SVM Learning and L^p Approximation by Gaussians on Riemannian Manifolds

Gui-Bo Ye

School of Mathematical Sciences, Fudan University

Shanghai 200433, China

yeguibo@hotmail.com

Ding-Xuan Zhou

Department of Mathematics, City University of Hong Kong

Kowloon, Hong Kong, China

mazhou@cityu.edu.hk

April 30, 2007

Abstract

We confirm by the multi-Gaussian support vector machine (SVM) classification that the intrinsic dimension of Riemannian manifolds improves the efficiency (learning rates) of learning algorithms. The essential analysis lies in the study of approximation in L^p $(1 \leq p < \infty)$ of L^p functions by their convolutions with the Gaussian kernel with variance $\sigma \to 0$. This covers the SVM case when the approximated function is the Bayes rule and is not continuous in general. The approximation error is estimated by imposing some regularity conditions on the approximated function to lie in some interpolation spaces. Then the learning rates for multi-Gaussian regularized classifiers with general classification loss functions are derived, and the rates depend on the intrinsic dimension of the Riemannian manifold, not the dimension of the underlying Euclidean space. Here the input space is assumed to be connected compact C^1 Riemannian submanifold of $\mathbb{R}^n$ which is isometrically embedded. The uniform normal neighborhoods of the Riemannian manifold and the radial basis form of Gaussian kernels play an important role.

Key words and phrases: Manifold learning, reproducing kernel Hilbert spaces, Gaussian kernels, approximation, multi-kernel regularized classifier, general loss function.

1 Introduction and Multi-Gaussian SVM

Manifold learning has become a hot topic in machine learning recently. It appears naturally from the observation or belief that high-dimensional data or functions arising from physical or biological systems can be effectively modeled or analyzed as being concentrated on a low-dimensional manifold. There have been many tasks for manifold learning demanded by many applications such as dimensionality reduction [4], feature selection [6, 17, 18], semi-supervised learning [3], and learning topological statistics [14]. Compared with vast practical observations and empirical testing, rigorous mathematical analysis in the topic of manifold learning is rather limited [2, 17, 14, 15, 18].

In [25] we consider the approximation of continuous functions on Riemannian manifolds by functions from reproducing kernel Hilbert spaces associated with Gaussian kernels. The obtained order of approximation is applied to the multi-kernel least-square regularization scheme generated by Gaussians with flexible variances. The derived learning rate is better than that in the setting of Euclidean space domains, which confirms the belief that the low dimensionality of manifolds improves the efficiency of learning algorithms.

Many problems in machine learning are about classification where an essential mathematical problem is the approximation of functions in spaces like $L^p(X)$, not in $C(X)$. So in this paper we study the **approximation in** $L^p(X)$ by reproducing kernel Hilbert spaces associated with Gaussian kernels K_σ with variances $\sigma \to 0$. Then we apply the approximation order to get learning rates of **multi-Gaussian regularized classifiers** with general classification loss functions. The obtained learning rates depend on the **intrinsic dimension** of the Riemannian manifold, not the dimension of the underlying Euclidean space.

Let us mention the setting of binary classification and the special example of support vector machines.

1.1 Binary classification

Two classes dealt with by binary classification learning algorithms can be represented by $Y = \{1, -1\}$. The events for which the prediction of classes is desired are points from a

metric space X (called the input space, maybe a subset of $\mathbb{R}^n$). A model used in learning theory is to assume a probability measure ρ on $Z := X \times Y$, then the conditional distribution of ρ at $x \in X$ is a probability distribution $\rho(\cdot|x)$ on Y. For $y = 1$ or -1 in Y, $P(y|x)$ stands for the probability for x to belong to the class y. The marginal distribution ρ_X of ρ on X measures how the events are distributed in X.

A **binary classifier** is a function C from X to Y. It gives a prediction of class $C(x) \in Y$ for each event $x \in X$. The **misclassification error** for the classifier C is defined as the probability of wrong prediction

$$R(f) := \text{Prob}\{C(x) = y\} = \int_X P(y = C(x)|x) d\rho_X. \tag{1.1}$$

By discussing for each event $x \in X$, we can easily see that a best classifier minimizing the misclassification error, called the **Bayes rule** (e.g. [11]), can be expressed as

$$f_c(x) = \begin{cases} 1, & \text{if } P(y = 1|x) \geq P(y = -1|x), \\ -1, & \text{if } P(y = 1|x) < P(y = -1|x). \end{cases} \tag{1.2}$$

The purpose of classification algorithms is to find good classifier approximations $C_{\mathbf{z}}$ of the Bayes rule from the random sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ drawn according to the probability measure ρ. We hope that the approximating classifier $C_{\mathbf{z}}$ will approach the Bayes rule when the number of samples increases, in the sense that the **excess misclassification error** $R(C_{\mathbf{z}}) - R(f_c)$ tends to zero with confidence as $m \to \infty$.

Most practical algorithms can be expressed mathematically as minimizers of some functionals over some spaces of continuous functions. A binary classifier can be derived from a continuous function $f : X \to \mathbb{R}$ as $\text{sgn}(f) : X \to Y$ given by $\text{sgn}(f)(x) = 1$ if $f(x) \geq 0$ and -1 otherwise. To measure how well $\text{sgn}(f)$ can be used for binary classification, we often use a loss function $\phi : \mathbb{R} \to \mathbb{R}_+$ and the value $\phi(yf(x))$ as the local error incurred in predicting y from $f(x)$. Define the **generalization error** of f with respect to the loss function ϕ as

$$E_\phi(f) = \int_Z \phi(yf(x)) d\rho \tag{1.3}$$

and the **empirical error** with respect to the loss function ϕ as

$$E_\phi^{\mathbf{z}}(f) = \frac{1}{n} \sum_{i=1}^n \phi(y_i f(x_i)). \tag{1.4}$$

Many learning algorithms for classification involve this empirical error. Their error analysis (for the misclassification error) can be done by estimating the excess generalization error

$\mathcal{E}_\phi(f) - \mathcal{E}_\phi(f_\rho^\phi)$ where f_ρ^ϕ is the target function defined by

$$f_\rho^\phi = \arg\min \ \mathcal{E}_\phi(f) : f \text{ is a measurable function from } X \text{ to } \mathbb{R} \ . \tag{1.5}$$

1.2 Support vector machines

Support vector machines form an important class of classification algorithms. As a special kernel method, they can be expressed in reproducing kernel Hilbert spaces associated with Mercer kernels.

We say that $K : X \times X \to \mathbb{R}$ is a **Mercer kernel** if it is continuous, symmetric and 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 $\langle K_x, K_y \rangle_K = K(x, y)$.

For example, the Gaussian kernel with variance $\sigma \in (0, \infty)$ defined by

$$K_\sigma(x, y) = \exp\left[-\frac{\|x - y\|^2}{2\sigma^2} \right], \qquad x, y \in X \subset \mathbb{R}^n \tag{1.6}$$

is a Mercer kernel (e. g. [9]).

The important role played by Mercer kernels in kernel methods can be seen from the **regularization scheme** for classification problem associated with the RKHS $\mathcal{H}_K$ and ϕ:

$$f_{\mathbf{z},\lambda,K} = \arg\min_{f \in \mathcal{H}_K} \left(\frac{1}{m} \sum_{i=1}^m \phi\left(y_i f(x_i) \right) + \lambda \|f\|_K^2 \right), \tag{1.7}$$

where $\lambda > 0$ is a constant called the **regularization parameter**. The classifier is given by the sign function $\operatorname{sgn}(f_{\mathbf{z},\lambda,K})$. The reproducing property of $\mathcal{H}_K$:

$$\langle K_x, f \rangle_K = f(x) \qquad x \in X, f \in \mathcal{H}_K \tag{1.8}$$

together with the orthogonal projection in the Hilbert space $\mathcal{H}_K$ tells us that the minimizer in (1.7) has the form $f_{\mathbf{z},\lambda,K} = \sum_{i=1}^m c_i^{\mathbf{z}} K_{x_i}$. The coefficients $(c_i^{\mathbf{z}})_{i=1}^m$ can be computed by solving an optimization problem which is convex when the loss function ϕ is admissible.

Definition 1. *We say that $\phi : \mathbb{R} \to \mathbb{R}_+$ is an* **admissible loss function** *if it is convex and differentiable at 0 with $\phi'(0) < 0$.*

A special setting is given by the hinge loss $\phi(x) = (1 - x)_+ := \max\{0, 1 - x\}$. The algorithm (1.7) with this special loss function ϕ is the support vector machine in the regularization form [13]. Its special form ensures that the convex optimization problem for finding $(c_i^{\mathbf{z}})_{i=1}^m$ in $f_{\mathbf{z},\lambda,K}$ is actually a convex quadratic programming one and many efficient interior point methods are available from optimization theory.

The minimizer f_ρ^ϕ for ϕ being the hinge loss is exactly the Bayes rule f_c. For the error analysis in this special case, Zhang [27] showed that $R(\mathrm{sgn}(f)) - R(f_c) \le E_\phi(f) - E_\phi(f_c)$ which in turn [8] can be bounded by $\|f - f_c\|_{L^1_{\rho_X}}$. Thus to estimate the excess misclassification error $R(\mathrm{sgn}(f_{\mathbf{z},\lambda,K})) - R(f_c)$ for the efficiency of the SVM algorithm, we need to consider $\|f_{\mathbf{z},\lambda,K} - f_c\|_{L^1_{\rho_X}}$, the approximation of the generally discontinuous function f_c in the function space $L^1_{\rho_X}$, not in the space $C(X)$ of all continuous functions on X. So the result from [25] cannot be used for SVM in the manifold setting. This is one motivation of our study in this paper.

1.3 Multi-kernel regularized classifiers

It was shown in [8, 19] that when f_ρ lies in the Sobolev space $H^s(X)$ with $s > 0$ and X is a domain of Euclidean space with nonempty interior, the learning rate of the algorithm (1.7) with a fixed Gaussian kernel and the least square loss $\phi(t) = (1 - t)^2$ is only $O((\log m)^{-s/2})$. If we allow flexible variances of Gaussian kernels, things are totally different and getting polynomial decay for the learning rate is possible [26, 25]. This confirms the usefulness of flexible variances in applying Gaussian kernels in practice.

In this paper we consider the multi-kernel regularized classifier $\mathrm{sgn}(f_{\mathbf{z},\lambda})$ generated by the regularization scheme associated with the general loss function ϕ and Gaussians (1.6) with flexible variances $\{K_\sigma : 0 < \sigma < \infty\}$ defined as

$$f_{\mathbf{z},\lambda} = \arg \min_{\sigma \in (0,+\infty)} \min_{f \in \mathcal{H}_{K_\sigma}} \left(\frac{1}{m} \sum_{i=1}^m \phi\left(y_i f(x_i)\right) + \lambda \|f\|_{K_\sigma}^2 \right). \tag{1.9}$$

Though multi-kernel algorithms for regression and classification have been applied extensively, their error analysis is well understood only if the input space X is a domain of $\mathbb{R}^n$ with nonempty interior and the learning rate is not as fast as expected when the dimension n of the Euclidean space is large. It was pointed out in [26] that when the input space X is a low-dimensional manifold embedded in the large-dimensional Euclidean space, the learning rates may be greatly improved. In such a manifold setting the Fourier transform technique in [26] can no longer be used and other methods are required.

We took a step toward this problem in [25], where we obtained satisfactory learning rates for the multi-kernel regression algorithm with the least square loss by means of uniformly normal neighborhoods in Riemannian manifolds. But the involved approximation scheme there cannot be directly used for the classification setting as mentioned in §1.2. The minimizing function f_ρ^ϕ associated with the loss function ϕ in the classification problem is often discontinuous and we need to consider the approximation problem in the space $L^p(X)$, not in $C(X)$.

2 Main Results on Riemannian Manifolds

In what follows we assume that X is a connected compact C^1 submanifold of $\mathbb{R}^n$ without boundary which is isometrically embedded and of dimension d. Under this assumption, X is a metric space with a 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^1). Here $\|\cdot\|$ is the norm in $\mathbb{R}^n$. Our assumption that the embedding map Φ is the inclusion map is essential. For a general embedding map (which always exits according to the Nash Embedding Theorem), we still do not know how to establish similar results.

For $1 \le p < \infty$, the space $L^p(X)$ on X consists of all measurable functions on X with the norm $\|f\|_{L^p(X)} = \left(\int_X |f(x)|^p dV(x)\right)^{\frac{1}{p}}$ finite, where V is the Riemannian volume measure of X.

To measure the regularity of functions on X, we need Sobolev spaces on the Riemannian manifold X. For an integer k and $f \in C^1(X)$, $\nabla^k f$ denotes the kth covariant derivative of f (with the convention $\nabla^0 f = f$). As an example, the components of ∇f in local coordinates are given by $(\nabla f)_i = \partial_i f$, while the components of $\nabla^2 f$ in local coordinates are given by $(\nabla^2 f)_{ij} = \partial_{ij} f - \sum_{\ell=1}^d \Gamma_{ij}^\ell \partial_\ell f$ where Γ_{ij}^ℓ is Christoffel symbols of X with respcet to $\{\partial_i\}_{i=1}^d$, see [16] for more details.

Definition 2. *Let $p \ge 1$ and $k \in \mathbb{N}$. The Sobolev space $H_k^p(X)$ is the completion of $C^1(X)$ with respect to the norm*

$$\|f\|_{H_k^p(X)} = \sum_{j=0}^k \left(\int_X |\nabla^j f|^p dV\right)^{1/p}.$$

Recall that the Bayes rule f_c is discontinuous in general. Its regularity may not be satisfactorily characterized by Sobolev spaces $H_k^p(X)$ with integer indices k. To get suitable

characterizations, we need interpolation spaces [5] from which one can define Sobolev spaces of arbitrary indices [12].

Definition 3. *Let $0 < \theta \leq 1$. The interpolation space $(L^p(X), H_2^p(X))_\theta$ between $L^p(X)$ and $H_2^p(X)$ consists of all functions $f \in L^p(X)$ such that the norm*

$$\|f\|_\theta := \sup_{t>0} \frac{\mathbb{K}(f,t)}{t^\theta} \tag{2.1}$$

is finite. Here $\mathbb{K}(f,t)$ is the $\mathbb{K}$-functional of the pair $(L^p(X), H_2^p(X))$ defined by

$$\mathbb{K}(f,t) = \inf_{g \in H_2^p(X)} \|f-g\|_{L^p(X)} + t\|g\|_{H_2^p(X)}, \quad t > 0. \tag{2.2}$$

The interpolation space $(L^p(X), H_2^p(X))_\theta$ is a Banach space between $L^p(X)$ (with $\theta = 0$) and $H_2^p(X)$ (with $\theta = 1$). It can be easily seen that the function $\mathbb{K}(f,t)$ of t is continuous, non-decreasing and bounded by $\|f\|_{L^p(X)}$ (take $g = 0$ in (2.2)). Since $H_2^p(X)$ is dense in $L^p(X)$, $\mathbb{K}(f,t)$ tends to zero as $t \to 0$. The requirement that the function f lies in $(L^p(X), H_2^p(X))_\theta$ is equivalent to the decay condition $\mathbb{K}(f,t) = O(t^\theta)$.

Example 1. *Let $X = S^1 = \{e^{iu} : u \in \mathbb{R}\} \subset \mathbb{R}^2, p = 2$ and $0 < a < b < 1$. Define a function $f \in L^2(X)$ by $f(e^{iu}) = 1$ for $a \leq u \leq b$ and zero for $u \in [0, 2\pi] \setminus [a,b]$. Then f is not in $H_1^2(X)$ since $f \in H_1^2(X)$ would imply $f \in C(X)$. But for any $0 < \theta \leq \frac{1}{4}$, we have $f \in (L^p(X), H_2^p(X))_\theta$. In fact, by choosing g_t as*

$$g_t(e^{iu}) = \begin{cases} 1 & \text{if } u \in [a,b], \\ \frac{2}{t}(u-(a-\sqrt{t}))^2 & \text{if } u \in [a-\sqrt{t}, a-\frac{\sqrt{t}}{2}), \\ 1-\frac{2}{t}(u-a)^2 & \text{if } u \in [a-\frac{\sqrt{t}}{2}, a), \\ 1-\frac{2}{t}(u-b)^2 & \text{if } u \in (b, b+\frac{\sqrt{t}}{2}], \\ \frac{2}{t}(u-(b+\sqrt{t}))^2 & \text{if } u \in (b+\frac{\sqrt{t}}{2}, b+\sqrt{t}], \\ 0 & \text{if } u \in [0,2\pi) \setminus [a-\sqrt{t}, b+\sqrt{t}], \end{cases}$$

we have $\|f-g_t\|_{L^2(X)} + t\|g_t\|_{H_2^p(X)} = O(t^{\frac{1}{4}})$.

Using the regularity condition imposed by interpolation spaces, we can state our first main result concerning the learning rates of the multi-kernel SVM with flexible Gaussians on Riemannian manifolds.

Theorem 1. *Let X be a connected compact C^1 submanifold of $\mathbb{R}^n$ without boundary which is isometrically embedded and of dimension d. Let $f_{z,\lambda}$ be defined by (1.9) and $\phi(x) = (1-x)_+$.*

If $f_c \quad (L^1(X), H_2^1(X))_\theta$ *for some* $0 < \theta \leq 1$, *then by taking* $\lambda = \left(\frac{\log^2 m}{m}\right)^{\frac{2\theta+d}{12\theta+2d}}$, *we have*

$$\mathbb{E}_{z2Z^m} \; R(sgn(f_{z,\lambda}) - R(f_c) \; \leq \mathcal{C} \; \left(\frac{\log^2 m}{m}\right)^{\frac{\theta}{6\theta+d}}, \tag{2.3}$$

where $\mathcal{C}$ *is a constant independent of* m.

Theorem 1 is exciting since the learning rate depends only on the intrinsic dimension d of the Riemannian manifold X, not on its extrinsic dimension n. As d is very small and much less than n in many practical problems, our learning rate is satisfactory and convincing theoretical studies in manifold learning. This is another motivation of our investigation in this paper.

Our second main result is about the approximation ability of Gaussians on Riemannian manifolds. This theorem is of importance on its own in approximation theory and it is the key step to prove Theorem 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 [12] and its application in error analysis has been well understood [19, 10, 22]. Things are totally different when X is a Riemannian submanifold of $\mathbb{R}^n$ and little is known. In [25], we considered the approximation ability of Gaussians on the space $C(X)$. Here we consider the approximation on the space $L^p(X)$.

Let $1 \leq p \leq \infty$. Define a family of linear operators $\{I_\sigma : L^p(X) \to L^p(X)\}_{\sigma>0}$ as

$$\begin{aligned} I_\sigma(f)(x) &= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^d} \int_X K_\sigma(x,y) f(y) dV(y) \\ &= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^d} \int_X \exp\left(-\frac{x-y}{2\sigma^2}\right) f(y) dV(y), \qquad x \quad X, \end{aligned} \tag{2.4}$$

where V is the Riemannian volume measure of X.

Note that a d-dimensional manifold is, roughly speaking, a topological space which is locally Euclidean of dimension d. That's why we use the scaling factor $\frac{1}{\left(\sqrt{2\pi}\sigma\right)^d}$.

Theorem 2. *Let* X *be a connected compact* C^1 *submanifold of* $\mathbb{R}^n$ *without boundary which is isometrically embedded and of dimension* d. *Let* $p \geq 1$. *Define* $I_\sigma : L^p(X) \to L^p(X)$ *for* $\sigma > 0$ *by (2.4). If* $f \quad H_2^p(X)$, *then we have*

$$I_\sigma(f) - f \; _{L^p(X} \leq C_X \; f \; _{H_2^p(X} \sigma^2 \qquad \sigma > 0, \tag{2.5}$$

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

The main difficulty in the proof of Theorem 2 (given in Section 3) lies in bounding the integrals over uniformly normal neighborhoods of the convolutions with Gaussians. This is different from the approximation in $C(X)$ where only function values need to be bounded [25].

Due to a saturation phenomenon in approximation theory, the order of approximation in (2.5) cannot be further increased for functions in higher order Sobolev spaces.

The methods of deriving learning rates in this paper can be extended to other learning algorithms on Riemannian manifolds such as online learning for classification [24] and other L^p problems on Riemannian manifolds.

3 L^p Approximation on Manifolds by Gaussians

In this section we prove Theorem 2 after some preparation and illustration.

3.1 Some knowledge on Riemannian manifolds

We start with a brief introduction to normal coordinates and uniform normal neighborhoods (see [14] and [25] in detail). These two basic concepts provide good coordinate systems on Riemannian manifolds and make computations easier. Denote the tangent space at $\mathbf{p} \quad X$ as $T_{\mathsf{p}}(X)$.

Definition 4. *For $\mathbf{p} \quad X$ and $v \quad T_{\mathsf{p}}(X)$, let $\gamma(t, \mathbf{p}, v), t > 0$, be the geodesic satisfying $\gamma(0, \mathbf{p}, v) = \mathbf{p}$ and $\gamma'(0, \mathbf{p}, v) = v$. The* **exponential map** *$\mathsf{E}_{\mathsf{p}} : T_{\mathsf{p}}(X) \to X$ is defined by $\mathsf{E}_{\mathsf{p}}(v) = \gamma(1, \mathbf{p}, v)$.*

By [7], we know that for each $\mathbf{p} \quad X$, there exists a strongly convex neighborhood U_{p} of $\mathbf{p}$, that is, for any two points $\mathbf{q}_1, \mathbf{q}_2$ in the closure $\overline{U}_{\mathsf{p}}$ of U_{p}, there exists a unique minimizing geodesic γ joining $\mathbf{q}_1$ and $\mathbf{q}_2$ whose interior is contained in U_{p}. Denote $B_\delta(0) = \{v \quad T(X) : |v| < \delta\}$ as the ball of $T(X)$ centered at 0 with radius δ.

Definition 5. *An open set $U \quad X$ is called* **uniformly normal** *if there exists some $\delta > 0$ such that $U \quad \mathsf{E}(B_\delta(0))$ for every $\mathbf{q} \quad U$.*

The following proposition (see [25]) tells us the existence of uniform normal neighborhoods having some good properties.

Proposition 1. *For every* $\mathbf{p}$ X *there exist neighborhoods* W_{p} *and* W_{p} *and a number* $\delta_{\mathsf{p}} > 0$ *such that the following conditions hold:*

(a) for every $\mathbf{q}$ W_{p}*, the map* E $: B_{\delta_{\mathsf{p}}}(0)$ $T(X) \to X$ *is a diffeomorphism on* $B_{\delta_{\mathsf{p}}}(0)$*;*

(b) W_{p} *is uniformly normal with respect to* δ_{p}*, that is,* W_{p} $E(B_{\delta_{\mathsf{p}}}(0))$ *for every* $\mathbf{q}$ W_{p}*;*

(c) The closure of W_{p} *is contained in* W_{p} *and* W_{p} U_{p}*.*

Choose an orthonormal basis $\{e_1, e_2, \cdots, e_d\}$ of $T_{\mathsf{p}}(X)$, then for each $\mathbf{q}$ U_{p}, the set of tangent vectors $\{e_1, e_2, \cdots, e_d\}$, moved by parallel transport from $\mathbf{p}$ to $\mathbf{q}$ along the unique minimizing geodesic, forms an orthonormal basis of $T(X)$. In addition, this frame depends smoothly on $\mathbf{q}$. According to (a) of Proposition 1, the map ϕ from $U = \{u$ $\mathbb{R}^d : u < \delta_{\mathsf{p}}\}$ $\mathbb{R}^d$ to X defined by $\phi(u^1, \cdots, u^d) = E(\sum_{i=1}^{d} u^i e_i)$ gives a system of coordinates around $\mathbf{q}$. We call such coordinates **q-normal coordinates**. Under these normal coordinates,

$$d_X(\mathbf{q}, E(\sum_{i=1}^{d} u^i e_i)) = u \qquad u < \delta_{\mathsf{p}}, \tag{3.1}$$

where d_X is the Riemannian metric of X.

In addition, the Riemannian structure g of the isometrically embedded manifold X under the $\mathbf{q}$-normal coordinate (U, ϕ) can be expressed as

$$g_{ij}(u^1, \cdots, u^d) := \langle d\Phi \frac{\partial}{\partial u^i}(\mathbf{q}), d\Phi \frac{\partial}{\partial u^j}(\mathbf{q}) \rangle_{\mathbb{R}^n}, \ i, j = 1, \cdots, d. \tag{3.2}$$

Here $\{\frac{\partial}{\partial u^i}(\mathbf{q})\}_{i=1}^{d}$ is a basis of $T(X)$ [7, 25] and $d\Phi$ is a map from $T(X)$ to $T_{\Phi(\cdot)}(\mathbb{R}^n)$ induced by the inclusion map Φ.

For each $\mathbf{q}$ W_{p} and u U, g_{ij} is well defined and is C^1 as a function on $W_{\mathsf{p}} \times U$. It satisfies $g_{ij}(0) = \delta_{ij}$ and furthermore, we have the following proposition.

Proposition 2. *For* $\mathbf{p}$ X*, choose* W_{p} *and* δ_{p} *as in Proposition 1. For each* $\mathbf{q}$ W_{p}*, choose the* $\mathbf{q}$*-normal coordinate* (U, ϕ) *and the corresponding local representation* g_{ij} *of the Riemannian structure as above. Then the following two bounds hold with a constant* C_{p} *independent of* $\mathbf{q}$ W_{p}*:*

$$\sqrt{det(g_{ij})}(u^1, \cdots, u^d) - 1 \le C_{\mathsf{p}} \|u\|^2 \qquad u \le \delta_{\mathsf{p}}, \tag{3.3}$$

$$d_X(\mathbf{q},x)^2 - \|\mathbf{q}-x\|^2 \le C_{\mathbf{p}}\, d_X(\mathbf{q},x)^4 \qquad x \in E(B_{\delta_{\mathbf{p}}}(0)). \tag{3.4}$$

This proposition is a slight variation of Proposition 2.2 in [14] and it is easy to give a self-contained proof as in [25]. So we omit the proof here.

In this paper, we have assumed that X is a Riemannian submanifold of $\mathbb{R}^n$. For each pair (x,y) of points on X, we have the distance $d_X(x,y)$ under the Riemannian metric and the distance $\|x-y\|$ under the Euclidean metric. The following lemma concerning a relationship between these two metrics was proved in detail in [25].

Lemma 1. *There exists a positive constant $C_0 \ge 1$ such that*

$$d_X(x,y) \le C_0\, \|x-y\| \qquad x,y \in X. \tag{3.5}$$

This lemma will be used frequently in the following since in learning processes we do not see the Riemannian metric d_X. We can only use the Euclidean norm $\|\cdot\|$. But in analysis, we can assume the existence of d_X and make good use of it.

3.2 An illustration of computing integrals on manifolds

In order to get some ideas of using $\mathbf{q}$-normal coordinates system to compute some integrals on the Riemannian manifold X, we prove the following lemma.

Lemma 2. *For the Gaussian kernel K_σ defined by* (1.6)*, we have*

$$\int_X K_\sigma(\mathbf{q},y)dV(y) \le \mathcal{C}_X\sigma^d \qquad \mathbf{q} \in X, \sigma > 0, \tag{3.6}$$

where $\mathcal{C}_X$ is a constant independent of σ or $\mathbf{q}$.

Proof. Let $W_{\mathbf{p}}, \delta_{\mathbf{p}}$ and $C_{\mathbf{p}}$ as in Proposition 1 and Proposition 2. Denote $W_{\mathbf{p}}^0 := W_{\mathbf{p}} \cap E_{\mathbf{p}}(B_{\delta_{\mathbf{p}}/2}(0))$. Since $X \subset \bigcup_{\mathbf{p}\in X} W_{\mathbf{p}}^0$ and X is compact, there exists a finite subset P of X such that $X \subset \bigcup_{\mathbf{p}\in P} W_{\mathbf{p}}^0$.

Let $\delta = \min_{\mathbf{p}\in P} \min\left\{\frac{1}{\sqrt{2C_{\mathbf{p}}}}, \delta_{\mathbf{p}}\right\} > 0$. Let the constant C_0 as in (3.5). Let $\mathbf{q} \in X$. It belongs to some $W_{\mathbf{p}}^0$ with $\mathbf{p} \in P$. Choose $B_\sigma := \{x \in X : d_X(\mathbf{q},x) < C_0\sqrt{2d+6}\sigma\sqrt{\log\frac{1}{\sigma}}\}$. Choose a constant $0 < \sigma_0 \le 1$ such that $C_0\sqrt{2d+6}\sigma_0\sqrt{\log\sigma_0^{-1}} < \delta$.

We first consider the case when $0 < \sigma < \sigma_0$. Since $\exp_{\mathbf{q}}$ is a diffeomorphism on $B_{\delta^*}(0)$, using the equality (3.1), we have $B_\sigma = \exp_{\mathbf{q}}(B_{\delta^*}(0))$ and $B_\sigma = \{\exp_{\mathbf{q}}(\sum_{i=1}^d u^i e_i) : u \in \widetilde{B}_\sigma\}$, where

$$\widetilde{B}_\sigma := \{u \in \mathbb{R}^d : \|u\| < C_0\sqrt{2d+6}\sigma\sqrt{\log\frac{1}{\sigma}}\}. \tag{3.7}$$

For $u = (u^1, \ldots, u^d) \in \mathbb{R}^d$, denote $\phi_{\mathbf{q}}(u) = \exp_{\mathbf{q}}(\sum_{i=1}^d u^i e_i)$, then $B_\sigma = \phi_{\mathbf{q}}(\widetilde{B}_\sigma)$.

Using (3.1) and the inequality (3.4) in Proposition 2, we have

$$d_X^2(\mathbf{q}, \phi_{\mathbf{q}}(u)) - \|\mathbf{q} - \phi_{\mathbf{q}}(u)\|^2 = \|u\|^2 - \|\mathbf{q} - \phi_{\mathbf{q}}(u)\|^2 \le C_{\mathbf{q}} d_X^4(\mathbf{q}, \phi_{\mathbf{q}}(u)) = C_{\mathbf{q}}\|u\|^4.$$

By the definition of δ, $d_X(\mathbf{q}, \phi_{\mathbf{q}}(u)) = \|u\| < \delta \le \min\{\frac{1}{\sqrt{2C_{\mathbf{q}}}}, \delta_{\mathbf{q}}\}$. Hence,

$$\frac{1}{2}\|u\|^2 \le \|\phi_{\mathbf{q}}(u) - \mathbf{q}\|^2 \le \frac{3}{2}\|u\|^2 \qquad u \in \widetilde{B}_\sigma. \tag{3.8}$$

In addition, by the inequality (3.3) in Proposition 2,

$$\frac{1}{2} \le \sqrt{\det(g_{ij})}(u^1, u^2, \cdots, u^d) \le \frac{3}{2} \qquad u \in \widetilde{B}_\sigma. \tag{3.9}$$

Decompose the domain X into two parts B_σ and $X \setminus B_\sigma$. We have

$$\int_X K_\sigma(\mathbf{q}, y)dV(y) = \int_{B_\sigma^{\mathbf{q}}} K_\sigma(\mathbf{q}, y)dV(y) + \int_{X \cap B_\sigma^{\mathbf{q}}} K_\sigma(\mathbf{q}, y)dV(y) := J_1(\mathbf{q}) + J_2(\mathbf{q}).$$

Using the local representation of the Riemannian volume measure under the $\mathbf{q}$-normal coordinates involving a measurable function h on B_σ :

$$\int_{B_\sigma^{\mathbf{q}}} h(y)dV(y) = \int_{\widetilde{B}_\sigma} h(\phi_{\mathbf{q}}(u))\sqrt{det(g_{ij})}(u)du, \tag{3.10}$$

the first term $J_1(\mathbf{q})$ is

$$J_1(\mathbf{q}) = \int_{\widetilde{B}_\sigma} \exp\left(-\frac{\|\mathbf{q} - \phi_{\mathbf{q}}(u)\|^2}{2\sigma^2}\right)\sqrt{\det(g_{ij})}(u)du.$$

By the inequalities (3.8) and (3.9), we have

$$J_1(\mathbf{q}) \le \frac{3}{2}\int_{\widetilde{B}_\sigma} \exp\left(-\frac{\|u\|^2}{4\sigma^2}\right)du \le \frac{3}{2}\int_{\mathbb{R}^d} \exp\left(-\frac{\|u\|^2}{4\sigma^2}\right)du.$$

Using the radial coordinates in $\mathbb{R}^d$, for any univariate function $\psi(r) : \mathbb{R}_+ \to \mathbb{R}$, we have the following equality for the radial function $\psi(\|y\|)$:

$$\int_{\mathbb{R}^d} \psi(\|y\|)dy = \frac{2\pi^{d/2}}{\Gamma(d/2)}\int_0^1 \psi(r)r^{d-1}dr, \tag{3.11}$$

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

Applying (3.11) to the function $\psi(r) = e^{-\frac{r^2}{4\sigma^2}}$, we have $J_1(\mathbf{q}) \le 3 \cdot 2^{d-1} \pi^{\frac{d}{2}} \sigma^d$.

As for the second term $J_2(\mathbf{q})$, we notice that for $y \in X \setminus B_\sigma$, the restriction $d_X(\mathbf{q}, y) \ge C_0 \sqrt{2d+6\sigma} \sqrt{\log \frac{1}{\sigma}}$ together with (3.5) implies $\mathbf{q} - y \ge \sqrt{2d+6\sigma} \sqrt{\log \frac{1}{\sigma}}$. Thus

$$\begin{aligned} J_2(\mathbf{q}) &\le \int_{X \setminus B_\sigma^{\mathbf{q}}} \exp\left(-\frac{(2d+6)\sigma^2 \log \frac{1}{\sigma}}{2\sigma^2}\right) dV(y) \\ &= \int_{X \setminus B_\sigma^{\mathbf{q}}} \sigma^{d+3} dV(y) \le \mathrm{Vol(X)} \sigma^d. \end{aligned}$$

Combining the estimates for $J_1(\mathbf{q})$ and $J_2(\mathbf{q})$, for $0 < \sigma < \sigma_0$, we have

$$\int_X K_\sigma(\mathbf{q}, y) dV(y) \le 3 \cdot 2^{d-1} \pi^{\frac{d}{2}} + \mathrm{Vol(X)} \sigma^d.$$

For the case $\sigma \ge \sigma_0$, it is easy to see from $e^{-r} \le 1$ for $r \ge 0$ that

$$\int_X K_\sigma(\mathbf{q}, y) dV(y) \le \mathrm{Vol(X)} \le \frac{\mathrm{Vol(X)}}{\sigma_0^d} \sigma^d.$$

This proves the desired result with the constant $C_X = 3 \cdot 2^{d-1} \pi^{\frac{d}{2}} + \mathrm{Vol(X)} + \frac{\mathrm{Vol(X)}}{\sigma_0^d}$. □

Remark 1. *In the proof of Lemma 2, we only need $W_{\mathbf{p}}$ instead of $W_{\mathbf{p}}^0$. To be consistent with later discussion, we use $W_{\mathbf{p}}^0$ here.*

3.3 Some ideas for proving Theorem 2

Since $\|I_\sigma(f) - f\|_{L^p(X)}^p = \int_X \int_X \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) f(y) dV(y) - f(x)|^p dV(x)$ involves two layers of integrals, we need to decompose it twice to make the integral computable in local coordinates.

Let $W_{\mathbf{p}}^0, \mathcal{P}$ and σ_0 be as in §3.2, we know that $X = \bigcup_{\mathbf{p} \in \mathcal{P}} W_{\mathbf{p}}^0$. Thus $\|I_\sigma(f) - f\|_{L^p(X)} \le \sum_{\mathbf{p} \in \mathcal{P}} \left(\int_{W'_{\mathbf{p}}} |I_\sigma(f)(x) - f(x)|^p dV(x) \right)^{\frac{1}{p}}$. It will be seen in the following Proposition 4 that the operator $I_\sigma : L^p(X) \to L^p(X)$ is uniformly bounded (the bound is independent of σ). Furthermore, $C^1(X)$ is dense in $H_2^p(X)$. Thus the problem becomes to estimate $\int_{W'_{\mathbf{p}}} |I_\sigma(f)(x) - f(x)|^p dV(x)$ for each $\mathbf{p} \in \mathcal{P}$, $\sigma \le \sigma_0$ and $f \in C^1(X)$.

Note that $I_\sigma(f)$ in the expression $\int_{W'_{\mathbf{p}}} |I_\sigma(f)(x) - f(x)|^p dV(x)$ still contains an integral over the whole manifold X. We need to decompose it further. Let $\mathbf{q} \in W_{\mathbf{p}}^0$. Choose B_σ and B'_σ as in §3.2.

Separating the domain X into two parts B_σ and $X \setminus B_\sigma$, we have

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

The second term of the above equation can be easily bounded due to the fast decay of $K_\sigma(\mathbf{q}, y)$. Using (3.10), the first term equals

$$\frac{1}{(\sqrt{2\pi}\sigma)^d} \int_{\tilde{B}_\sigma} \exp\left(-\frac{\|\mathbf{q} - \phi(u)\|^2}{2\sigma^2}\right) f(\phi(u)) \sqrt{\det(g_{ij})(u)}\, du. \tag{3.12}$$

For approximation in $C(X)$, the quantity $|f(\phi(u))|$ can be easily bounded by the uniform norm $\|f\|_{C(X)}$.

For L^p approximation, we need to tackle the following problem.

Question 1: How can one bound the expression (3.12) in terms of $\|f\|_{H_2^p(X)}$ by treating $f(\phi(u))$ properly?

In the further decompositions, the term $f(\phi(u)) - f(\mathbf{q})$ naturally appears. Since $f \in C^1(X) \subset H_2^p(X)$, it reminds us of the Taylor expansion of $f(\phi(u))$ in its integral form. We denote by $D^k h$ the kth derivative of a function h on the Euclidean space $\mathbb{R}^d$. That is, the components of $D^k h$ are given by $(D^k h)_{i_1 \cdots i_d} = \partial_{i_1 \cdots i_d} h$, where $(i_1, \cdots, i_d) \in \mathbb{R}^d$ and $i_1 + \cdots + i_d = k$. Then

$$f(\phi(u)) - f(\mathbf{q}) = \left\langle f(\mathbf{q}), \sum_{i=1}^d e_i u^i \right\rangle + \int_0^1 (1-y) D^2(f(\phi(yu)))(u, u) dy. \tag{3.13}$$

Question 2: How can one bound (3.13) in terms of $\|f\|_{H_2^p(X)}$ by handling $D^2(f(\phi(yu)))$ properly?

The above two problems will essentially be solved by the following Proposition 3. It gives us bounds of $\int_{W_{\mathbf{p}}'} |f(\phi(u))|^p dV(\mathbf{q})$ and $\int_{W_{\mathbf{p}}'} \|D^2(f \circ \phi)(u)\|^p dV(\mathbf{q})$ in terms of $\|f\|_{H_2^p(X)}$.

Proposition 3. *Let $\mathbf{p} \in X$ and $\phi(u) = \mathbf{p} + \sum_{i=1}^d u^i e_i$ for $u = (u^1, u^2, \cdots u^d) \in \mathbb{R}^d$ and $\mathbf{q} \in W_{\mathbf{p}}^0 = W_{\mathbf{p}} \cap \mathbf{p}(B_{\delta_{\mathbf{p}}/2}(0))$. Then there exists a constant $\delta_{\mathbf{p}}^0$ satisfying $0 < \delta_{\mathbf{p}}^0 \le \frac{\delta_{\mathbf{p}}}{2}$ such that for all $\|u\| \le \delta_{\mathbf{p}}^0$, we have*

$$\int_{W_{\mathbf{p}}'} |f(\phi(u))|^p dV(\mathbf{q}) \le C_{\mathbf{p}}^0 \int_X |f(\mathbf{q})|^p dV(\mathbf{q}) \qquad f \in L^p(X) \tag{3.14}$$

$$\int_{W_{\mathbf{p}}'} \|D^2(f \circ \phi)(u)\|^p dV(\mathbf{q}) \le C_{\mathbf{p}}^{00} \|f\|^p_{H_2^p(X)} \qquad f \in C_2^p(X), \tag{3.15}$$

where $C_{\mathbf{p}}^0$ and $C_{\mathbf{p}}^{00}$ are two constants independent of f or u.

Proof. Let $W_{\mathbf{p}}$ be as in Proposition 1. Let $\mathbf{q} = \phi^{\mathbf{p}}(x) \quad W_{\mathbf{p}}$ and $\tilde{h}(x,u) = h(\phi^{\mathbf{p}}(x),u) = \phi\ (u)$. Write $\int_{W'_{\mathbf{p}}} f(\phi\ (u))^{\ p} dV(\mathbf{q})$ in the $\mathbf{p}$-normal coordinate, we know that

$$\int_{W'_{\mathbf{p}}} |f(\phi\ (u))|^p dV(\mathbf{q}) = \int_{\{x \in \mathbb{R}^d : \phi^{\mathbf{p}}(x) \in W'_{\mathbf{p}}\}} f(\tilde{h}(x,u))^{\ p} \sqrt{\det g^{\mathbf{p}}_{ij}}(x)dx. \tag{3.16}$$

From the definition of $g^{\mathbf{p}}_{ij}(x)$, we know that $\sqrt{\det g^{\mathbf{p}}_{ij}}(x) > 0$ for all $\quad x\ < \delta_{\mathbf{p}}$. On the other hand, $d_X(\mathbf{p}, \tilde{h}(x,u)) \leq d_X(\mathbf{p}, \phi^{\mathbf{p}}(x)) + d_X(\mathbf{q}, \phi\ (u)) \leq \frac{\delta_{\mathbf{p}}}{2} + \quad u\ \ \leq \delta_{\mathbf{p}}$ for all $(x,u) \quad B_1 := \{(x,u) : \phi^{\mathbf{p}}(x) \quad \overline{W^0_{\mathbf{p}}}, \quad u\ \ \leq \frac{\delta_{\mathbf{p}}}{2}\}$. Thus $\sqrt{\det g^{\mathbf{p}}_{ij}} \circ (\phi^{\mathbf{p}})^{\ -1} \circ \tilde{h}(x,u) > 0$ for all $(x,u) \quad B_1$. Let $g(x,u) = (\phi^{\mathbf{p}})^{\ -1} \circ \tilde{h}(x,u)$ and $J(x,u)$ be the Jacobian of $g(x,u)$ with respect to the variable x, i.e.,

$$J(x,u) = \begin{vmatrix} \frac{\partial g_1(x,u}{\partial x^1} & \frac{\partial g_1(x,u}{\partial x^2} & \cdots & \frac{\partial g_1(x,u}{\partial x^d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_d(x,u}{\partial x^1} & \frac{\partial g_d(x,u}{\partial x^2} & \cdots & \frac{\partial g_d(x,u}{\partial x^d} \end{vmatrix}.$$

It is easy to see that $g(x,0) = (\phi^{\mathbf{p}})^{\ -1} \circ \tilde{h}(x,0) = x$. Hence $J(x,0) = 1$. Since $J(x,u)$ is continuous on the compact set B_1, there exists a constant $0 < \delta^0_{\mathbf{p}} \leq \frac{\delta_{\mathbf{p}}}{2}$ such that $J(x,u) > 0$ for all $(x,u) \quad B_2 := \{(x,u) : \phi^{\mathbf{p}}(x) \quad \overline{W^0_{\mathbf{p}}}, \quad u\ \ \leq \delta^0_{\mathbf{p}}\}$.

Therefore, the function $\frac{\sqrt{\det g^{\mathbf{p}}_{ij}}(x}{\sqrt{\det g^{\mathbf{p}}_{ij}}\ (\phi^{\mathbf{p}}}{}^{-1}\ \tilde{h}(x,u}J(x,u}$ is well defined and continuous on the compact set B_2. So the constant

$$C^0_{\mathbf{p}} := \max_{(x,u) \in B_2} \frac{\sqrt{\det g^{\mathbf{p}}_{ij}}(x)}{\sqrt{\det g^{\mathbf{p}}_{ij}} \circ (\phi^{\mathbf{p}})^{\ -1} \circ \tilde{h}(x,u)J(x,u)}$$

is positive and finite. It follows that (3.16) can be bounded as

$$\begin{aligned} & \int_{W'_{\mathbf{p}}} |f(\phi\ (u))|^p dV(\mathbf{q}) \\ \leq\ & C^0_{\mathbf{p}} \int_{\{x \in \mathbb{R}^d : \phi^{\mathbf{p}}(x) \in W'_{\mathbf{p}}\}} f(\tilde{h}(x,u))^{\ p} \sqrt{\det g^{\mathbf{p}}_{ij}} \circ (\phi^{\mathbf{p}})^{\ -1} \circ \tilde{h}(x,u)J(x,u)dx \\ =\ & C^0_{\mathbf{p}} \int_{W'_{\mathbf{p}}} |f(\phi\ (u))|^p dV(\phi\ (u)) \leq C^0_{\mathbf{p}} \int_X |f(\mathbf{q})|^p dV(\mathbf{q}) \end{aligned} \tag{3.17}$$

by a change of variables $x \quad \phi\ (u)$. This proves the inequality (3.14).

As to the second inequality (3.15), denote $\tilde{f}(u) := f \circ \phi^{\mathbf{p}}(u)$ and $g(\mathbf{q},u) := (\phi^{\mathbf{p}})^{\ -1} \circ \phi\ (u)$, then

$$f(\phi\ (u)) = f \circ \phi^{\mathbf{p}} \circ (\phi^{\mathbf{p}})^{\ -1} \circ \phi\ (u) = \tilde{f} \circ g(\mathbf{q},u).$$

Let $y = g(\mathbf{q}, u)$, by the chain rule,

$$D(f \circ \phi\ )(u)\ _i = \sum_{k=1}^{d} \frac{\partial f(y)}{\partial y^k} \cdot \frac{\partial g_k(\mathbf{q}, u)}{\partial u^i}$$

and

$$D^2(f \circ \phi\ )(u)\ _{ij} = \sum_{k,\ell=1}^{d} \frac{\partial^2 f(y)}{\partial y^k \partial y^\ell} \cdot \frac{\partial g_k(\mathbf{q}, u)}{\partial u^i} \frac{\partial g_\ell(\mathbf{q}, u)}{\partial u^j} + \sum_{k=1}^{d} \frac{\partial f(y)}{\partial y^k} \frac{\partial^2 g_k(\mathbf{q}, u)}{\partial u^i \partial u^j}.$$

Since the function $g(\mathbf{q}, u)$ is C^1 on $\{(\mathbf{q}, u) : \mathbf{q} \in \overline{W_{\mathbf{p}}^0},\ \|u\| \le \delta_{\mathbf{p}}\}$, we have

$$C_1 := \sup_{\mathbf{q} \in \overline{W_{\mathbf{p}}^r}, \|u\| \le \delta_{\mathbf{p}}, 1 \le k,\ell,i,j \le d} \left| \frac{\partial g_k(\mathbf{q}, u)}{\partial u^i} \cdot \frac{\partial g_\ell(\mathbf{q}, u)}{\partial u^j} \right| < \infty,$$

$$C_2 := \sup_{\mathbf{q} \in \overline{W_{\mathbf{p}}^r}, \|u\| \le \delta_{\mathbf{p}}, 1 \le k,i,j \le d} \left| \frac{\partial^2 g_k(\mathbf{q}, u)}{\partial u^i \partial u^j} \right| < \infty.$$

Hence

$$D^2(f \circ \phi\ )(u)\ _{ij} \le C_1 \sum_{k,\ell=1}^{d} \left| \frac{\partial^2 f(y)}{\partial y^k \partial y^\ell} \right| + C_2 \sum_{k=1}^{d} \left| \frac{\partial f(y)}{\partial y^k} \right|. \tag{3.18}$$

Applying the local representation of f and $\nabla^2 f$ under the $\mathbf{p}$-normal coordinate, we know that

$$\frac{\partial^2 f(y)}{\partial y^k \partial y^\ell} = \nabla^2 f\ \phi\ (u)\ _{k\ell} + \sum_{m=1}^{d} \Gamma_{k\ell}^m\ g(\mathbf{q}, u)\ \nabla f\ \phi\ (u)\ _m, \quad \frac{\partial f(y)}{\partial y^k} = \nabla f\ \phi\ (u)\ _k.$$

Denote $C_3 := \sup_{\mathbf{q} \in \overline{W_{\mathbf{p}}^r}, \|u\| \le \delta_{\mathbf{p}}, 1 \le k,\ell,m \le d} \left| \Gamma_{k\ell}^m\ g(\mathbf{q}, u) \right|$, we can bound (3.18) further as

$$D^2(f \circ \phi\ )(u)\ _{ij} \le C_1 \sum_{k,\ell=1}^{d} \nabla^2 f\ \phi\ (u)\ _{k\ell} + (C_2 + d^2 C_1 C_3) \sum_{k=1}^{d} \nabla f\ \phi\ (u)\ _k.$$

This together with the elementary inequality $(\sum_{i=1}^m |a_i|)^p \le m^p \sum_{i=1}^m |a_i|^p$ implies that

$$D^2(f \circ \phi\ )(u)\ ^p \le 3^p d^{p+2} (C_1 + C_2 + d^2 C_1 C_3)^p \left(\nabla^2 f\ \phi\ (u)\ ^p + \nabla f\ \phi\ (u)\ ^p \right). \tag{3.19}$$

Integrating over $W_{\mathbf{p}}^0$ with respect to $\mathbf{q}$ and using the inequality (3.14), we get the desired result. □

Proposition 3 yields the following lemma that will be frequently used in proving Theorem 2.

Lemma 3. *Let* $W_{\mathsf{p}}^0 = W_{\mathsf{p}} \cap E_{\mathsf{p}}$ $B_{\delta_{\mathsf{p}}/2}(0)$ *and* $\widetilde{B}_\sigma$ *be given by* (3.7). *Let* δ_{p}^0 *and* C_{p}^0 *be the constants in Proposition 3. Set* $\delta = \min_{\mathsf{p}\in\mathsf{P}} \min\{\delta_{\mathsf{p}}^0, \frac{1}{\sqrt{2C_{\mathsf{p}}}}\}$ *and* $\sigma_0 \in (0,1]$ *satisfying* $C_0\sqrt{2d+6\sigma_0}\prod_{\mathsf{p}\in\mathsf{P}}\sqrt{\log\sigma_0^{-1}} < \delta$. *Assume* h *is a measurable function on* $\mathbb{R}^d$ *and* $f \in L^p(X)$ *with* $p \geq 1$. *Then*

$$\int_{W_{\mathsf{p}}'}\left(\int_{\widetilde{B}_\sigma} h(u)f(\phi_{\mathbf{q}}(u))du\right)^p dV(\mathbf{q})\Big)^{\frac{1}{p}} \leq (C_{\mathsf{p}}^0)^{\frac{1}{p}}\|f\|_{L^p(X)}\int_{\widetilde{B}_\sigma}|h(u)|du. \tag{3.20}$$

Proof. By the definition of $\widetilde{B}_\sigma$, we know that for $u \in \widetilde{B}_\sigma$, $|u| < \delta \leq \delta_{\mathsf{p}}^0$. Then the case $p = 1$ follows from (3.14) and a change of order of integrals.

When $p > 1$, let $q = \frac{p}{p-1}$ and write $|h(u)| = |h(u)|^{\frac{1}{p}+\frac{1}{q}}$. Then using the Hölder inequality and (3.14), we have

$$\begin{aligned}
&\int_{W_{\mathsf{p}}'}\left(\int_{\widetilde{B}_\sigma} h(u)f(\phi_{\mathbf{q}}(u))du\right)^p dV(\mathbf{q})\Big)^{\frac{1}{p}} \\
&\leq \int_{W_{\mathsf{p}}'}\left(\int_{\widetilde{B}_\sigma}|h(u)|du\right)^{\frac{p}{q}}\int_{\widetilde{B}_\sigma}|h(u)||f(\phi_{\mathbf{q}}(u))|^p du dV(\mathbf{q})\Big)^{\frac{1}{p}} \\
&= \left(\int_{\widetilde{B}_\sigma}|h(u)|du\right)^{\frac{p}{q}}\int_{\widetilde{B}_\sigma}|h(u)|\int_{W_{\mathsf{p}}'}|f(\phi_{\mathbf{q}}(u))|^p dV(\mathbf{q})du\Big)^{\frac{1}{p}} \\
&\leq (C_{\mathsf{p}}^0)^{\frac{1}{p}}\|f\|_{L^p(X)}\int_{\widetilde{B}_\sigma}|h(u)|du.
\end{aligned}$$

This proves (3.20) in the case $p > 1$. □

3.4 Uniform boundedness of linear operators

We give the uniform boundedness of the operator $I_\sigma : L^p(X) \to L^p(X)$ defined by (2.4). It will be used not only for the proof of Theorem 2, but also for deriving Theorem 1.

Proposition 4. *Let* $I_\sigma : L^p(X) \to L^p(X)$ *be defined by* (2.4). *Then*

$$\|I_\sigma(f)\|_{L^p(X)} \leq C_X^0\|f\|_{L^p(X)} \qquad \sigma > 0, f \in L^p(X), \tag{3.21}$$

where C_X^0 *is a constant independent of* σ *or* f.

Proof. Let $W_{\mathsf{p}}^0, \delta_{\mathsf{p}}^0, \delta$ and σ_0 be given in Lemma 3 and P be a finite subset of X such that

X $\mathsf{p2P}\, W_{\mathsf{p}}^0$. For $\sigma \geq \sigma_0$,

$$\begin{aligned} I_\sigma(f)\ _{L^p(X} &= \frac{1}{(\sqrt{2\pi}\sigma)^d} \int_X \Big(\int_X \exp\Big\{-\frac{x-y\ ^2}{2\sigma^2}\Big\} f(y)dV(y) \Big)^p dV(x) \Big)^{\frac{1}{p}} \\ &\leq \frac{1}{(\sqrt{2\pi}\sigma)^d} \int_X \Big(\int_X |f(y)|dV(y) \Big)^p dV(x) \Big)^{\frac{1}{p}} \\ &\leq \frac{(\mathrm{Vol(X)})^{\frac{1}{p}}}{(\sqrt{2\pi}\sigma)^d} \int_X |f(y)|dV(y) \leq \frac{\mathrm{Vol(X)}}{(\sqrt{2\pi}\sigma_0)^d}\ f\ _{L^p(X}. \end{aligned}$$

For $0 < \sigma < \sigma_0$, we know from the inequality

$$I_\sigma(f)\ _{L^p(X} \leq \sum_{\mathsf{p2P}} \Big(\int_{W'_{\mathbf{p}}} |I_\sigma(f)(\mathbf{q})|^p dV(\mathbf{q}) \Big)^{\frac{1}{p}}$$

that it is sufficient to prove for each $\mathbf{p}$ P,

$$\int_{W'_{\mathbf{p}}} |I_\sigma(f)(\mathbf{q})|^p dV(\mathbf{q}) \Big)^{\frac{1}{p}} \leq C^0_{X,\mathsf{p}}\ f\ _{L^p(X} \qquad 0 < \sigma < \sigma_0. \tag{3.22}$$

In fact we get (3.21) by setting $C^0_X = \max\ \frac{\mathrm{Vol(X)}}{(\sqrt{2\pi}\sigma_0)^d}, \sum_{\mathsf{p2P}} C^0_{X,\mathsf{p}}$.

Now we prove (3.22) for each $\mathbf{p}$ P. Let $0 < \sigma < \sigma_0$.

For any $\mathbf{q}$ W^0_{p}, choose $B_\sigma, \widetilde{B}_\sigma$ as in §3.2. Decompose the domain X into two parts B_σ and $X \setminus B_\sigma$. We have from (3.10)

$$\begin{aligned} I_\sigma(f)(\mathbf{q}) &= \frac{1}{(\sqrt{2\pi}\sigma)^d} \int_{\widetilde{B}_\sigma} \exp\Big\{-\frac{\mathbf{q}-\phi\ (u)\ ^2}{2\sigma^2}\Big\} f(\phi\ (u)) \sqrt{\det(g_{ij})(u)}du \\ &\quad + \frac{1}{(\sqrt{2\pi}\sigma)^d} \int_{X \setminus B^{\mathbf{q}}_\sigma} K_\sigma(\mathbf{q}, y) f(y) dV(y) \\ &:= J_1(\mathbf{q}) + J_2(\mathbf{q}). \end{aligned}$$

For the first term $J_1(\mathbf{q})$, we see from (3.8) and (3.9) that

$$J_1(\mathbf{q}) \leq \frac{3}{2(\sqrt{2\pi}\sigma)^d} \int_{\widetilde{B}_\sigma} \exp\Big\{-\frac{u\ ^2}{4\sigma^2}\Big\} |f(\phi\ (u))|\, du. \tag{3.23}$$

By inequality (3.20) in Lemma 3 with $h(u) = \frac{3}{2(\sqrt{2\pi}\sigma)^d} \exp\Big\{-\frac{\|u\|^2}{4\sigma^2}\Big\}$, we have

$$\int_{W'_{\mathbf{p}}} |J_1(\mathbf{q})|^p dV(\mathbf{q}) \Big)^{\frac{1}{p}} \leq (C^0_{\mathsf{p}})^{\frac{1}{p}}\ f\ _{L^p(X} \int_{\widetilde{B}_\sigma} h(u)du.$$

By a change of variables $\frac{u}{\sigma}$ and the equation (3.11), we have

$$\int_{\widetilde{B}_\sigma} h(u)du \le \frac{3}{2}(2\pi)^{\frac{d}{2}} \int_{\mathbb{R}^d} \exp\left(-\frac{u^2}{4}\right) du = 3\cdot 2^{\frac{d}{2}-1}.$$

Therefore,

$$\left(\int_{W'_{\mathbf{p}}} |J_1(\mathbf{q})|^p dV(\mathbf{q})\right)^{\frac{1}{p}} \le 3\cdot 2^{\frac{d}{2}-1}(C^0_{\mathbf{p}})^{\frac{1}{p}} \|f\|_{L^p(X)}. \tag{3.24}$$

As for the second term $J_2(\mathbf{q})$, we notice that for $y \in X\setminus B_\sigma$, the restriction $d_X(\mathbf{q},y) \ge C_0\sqrt{2d+6}\sigma\sqrt{\log\frac{1}{\sigma}}$ yields $\|\mathbf{q}-y\| \ge \sqrt{2d+6}\sigma\sqrt{\log\frac{1}{\sigma}}$. Thus

$$\begin{aligned}
|J_2(\mathbf{q})| &= \frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{X\cap B^{\mathbf{q}}_\sigma} K_\sigma(\mathbf{q},y)f(y)dV(y)\\
&\le \frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{X\cap B^{\mathbf{q}}_\sigma} \exp\left(-\frac{(2d+6)\sigma^2\log\frac{1}{\sigma}}{2\sigma^2}\right)|f(y)|dV(y)\\
&\le (2\pi)^{-d/2}\sigma^3\int_X |f(y)|dV(y).
\end{aligned}$$

Hence

$$\begin{aligned}
\left(\int_{W'_{\mathbf{p}}} |J_2(\mathbf{q})|^p dV(\mathbf{q})\right)^{\frac{1}{p}} &\le (2\pi)^{-d/2}\sigma^3 \left(\int_{W'_{\mathbf{p}}}\left(\int_X |f(y)|dV(y)\right)^p dV(\mathbf{q})\right)^{\frac{1}{p}}\\
&\le (2\pi)^{-d/2}\mathrm{Vol}(X)\,\|f\|_{L^p(X)}\sigma^3\\
&\le (2\pi)^{-d/2}\mathrm{Vol}(X)\,\|f\|_{L^p(X)}. \qquad (3.25)
\end{aligned}$$

Combining this with the inequality (3.24), we get the desired bound (3.22). □

3.5 Proof of Theorem 2

Now we can prove Theorem 2.

Proof of Theorem 2: Let $W^0_{\mathbf{p}}, \delta^0_{\mathbf{p}}, \delta$ and σ_0 be given in Lemma 3 and P be a finite subset of X such that $X \subset \bigcup_{\mathbf{p}\in P} W^0_{\mathbf{p}}$. We have $\|I_\sigma(f)-f\|_{L^p(X)} \le \sum_{\mathbf{p}\in P}\left(\int_{W'_{\mathbf{p}}} |I_\sigma(f)(\mathbf{q})-f(\mathbf{q})|^p dV(\mathbf{q})\right)^{\frac{1}{p}}$. Note that $C^1(X)$ is dense in $H^p_2(X)$. Together with the inequality (3.21) in Proposition 4, the inequality (2.5) is verified with $C_X = \max\left((C^0_X+1)\sigma_0^2, \sum_{\mathbf{p}\in P} C_{X,\mathbf{p}}\right)$ if we can prove that for $f \in C^1(X)$ and $\mathbf{p} \in P$,

$$\left(\int_{W'_{\mathbf{p}}} |I_\sigma(f)(x)-f(x)|^p dV(x)\right)^{\frac{1}{p}} \le C_{X,\mathbf{p}}\,\|f\|_{H^p_2(X)}\sigma^2 \qquad 0<\sigma<\sigma_0. \tag{3.26}$$

We prove (3.26) in three steps. Let $0 < \sigma < \sigma_0$.

Step 1: Decomposition. Let $\mathbf{q} \in W_p^0$. Choose $B_\sigma, \widetilde{B}_\sigma$ as in §3.2. By the identity $\frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{\mathbb{R}^d}\exp\{-\frac{|u|^2}{2\sigma^2}\}du = 1$, we can decompose $f(\mathbf{q})$ as

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

Separate the integral on X for $I_\sigma(f)(\mathbf{q})$ to two parts on B_σ and $X\setminus B_\sigma$, we have

$$I_\sigma(f)(\mathbf{q}) - f(\mathbf{q}) = J_1(\mathbf{q}) + J_2(\mathbf{q}) \tag{3.27}$$

where

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

Step 2: Estimation of $\left(\int_{W_p'}|J_1(\mathbf{q})|^p dV(\mathbf{q})\right)^{\frac{1}{p}}$. We separate the error further as

$$\begin{aligned}
J_1(\mathbf{q}) &= \frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{\widetilde{B}_\sigma}\exp\left\{-\frac{|\mathbf{q}-\phi_{\mathbf{q}}(u))|^2}{2\sigma^2}\right\} - \exp\left\{-\frac{|u|^2}{2\sigma^2}\right\}f(\phi_{\mathbf{q}}(u))du \\
&\quad + \frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{\widetilde{B}_\sigma}\exp\left\{-\frac{|\mathbf{q}-\phi_{\mathbf{q}}(u)|^2}{2\sigma^2}\right\}f(\phi_{\mathbf{q}}(u))\sqrt{\det g_{ij}(u)} - 1\ du \\
&\quad + \frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{\widetilde{B}_\sigma}\exp\left\{-\frac{|u|^2}{2\sigma^2}\right\}f(\phi_{\mathbf{q}}(u)) - f(\mathbf{q})\ du \\
&:= J_{11}(\mathbf{q}) + J_{12}(\mathbf{q}) + J_{13}(\mathbf{q}).
\end{aligned}$$

For $J_{11}(\mathbf{q})$, we use (3.4) and the elementary inequality $|e^{-a} - e^{-b}| \le |a-b|\max\{e^{-a}, e^{-b}\}$ (valid for any $a, b > 0$) and find that

$$\begin{aligned}
|J_{11}(\mathbf{q})| &\le \frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{\widetilde{B}_\sigma}\exp\left\{-\frac{|\mathbf{q}-\phi_{\mathbf{q}}(u)|^2}{2\sigma^2}\right\} - \exp\left\{-\frac{|u|^2}{2\sigma^2}\right\}|f(\phi_{\mathbf{q}}(u))|du \\
&\le \frac{1}{(\sqrt{2\pi}\sigma)^d}\int_{\widetilde{B}_\sigma}\max\left\{\exp\left\{-\frac{|\mathbf{q}-\phi_{\mathbf{q}}(u)|^2}{2\sigma^2}\right\}, \exp\left\{-\frac{|u|^2}{2\sigma^2}\right\}\right\}\frac{C_p|u|^4}{2\sigma^2}|f(\phi_{\mathbf{q}}(u))|du.
\end{aligned}$$

So by (3.8),

$$|J_{11}(\mathbf{q})| \le \int_{\widetilde{B}_\sigma}\frac{1}{(\sqrt{2\pi}\sigma)^d}\exp\left\{-\frac{|u|^2}{4\sigma^2}\right\}\frac{C_p|u|^4}{2\sigma^2}|f(\phi_{\mathbf{q}}(u))|du.$$

It follows from (3.20) in Lemma 3 with $h(u) := \frac{\sqrt{C_{\mathbf{p}}}}{(\sqrt{2\pi}\sigma)^d} \exp\left(-\frac{\|u\|^2}{4\sigma^2}\right) \frac{C_{\mathbf{p}}\|u\|^4}{2\sigma^2}$ that

$$\int_{W'_{\mathbf{p}}} |J_{11}(\mathbf{q})|^p dV(\mathbf{q}) \Big)^{\frac{1}{p}} \le (C_{\mathbf{p}}^0)^{\frac{1}{p}} \|f\|_{L^p(X)} \int_{\widetilde{B}_\sigma} h(u)du.$$

By a change of variables $\frac{u}{\sigma}$ and the equation (3.11), we have

$$\int_{\widetilde{B}_\sigma} h(u)du \le \frac{C_{\mathbf{p}}}{2}(2\pi)^{\frac{d}{2}}\sigma^2 \int_{\mathbb{R}^d} \exp\left(-\frac{\|u\|^2}{4}\right) \|u\|^4 du = \frac{2^{\frac{d}{2}+3} C_{\mathbf{p}} \Gamma(\frac{d+4}{2})}{\Gamma(\frac{d}{2})}\sigma^2.$$

Hence

$$\int_{W'_{\mathbf{p}}} |J_{11}(\mathbf{q})|^p dV(\mathbf{q}) \Big)^{\frac{1}{p}} \le \frac{2^{\frac{d}{2}+3} C_{\mathbf{p}} (C_{\mathbf{p}}^0)^{\frac{1}{p}} \Gamma(\frac{d+4}{2})}{\Gamma(\frac{d}{2})} \|f\|_{L^p(X)}\sigma^2. \tag{3.28}$$

For $J_{12}(\mathbf{q})$, we use (3.3) and (3.8) and obtain

$$|J_{12}(\mathbf{q})| \le \frac{\sqrt{C_{\mathbf{p}}}}{(\sqrt{2\pi}\sigma)^d} \int_{\widetilde{B}_\sigma} \exp\left(-\frac{\|u\|^2}{4\sigma^2}\right) \|u\|^2 |f(\phi_{\mathbf{q}}(u))| du.$$

Thus using (3.20) in Lemma 3 with $h(u) = \frac{\sqrt{C_{\mathbf{p}}}}{(\sqrt{2\pi}\sigma)^d} \exp\left(-\frac{\|u\|^2}{4\sigma^2}\right) \|u\|^2$, we obtain

$$\int_{W'_{\mathbf{p}}} |J_{12}(\mathbf{q})|^p dV(\mathbf{q}) \Big)^{\frac{1}{p}} \le \frac{2^{\frac{d}{2}+2} C_{\mathbf{p}} (C_{\mathbf{p}}^0)^{\frac{1}{p}} \Gamma(\frac{d+2}{2})}{\Gamma(\frac{d}{2})} \|f\|_{L^p(X)}\sigma^2. \tag{3.29}$$

For the last term $J_{13}(\mathbf{q})$ of $J_1(\mathbf{q})$, we apply the Taylor expansion (3.13) and get the following further decomposition:

$$\begin{aligned} J_{13}(\mathbf{q}) &= \frac{\sqrt{1}}{(\sqrt{2\pi}\sigma)^d} \int_{\widetilde{B}_\sigma} \exp\left(-\frac{\|u\|^2}{2\sigma^2}\right) \langle \nabla f(\mathbf{q}), \sum_{i=1}^d e_i u_i \rangle du \\ &\quad + \frac{\sqrt{1}}{(\sqrt{2\pi}\sigma)^d} \int_{\widetilde{B}_\sigma} \exp\left(-\frac{\|u\|^2}{2\sigma^2}\right) \int_0^1 (1-y) D^2(f(\phi_{\mathbf{q}}(yu)))(u,u) dy du \\ &:= J_{13}^0(\mathbf{q}) + J_{13}^{00}(\mathbf{q}). \end{aligned} \tag{3.30}$$

Since

$$\frac{\sqrt{1}}{(\sqrt{2\pi}\sigma)^d} \int_{\mathbb{R}^d} \exp\left(-\frac{\|u\|^2}{2\sigma^2}\right) \langle \nabla f(\mathbf{q}), \sum_{i=1}^d e_i u^i \rangle du = 0,$$

we have

$$|J_{13}^0(\mathbf{q})| \le \frac{\sqrt{1}}{(\sqrt{2\pi}\sigma)^d} \int_{\mathbb{R}^d \setminus \widetilde{B}_\sigma} \exp\left(-\frac{\|u\|^2}{2\sigma^2}\right) |\nabla f(\mathbf{q})| \|u\| du. \tag{3.31}$$

Let $q = \frac{p}{p-1}$ when $p > 1$ and denote $h_\sigma(u) := \frac{1}{(\sqrt{2\pi}\sigma)^d} \exp\left(-\frac{\|u\|^2}{2\sigma^2}\right)$ u . Then using the Hölder inequality, we get

$$
\begin{aligned}
&\int_{W'_p} |J_{13}^0(\mathbf{q})|^p dV(\mathbf{q}) \\
&\le \int_{W'_p} \left(\int_{\mathbb{R}^d \cap \widetilde{B}_\sigma} h_\sigma(u) du \right)^{\frac{p}{q}} \int_{\mathbb{R}^d \cap \widetilde{B}_\sigma} h_\sigma(u) |f(\mathbf{q})|^p du dV(\mathbf{q}) \Big)^{\frac{1}{p}} \\
&\le \|f\|_{H_2^p(X)} \int_{\mathbb{R}^d \cap \widetilde{B}_\sigma} \frac{1}{(\sqrt{2\pi}\sigma)^d} \exp\left(-\frac{\|u\|^2}{2\sigma^2}\right) \|u\| \, du \\
&= \|f\|_{H_2^p(X)} \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \int_0^{C_0\sqrt{2d+6}} \frac{1}{\sqrt{\log\frac{1}{\sigma}}} \exp\left(-\frac{r^2}{2}\right) r^d dr \\
&\le \|f\|_{H_2^p(X)} \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \int_0^{C_0\sqrt{2d+6}} \frac{1}{\sqrt{\log\frac{1}{\sigma}}} \exp\left(-\frac{C_0^2(2d+6)(\log\sigma^{-1})}{4}\right) \exp\left(-\frac{r^2}{4}\right) r^d dr \\
&\le \frac{2\pi^{\frac{d}{2}} \|f\|_{H_2^p(X)} \sigma^{1+\frac{d+3}{2}C_0^2}}{\Gamma(\frac{d}{2})} \int_0^1 \exp\left(-\frac{r^2}{4}\right) r^d dr \le \frac{2\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})} \|f\|_{H_2^p(X)} \sigma^2. \qquad (3.32)
\end{aligned}
$$

The proof for (3.32) in the case $p = 1$ follows directly from a change of order of integrals.

As for the term $J_{13}^{00}(\mathbf{q})$, it is easy to see that

$$
|J_{13}^{00}(\mathbf{q})| \le \frac{1}{(\sqrt{2\pi}\sigma)^d} \int_{\widetilde{B}_\sigma} \exp\left(-\frac{\|u\|^2}{2\sigma^2}\right) \int_0^1 (1-y) \|D^2(f \circ \phi_{\mathbf{q}})(yu)\| \, \|u\|^2 dy du.
$$

For $u \in \widetilde{B}_\sigma$, by (3.15) and the Hölder inequality when $p > 1$, we get

$$
\begin{aligned}
&\int_{W'_p} \left(\int_0^1 (1-y) \|D^2(f \circ \phi_{\mathbf{q}})(yu)\| \, dy \right)^p dV(\mathbf{q}) \\
&\le \int_{W'_p} \left(\int_0^1 (1-y) dy \right)^{\frac{p}{q}} \int_0^1 (1-y) \|D^2(f \circ \phi_{\mathbf{q}})(yu)\|^p dy dV(\mathbf{q}) \\
&\le \left(\int_0^1 (1-y) dy \right)^p C_p^{00} \|f\|^p_{H_2^p(X)} = \frac{C_p^{00}}{2^p} \|f\|^p_{H_2^p(X)}.
\end{aligned}
$$

Denote $\widetilde{\mathbb{R}}_\sigma(u) := \frac{1}{(\sqrt{2\pi}\sigma)^d} \exp\left(-\frac{\|u\|^2}{2\sigma^2}\right) \|u\|^2$. Then using the Hölder inequality, we know that

$$\begin{aligned}
&\int_{W'_{\mathbf{p}}} |J_{13}^{00}(\mathbf{q})|^p dV(\mathbf{q}) \Big)^{\frac{1}{p}} \\
&\leq \int_{W'_{\mathbf{p}}} \Big(\int_{\widetilde{B}_\sigma} \widetilde{\mathbb{R}}_\sigma(u) du \Big)^{\frac{p}{q}} \int_{\widetilde{B}_\sigma} \widetilde{\mathbb{R}}_\sigma(u) \int_0^1 (1-y) |D^2(f \circ \phi_{\mathbf{q}})(yu)| dy \Big)^p du dV(\mathbf{q}) \Big)^{\frac{1}{p}} \\
&\leq \frac{1}{2}(C_{\mathbf{p}}^{00})^{\frac{1}{p}} \|f\|_{H_2^p(X)} \int_{\widetilde{B}_\sigma} \frac{1}{(\sqrt{2\pi}\sigma)^d} \exp\left(-\frac{\|u\|^2}{2\sigma^2}\right) \|u\|^2 du \\
&\leq (C_{\mathbf{p}}^{00})^{\frac{1}{p}} \frac{\Gamma(\frac{d+2}{2})}{\Gamma(\frac{d}{2})} \|f\|_{H_2^p(X,\sigma^2)}.
\end{aligned}$$

This together with (3.32) yields

$$\int_{W'_{\mathbf{p}}} |J_{13}(\mathbf{q})|^p dV(\mathbf{q}) \Big)^{\frac{1}{p}} \leq ((C_{\mathbf{p}}^{00})^{\frac{1}{p}} + 2) \frac{\Gamma(\frac{d+2}{2})}{\Gamma(\frac{d}{2})} \|f\|_{H_2^p(X,\sigma^2)}.$$

Combining the estimates for $J_{11}(\mathbf{q}), J_{12}(\mathbf{q})$ and $J_{13}(\mathbf{q})$, we have

$$\Big(\int_{W'_{\mathbf{p}}} |J_1(\mathbf{q})|^p dV(\mathbf{q}) \Big)^{\frac{1}{p}} \leq 2^{\frac{d}{2}+4} C_{\mathbf{p}} (C_{\mathbf{p}}^0)^{\frac{1}{p}} + (C_{\mathbf{p}}^{00})^{\frac{1}{p}} + 2 \Big) \frac{\Gamma(\frac{d+4}{2})}{\Gamma(\frac{d}{2})} \|f\|_{H_2^p(X,\sigma^2)}. \tag{3.33}$$

Step 3: Estimation of $\int_{W'_{\mathbf{p}}} |J_2(\mathbf{q})|^p dV(\mathbf{q}) \Big)^{\frac{1}{p}}$. Denote the two terms in the expression of $J_2(\mathbf{q})$ as $J_2^0(\mathbf{q})$ and $J_2^{00}(\mathbf{q})$. The first term $J_2^0(\mathbf{q})$ of $J_2(\mathbf{q})$ has been estimated in the proof of Proposition 4 as (3.25). Hence

$$\int_{W'_{\mathbf{p}}} |J_2^0(\mathbf{q})|^p dV(\mathbf{q}) \Big)^{\frac{1}{p}} \leq (2\pi)^{d/2} \mathrm{Vol}(X) \|f\|_{L^p(X,\sigma^2)}. \tag{3.34}$$

Now we bound the second term of $J_2(\mathbf{q})$. Using (3.11) again, we have

$$\begin{aligned}
|J_2^{00}(\mathbf{q})| &= \frac{1}{(\sqrt{2\pi}\sigma)^d} \int_{\mathbb{R}^d \cap \widetilde{B}_\sigma} \exp\left(-\frac{\|u\|^2}{2\sigma^2}\right) f(\mathbf{q}) du \\
&\leq \frac{|f(\mathbf{q})|}{(\sqrt{2\pi}\sigma)^d} \int_{\|u\| \leq C_0 \sqrt{2d+6}\sigma(\log \sigma^{-1})^{1/2}} \exp\left(-\frac{\|u\|^2}{2\sigma^2}\right) du \\
&= \frac{2^{1-\frac{d}{2}}}{\Gamma(\frac{d}{2})} |f(\mathbf{q})| \int_0^{C_0 \sqrt{2d+6}\sigma(\log \sigma^{-1})^{1/2}} \exp\left(-\frac{r^2}{2}\right) r^{d-1} dr \\
&\leq \frac{2^{1-\frac{d}{2}}}{\Gamma(\frac{d}{2})} |f(\mathbf{q})| \int_0^{C_0 \sqrt{2d+6}\sigma(\log \sigma^{-1})^{1/2}} \exp\left(-\frac{C_0^2(2d+6)(\log \sigma^{-1})}{4}\right) \exp\left(-\frac{r^2}{4}\right) r^{d-1} dr \\
&\leq \frac{2^{1-\frac{d}{2}}}{\Gamma(\frac{d}{2})} |f(\mathbf{q})| \int_0^1 \sigma^{\frac{d+3}{2} C_0^2} \exp\left(-\frac{r^2}{4}\right) r^{d-1} dr = 2^{\frac{d}{2}} |f(\mathbf{q})| \sigma^{\frac{d+3}{2} C_0^2}.
\end{aligned}$$

But $C_0 \geq 1$ and $d \geq 1$, so there holds

$$\int_{W_{\mathbf{p}}'} |J_2^{00}(\mathbf{q})| dV(\mathbf{q}) \Big)^{\frac{1}{p}} \leq 2^{\frac{d}{2}} \sigma^2 \int_{W_{\mathbf{p}}'} |f(\mathbf{q})|^p dV(\mathbf{q}) \Big)^{\frac{1}{p}} \leq 2^{\frac{d}{2}} \ f\ _{L^p(X)} \sigma^2.$$

Combining this with (3.34), we get

$$\int_{W_{\mathbf{p}}'} |J_2(\mathbf{q})| dV(\mathbf{q}) \Big)^{\frac{1}{p}} \leq \ (2\pi)^{\frac{d}{2}} \mathrm{Vol}(X) + 2^{\frac{d}{2}} \ \ f\ _{L^p(X)} \sigma^2.$$

This together with (3.33) yields the desired estimate (3.26). □

4 Learning Rates

In this section, we derive learning rates for the multi-kernel classification algorithm (1.9) in the manifold setting, especially for the case of SVM. This is done by balancing the sample error and the regularization error [SZ2, WYZ, WZ].

We need the following result from [26] where we have changed some notation in order to make it consistent with this paper.

The regularization error of the algorithm (1.9) is defined as

$$D(\lambda) = \min_{\sigma \in (0,1]} \min_{f \in H_{K_\sigma}} \ E_\phi(f) - E_\phi(f_\rho^\phi) + \lambda \ f\ _{K_\sigma}^2 \ , \quad \lambda > 0. \tag{4.1}$$

Denote

$$C_\lambda = \sup \ \max\{|\phi^0(t)|, |\phi_+^0(t)|\} : |t| \leq \sqrt{\frac{|\phi(0)|}{\lambda}} \ . \tag{4.2}$$

Proposition 5. *Let X be a subset of $\mathbb{R}^n$ and ϕ be admissible with $C_\lambda < \infty$. Define $f_{z,\lambda}$ by (1.9). Then we have*

$$\mathbb{E}\ E_\phi(f_{z,\lambda}) - E_\phi(f_\rho^\phi) \ \leq C_\lambda \sqrt{\frac{C^0 |\phi(0)|}{\lambda}} \ \frac{\log^2 m}{m} \Big)^{1/4} + \frac{2|\phi(0)|}{\sqrt{m}} + D(\lambda), \tag{4.3}$$

where C^0 is a constant independent of m or λ.

For the SVM case, we have a simplified version of Proposition 5.

Proposition 6. *Let* $\phi(x) = (1 - x)_+ = \max\{0, 1 - x\}$ *and* $X \subset \mathbb{R}^n$. *Define* $f_{z,\lambda}$ *by* (1.9). *If* $0 < \lambda < 1$, *then there exists a constant* C' *independent of* m *or* λ *such that*

$$\mathbb{E}\, E_\phi(f_{z,\lambda}) - E_\phi(f_c) \le \left(\frac{C'}{\lambda} \frac{\log^2 m}{m} \right)^{1/4} + D(\lambda). \tag{4.4}$$

Thus if we can estimate the regularization error $D(\lambda)$ in the manifold setting, the excess misclassification error can be easily derived using the comparison theorem between the excess misclassification error and excess generalization error.

4.1 Learning rates for the multi-Gaussian SVM classifier

We begin with the special case of multi-Gaussian SVM classifier. Since $f_\rho^\phi = f_c$ and $\phi(x) - \phi(y) \le |x - y|$ for the hinge loss, we can bound the regularization error $D(\lambda)$ as

$$D(\lambda) \le \inf_{\sigma \in (0,1]} \inf_{f \in H_{K_\sigma}} \|f - f_c\|_{L^1_{\rho_X}(X)} + \lambda \|f\|^2_{K_\sigma}.$$

Thus we only need to bound $\inf_{\sigma \in (0,1]} \inf_{f \in H_{K_\sigma}} \|f - f_c\|_{L^1_{\rho_X}(X)} + \lambda \|f\|^2_{K_\sigma}$ for the SVM case.

We shall choose $f = I_\sigma(f_c)$ to bound $D(\lambda)$. So we need to estimate $\|I_\sigma(f_c)\|_{K_\sigma}$.

Lemma 4. *Let* $1 \le p < \infty$ *and* $f \in L^p(X)$. *Then the function* $I_\sigma(f)$ *is in* $H_{K_\sigma}(X)$ *and*

$$\begin{aligned} \|I_\sigma(f)\|_{K_\sigma} &\le \begin{cases} \left(\frac{1}{\sqrt{2\pi}}\right)^d (\mathcal{C}_X)^{\frac{p-1}{p}} \|f\|_{L^p(X)} \sigma^{-\frac{d}{p}}, & \text{if } 1 \le p \le 2, \\ \left(\frac{1}{\sqrt{2\pi}}\right)^d (\mathcal{C}_X)^{\frac{1}{2}} (Vol(X))^{\frac{1}{2} - \frac{1}{p}} \|f\|_{L^p(X)} \sigma^{-\frac{d}{2}}, & \text{if } p > 2 \end{cases} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^d (\mathcal{C}_X)^{1 - \frac{1}{\min\{p,2\}}} (Vol(X))^{\frac{1}{2} - \frac{1}{\max\{p,2\}}} \|f\|_{L^p(X)} \sigma^{-\frac{d}{\min\{p,2\}}} \end{aligned} \tag{4.5}$$

where $\mathcal{C}_X$ *is the constant given by Lemma 2.*

Proof. By the definition of $I_\sigma(f)$ and the equation

$$\langle K_\sigma(\cdot, y), K_\sigma(\cdot, z) \rangle_{K_\sigma} = K_\sigma(y, z), \tag{4.6}$$

we have

$$\|I_\sigma(f)\|^2_{K_\sigma} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{2d} \int_X \int_X K_\sigma(y, z) f(y) f(z) dV(y) dV(z).$$

When $p = 1$, we get from $K_\sigma(y,z) \le 1$ that

$$\|I_\sigma(f)\|_{K_\sigma}^2 \le \frac{1}{(\sqrt{2\pi}\sigma)^{2d}} \int_X \int_X |f(y)||f(z)|dV(y)dV(z) = \frac{1}{(2\pi)^d}\sigma^{2d}\|f\|_{L^1(X)}^2,$$

which proves (4.5) in this case.

When $1 < p \le 2$, we set $q = \frac{p}{p-1}$ and apply the Hölder inequality to the function $(K_\sigma(y,z))^{\frac{1}{q}}|f(y)|^{\frac{p}{q}} \cdot (K_\sigma(y,z))^{\frac{1}{p}}|f(y)|^{1-\frac{p}{q}}|f(z)|$. Then

$$\|I_\sigma(f)\|_{K_\sigma}^2 \le \frac{1}{(\sqrt{2\pi}\sigma)^{2d}} \int_X \int_X K_\sigma(y,z)|f(y)|^p dV(y)dV(z)^{\frac{1}{q}} \int_X \int_X K_\sigma(y,z)|f(y)|^{p(1-\frac{p}{q})}|f(z)|^p dV(y)dV(z)^{\frac{1}{p}}. \quad (4.7)$$

Lemma 2 tells us that $\int_X K_\sigma(y,z)dV(z) \le \mathcal{C}_X\sigma^d$ for each $y \in X$. So

$$\int_X \int_X K_\sigma(y,z)|f(y)|^p dV(y)dV(z) \le \mathcal{C}_X\sigma^d\|f\|_{L^p(X)}^p.$$

On the other hand, for $z \in X$, we apply the Hölder inequality and find

$$\int_X K_\sigma(y,z)|f(y)|^{p(1-\frac{p}{q})}dV(y) \le \int_X |f(y)|^p dV(y)^{1-\frac{p}{q}} \int_X (K_\sigma(y,z))^{\frac{q}{p}}dV(y)^{\frac{p}{q}}.$$

But $\frac{q}{p} > 1$ and $K_\sigma(y,z) \le 1$. So $\int_X (K_\sigma(y,z))^{\frac{q}{p}}dV(y) \le \int_X K_\sigma(y,z)dV(y) \le \mathcal{C}_X\sigma^d$. Hence

$$\int_X K_\sigma(y,z)|f(y)|^{p(1-\frac{p}{q})}dV(y) \le \|f\|_{L^p(X)}^{p(1-\frac{p}{q})}(\mathcal{C}_X\sigma^d)^{\frac{p}{q}}.$$

It follows that

$$\begin{aligned}\|I_\sigma(f)\|_{K_\sigma}^2 &\le \frac{1}{(\sqrt{2\pi}\sigma)^{2d}}\left(\mathcal{C}_X\sigma^d\|f\|_{L^p(X)}^p\right)^{\frac{1}{q}}\left(\|f\|_{L^p(X)}^{p+p(1-\frac{p}{q})}(\mathcal{C}_X\sigma^d)^{\frac{p}{q}}\right)^{\frac{1}{p}} \\ &= \frac{\mathcal{C}_X^{\frac{2}{p}}}{(2\pi)^d}\|f\|_{L^p(X)}^2\sigma^{\frac{2d}{p}}.\end{aligned}$$

That is,

$$\|I_\sigma(f)\|_{K_\sigma} \le \frac{(\mathcal{C}_X)^{\frac{p-1}{p}}}{(\sqrt{2\pi})^d}\|f\|_{L^p(X)}\sigma^{\frac{d}{p}}.$$

When $p > 2$, each function in $L^p(X)$ lies in $L^2(X)$:

$$\begin{aligned}\|f\|_{L^2(X)} &= \left(\int_X |f(x)|^2 dV(x)\right)^{\frac{1}{2}} \\ &\le \left(\int_X |f(x)|^p dV(x)^{\frac{2}{p}} \int_X dV(x)^{1-\frac{2}{p}}\right)^{\frac{1}{2}} \\ &= (\mathrm{Vol(X)})^{\frac{1}{2}-\frac{1}{p}}\|f\|_{L^p(X)}.\end{aligned}$$

So the desired bound follows from the case $p = 2$:

$$I_\sigma(f)\ _{K_\sigma} \le \frac{(\mathcal{C}_X)^{\frac{1}{2}}}{(\sqrt{2\pi})^d}\ f\ _{L^2(X,\sigma)}^{\frac{d}{2}} \le \frac{(\mathcal{C}_X)^{\frac{1}{2}}(\mathrm{Vol(X)})^{\frac{1}{2}-\frac{1}{p}}}{(\sqrt{2\pi})^d}\ f\ _{L^p(X,\sigma)}^{\frac{d}{2}}.$$

This proves the desired inequality (4.5). □

Proposition 7. *Let X be a connected compact C^1 submanifold of $\mathbb{R}^n$ without boundary which is isometrically embedded and of dimension d. If $f_c \in (L^1(X), H_2^1(X))_\theta$ for some $0 < \theta \le 1$, then*

$$D(\lambda) \le \left((C_X^0 + 1 + C_X)\ f_c\ _\theta + \frac{\mathcal{C}_X\,(Vol(X))^2}{(2\pi)^d}\right)\lambda^{\frac{2\theta}{2\theta+d}}.$$

Proof. Let $\sigma \in (0, \infty)$. For any $g \in H_2^1(X)$, we get from Theorem 2 that

$$I_\sigma(g) - g\ _{L^1(X,\sigma)} \le C_X\ g\ _{H_2^1(X,\sigma^2)}.$$

By Proposition 4, we also have

$$I_\sigma(f_c) - I_\sigma(g)\ _{L^1(X,\sigma)} = \ I_\sigma(f_c - g)\ _{L^1(X,\sigma)} \le C_X^0\ g - f_c\ _{L^1(X,\sigma)}.$$

Since $f_c \in (L^1(X), H_2^1(X))_\theta$, we know that

$$\begin{aligned} I_\sigma(f_c) - f_c\ _{L^1(X,\sigma)} &\le \ I_\sigma(f_c) - I_\sigma(g)\ _{L^1(X,\sigma)} + \ I_\sigma(g) - g\ _{L^1(X,\sigma)} + \ g - f_c\ _{L^1(X,\sigma)} \\ &\le (C_X^0 + 1)\ g - f_c\ _{L^1(X,\sigma)} + C_X\ g\ _{H_2^1(X,\sigma^2)}. \end{aligned}$$

Taking infimum over $g \in H_2^1(X)$, we get

$$\begin{aligned} I_\sigma(f_c) - f_c\ _{L^1(X,\sigma)} &\le (C_X^0 + 1 + C_X) \inf_{g \in H_2^1(X,\sigma)} \left\{ g - f_c\ _{L^1(X,\sigma)} + \sigma^2\ g\ _{H_2^1(X,\sigma)} \right\} \\ &= (C_X^0 + 1 + C_X)\mathbb{K}(f_c, \sigma^2) \le (C_X^0 + 1 + C_X)\ f_c\ _\theta \sigma^{2\theta}. \end{aligned}$$

Since f_c is the Bayes rule, $|f_c(x)| \le 1$ for all $x \in X$. Thus $\|f_c\|_{L^2(X,\sigma)} \le \sqrt{\mathrm{Vol(X)}}$ and the regularization error can be bounded as

$$\begin{aligned} D(\lambda) &\le \inf_{\sigma \in (0,1]} \left\{ I_\sigma(f_c) - f_c\ _{L^1(X,\sigma)} + \lambda\ I_\sigma(f_c)\ _{K_\sigma}^2 \right\} \\ &\le \inf_{\sigma \in (0,1]} \left\{ (C_X^0 + 1 + C_X)\ f_c\ _\theta \sigma^{2\theta} + \lambda \cdot \frac{\mathcal{C}_X \mathrm{Vol(X)}}{(2\pi)^d} \sigma^{-d} \right\} \\ &\le \left((C_X^0 + 1 + C_X)\ f_c\ _\theta + \frac{\mathcal{C}_X \mathrm{Vol(X)}}{(2\pi)^d} \right) \lambda^{\frac{2\theta}{2\theta+d}}, \end{aligned}$$

where we have chosen $\sigma = \lambda^{\frac{1}{2\theta+d}}$. □

Proof of Theorem 1: An important relation between the excess misclassification error and the excess generalization error for the hinge loss asserts that [27] for any measurable function $f : X \to \mathbb{R}$,

$$R(\operatorname{sgn}(f)) - R(f_c) \le E_\phi(f) - E_\phi(f_c). \tag{4.8}$$

This together with Proposition 6 and Proposition 7 yields the desired results. □

4.2 Learning rates for general loss functions

We need the following relationship between $E_\phi(f) - E_\phi(f_\rho^\phi)$ and $\|f - f_\rho^\phi\|_{L^p(X}$ which can be derived as Theorem 17 in [21].

Proposition 8. *Let X be a subset of $\mathbb{R}^n$ and ϕ be admissible.*

(a)If ϕ is a Lipschitz s classification loss function on $\mathbb{R}$ with Lipschitz constant C, then for any measurable function $f : X \to \mathbb{R}$,

$$E_\phi(f) - E_\phi(f_\rho^\phi) \le C \int_X |f - f_\rho^\phi|^s d\rho_X \le C \|f - f_\rho^\phi\|_{L^1_{\rho_X}}^s \le C \|f - f_\rho^\phi\|_{L^2_{\rho_X}}^{\frac{s}{2}}.$$

(b) If ϕ is C^1 and its derivative is Lipschitz s on $\mathbb{R}$ with Lipschitz constant C, then

$$E_\phi(f) - E_\phi(f_\rho^\phi) \le C \|f - f_\rho^\phi\|_{L^{1+s}_{\rho_X}}^{1+s} \le C \|f - f_\rho^\phi\|_{L^2_{\rho_X}}^{\frac{1+s}{2}}.$$

Motivated by this result, in the following, we assume that for any measurable function $f : X \to \mathbb{R}$, the excess generalization error satisfies

$$E_\phi(f) - E_\phi(f_\rho^\phi) \le C \|f - f_\rho^\phi\|_{L^p(X}^\alpha, \tag{4.9}$$

where $C, p \ge 1$ and $\alpha \in \mathbb{R}_+$ are constants independent of f.

Recall C_λ defined by (4.2). As in [24], we assume that

$$C_\lambda \le C_0 \lambda^{-\beta}, \text{ for some } \beta \in \mathbb{R}_+, \tag{4.10}$$

where C_0 is a constant.

Theorem 3. *Let ϕ be an admissible loss function with $\phi''(0) > 0$ satisfying (4.9) and (4.10). Define $f_{z,\lambda}$ by (1.9) and f_ρ^ϕ by (1.5). Assume $f_\rho^\phi \in (L^p(X), H_2^p(X))_\theta$ for some $0 < \theta \le 1$ and $p \ge 1$. Then by taking $\lambda = \left(\frac{\log^2 m}{m}\right)^{\frac{\theta\alpha\min\{p,2\}+d}{2(3+2\beta)\theta\alpha\min\{p,2\}+2(2\beta+1)d}}$, we have*

$$\mathbb{E}_{z \in Z^m} R(sgn(f_{z,\lambda}) - R(f_c) = O\left(\frac{\log^2 m}{m}\right)^{\frac{\theta\alpha\min\{p,2\}}{4(3+2\beta)\theta\alpha\min\{p,2\}+4(2\beta+1)d}}!. \tag{4.11}$$

Proof. As in the proof of Proposition 7, we have

$$I_\sigma(f_\rho^\phi) - f_\rho^\phi \ _{L^p(X} \le \{C_X^0 + 1 + C_X\} \ f_\rho^\phi \ _\theta \sigma^{2\theta},$$

It follows from inequalities (4.9) and (4.5) that

$$\begin{aligned} D(\lambda) &\le \inf_{\sigma 2 (0,1} \Big\{ C \ I_\sigma(f_\rho^\phi) - f_\rho^\phi \ _{L^p(X}^{\alpha} + \lambda \ I_\sigma(f_\rho^\phi) \ _{K_\sigma}^{2} \\ &\le \inf_{\sigma 2 (0,1} \Big\{ C\{C_X^0 + 1 + C_X\}^\alpha \ f_\rho^\phi \ _\theta^\alpha \sigma^{2\alpha\theta} + \lambda \cdot C_X^0 \sigma^{-\frac{2d}{\min\{p,2\}}} \Big\}, \end{aligned}$$

where $C_X^0 = \frac{1}{(2\pi)^d}(C_X)^{2-\frac{2}{\min\{p,2\}}} (\mathrm{Vol}(X))^{1-\frac{2}{\max\{p,2\}}} \ f_\rho^\phi \ _{L^p(X}$. Taking $\sigma = \lambda^{\frac{\min\{p,2\}}{2\alpha\theta\min\{p,2\}+2d}}$, we have

$$D(\lambda) \le \Big\{ C\{C_X^0 + 1 + C_X\}^\alpha \ f_\rho^\phi \ _\theta^\alpha + C_X^0 \Big\} \lambda^{\frac{\alpha\theta\min\{p,2\}}{\alpha\theta\min\{p,2\}+d}}. \tag{4.12}$$

Since ϕ is an admissible loss function with $\phi'(0) > 0$, it is shown in [8] that there exists a constant C_ϕ depending only on ϕ such that for any measurable function $f : X \to \mathbb{R}$.

$$R(\mathrm{sgn}(f)) - R(f_c) \le c_\phi \sqrt{E_\phi(f) - E_\phi(f_\rho^\phi)}. \tag{4.13}$$

Then the stated error bound follows from Proposition 5 and (4.10). This proves Theorem 3. □

Now we apply Theorem 3 to q-norm hinge loss $\phi(x) = (1-x)_+^q$ with $q > 1$.

Corollary 1. *Let $\phi(x) = (1-x)_+^q$ with $q > 1$. Define $f_{z,\lambda}$ by (1.9) and f_ρ^ϕ by (1.5). Suppose that there exists a positive constant C_ρ such that $d\rho_X \le C_\rho dV$. Assume f_ρ^ϕ $(L^q(X), H_2^q(X))_\theta$ for some $0 < \theta \le 1$. If $1 \le q \le 2$, then by taking $\lambda = \left(\frac{\log^2 m}{m}\right)^{\frac{q^2\theta+d}{2(2+q)q^2\theta+2qd}}$, we have*

$$\mathbb{E}_{z2Z^m} \ R(sgn(f_{z,\lambda}) - R(f_c) \ = O\left(\frac{\log^2 m}{m}\right)^{\frac{q^2\theta}{4(2+q)q^2\theta+4qd}}. \tag{4.14}$$

If $q > 2$, then by taking $\lambda = \left(\frac{\log^2 m}{m}\right)^{\frac{2\theta+d}{4(2+q)\theta+2qd}}$, we have

$$\mathbb{E}_{z2Z^m} \ R(sgn(f_{z,\lambda}) - R(f_c) \ = O\left(\frac{\log^2 m}{m}\right)^{\frac{\theta}{4(2+q)\theta+2qd}}. \tag{4.15}$$

Proof. For $\phi(x) = (1 - x)_+^q (q > 1)$, we have the following estimate for the excess generalization error by Theorem 25 in [8].

$$\mathcal{E}_\phi(f) - \mathcal{E}_\phi(f_\rho^\phi) \leq \begin{cases} \|f - f_\rho^\phi\|_{L^q_{\rho_X}}, & \text{if } 1 < q \leq 2, \\ q 2^{q-1} \|f - f_\rho^\phi\|_{L^q_{\rho_X}} (2^{q-1} + \|f - f_\rho^\phi\|^{q-1}_{L^q_{\rho_X}}), & \text{if } q > 2. \end{cases}$$

Together with the assumption that $d\rho_X \leq C_\rho dV$, for the case $1 < q \leq 2$, we can get the desired results by using Theorem 3 with $\alpha = q$ and $\beta = \frac{q-1}{2}$. As to the case $q > 2$, choose $\alpha = 1$ and $\beta = \frac{q-1}{2}$. This is the end of the proof. □

Acknowledgments

This work is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 103206], City University of Hong Kong [Project No. 7002126], National Science Fund for Distinguished Young Scholars of China [Project No. 10529101], and National Basic Research Program of China [Project No. 973-2006CB303102]. The corresponding author is Ding-Xuan Zhou.

References

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

[2] M. Belkin and P. Niyogi, Towards a theoretical foundation for Laplacian-Based manifold methods, *COLT* (2005), P. Auer and R. Meir (eds.), pp. 486–500, 2005.

[3] M. Belkin and P. Niyogi, Semi-supervised learning on Riemannian manifolds, *Mach. Learning* **56** (2004), 209–239.

[4] M. Belkin and P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Comput.* **15** (2003), 1373–1396.

[5] J. Bergh and J. Löfström, Interpolation Spaces, Springer-Verlag, 1976.

[6] O. Bousquet, O. Chapelle, and M. Hein, Measure based regularization, *NIPS* (2003).

[7] M. do Carmo, Riemannian Geometry, Birkhäuser, Boston, 1992.

[8] D. R. Chen, Q. Wu, Y. Ying and D. X. Zhou, Support vector machine soft margin classifiers: error analysis, *J. Mach. Learning Res.* **5** (2004), 1143–1175.

[9] F. Cucker and D. X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, 2007.

[10] E. De Vito, A. Caponnetto, and L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, *Found. Comput. Math.* **5** (2005), 59–85.

[11] L. Devroye, L. Györfi, and G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer-Verlag, New York, 1997.

[12] Z. Ditzian and V. Totik, Moduli of Smoothness, Springer-Verlag, New York, 1987.

[13] T. Evgeniou, M. Pontil and T. Poggio, Regularization networks and suport vector machines, *Adv. Comput. Math.* **13**(2000),1-50

[14] E. Gine and V. Koltchinskii, Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results, to appear in the Proceedings of the 4th International Conference on High Dimensional Probability.

[15] D. Hardin, I. Tsamardinos, and C. F. Aliferis, A theoretical characterization of linear SVM-based feature selection, *Proc. of the 21st Int. Conf. on Machine Learning*, Banff, Canada, 2004.

[16] E. Hebey, Sobolev Spaces on Riemannian Manifolds, Lecture Notes in Mathematics **1635**, Springer-Verlag, 1996.

[17] U. von Luxburg, M. Belkin, and O. Bousquet, Consistency of spectral clustering, preprint, 2004.

[18] S. Mukherjee, Q. Wu, and D. X. Zhou, Learning gradients and feature selection on manifolds, preprint, 2006.

[19] S. Smale and D. X. Zhou, Estimating the approximation error in learning theory, *Anal. Appl.* **1** (2003), 17–41.

[20] S. Smale and D. X. Zhou, Learning theory estimates via integral operators and their applications, *Constr. Approx.* 2007.

[21] Q. Wu, Y. Ying and D. X. Zhou, Learning theory: from regression to classification, Topics in Multivariate Approximation and Interpolation edited by K. Jetter et al., 2005.

[22] Q. Wu, Y. Ying and D. X. Zhou, Multi-kernel regularized classifiers, *J. Complexity*, **23** (2007), 108-134.

[23] Q. Wu and D. X. Zhou, SVM soft margin classifiers: linear programming versus quadratic programming, *Neural Comp.* **17** (2005), 1160–1187.

[24] G. B. Ye and D. X. Zhou, Fully online classification by regularization, *Appl. Comput. Harmonic Anal.*, in press. http://dx.doi. org/10.1016/j.acha.2006.12.001.

[25] G. B. Ye and D. X. Zhou, Learning and approximation by Gaussians on Riemannian manifolds, *Adv. Comput. Math.*, to appear.

[26] Y. Ying and D. X. Zhou, Learnability of Gaussians with flexible variances, *J. Machine Learning Res.* **8** (2007), 249–276.

[27] T. Zhang, Statistical behavior and consistency of classification methods based on convex risk minimization, *Ann. Stat.* **32** (2004), 56–85.

[28] D. X. Zhou, The covering number in learning theory, *J. Complexity* **18** (2002), 739–767.

[29] D. X. Zhou, Capacity of reproducing kernel spaces in learning theory, *IEEE Trans. Inform. Theory* **49** (2003), 1743–1752.