Z^m}\{\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2(X)}\} = O\big((\log m)^{1/2} m^{-\frac{s}{8s+2n}}\big)$ in [25] and learning rates with confidence in [17]. These learning rates are better than (7) in the case $d = n$. However, in our current setting the input space X is a Riemannian submanifold of $\mathbb{R}^n$ with dimension d. The Fourier transform technique cannot be used for our estimates of approximation orders, which is the key difficulty we overcome. So the learning rate stated in Theorem 2 is completely new. When the manifold dimension d is much smaller than n, the index $\frac{s}{8s+4d}$ appearing in (7) is much better than the index $\frac{s}{8s+2n}$ in [17, 25] if we regard X as a subset of $\mathbb{R}^n$. As in the case of domains ($d = n$), it would be interesting to derive optimal or almost optimal rates for the learning algorithm (6).

Our ideas for deriving quantitative analysis for learning algorithms on Riemannian manifolds can be extended to other problems such as online learning [24] and feature selection [15].

2 Ideas and knowledge on Riemannian manifolds

In this section, we give some ideas on the approximation by Gaussians for the proof of Theorem 1 and then introduce two important concepts for Riemannian manifolds: exponential map and uniform normal neighborhoods which will play an important role in our proof.

2.1 Some ideas from Gaussian approximation on $\mathbb{R}^n$

Let us recall the following standard and well-known result [11] for approximation by Gaussians on the whole Euclidean space $\mathbb{R}^n$. We state it and give its proof in order to illustrate general ideas for deriving our error bounds.

Proposition 1 *If $f \in Lip(s, \mathbb{R}^n)$ for some $0 < s \le 1$, then for every $\sigma > 0$, we have*

$$\sup_{x\in\mathbb{R}^n}\left|\frac{1}{(\sqrt{2\pi}\sigma)^n}\int_{\mathbb{R}^n}\exp\left\{-\frac{\|x-y\|^2}{2\sigma^2}\right\}f(y)dy - f(x)\right| \le \left\{2^{s/2}\frac{\Gamma(\frac{n+s}{2})}{\Gamma(\frac{n}{2})}|f|_{Lip(s)}\right\}\sigma^s,$$

where Γ is the Gamma function given for $\alpha \in (0,\infty)$ by $\Gamma(\alpha) = \int_0^\infty r^{\alpha-1}e^{-r}dr$.

Proof Let $x \in \mathbb{R}^n$. Since

$$\frac{1}{(\sqrt{2\pi}\sigma)^n}\int_{\mathbb{R}^n}\exp\left\{-\frac{\|x-y\|^2}{2\sigma^2}\right\}dy = 1, \tag{8}$$

we can express $f(x)$ as

$$f(x) = \frac{1}{(\sqrt{2\pi}\sigma)^n}\int_{\mathbb{R}^n}\exp\left\{-\frac{\|x-y\|^2}{2\sigma^2}\right\}f(x)dy.$$

It follows from the definition of the Lipschitz seminorm $|f|_{Lip(s)}$ that

$$\begin{aligned}
&\left|\frac{1}{(\sqrt{2\pi}\sigma)^n}\int_{\mathbb{R}^n}\exp\left\{-\frac{\|x-y\|^2}{2\sigma^2}\right\}f(y)dy - f(x)\right| \\
&\quad\le \frac{1}{(\sqrt{2\pi}\sigma)^n}\int_{\mathbb{R}^n}\exp\left\{-\frac{\|x-y\|^2}{2\sigma^2}\right\}|f(y)-f(x)|dy \\
&\quad\le \frac{1}{(\sqrt{2\pi}\sigma)^n}\int_{\mathbb{R}^n}\exp\left\{-\frac{\|x-y\|^2}{2\sigma^2}\right\}\|x-y\|^s dy|f|_{Lip(s)}.
\end{aligned}$$

Recall that by spherical coordinates for $\mathbb{R}^n$, for a radial function $h(\|y\|)$ on $\mathbb{R}^n$ with a univariate function $h:\mathbb{R}\to\mathbb{R}$, there holds

$$\int_{\mathbb{R}^n} h(\|y\|)dy = \frac{2\pi^{n/2}}{\Gamma(n/2)}\int_0^\infty h(r)r^{n-1}dr. \tag{9}$$

Applying (9) to the function $h(r)=r^s\exp\{-\frac{r^2}{2\sigma^2}\}$, we have

$$\begin{aligned}
&\left|\frac{1}{(\sqrt{2\pi}\sigma)^n}\int_{\mathbb{R}^n}\exp\left\{-\frac{\|x-y\|^2}{2\sigma^2}\right\}f(y)dy - f(x)\right| \\
&\quad\le \left\{\frac{2\pi^{n/2}}{\Gamma(n/2)}\frac{1}{(\sqrt{2\pi}\sigma)^n}\int_0^\infty r^{n-1+s}\exp\left\{-\frac{r^2}{2\sigma^2}\right\}dr\right\}|f|_{Lip(s)} \\
&\quad= \frac{\Gamma(\frac{n+s}{2})}{\Gamma(\frac{n}{2})}2^{s/2}\sigma^s|f|_{Lip(s)}.
\end{aligned}$$

This bound is independent of $x\in\mathbb{R}^n$, so the proposition is proved. □

If we divide $\mathbb{R}^n$ into two parts: a neighborhood $B_\sigma(x)$ of x with suitable radius and its complement $\mathbb{R}^n\setminus B_\sigma(x)$, we see two key features caused by the fast decay of the Gaussian: the first is the fast decay of the kernel $K_\sigma(x,y)$ when $\|x-y\|$ becomes large making $\int_{\mathbb{R}^n\setminus B_\sigma(x)}K_\sigma(x,y)\|x-y\|^s dy = O(\sigma^s)$; the second is that

$$\frac{1}{(\sqrt{2\pi}\sigma)^n}\int_{B_\sigma(x)}\exp\left\{-\frac{\|x-y\|^2}{2\sigma^2}\right\}dy \approx 1. \tag{10}$$

Such a neighborhood can be chosen as $B_\sigma(x)=\{y\in\mathbb{R}^n:\|x-y\|\le\sqrt{2n+2}\sigma\sqrt{\log(1/\sigma)}\}$. Two key features allow us to adapt the proof of Proposition 1 to a manifold setting: choosing suitable appropriate atlas (U_i,ϕ_i) on which the Riemannian volume measure corresponds to a good approximation of the Lebesgue measure on the mapped region on $\mathbb{R}^d$; then computing approximately in the small region of $\mathbb{R}^d$. This procedure requires some good properties of the atlas (U_i,ϕ_i) of the manifold, which leads

to two important concepts in Riemannian manifold: exponential map and uniform normal neighborhoods introduced into learning theory in e.g. [2, 14].

2.2 Riemannian structures

A tangent vector v at $p \in X$ is a linear functional on $C^\infty(X)$ satisfying $v(fg) = v(f)g(p) + f(p)v(g)$. Denote by $T_p(X)$ the tangent space consisting of all tangent vectors of X at p. The C^∞ map $\Phi : X \mapsto \mathbb{R}^n$ induces a map $d\Phi_p : T_p(X) \mapsto T_{\Phi(p)}(\mathbb{R}^n)$ for each $p \in X$ expressed as

$$\big(d\Phi_p(v)\big)(f) = v(f \circ \Phi), \qquad v \in T_p(X),\ f \in C^\infty(\mathbb{R}^n).$$

Many quantities of Riemannian manifolds (such as area and length) are determined by Riemannian structures.

Definition 2 A **Riemannian structure** g on a C^∞ manifold X is an inner product g_p on each $T_p(X)$ such that, for each pair of C^∞ vector fields Y and Z, the function from X to $\mathbb{R}$ given by

$$p \mapsto g_p(Y_p, Z_p)$$

is C^∞. Here for a vector field $Y : C^\infty(X) \mapsto C^\infty(X)$, Y_p is the tangent vector defined by $Y_p f = Yf(p)$ for $f \in C^\infty(X)$. If X is isometrically embedded in $\mathbb{R}^n$ by the inclusion map $\Phi : X \mapsto \mathbb{R}^n$, then g_p has a special form given by

$$g_p(Y_p, Z_p) = \langle d\Phi_p(Y_p), d\Phi_p(Z_p)\rangle_{\mathbb{R}^n}. \tag{11}$$

The special feature of the isometrically embedded manifold $X \hookrightarrow \mathbb{R}^n$ is that the inner product on the tangent space $T_{\Phi(p)}(\mathbb{R}^n)$ is identical with that in $\mathbb{R}^n$.

Let us illustrate the concepts of tangent space and Riemannian structure in terms of local coordinates.

Let $\phi : U \subset \mathbb{R}^d \to X$ be a system of coordinates around p and $q = \phi(x^1, x^2, \cdots, x^d) \in \phi(U) \subset X$ with $(x^1, x^2, \cdots, x^d) \in U$. Let $\psi_i(t) = \phi(x^1, \cdots, x^{i-1}, x^i + t, x^{i+1}, \cdots, x^d)$ for $t \in (-\epsilon, \epsilon)$ with sufficiently small $\epsilon > 0$, then $\big(\frac{\partial}{\partial x^i}(q)\big)(f) := \frac{d(f\circ\psi_i)}{dt}\Big|_{t=0}$ gives a tangent vector $\frac{\partial}{\partial x^i}(q)$ in the tangent space $T_q(X)$ and $\big\{\frac{\partial}{\partial x^i}(q)\big\}_{i=1}^d$ forms a basis of $T_q(X)$. Under this basis, the map $d\Phi_q$ can be determined by

$$d\Phi_q\left(\frac{\partial}{\partial x^i}(q)\right) = \frac{d(\Phi \circ \psi_i)}{dt}\bigg|_{t=0}, \quad i = 1, \cdots, d. \tag{12}$$

The Riemannian structure g of the isometrically embedded manifold X can be expressed under this basis as

$$g_{ij}^{\phi}(x^1, \cdots, x^d) := \left\langle d\Phi_q\left(\frac{\partial}{\partial x^i}(q)\right), d\Phi_q\left(\frac{\partial}{\partial x^j}(q)\right)\right\rangle_{\mathbb{R}^n}, \quad i, j = 1, \cdots, d. \tag{13}$$

In fact, any pair of tangent vectors $v, w \in T_q(X)$ can be written as $v = \sum_{i=1}^d v^i \frac{\partial}{\partial x^i}(q)$ and $w = \sum_{i=1}^d w^i \frac{\partial}{\partial x^i}(q)$. In this form, the inner product (11) equals

$$\begin{aligned} g_q(v, w) &= \sum_{i,j=1}^d v^i w^j g_q\left(\frac{\partial}{\partial x^i}(q), \frac{\partial}{\partial x^j}(q)\right) \\ &= \sum_{i,j=1}^d v^i w^j \left\langle d\Phi_q\left(\frac{\partial}{\partial x^i}(q)\right), d\Phi_q\left(\frac{\partial}{\partial x^j}(q)\right)\right\rangle_{\mathbb{R}^n} \\ &= \sum_{i,j=1}^d v^i w^j g_{ij}^{\phi}(x^1, \cdots, x^d). \end{aligned}$$

The function g_{ij}^{ϕ} which is C^∞ on U is called the local representation of the Riemannian structure in the coordinate system (U, ϕ).

2.3 Exponential map and uniform normal neighborhoods

To understand the local structure we use the concept of exponential map based on geodesics which are special curves satisfying some ordinary differential equations, for details, see [7].

Definition 3 For $p \in X$ and $v \in T_p(X)$, let $\gamma(t, p, v), t > 0$, be the geodesic satisfying $\gamma(0, p, v) = p$ and $\gamma'(0, p, v) = v$. The **exponential map** $\mathcal{E}_p : T_p(X) \to X$ is defined by $\mathcal{E}_p(v) = \gamma(1, p, v)$.

Recall that a Riemannian structure gives an inner product g_q hence the metric on $T_q(X)$: $|v| = \sqrt{g_q(v, v)}$. A minimizing geodesic joining two points $p, q \in X$ is a geodesic $\gamma(t), t_0 \le t \le t_1$, having the minimum length $\int_{t_0}^{t_1} |\gamma'(t)| dt$ among all geodesics joining p and q. It carries the tangent space $T_p(X)$ at p to the tangent space $T_{\gamma(t)}(X)$ at $\gamma(t) \in X$ smoothly by parallel transport [7]: $v \in T_p(X) \mapsto v^{\gamma(t)} \in T_{\gamma(t)}(X)$. A special feature of this parallel transport is that it keeps the inner product: $g_{\gamma(t)}(v^{\gamma(t)}, w^{\gamma(t)}) = g_p(v, w), \forall\, v, w \in T_p(X)$. In particular, $g_q(v^q, w^q) = g_p(v, w)$.

By [7], we know that for each $p \in X$, there exists a strongly convex neighborhood U_p of p, that is, for any two points q_1, q_2 in the closure $\overline{U_p}$ of U_p, there exists a unique minimizing geodesic γ joining q_1 and q_2 whose interior is contained in U_p.

Choose an orthonormal basis $\{e_1, e_2, \cdots, e_d\}$ of $T_p(X)$, then for each $q \in U_p$, the set of tangent vectors $\{e_1^q, e_2^q, \cdots, e_d^q\}$, moved by parallel transport from p to q along the unique minimizing geodesic, forms an orthonormal basis of $T_q(X)$. In addition, this frame depends smoothly on q.

In order to study the structure in a small neighborhood of each $q \in U_p$, we need the concept of uniform normal neighborhood of p. Denote $B_\delta(0) = \{v \in T_q(X) : |v| < \delta\}$ as the ball of $T_q(X)$ centered at 0 with radius δ.

Definition 4 An open set $U \subset X$ is called **uniformly normal** if there exists some $\delta > 0$ such that $U \subseteq \mathcal{E}_q(B_\delta(0))$ for every $q \in U$.

The following proposition which will be proved in the Appendix tells us the existence of a special uniformly normal neighborhood.

Proposition 2 *For every $p \in X$ there exist a neighborhood W_p and a number $\delta_p > 0$ such that the following conditions hold:*

(a) *For every $q \in W_p$, the map $\mathcal{E}_q : B_{\delta_p}(0) \subset T_q(X) \to X$ is a diffeomorphism on $B_{\delta_p}(0)$;*
(b) *W_p is uniformly normal with respect to δ_p, that is, $W_p \subseteq \mathcal{E}_q(B_{\delta_p}(0))$ for every $q \in W_p$;*
(c) *The closure of W_p is contained in a strongly convex neighborhood U_p of p.*

Since we have an orthonormal basis $\{e_1^q, \cdots, e_d^q\}$ of $T_q(X)$ for each $q \in W_p$, according to (a) of Proposition 2, the map ϕ^q from $U = \{u \in \mathbb{R}^d : \|u\| < \delta_p\} \subset \mathbb{R}^d$ to X defined by $\phi^q(u^1, \cdots, u^d) = \mathcal{E}_q(\sum_{i=1}^d u^i e_i^q)$ gives a system of coordinates around q. We call such coordinates **q-normal coordinates**. Under these normal coordinates, $g_{ij}^q(u) := g_{ij}^{\phi^q}(u)$ is well defined for $q \in W_p$ and $u \in U$, and is C^∞ as a function on $W_p \times U$. It satisfies $g_{ij}^q(0) = \delta_{ij}$: according to the definition of $\frac{\partial}{\partial u^i}(q)$ by means of the local coordinates (U, ϕ^q) and $\phi^q(0, \cdots, 0, t, 0, \cdots, 0) = \mathcal{E}_q(te_i^q)$, for $f \in C^\infty(X)$ there holds $\big(\frac{\partial}{\partial u^i}(q)\big)(f) = \frac{d(f \circ \mathcal{E}_q(te_i^q))}{dt}\big|_{t=0}$, but $\mathcal{E}_q(te_i^q) = \gamma(1, q, te_i^q) = \gamma(t, q, e_i^q)$, so we have $\frac{d(f \circ \mathcal{E}_q(te_i^q))}{dt}\big|_{t=0} = \frac{d(f \circ \gamma(t, q, e_i^q))}{dt}\big|_{t=0} = e_i^q(f)$. Thus $\{\frac{\partial}{\partial u^i}(q) = e_i^q\}_{i=1}^d$ is an orthonormal basis on $T_q(X)$. Hence $g_{ij}^q(0) = g_q(\frac{\partial}{\partial u^i}(q), \frac{\partial}{\partial u^j}(q)) = g_q(e_i^q, e_j^q) = \delta_{ij}$.

For $u \in U$, let $v = \sum_{i=1}^d u^i e_i^q$. By (a) of Proposition 2, there exists a minimizing geodesic $\gamma(t, q, v)$, $0 \le t \le 1$, such that $\gamma(0, q, v) = q$, $\gamma'(0, q, v) = v$ and $\gamma(1, q, v) = \mathcal{E}_q(v)$. Hence $d_X(q, \mathcal{E}_q(v)) = \int_0^1 |\gamma'(t, q, v)| dt = \int_0^1 |\gamma'(0, q, v)| dt = \int_0^1 |v| dt = |v| = \|u\|$. That is,

$$d_X\left(q, \mathcal{E}_q\left(\sum_{i=1}^d u^i e_i^q\right)\right) = \|u\|, \qquad \forall \|u\| < \delta_p. \tag{14}$$

From the above discussion we have the following proposition.

Proposition 3 *For $p \in X$, choose W_p and δ_p as in Proposition 2. For each $q \in W_p$, choose q-normal coordinates (U, ϕ^q) and the corresponding local representation g_{ij}^q of the Riemannian structure as above. Then the following two bounds hold with a constant C_p independent of $q \in W_p$:*

$$\left|\sqrt{det(g_{ij}^q)}(u^1, \cdots, u^d) - 1\right| \le C_p \|u\|, \quad \forall \|u\| \le \delta_p/2, \tag{15}$$

$$\left|\big(d_X(q, x)\big)^2 - \|q - x\|^2\right| \le C_p \big(d_X(q, x)\big)^3, \quad \forall x \in \mathcal{E}_q(B_{\delta_p/2}(0)). \tag{16}$$

This result has been proved with $\delta_p/2$ replaced by δ_p as Proposition 2.2 in [14]. For completeness, we will give a proof in the appendix.

Now we can give the Riemannian volume measure. It is a standard measure on the Riemannian manifold which generalizes the Lebesgue measure of Euclidean spaces and has a clear geometric meaning: for any $U \subset X$, $\int_U dV = \mathrm{Vol}(U)$. Moreover [7],

if (U, ϕ) is a system of coordinates with $\phi : U \subset \mathbb{R}^d \mapsto X$, then for any measurable function f,

$$\int_{\phi(U)} f dV = \int_U f(\phi(u))\sqrt{det(g_{ij}^{\phi})}(u^1, \cdots, u^d)du^1 \cdots du^d. \tag{17}$$

3 Approximation by Gaussians

To prove our first main result, we need the following lemma.

Lemma 1 *Let X be a connected compact C^∞ submanifold of $\mathbb{R}^n$ which is isometrically embedded. Then there exists a positive constant $C_0 \geq 1$ such that*

$$d_X(x, y) \leq C_0\|x - y\|, \quad \forall x, y \in X. \tag{18}$$

Now we are in a position to prove Theorem 1.

Proof of Theorem 1 Let W_p, δ_p and C_p as in Proposition 3. Since $X \subseteq \cup_{p\in X} W_p$ and X is compact, there exists a finite subset $\mathcal{P}$ of X such that $X \subseteq \cup_{p\in\mathcal{P}} W_p$. Then $\|I_\sigma(f) - f\|_{C(X)} = \max_{p\in\mathcal{P}} \|I_\sigma(f) - f\|_{C(W_p)}$. Also, for $\sigma \geq \sigma_0$,

$$\|I_\sigma(f)\|_{C(X)} \leq \frac{1}{(\sqrt{2\pi}\sigma)^d}\int_X \|f\|_{C(X)}dV = \frac{\mathrm{Vol(X)}}{(\sqrt{2\pi}\sigma)^d}\|f\|_{C(X)} \leq \frac{\mathrm{Vol(X)}}{(\sqrt{2\pi}\sigma_0)^d}\|f\|_{C(X)}.$$

So (3) is verified with $C_X = \max\left\{\left(\frac{\mathrm{Vol(X)}}{(\sqrt{2\pi}\sigma_0)^d} + 1\right)\sigma_0^{-s}, \max_{p\in\mathcal{P}} C_{X,p}\right\}$ if we can prove for some $\sigma_0 > 0$ that

$$\|I_\sigma(f) - f\|_{C(W_p)} \leq C_{X,p}\|f\|_{Lip(s)}\sigma^s, \quad \forall 0 < \sigma < \sigma_0, \quad p \in \mathcal{P}. \tag{19}$$

Take $\delta^* = \min_{p\in\mathcal{P}}\left\{\min\{\frac{\delta_p}{2}, \frac{1}{2C_p}\}\right\} > 0$ and C_0 as in (18). Take $0 < \sigma_0 < \frac{1}{2}$ such that $C_0\sqrt{2d + 2}\sigma_0\sqrt{\log\sigma_0^{-1}} < \delta^*$. We prove (19) in three steps. Let $p \in \mathcal{P}$ and $0 < \sigma < \sigma_0$.

Step 1: Decomposition. Let $q \in W_p$. Choose

$$B_\sigma^q := \left\{x \in X : d_X(q, x) < C_0\sqrt{2d + 2}\sigma\sqrt{\log\frac{1}{\sigma}}\right\}.$$

Since $\mathcal{E}_q$ is a diffeomorphism on $B_{\delta^*}(0)$, we see from (14) that $B_\sigma^q \subset \mathcal{E}_q(B_{\delta^*}(0))$ and $B_\sigma^q = \{\mathcal{E}_q(\sum_{i=1}^d u^i e_i^q) : u \in \widetilde{B}_\sigma\}$ where

$$\widetilde{B}_\sigma := \left\{u \in \mathbb{R}^d : \|u\| < C_0\sqrt{2d + 2}\sigma\sqrt{\log\frac{1}{\sigma}}\right\}. \tag{20}$$

Denote $\phi^q(u) = \mathcal{E}_q(\sum_{i=1}^d u^i e_i^q)$ for $u = (u^1, \ldots, u^d) \in \mathbb{R}^d$. Then $B_\sigma^q = \phi^q(\widetilde{B}_\sigma^q)$. Separating the domain X into two parts, we have

$$I_\sigma(f)(q) = \frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{B_\sigma^q} K_\sigma(q, y)f(y)dV(y) + \frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{X\setminus B_\sigma^q} K_\sigma(q, y)f(y)dV(y).$$

By (17), the first term on the above right hand side equals

$$\frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{\widetilde{B}_\sigma}\exp\left\{-\frac{\|q-\phi^q(u)\|}{2\sigma^2}\right\}f(\phi^q(u))\sqrt{\det(g_{ij}^q)}(u)du.$$

Using (8), we can decompose $f(q)$ as

$$\begin{aligned}f(q)=&\frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{\widetilde{B}_\sigma}\exp\left\{-\frac{\|u\|^2}{2\sigma^2}\right\}f(q)du\\&+\frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{\mathbb{R}^d\setminus\widetilde{B}_\sigma}\exp\left\{-\frac{\|u\|^2}{2\sigma^2}\right\}f(q)du.\end{aligned}$$

Combining the above decomposition, we have

$$I_\sigma(f)(q)-f(q)=J_1+J_2 \tag{21}$$

where

$$\begin{aligned}J_1=&\frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{\widetilde{B}_\sigma}\left\{\exp\left\{-\frac{\|q-\phi^q(u)\|^2}{2\sigma^2}\right\}f(\phi^q(u))\sqrt{\det(g_{ij}^q)}(u)\right.\\&\left.-\exp\left\{-\frac{\|u\|^2}{2\sigma^2}\right\}f(q)\right\}du,\\J_2=&\frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{X\setminus B_\sigma^q}K_\sigma(q,y)f(y)dV(y)\\&-\frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{\mathbb{R}^d\setminus\widetilde{B}_\sigma}\exp\left\{-\frac{\|u\|^2}{2\sigma^2}\right\}f(q)du.\end{aligned}$$

Step 2: *Estimation of J_1 for the error in a neighborhood*. We separate the error J_1 further as

$$\begin{aligned}J_1=&\frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{\widetilde{B}_\sigma}\exp\left\{-\frac{\|u\|^2}{2\sigma^2}\right\}\left[f(\phi^q(u))-f(q)\right]du\\&+\frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{\widetilde{B}_\sigma}\left[\exp\left\{-\frac{\|q-\phi^q(u)\|^2}{2\sigma^2}\right\}\right.\\&\left.-\exp\left\{-\frac{\|u\|^2}{2\sigma^2}\right\}\right]f(\phi^q(u))du\\&+\frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{\widetilde{B}_\sigma}\exp\left\{-\frac{\|q-\phi^q(u)\|^2}{2\sigma^2}\right\}f(\phi^q(u))\left(\sqrt{\det g_{ij}^q}(u)-1\right)du\\:=&J_1'+J_1''+J_1'''.\end{aligned}$$

Since $f\in Lip(s)$, We know that $|f(\phi^q(u))-f(q)|\le|f|_{Lip(s)}\,(d_X(\phi^q(u),q))^s$. By (14), for $u\in\widetilde{B}_\sigma$, $d_X(q,\phi^q(u))=\|u\|<\delta^*\le\frac{1}{2C_p}$. So by (16),

$$\left|\|\phi^q(u)-q\|^2-\|u\|^2\right|\le\frac{1}{2}\|u\|^2,\qquad\forall u\in\widetilde{B}_\sigma^q. \tag{22}$$

It follows that $\|\phi^q(u) - q\| \le 2\|u\|$. Thus

$$|J_1'| \le \frac{|f|_{Lip(s)}}{(\sqrt{2\pi}\sigma)^d} \int_{\widetilde{B}_\sigma} \exp\left\{ - \frac{\|u\|^2}{2\sigma^2} \right\} 2^s \|u\|^s du.$$

By a change of variables $\frac{u}{\sigma}$ and (9), we have

$$|J_1'| \le (2\pi)^{-d/2} |f|_{Lip(s)} 2^s \sigma^s \int_{\mathbb{R}^d} \exp\left\{ - \frac{\|u\|^2}{2} \right\} \|u\|^s du = \frac{2^{\frac{3s}{2}} \Gamma(\frac{s+d}{2})}{\Gamma(\frac{d}{2})} |f|_{Lip(s)} \sigma^s.$$

Consider the term J_1''. Applying (14) and (16) of Proposition 3, we know that for $u \in \widetilde{B}_\sigma$, $\phi^q(u) \in \mathcal{E}_q(B_{\delta_p/2}(0))$ and

$$\begin{aligned} \left| d_X^2(q, \phi^q(u)) - \|q - \phi^q(u)\|^2 \right| &= \left| \|u\|^2 - \|q - \phi^q(u)\|^2 \right| \\ &\le C_p d_X^3(q, \phi^q(u)) = C_p \|u\|^3. \end{aligned}$$

It follows from (16) and the elementary inequality $|e^{-a} - e^{-b}| \le |a - b| \max\{e^{-a}, e^{-b}\}$ (valid for any $a, b > 0$) that

$$\begin{aligned} |J_1''| &\le \frac{\|f\|_{C(X)}}{(\sqrt{2\pi}\sigma)^d} \int_{\widetilde{B}_\sigma} \left| \exp\left\{ - \frac{\|q - \phi^q(u)\|^2}{2\sigma^2} \right\} - \exp\left\{ - \frac{\|u\|^2}{2\sigma^2} \right\} \right| du \\ &\le \frac{\|f\|_{C(X)}}{(\sqrt{2\pi}\sigma)^d} \int_{\widetilde{B}_\sigma} \max\left\{ \exp\left\{ - \frac{\|q - \phi^q(u)\|^2}{2\sigma^2} \right\}, \exp\left\{ - \frac{\|u\|^2}{2\sigma^2} \right\} \right\} \frac{C_p \|u\|^3}{2\sigma^2} du. \end{aligned}$$

This in connection with (22) implies

$$\begin{aligned} |J_1''| &\le \frac{\|f\|_{C(X)}}{(\sqrt{2\pi}\sigma)^d} \int_{\widetilde{B}_\sigma} \exp\left\{ - \frac{\|u\|^2}{4\sigma^2} \right\} \frac{C_p \|u\|^3}{2\sigma^2} du \\ &\le \frac{C_p}{2} (2\pi)^{-d/2} \|f\|_{C(X)} \sigma \int_{\mathbb{R}^d} \exp\left\{ - \frac{\|u\|^2}{4} \right\} \|u\|^3 du \\ &= \frac{2^{\frac{d}{2}+2} C_p \Gamma(\frac{d+3}{2})}{\Gamma(\frac{d}{2})} \|f\|_{C(X)} \sigma. \end{aligned}$$

As for J_1''', we use (15) and (22) and obtain

$$\begin{aligned} |J_1'''| &\le \frac{C_p \|f\|_{C(X)}}{(\sqrt{2\pi}\sigma)^d} \int_{\widetilde{B}_\sigma} \exp\left\{ - \frac{\|u\|^2}{4\sigma^2} \right\} \|u\| du \\ &\le C_p (2\pi)^{-d/2} \|f\|_{C(X)} \sigma \int_{\mathbb{R}^d} \exp\left\{ - \frac{\|u\|^2}{4} \right\} \|u\| du \le \frac{2^{\frac{d}{2}+2} \Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})} C_p \|f\|_{C(X)} \sigma. \end{aligned}$$

Combining the above estimates for J_1', J_1'', J_1''', we have

$$|J_1| \le \frac{\|f\|_{Lip(s)}}{\Gamma(\frac{d}{2})} \left\{ 4\Gamma\left(\frac{s+d}{2}\right) + 2^{\frac{d}{2}+2} C_p \Gamma\left(\frac{d+3}{2}\right) + 2^{\frac{d}{2}+2} C_p \Gamma\left(\frac{d+1}{2}\right) \right\} \sigma^s. \tag{23}$$

Step 3: *Estimation of J_2 away from the point.* Bounding the first term of J_2 is easy: we use (18). For $y \in X\setminus B_\sigma^q$, $d_X(q,y) \ge C_0\sqrt{2d+2}\sigma\sqrt{\log\frac{1}{\sigma}}$ implies $\|q-y\| \ge \sqrt{2d+2}\sigma\sqrt{\log\frac{1}{\sigma}}$. It follows that

$$\begin{aligned}|J_2'| &:= \left|\frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{X\setminus B_\sigma^q} K_\sigma(q,y)\,f(y)dV(y)\right| \\ &\le \frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{X\setminus B_\sigma^q}\exp\left\{-\frac{(2d+2)\sigma^2\log\frac{1}{\sigma}}{2\sigma^2}\right\}|f(y)|dV(y) \\ &\le (2\pi)^{-d/2}\mathrm{Vol}(X)\|f\|_{C(X)}\sigma. \end{aligned} \tag{24}$$

Now we bound the second term of J_2. Using (9) again, we have

$$\begin{aligned}|J_2''| &:= \left|\frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{\mathbb{R}^d\setminus\widetilde{B}_\sigma}\exp\left\{-\frac{\|u\|^2}{2\sigma^2}\right\}f(q)du\right| \\ &\le \frac{\|f\|_{C(X)}}{(\sqrt{2\pi}\sigma)^d}\int_{\|u\|\ge C_0\sqrt{2d+2}\sigma(\log\sigma^{-1})^{1/2}}\exp\left\{-\frac{\|u\|^2}{2\sigma^2}\right\}du \\ &= \frac{2^{1-\frac{d}{2}}}{\Gamma(\frac{d}{2})}\|f\|_{C(X)}\int_{r\ge C_0\sqrt{2d+2}(\log\sigma^{-1})^{1/2}}\exp\left\{-\frac{r^2}{2}\right\}r^{d-1}dr \\ &\le \frac{2^{1-\frac{d}{2}}}{\Gamma(\frac{d}{2})}\|f\|_{C(X)}\int_{r\ge C_0\sqrt{2d+2}(\log\sigma^{-1})^{1/2}}\exp\left\{-\frac{C_0^2(2d+2)(\log\sigma^{-1})}{4}\right\} \\ &\quad\times\exp\left\{-\frac{r^2}{4}\right\}r^{d-1}dr \\ &\le \frac{2^{1-\frac{d}{2}}}{\Gamma(\frac{d}{2})}\|f\|_{C(X)}\int_0^\infty\sigma^{\frac{d+1}{2}C_0^2}\exp\left\{-\frac{r^2}{4}\right\}r^{d-1}dr = 2^{\frac{d}{2}}\|f\|_{C(X)}\sigma^{\frac{d+1}{2}C_0^2}. \end{aligned}$$

But $C_0 \ge 1$, so there holds

$$|J_2''| \le 2^{\frac{d}{2}}\|f\|_{C(X)}\sigma.$$

Combining this with (24), we get

$$|J_2| \le \left\{(2\pi)^{-\frac{d}{2}}\mathrm{Vol}(X) + 2^{\frac{d}{2}}\right\}\|f\|_{Lip(s)}\sigma.$$

This together with (23) yields the desired result. □

4 Learnability of Gaussians with flexible variances

The generalization error associated with the probability measure ρ on Z is defined for $f: X \to Y$ as

$$\mathcal{E}(f) = \int_Z (f(x)-y)^2 d\rho.$$

A special feature of the generalization error with the least square loss function is $\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{L^2_{\rho_X}}$. It tells us that the regression function (4) minimizes the generalization error.

The efficiency of the learning algorithm (6) for most purposes is measured by the excess generalization error $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)$ which can be decomposed into the sample error and the regularization error, see [22, 25] (and [8, 23] for more general schemes). The sample error has been well understood. As to the regularization error, when the input space is a domain of $\mathbb{R}^n$ (with nonempty interior) and ρ_X is the Lebesgue measure, it has a polynomial decay for $f_\rho \in \mathrm{Lip}(s)$. But the constraint to X (and ρ_X) is too strong, which excludes the manifold setting. Now by applying Theorem 1, we can give a satisfactory decay rate for the regularization error in the manifold setting.

Let

$$f_\lambda := \arg \min_{\sigma\in(0,+\infty)} \min_{f\in\mathcal{H}_{K_\sigma}} \left\{\mathcal{E}(f) + \lambda\|f\|^2_{\mathcal{H}_{K_\sigma}}\right\}. \tag{25}$$

As in [22, 25], we can bound the excess generalization error $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)$ as

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \le \left\{\left\{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}^{\mathbf{z}}(f_{\mathbf{z},\lambda})\right\} + \left\{\mathcal{E}^{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\right\}\right\} + \mathcal{D}(\lambda) \tag{26}$$

where $\mathcal{E}^{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^m \left(f(x_i) - y_i\right)^2$ and $\mathcal{D}(\lambda)$ is the regularization error of the scheme (6).

Definition 5 The regularization error of the scheme (6) is defined as

$$\mathcal{D}(\lambda) = \inf_{\sigma\in(0,\infty)} \inf_{f\in\mathcal{H}_\sigma} \left\{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda\|f\|^2_{K_\sigma}\right\}, \quad \lambda > 0.$$

We need to estimate $\mathcal{D}(\lambda)$ for our error analysis. Our decay rate will only be related to the manifold dimension d. Hence when d is much smaller than n, our estimate is much better than those in the domain setting given in [17, 25].

Theorem 3 *Let X be a connected compact C^∞ submanifold of $\mathbb{R}^n$ which is isometrically embedded and of dimension d. If $f_\rho \in Lip(s)$ for some $0 < s \le 1$, then*

$$\mathcal{D}(\lambda) \le \left\{C_X^2 + (2\pi)^{-d}(Vol(X))^2\right\}\|f_\rho\|^2_{Lip(s)}\lambda^{\frac{s}{s+d}} \qquad \forall\lambda > 0. \tag{27}$$

Proof For $\sigma \in (0, +\infty)$, we take functions $f_{\rho,\sigma} = I_\sigma(f_\rho)$. Since $f_\rho \in Lip(s)$, we know from Theorem 1 that

$$\|f_{\rho,\sigma} - f_\rho\|_{C(X)} \le C_X\|f_\rho\|_{Lip(s)}\sigma^s.$$

Thus

$$\mathcal{E}(f_{\rho,\sigma}) - \mathcal{E}(f_\rho) = \|f_{\rho,\sigma} - f_\rho\|^2_{L^2_{\rho_X}} \le C_X^2\|f_\rho\|^2_{Lip(s)}\sigma^{2s}.$$

By the definition of $f_{\rho,\sigma}$ and the equation

$$\langle K_\sigma(\cdot, y), K_\sigma(\cdot, z)\rangle_{K_\sigma} = K_\sigma(y, z),$$

we have

$$\|f_{\rho,\sigma}\|_{K_\sigma}^2 = \frac{1}{(\sqrt{2\pi}\sigma)^{2d}} \int_X \int_X K_\sigma(y,z) f_\rho(y) f_\rho(z) dV(y) dV(z)$$
$$\le \frac{1}{(\sqrt{2\pi}\sigma)^{2d}} \big(\mathrm{Vol}(X)\big)^2 \|f_\rho\|_{C(X)}^2. \tag{28}$$

These yield

$$\mathcal{D}(\lambda) \le \inf_{\sigma\in(0,\infty)} \left\{ \|f_{\rho,\sigma} - f_\rho\|_{L^2_{\rho_X}}^2 + \lambda \|f_{\rho,\sigma}\|_{K_\sigma}^2 \right\}$$
$$\le \inf_{\sigma\in(0,\infty)} \left\{ C_X^2 \|f_\rho\|_{Lip(s)}^2 \sigma^{2s} + \lambda (2\pi)^{-d} \big(\mathrm{Vol(X)}\big)^2 \|f_\rho\|_{C(X)}^2 \sigma^{-2d} \right\}.$$

Taking $\sigma = \lambda^{\frac{1}{2s+2d}}$, we have $\mathcal{D}(\lambda) \le \big\{C_X^2 \|f_\rho\|_{Lip(s)}^2 + (2\pi)^{-d}\big(\mathrm{Vol}(X)\big)^2 \|f_\rho\|_{C(X)}^2\big\}\lambda^{\frac{s}{s+d}}$. This proves Theorem 3. □

To get some error estimates for the algorithm (6), we use a result from [25] which is stated as follows.

Proposition 4 *Let $f_{\mathbf{z},\lambda}$ be given by (6). If $0 < \lambda \le 1$ and for some $M > 0$, $\rho(\cdot|x)$ is supported on $[-M, M]$ for almost every $x \in X$, then there exists a constant $\widetilde{C}$ independent of m or λ such that*

$$\mathbb{E}\big(\|f_{\mathbf{z},\lambda} - f_\rho\|_{L^2_{\rho_X}}^2\big) \le \widetilde{C}\left(\frac{\log^2 m}{m\lambda^4}\right)^{1/4} + \mathcal{D}(\lambda). \tag{29}$$

Choose $\lambda = \big(\frac{\log^2 m}{m}\big)^{\frac{s+d}{8s+4d}}$, we get from Theorem 3 and Proposition 4 that

$$\mathbb{E}\big(\|f_{\mathbf{z},\lambda} - f_\rho\|_{L^2_{\rho_X}}^2\big) \le \left\{\widetilde{C} + \big[C_X^2 + (2\pi)^{-d}\big(\mathrm{Vol}(X)\big)^2\big]\|f_\rho\|_{Lip(s)}^2\right\}\left(\frac{\log^2 m}{m}\right)^{\frac{s}{8s+4d}}. \tag{30}$$

This proves Theorem 2.

5 Approximation ability of a single Gaussian

In this section, we show that the approximation ability of a single Gaussian kernel is rather weak. This will be proved for a general C^∞ Mercer kernel K.

When X is a domain of $\mathbb{R}^n$ with non-empty interior and ρ_X dominates the Lebesgue measure in the sense $d\rho_X(x) \ge c_\rho dx$ for some $c_\rho > 0$, the following result was proved in [9, 19]. The current theorem is stated in a manifold setting. It shows that to get a polynomial decay of the regularization error with one Gaussian kernel, the regression function must belong to $C^\infty(X)$. This in connection with Theorem 3 demonstrates that for many applications the learning algorithm using Gaussian kernels with flexible variances has advantages over that with a single Gaussian.

Theorem 4 *Let X be a connected compact C^∞ submanifold of $\mathbb{R}^n$ which is isometrically embedded and of dimension d. Assume the marginal distribution ρ_X satisfies $d\rho_X \geq c_\rho dV$ for some positive constant c_ρ and K is a C^∞ Mercer kernel. If*

$$\mathcal{D}(\lambda) = \inf_{f\in\mathcal{H}_K} \left\{\|f - f_\rho\|^2_{L^2_{\rho_X}} + \lambda\|f\|^2_K\right\} = O(\lambda^\beta), \tag{31}$$

for some $0 < \beta \leq 1$, then $f_\rho \in C^\infty(X)$.

Proof Let $p \in X$. By Proposition 3, there exists a $\delta'_p > 0$ such that $\sqrt{\det(g^p_{ij})}(u) > \frac{1}{2}$ for all $u \in B = \{x \in \mathbb{R}^d : \|u\| \leq \delta'_p\}$ and $\phi^p(B)$ is a neighborhood of p in X. By the assumption $d\rho_X \geq c_\rho dV$, for any $f \in \mathcal{H}_K$ we have

$$\|f - f_\rho\|^2_{L^2_{\rho_X}} \geq \int_{\phi^p(B)} \big(f(x) - f_\rho(x)\big)^2 d\rho_X(x) \geq c_\rho \int_{\phi^p(B)} \big(f(x) - f_\rho(x)\big)^2 dV(x).$$

Then by using the formula (17), we obtain

$$\|f - f_\rho\|^2_{L^2_{\rho_X}} \geq c_\rho \int_B \big(f(\phi^p(u)) - f_\rho(\phi^p(u))\big)^2 \sqrt{det(g^\phi_{ij})}(u)du \geq \frac{c_\rho}{2}\|f\circ\phi^p - f_\rho\circ\phi^p\|^2_{L^2(B)}.$$

This in connection with (31) implies that

$$\inf_{f\in\mathcal{H}_K} \left\{\|f\circ\phi^p - f_\rho\circ\phi^p\|^2_{L^2(B)} + \lambda\|f\|^2_K\right\} = O(\lambda^\beta). \tag{32}$$

Now we restrict K onto $\phi^p(B)$ and set $\widetilde{K} = K|_{\phi^p(B)\times\phi^p(B)}$. We know from [1] that $\widetilde{K}$ is a Mercer kernel on $\phi^p(B) \subset X$ satisfying

$$\|g\|_{\widetilde{K}} = \inf\{\|f\|_K : f \in \mathcal{H}_K,\ f|_{\phi^p(B)} = g\} \qquad \forall g \in \mathcal{H}_{\widetilde{K}}.$$

In particular, $\|f|_{\phi^p(B)}\|_{\widetilde{K}} \leq \|f\|_K$ for any $f \in \mathcal{H}_K$.

Define a Mercer kernel $\widehat{K}$ on B by $\widehat{K}(u, v) = \widetilde{K}(\phi^p(u), \phi^p(v))$. Then

$$\mathcal{H}_{\widehat{K}} = \{g\circ\phi^p : g \in \mathcal{H}_{\widetilde{K}}\} = \{f\circ\phi^p : f \in \mathcal{H}_K\} \quad \text{and} \quad \|g\|_{\widetilde{K}} = \|g\circ\phi^p\|_{\widehat{K}} \quad \forall g \in \mathcal{H}_{\widetilde{K}}.$$

Hence $\|f\circ\phi^p\|_{\widehat{K}} = \|f|_{\phi^p(B)}\|_{\widetilde{K}} \leq \|f\|_K$ for any $f \in \mathcal{H}_K$. Combining this with (32), we find that

$$\inf_{h\in\mathcal{H}_{\widehat{K}}} \left\{\|h - f_\rho\circ\phi^p\|^2_{L^2(B)} + \lambda\|h\|^2_{\widehat{K}}\right\} \leq \inf_{f\in\mathcal{H}_K} \left\{\|f\circ\phi^p - f_\rho\circ\phi^p\|^2_{L^2(B)} + \lambda\|f\|^2_K\right\} = O(\lambda^\beta).$$

By Theorem 6.2 in [9], this implies that $f_\rho\circ\phi^p \in C^\infty(B)$. Since (ϕ^p, B) is a system of local coordinates around an arbitrary point $p \in X$, we conclude that $f_\rho \in C^\infty(X)$. □

Acknowledgements The authors would like to thank the referees for valuable comments and suggestions. In particular, a simplified proof of Theorem 4 appearing in the revised version was provided by one referee.

Appendix

In this appendix, we prove Lemma 1, Proposition 2 and Proposition 3.

Lemma 1 might exist in the literature which is not available to the authors. So we give a complete proof here.

Proof of Lemma 1 Suppose to the contrary that there is a sequence of pairs of distinct points $\{(x_k, y_k)\}_{k=1}^{\infty}$ such that

$$d_X(x_k, y_k) > k\|x_k - y_k\|, \qquad \forall k \in \mathbb{N}.$$

Then

$$\|x_k - y_k\| < \frac{1}{k} d_X(x_k, y_k) \le \frac{1}{k}\mathrm{diam}(X),$$

where $\mathrm{diam}(X) := \sup_{x\neq y\in X} d_X(x, y) < \infty$. Since X is compact, the sequences $\{x_k\}_{k=1}^{\infty}$ and $\{y_k\}_{k=1}^{\infty}$ have convergent subsequences $\{x_{k_j}\}_{j=1}^{\infty}$,$\{y_{k_j}\}_{j=1}^{\infty}$ converging in (X, d_X), hence in $\mathbb{R}^n$, to p and p^*. But $\|x_{k_j} - y_{k_j}\| \le \frac{\mathrm{diam}(X)}{k_j} \to 0$, we must have $p = p^*$.

Now let us derive a contradiction. Take W_p, δ_p and C_p as in Proposition 3. Find some $2 \le J \in \mathbb{N}$ such that $y_{k_j} \in W_p$ and $d_X(x_{k_j}, y_{k_j}) < \delta_p/2$ for all $j \ge J$. Take $q = y_{k_j}$. Then $x_{k_j} \in B_q(B_{\delta_p/2}(0))$. For any $j \ge J$, $\|x_{k_j} - y_{k_j}\| \le \frac{1}{2} d_X(x_{k_j}, y_{k_j})$. Putting his into (16), we see that

$$\frac{3}{4}(d_X(x_{k_j}, y_{k_j}))^2 \le C_p(d_X(x_{k_j}, y_{k_j}))^3.$$

and hence

$$d_X(x_{k_j}, y_{k_j}) \ge \frac{3}{4C_p}$$

which is a contradiction since $d_X(x_{k_j}, y_{k_j}) \to 0$ as $j \to \infty$. This proves Lemma 1. □

To illustrate some knowledge on manifolds, we prove Proposition 2 and Proposition 3 here for completeness.

Proof of Proposition 2 The existence of a strongly convex neighborhood U_p is proved in Chapter 3 of [7] (as Proposition 4.2). By Theorem 3.7 there, there are another neighborhood $\widetilde{W}_p$ and a number $\delta_p > 0$ such that (a) and (b) hold for $\widetilde{W}_p$.

Since $\mathcal{E}_p$ is a diffeomorphism of $B_{\delta_p}(0)$ onto an open subset $\mathcal{E}_p(B_{\delta_p}(0))$ of X, we can find some $0 < \delta^* < \delta_p$ such that the neighborhood $\mathcal{E}_p(B_{\delta^*}(0))$ of p is contained in the open set $U_p \cap \widetilde{W}_p$. Take $W_p = \mathcal{E}_p\left(B_{\delta^*/2}(0)\right)$. It is a neighborhood of p and, as a subset of $\widetilde{W}_p$, satisfies (a) and (b). Moreover, its closure equals $\mathcal{E}_p\left(\overline{B_{\delta^*/2}(0)}\right) \subset \mathcal{E}_p\left(B_{\delta^*}(0)\right) \subseteq U_p$. Hence (c) also holds. This proves Proposition 2. □

Proof of Proposition 3 Recall that $U = \{u \in \mathbb{R}^d : \|u\| < \delta_p\}$ and the functions $g_{ij}^q(u) = g_{ij}^{\phi^q}(u)$ are well defined and C^∞ on $W_p \times U$ satisfying $g_{ij}^q(0) = \delta_{ij}$. Then the function $h(q, u) := det(g_{ij}^q)(u)$ is nonnegative, C^∞ on $W_p \times U$ and satisfies $h(q, 0) = 1$ for each $q \in W_p$. Now both $\overline{W}_p$ and $\widetilde{B} := \{u \in \mathbb{R}^d : \|u\| \le \delta_p/2\} \subset U$ are compact sets. So for every $q \in \overline{W}_p$ and every $u \in \widetilde{B}$ there holds

$$|h(q, u) - 1| = |h(q, u) - h(q, 0)| \le \left\{\sum_{i=1}^{d} \left\| \frac{\partial h}{\partial u^i} \right\|^2_{C(\overline{W}_p \times \widetilde{B})}\right\}^{1/2} \|u\|$$

which implies

$$\left|\sqrt{h(q,u)}-1\right| = \frac{|h(q,u)-1|}{\sqrt{h(q,u)}+1} \le |h(q,u)-1| \le C_p'\|u\|,$$

where $C_p' = \left\{\sum_{i=1}^d \left\|\frac{\partial h}{\partial u^i}\right\|^2_{C(\overline{W}_p\times\widetilde{B})}\right\}^{1/2}$ is a constant independent of $q \in W_p$. This proves the inequality (15).

As to the second inequality (16), we do the same and define a vector-valued function $F : W_p \times U \to \mathbb{R}^n$ as

$$F(q,u) = (F_1(q,u),\cdots,F_n(q,u)) := \Phi \circ \mathcal{E}_q\left(\sum_{i=1}^d u^i e_i^q\right).$$

This function is C^∞ on $W_p \times U$. Now for $q \in W_p$ and $u \in U$, denote $x = \mathcal{E}_q(\sum_{i=1}^d u^i e_i^q)$. Then $\|q-x\|^2 = \|F(q,0)-F(q,u)\|^2 = \sum_{\alpha=1}^n (F_\alpha(q,u)-F_\alpha(q,0))^2$. So using (14), we have

$$d_X^2(q,x) - \|x-q\|^2 = \|u\|^2 - \sum_{\alpha=1}^n \left(\sum_{i=1}^d \frac{\partial F_\alpha}{\partial u^i}(q,0)u^i + R_\alpha(q,u)\right)^2 \tag{33}$$

where $R_\alpha(q,u)$ is a remainder term in the Taylor expansion which can be bounded as

$$|R_\alpha(q,u)| \le \left\{\sum_{i,j=1}^n \left\|\frac{\partial^2 F_\alpha}{\partial u^i \partial u^j}(q,u)\right\|^2_{C(\overline{W}_p\times\widetilde{B})}\right\}^{1/2}\|u\|^2, \qquad \forall q \in \overline{W}_p, u \in \widetilde{B}. \tag{34}$$

To analyze (33), we need to find $\frac{\partial F_\alpha}{\partial u^i}(q,0)$. Towards this end, for $i = 1,\ldots,d$, choose the curve $\gamma(t) = \mathcal{E}_q(e_i^q t)$. Using (12), we have

$$d\Phi_q\left(\frac{\partial}{\partial u^i}(q)\right) = \left.\frac{d\Phi\circ\gamma(t)}{dt}\right|_{t=0} = \left.\frac{d\Phi\circ\mathcal{E}_q(e_i^q t)}{dt}\right|_{t=0} = \left(\frac{\partial F_1}{\partial u^i}(q,0),\ldots,\frac{\partial F_n}{\partial u^i}(q,0)\right).$$

Hence

$$g_{ij}^q(0) = \left\langle d\Phi_q\left(\frac{\partial}{\partial u^i}(q)\right), d\Phi_q\left(\frac{\partial}{\partial u^j}(q)\right)\right\rangle_{\mathbb{R}^n} = \sum_{\alpha=1}^n \frac{\partial F_\alpha}{\partial u^i}(q,0)\frac{\partial F_\alpha}{\partial u^j}(q,0).$$

But $g_{ij}^q(0) = \delta_{i,j}$, we obtain

$$\sum_{i,j=1}^d \sum_{\alpha=1}^n \frac{\partial F_\alpha}{\partial u^i}(q,0)\frac{\partial F_\alpha}{\partial u^i}(q,0)u^i u^j = \|u\|^2$$

and hence

$$\sum_{\alpha=1}^n \left(\sum_{i=1}^d \frac{\partial F_\alpha}{\partial u^i}(q,0)u^i + R_\alpha(q,u)\right)^2$$

$$= \|u\|^2 + 2\sum_{\alpha=1}^n\sum_{i=1}^d \frac{\partial F_\alpha}{\partial u^i}(q,0)u^i R_\alpha(q,u) + \sum_{\alpha=1}^n (R_\alpha(q,u))^2$$

This in connection with (33) tells us that or $q \in \overline{W}_p$ and $u \in \widetilde{B}$ there holds

$$\left|d_X^2(q,x) - \|x-q\|^2\right| \le \left|2\sum_{\alpha=1}^{n}\sum_{i=1}^{d}\frac{\partial F_\alpha}{\partial u^i}(q,0)u^i R_\alpha(q,u) + \sum_{\alpha=1}^{n}(R_\alpha(q,u))^2\right|$$

$$\le 2\sum_{\alpha=1}^{n}\left\{\sum_{i=1}^{d}\left\|\frac{\partial F_\alpha}{\partial u^i}(q,0)\right\|^2_{C(\overline{W}_p)}\right\}^{1/2}\|u\|\,|R_\alpha(q,u)| + \sum_{\alpha=1}^{n}(R_\alpha(q,u))^2 .$$

Together with (34) this verifies the inequality (16) with the constant

$$2\left\{\sum_{\alpha=1}^{n}\sum_{i=1}^{d}\left\|\frac{\partial F_\alpha}{\partial u^i}(q,0)\right\|^2_{C(\overline{W}_p)}\right\}^{1/2}\left\{\sum_{\alpha=1}^{n}\sum_{i,j=1}^{n}\left\|\frac{\partial^2 F_\alpha}{\partial u^i \partial u^j}(q,u)\right\|^2_{C(\overline{W}_p\times\widetilde{B})}\right\}^{1/2}$$

$$+\sum_{\alpha=1}^{n}\sum_{i,j=1}^{n}\left\|\frac{\partial^2 F_\alpha}{\partial u^i \partial u^j}(q,u)\right\|^2_{C(\overline{W}_p\times\widetilde{B})}\frac{\delta_p}{2}$$

independent of $q \in W_p \subset \overline{W}_p$. This proves Proposition 3. □

References

1. Aronszajn, N.: Theory of reproducing kernels. Trans. Amer. Math. Soc. **68**, 337–404 (1950)
2. Belkin, M., Niyogi, P.: Towards a theoretical foundation for Laplacian-based manifold methods. In: Auer, P., Meir, R. (eds.) COLT 2005, pp. 486–500 (2005)
3. Belkin, M., Niyogi, P.: Semi-supervised learning on Riemannian manifolds. Mach. Learn. **56**, 209–239 (2004)
4. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **15**, 1373–1396 (2003)
5. Boothby, W.M.: An Introduction to Differentiable Manifolds and Riemannian Geometry. Academic, New York (1986)
6. Bousquet, O., Chapelle, O., Hein, M.: Measure based regularization. In: NIPS (2003)
7. do Carmo, M.: Riemannian Geometry. Birkhäuser, Boston (1992)
8. Chen, D.R., Wu, Q., Ying, Y., Zhou, D.X.: Support vector machine soft margin classifiers: error analysis. J. Mach. Learn. Res. **5**, 1143–1175 (2004)
9. Cucker, F., Zhou, D.X.: Learning Theory: An Approximation Theory Viewpoint. Cambridge University Press, Cambridge (2007)
10. De Vito, E., Caponnetto, A., Rosasco, L.: Model selection for regularized least-squares algorithm in learning theory. Found. Comput. Math. **5**, 59–85 (2005)
11. Ditzian, Z., Totik, V.: Moduli of Smoothness. Springer, New York (1987)
12. Donoho, D., Grimes, C.: Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. Proc. Natl. Acad. Sci. USA **100**, 5591–5596 (2003)
13. Evgeniou, T., Pontil, M., Poggio, T.: Regularization networks and suport vector machines. Adv. Comput. Math. **13**, 1–50 (2000)
14. Gine, E., Koltchinskii, V.: Empirical graph Laplacian approximation of Laplace–Beltrami operators: large sample results. IMS Lecture Notes-monograph Series High Dimensional Probability **51**, 238–259 (2006)
15. Hardin, D., Tsamardinos, I., Aliferis, C.F.: A theoretical characterization of linear SVM-based feature selection. Proc. of the 21st Int. Conf. on Machine Learning, Banff, Canada (2004)
16. Lee, J.M.: Riemannian Manifolds, Springer, New York (1997)
17. Micchelli, C.A., Pontil, M., Wu, Q. Zhou, D.X.: Error bounds for learning the kernel. In: Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL), Research note no. 05/09, pp. 1–14 (2005)
18. Mukherjee, S., Wu, Q., Zhou, D.X.: Learning gradients and feature selection on manifolds. Preprint (2007)

19. Smale, S., Zhou, D.X.: Estimating the approximation error in learning theory. Anal. Appl. **1**, 17–41 (2003)
20. Smale, S., Zhou, D.X.: Learning theory estimates via integral operators and their applications. Constr. Approx. (2007) doi:10.1007/s00365-006-0659-y
21. von Luxburg, U., Belkin, M., Bousquet, O.: Consistency of spectral clustering. Ann. Statist. (2007) (in press)
22. Wu, Q., Ying, Y., Zhou, D.X.: Multi-kernel regularized classifiers. J. Complexity **23**, 108–134 (2007)
23. Wu, Q., Zhou, D.X.: SVM soft margin classifiers: linear programming versus quadratic programming. Neural Comput. **17**, 1160–1187 (2005)
24. Ye, G.B., Zhou, D.X.: Fully online classification by regularization. Appl. Comput. Harmon. Anal. (2007) doi:10.1016/j.acha.2006.12.001
25. Ying, Y., Zhou, D.X.: Learnability of Gaussians with flexible variances. J. Mach. Learn. Res. **8**, 249–276 (2007)
26. Zhou, D.X.: The covering number in learning theory. J. Complexity **18**, 739–767 (2002)
27. Zhou, D.X.: Capacity of reproducing kernel spaces in learning theory. IEEE Trans. Inform Theory **49**, 1743–1752 (2003)