Appl. Comput. Harmon. Anal. 30 (2011) 319–336

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Consistency of regularized spectral clustering ☆

Ying Cao, Di-Rong Chen *

Department of Mathematics, LMIB, Beijing University of Aeronautics and Astronautics, Beijing 100083, PR China

ARTICLE INFO

Article history:
Received 10 November 2009
Revised 12 July 2010
Accepted 9 September 2010
Available online 16 September 2010
Communicated by Dominique Picard

Keywords:
Regularized spectral clustering
Graph Laplacian
Target function
Learning rate

ABSTRACT

Clustering is a widely used technique in machine learning, however, relatively little research in consistency of clustering algorithms has been done so far. In this paper we investigate the consistency of the regularized spectral clustering algorithm, which has been proposed recently. It provides a natural out-of-sample extension for spectral clustering. The presence of the regularization term makes our situation different from that in previous work. Our approach is mainly an elaborate analysis of a functional named the clustering objective. Moreover, we establish a convergence rate. The rate depends on the approximation property and the capacity of the reproducing kernel Hilbert space measured by covering numbers. Some new methods are exploited for the analysis since the underlying setting is much more complicated than usual. Some new methods are exploited for the analysis since the underlying setting is much more complicated than usual.

1. Introduction

Clustering is wildly used in statistics, computer science and various data analysis applications. Clustering algorithms partition a given data set $\mathbf{x} = \{x_i\}_{i=1}^n \subset X$ into several groups based on notions of similarity among the data points. Very often we assume the data points are drawn from an underlying probability distribution on X. The most fundamental issue is whether the clustering algorithm is consistent: do the clusterings constructed by the given algorithm converge to a useful partition of the whole data space as the sample size increases? Consistency is a key property of statistical learning. While extensive literature exists on clustering and partition, very few results on their consistency have established, for exceptions being only k-centers [9], linkage algorithm [6] and spectral clustering [17].

Spectral clustering has attracted a considerable amount of attention recently. With the similarity $K(x_i, x_j)$ among the data points, one can construct graph Laplacian matrices. The simplest form of spectral clustering uses the second eigenvector of graph Laplacian to partition the data points into two groups. In general the first several eigenvectors are all used. We refer to Spielman and Teng [14] for a survey of spectral clustering and von Luxburg [16] for a tutorial to spectral clustering. The consistency of spectral clustering is essentially the convergence of spectral properties of graph Laplacian matrices. Von Luxburg et al. [17] prove that under some mild assumptions the normalized spectral clustering is consistent. However, the unnormalized spectral clustering is consistent only under very specific conditions. Their approach is based on perturbation theory on linear operators in Banach spaces. The problem of out-of-sample extension is considered via Nyström approximation argument, that is, by relating the eigenvectors of the graph Laplacian to the eigenfunctions of a linear operator U_n' on $C(X)$. They establish the convergence of the eigenfunctions of U_n' to those of the limit operator. In either case, von Luxburg et al. [17] assume that, for some positive constant l, the similarity function $K(x, y) \geqslant l$.

☆ Research supported in part by NSF of China under grant 10871015, National Basic Research Program of China under grant 973-2010CB731900 and National Technology Support Program under grant 2008BAC34B03.

* Corresponding author.
E-mail addresses: caoying@ss.buaa.edu.cn (Y. Cao), drchen@buaa.edu.cn (D.-R. Chen).

1063-5203/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.acha.2010.09.002

The regularized spectral clustering algorithm is developed from a general framework proposed by Belkin et al. [3]. It also provides a natural out-of-sample extension for clustering data points not in the original data set. In this algorithm, a regularization term is required to control the smoothness of the target function on X. The experiments in [3, Section 6.1] show the impact of different values of the regularization parameter γ on clustering results. We note that a similar regularized method has also been proposed in the context of graph inference in Vert and Yamanishi [15]. The framework of Belkin et al. [3] actually spans the range from unsupervised to fully supervised learning. It brings together three distinct concepts: manifold learning, spectral graph theory and kernel based learning algorithms. Although plentiful experiments were performed with the proposed algorithms and comparisons were made with inductive methods (SVM, regularized least squares) in Belkin et al. [3], the consistency of algorithms has not been addressed yet.

This paper investigates the consistency of the regularized spectral clustering algorithm. We prove that, as the number of samples tends to infinity, the sequence of the target functions (modulo signs) converges to a function f^*, which corresponds to the limit clustering, essentially the same as in [17]. Moreover, a convergence rate is established. It should be noted that our arguments require the similarity function to be a positive semi-definite kernel function. In addition, the linear operators in this paper are defined on $\mathcal{L}^2_\rho(X)$ rather than $C(X)$. This might also make our results different from those in [17].

It should be pointed out that, due to the presence of regularization term, the situation is different from that in [17]. In particular, the target function is no longer given as an eigenfunction of some linear operator approaching to the limit operator. Therefore, the methods of [17] do not work well in our setting. Our approach is mainly an elaborate analysis of a functional $\varepsilon(f)$ (see Section 3 below for the definition) measuring the quality of the partition induced by f. This functional plays the same role as the functional of generalization error does in supervised learning. Adopting the arguments for estimation of generalization error, we prove that the functional of the target functions tends to the minimum $\varepsilon(f^*)$. This is interesting in its own right, and it in turn yields the convergence of the regularized spectral clustering. Moreover, we use the results of the perturbation theory on the eigenvalues of a integral operator to bound the norm of $f_{\mathbf{x},\gamma}$. A concentration inequality for random variables with values in a Hilbert space is also required to give the convergence rate (usually referred to as learning rate). We note that similar concentration inequalities and perturbation results are used in Rosasco et al. [10] and Smale and Zhou [12]. In the former paper, the authors used a technique based on a concentration inequality for Hilbert spaces to simplify proofs of many results in spectral approximation. Using this method they also provided several new results on spectral properties of the graph Laplacian extending and strengthening results from von Luxburg et al. [17]. In the latter, results similar to Rosasco et al. [10] are derived. Yet the similarity function is required to be a positive definite kernel function.

We would like to compare the analysis of clustering objective $\varepsilon(f)$ with previous work on estimation of generalization error. There are a few new features in our analysis since the setting of regularized spectral clustering is more complicated. In regularized spectral clustering, orthogonality and normalization restrictions are required on the target function and the limit function, and yet there are no such restrictions in other cases, such as SVM and regularized least squares, etc. Moreover, the restrictions on the target function are different from those on the limit function. In fact, the former are empirical versions of the latter. The restrictions (and the difference of the restrictions) make the analysis of consistency much more involved than that in other cases. We mention two points. One is the estimation of the so-called space error. There is no such a step in previous work. For the consideration of space error, some auxiliary functions and new techniques are needed. The other is the estimation of the norm of target functions. Only with a great deal of efforts can we establish even a rough bound. In this regard, the estimation of eigenvalues of some random matrices plays an important role. But it is quite straight in the cases of SVM and regularized least squares to establish such bounds of norms of target functions.

The paper is organized as follows. In Section 2, we introduce the regularized spectral clustering algorithm described in Belkin et al. [3]. Due to the normalization consideration, our restrictions on hypothesis spaces are somewhat different from those of [3]. The clustering objective for measuring the quality of algorithms is defined in Section 3. After proving the existence of its minimizer f^*, we state our main results in Theorem 3.3. In Section 4, the convergence of $\varepsilon(f_{\mathbf{x},\gamma})$ to $\varepsilon(f^*)$ is investigated, and its convergence rate is derived in Theorem 4.1. The space error and the norm of target functions are both considered. Finally, the proof of Theorem 3.3 is given via Theorem 4.1 in Section 5.

2. Regularized spectral clustering

In this section, regularized spectral clustering algorithm and its limit version will be introduced. In the rest of the paper we assume the data space $X \subset \mathbb{R}^d$ is a compact metric space, and ρ is a probability measure on X. The function $K(x, y) : X \times X \to R$ measures the similarities between pairs of points $x, y \in X$.

2.1. Spectral clustering

Given a set of samples $\mathbf{x} = \{x_i\}_{i=1}^n \subset X$ drawn independently according to ρ, we construct a weighted undirected graph $G = (V, E)$ with vertex set $V = \mathbf{X}$. Each edge carries a non-negative weights $K_{ij} = K(x_i, x_j) \geqslant 0$.

Let $\mathbf{K} = (K_{ij})_{i,j=1}^n$ be the similarity matrix and D be a diagonal matrix with diagonal entries $D_{ii} = \sum_{j=1}^n K_{ij}$. The unnormalized graph Laplacian of G is defined as $L_n = D - \mathbf{K}$ and the normalized graph Laplacians are defined as

$$L_n' = I - D^{-1/2}\mathbf{K}D^{-1/2}, \qquad L_n'' = I - D^{-1}\mathbf{K}.$$

Note that L_n and L'_n are symmetric and positive semi-definite. It is easy to see that L_n, L'_n and L'_n have 0 as the smallest eigenvalue.

The normalized spectral clustering uses the first several eigenvectors to obtain a partition of the set $\mathbf{x}$. It is known that v is an eigenvector of L''_n with eigenvalue λ if and only if $w = D^{1/2}v$ is an eigenvector of L'_n with eigenvalue λ. Hence the two normalized graph Laplacians are equivalent from a spectral point of view. The main result of von Luxburg et al. [17] for normalized spectral clustering is that under mild assumptions, among which $K(x,y) \geqslant l$ for some positive constant l, the first several eigenvectors of L'_n converge to those of a limit operator.

In the simplest form, only the eigenvector v with the smallest positive eigenvalue of L'_n is needed. By

$$\frac{1}{2n^2}\sum_{i,j=1}^{n}(u_i-u_j)^2K_{ij} = \frac{1}{n^2}u^TL_nu, \tag{2.1}$$

we can deduce

$$v = \mathop{\arg\min}_{\frac{1}{n}u^TDu=1,\,u^TD\mathbf{1}=0} \frac{1}{2n^2}\sum_{i,j=1}^{n}(u_i-u_j)^2K_{ij}. \tag{2.2}$$

The restrictions on u in (2.2) remove an arbitrary scale factor and a translation invariance in u. The clusters of $\mathbf{x}$ are given by v, e.g. $C = \{x_i \mid v_i \geqslant 0\}$ and $\bar{C} = \{x_i \mid v_i < 0\}$. For details we refer to von Luxburg [16].

2.2. Regularized spectral clustering algorithm

The spectral clustering algorithm (2.2) provides a vector dealing with the sample data only. We construct a function f, defined on whole space X, to give an out-of-sample extension for clustering points that are not in the original data set. By regarding u in (2.2) as the sample $\hat{f} = (f(x_1),\dots,f(x_n))^T$ of f on $\mathbf{x}$, we consider the following optimization problem over a set of functions

$$f = \mathop{\arg\min}_{\frac{1}{n}\hat{f}^TD\hat{f}=1,\,\hat{f}^TD\mathbf{1}=0} \frac{1}{n^2}\hat{f}^TL_n\hat{f}. \tag{2.3}$$

It is understood that a regularization term is necessary to ensure the resulting function is smooth on X. In Belkin et al. [3], the set of functions is a reproducing kernel Hilbert space (RKHS) and the regularizer is the squared norm of f.

Recall that there is a one-to-one correspondence between RKHSs and Mercer kernels. A function $K(x,y)$ is a Mercer kernel, if $K(x,y) = K(y,x)$, and, given an arbitrary finite set of points $\{x_1,\dots,x_n\} \subset X$, the matrix $\mathbf{K} = (K(x_i,x_j))_{i,j=1}^n$ is positive semi-definite. Such an example is the Gaussian kernel $K(x,y) = \exp\{-\frac{|x-y|^2}{2\sigma^2}\}$. The RKHS $\mathcal{H}_K$ associated with the kernel K is the completion of $\operatorname{span}\{K_x = K(x,\cdot)\colon x \in X\}$, with respect to the inner product given by $\langle K_x, K_y\rangle_K = K(x,y)$. See Aronszajn [1] and [5, Chapter 4] for details. Let $\kappa = \sup_{x\in X}\sqrt{K(x,x)}$. It follows that $\kappa = \sup_{x,y\in X}\sqrt{|K(x,y)|}$. Then by $f(x) = \langle f, K_x\rangle_K$, $f \in \mathcal{H}_K$ we have

$$\left|f(x)\right| \leqslant \kappa\|f\|_K, \quad \forall f \in \mathcal{H}_K,\ x \in X. \tag{2.4}$$

Precisely, the regularized spectral clustering algorithm of [3] computes the target function

$$f_{\mathbf{x},\gamma} = \mathop{\arg\min}_{f\in\mathcal{A}\cap\mathcal{H}_K} \frac{1}{n^2}\hat{f}^TL_n\hat{f} + \gamma\|f\|_K^2, \tag{2.5}$$

where

$$\mathcal{A} = \left\{ f \in \mathcal{L}^2_\rho \;\middle|\; \sum_{i=1}^{n} f(x_i)p_{\mathbf{x}}(x_i) = 0,\ \frac{1}{n}\sum_{i=1}^{n} f^2(x_i)p_{\mathbf{x}}(x_i) = 1 \right\}, \tag{2.6}$$

with $p_{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n K_{x_i}$. The set of constrains (2.6) is slightly different from that adopted by Belkin et al. in [3]. They consider the unnormalized spectral clustering, while we choose a set of normalized constrains that is also used in Belkin and Niyogi [2]. Different kinds of constrains in spectral clustering are discussed in [16].

The regularization term $\gamma\|f\|_K^2$ in (2.5) controls the smoothness of the resulting function in the ambient space. The experiments in [3, Section 6.1] show the impact of different values of γ on clustering results.

Although the constrains in (2.6) are a bit different from those in [3], we can also conclude that the target function $f_{\mathbf{x},\gamma}$ admits the representation of the form

$$f_{\mathbf{x},\gamma} = \sum_{i=1}^{n}\alpha_i^{\mathbf{x}}K(x_i,\cdot),$$

where $\alpha^{\mathbf{x}} = (\alpha_1^{\mathbf{x}}, \ldots, \alpha_n^{\mathbf{x}})^T \in \mathbb{R}^n$ is determined by the optimization problem

$$\alpha^{\mathbf{x}} = \arg\min_{\alpha \in \mathbb{R}^n} \frac{1}{n^2}\alpha^T \mathbf{K} L_n \mathbf{K}\alpha + \gamma \alpha^T \mathbf{K}\alpha, \quad \text{s.t. } \frac{1}{n^2}\alpha^T \mathbf{K} D \mathbf{K}\alpha = 1, \quad \alpha^T \mathbf{K} D \mathbf{1} = 0. \tag{2.7}$$

Please refer to Belkin et al. [3] for details.

3. The limit version of spectral clustering and main result

In this section, we will give the notions of the clustering objective and the limit version of graph Laplacian, provide the existence of the limit function, and state our main results on the consistency.

To measure the quality of clustering by a function, the *clustering objective* is defined by

$$\varepsilon(f) = \frac{1}{2}\iint \big(f(x) - f(y)\big)^2 K(x, y)\, d\rho(y)\, d\rho(x).$$

Hereinafter, $\iint = \int_{X^2}$ and $\int = \int_X$. Moreover, $\|\cdot\|_2$ is used instead of $\|\cdot\|_{\mathcal{L}^2_\rho}$ for simplicity. In fact, the clustering objective is a limit version of the quadratic form $\frac{1}{n^2}\hat{f}^T L_n \hat{f}$ in (2.3). By transforming the constrains in (2.3) into the limit form, we minimize $\varepsilon(f)$ over the set

$$\mathcal{B} = \left\{ f \in \mathcal{L}^2_\rho \,\middle|\, \int fp\, d\rho = 0, \int f^2 p\, d\rho = 1 \right\},$$

where $p(x) = \int K(x, y)\, d\rho(y)$. The existence of its minimizer over $\mathcal{B}$ will be discussed later, with the help of spectral properties of the operator $T_K : \mathcal{L}^2_\rho(X) \to \mathcal{L}^2_\rho(X)$ defined as

$$T_K f(x) = \int f(x) \frac{K(x, y)}{\sqrt{p(x)p(y)}}\, d\rho(y).$$

The limit version of graph Laplacians L_n and L'_n, denoted by L and L' respectively are two operators on $\mathcal{L}^2_\rho(X)$, given by

$$Lf(x) = f(x)p(x) - \int f(y)K(x, y)\, d\rho(y),$$
$$L'f(x) = f(x) - \int f(y)\frac{K(x, y)}{\sqrt{p(x)p(y)}}\, d\rho(y).$$

Clearly, L satisfies a limit version of (2.1)

$$\varepsilon(f) = \int f(x)Lf(x)\, d\rho(x), \tag{3.1}$$

and $L' = I - T_K$. Moreover, two operators L and L' are associated by the equality

$$\int f(x)Lf(x)\, d\rho(x) = \int g(x)L'g(x)\, d\rho(x) \quad \text{with } g = f\sqrt{p}. \tag{3.2}$$

From now on, we make the following assumption.

General assumption. There exists a constant $l > 0$ such that $p(x) \geqslant l$ for any $x \in X$.

As it is known, $K(x, y) \leqslant \kappa^2$. Hence, under the General assumption, p satisfies

$$l \leqslant p(x) \leqslant \kappa^2, \quad x \in X. \tag{3.3}$$

Note that T_K is a compact and positive operator under the General assumption. Its eigenvalues can be listed in non-increasing $\mu_1 \geqslant \mu_2 \geqslant \cdots \geqslant 0$ counting multiplicities. Moreover, $\{1 - \mu_i\}_{i \geqslant 1}$ is the set of eigenvalues of L'. By (3.1) and (3.2) we know L' is a positive semi-definite operator. It follows that $\mu_i \leqslant 1$ for all $i \geqslant 1$. Consequently, 1 is the largest eigenvalue of T_K with $\sqrt{p}$ being a corresponding eigenfunction.

A sufficient condition for the existence of the minimizer of $\varepsilon(f)$ over $\mathcal{B}$ is established in the following proposition. Although the proof is elementary, we provide it for readers' convenience.

Proposition 3.1. *Suppose $\mu_1 = 1$ is a simple eigenvalue of T_K. Then there exists a function $f^* \in \mathcal{B}$ such that*

$$0 \leqslant \varepsilon(f) - \varepsilon(f^*) \leqslant \mu_2 \kappa^2 \|f - f^*\|_2^2, \quad \forall f \in \mathcal{B}, \tag{3.4}$$

where μ_2 is the second largest eigenvalue of T_K.

Assume in addition that μ_2 is a simple eigenvalue of T_K. Then for any $f \in \mathcal{B}$ with $\langle f\sqrt{p}, f^*\sqrt{p}\rangle \geqslant 0$, there holds

$$\|f - f^*\|_2^2 \leqslant \frac{2(\varepsilon(f) - \varepsilon(f^*))}{(\mu_2 - \mu_3)l}. \tag{3.5}$$

Proof. Let ϕ_i be normalized eigenfunctions of T_K associated with μ_i, $i = 1, 2, \ldots$. Then $\{\phi_i\}_{i \geqslant 1}$ is an orthonormal basis of $\mathcal{L}^2_\rho$. Set $\phi_1 = \sqrt{p}/\|\sqrt{p}\|_2$, $\int \phi_2 \sqrt{p}\, d\rho = 0$ and, consequently, $f^* = \phi_2/\sqrt{p} \in \mathcal{B}$. We claim that the function f^* is a minimizer of $\varepsilon(f)$ over $\mathcal{B}$ and $\varepsilon(f^*) = 1 - \mu_2$.

In fact, for any $f \in \mathcal{B}$, the function $g = f\sqrt{p}$ can be represented as a series $g = \sum_{i \geqslant 2} a_i \phi_i$ with $\sum_{i \geqslant 2} a_i^2 = 1$. It follows from (3.1) and (3.2) that

$$\varepsilon(f) - \varepsilon(f^*) = -(1 - \mu_2) + \sum_{i \geqslant 2} a_i^2 (1 - \mu_i) = \sum_{i \geqslant 3} a_i^2 (\mu_2 - \mu_i) \geqslant 0, \tag{3.6}$$

confirming f^* is a minimizer of $\varepsilon(f)$ over $\mathcal{B}$.

With notations as above, (3.6) also tells $\varepsilon(f) - \varepsilon(f^*) \leqslant \mu_2(1 - a_2^2) \leqslant 2\mu_2(1 - a_2)$. On the other hand, since $(f - f^*)\sqrt{p} = g - \phi_2$, (3.3) yields $\|g - \phi_2\|_2 \leqslant \kappa \|f - f^*\|_2$. Now (3.4) follows from

$$\|g - \phi_2\|_2^2 = (1 - a_2)^2 + \sum_{i \geqslant 3} a_i^2 = 2(1 - a_2). \tag{3.7}$$

Now suppose furthermore μ_2 is simple, i.e., $\mu_3 < \mu_2$. By (3.6), we find

$$\varepsilon(f) - \varepsilon(f^*) \geqslant (\mu_2 - \mu_3) \sum_{i \geqslant 3} a_i^2 = (\mu_2 - \mu_3)(1 - a_2^2) \geqslant (\mu_2 - \mu_3)(1 - a_2), \tag{3.8}$$

where the last inequality holds due to

$$a_2 = \langle f\sqrt{p}, \phi_2\rangle = \langle f\sqrt{p}, f^*\sqrt{p}\rangle \geqslant 0.$$

Under the General assumption, $\sqrt{l}\|f - f^*\|_2 \leqslant \|g - \phi_2\|_2$, where, as above, $g = f\sqrt{p}$. This, in connection with (3.7) and (3.8), implies (3.5). The proof is complete. □

Proposition 3.1 tells us that f^* is the minimizer of $\varepsilon(f)$ over $\mathcal{B}$ and $f^*\sqrt{p}$ is an eigenfunction of L'. In von Luxburg et al. [7], it is known that the limit partition, given by f^*, segments the data space into sets such that the similarity within the sets is high and the similarity between the sets is low. That intuitively is what clustering is supposed to do. Hence our task is to prove the sequence of target function $f_{\mathbf{x},\gamma}$ converges to f^* in a certain sense, modulo signs (for the sign of a function has no effect on the final partition).

To state the main results, some notations should be made. For a kernel $K(x, y)$, the integral operator $S_K : \mathcal{L}^2_\rho \to \mathcal{H}_K$ is defined by

$$S_K f(x) = \int_X K(x, y) f(y)\, d\rho(y), \quad x \in X. \tag{3.9}$$

Clearly, as K is a Mercer kernel, S_K is a self-adjoint, positive semi-definite and compact operator. Therefore, S_K^α is well defined for any $\alpha > 0$. It is well known that $\mathcal{H}_K = S_K^{1/2}(\mathcal{L}^2_\rho)$. See [5, Chapter 4] for details.

Since our approach to estimate the approximation of f^* by $f_{\mathbf{x},\gamma}$ involves the capacity of the function space $\mathcal{H}_K$, we measure the capacity by means of the covering number of the balls $B_R = \{f \in \mathcal{H}_K;\ \|f\|_K \leqslant R\}$.

Definition 3.2. For a subset $\mathcal{S}$ of a metric space and $\eta > 0$. The covering number $\mathcal{N}(\mathcal{S}, \eta)$ is defined to be the minimal $l \in \mathbb{N}$ such that there exist l disks with radius η covering $\mathcal{S}$.

When $\mathcal{S}$ is compact this number is finite. Denote the covering number of B_1 in $C(X)$ with the metric $\|\cdot\|_\infty$ by $\mathcal{N}(\eta)$. We refer to [19,20] for more details about the covering number.

Our main result establishes a convergence rate of $f_{\mathbf{x},\gamma}$ to f^*, modulo signs.

Theorem 3.3. *Under the General assumption assume that K and ρ satisfy*

(i) $\log\mathcal{N}(B_1,\eta) \leqslant C_0(1/\eta)^s$ *for some* $s > 0$;
(ii) $S_K^{-\alpha/2} f^* \in \mathcal{L}^2_\rho$ *for some* $\alpha \in (0,1]$;
(iii) $\mu_1 = 1$ *and* μ_2 *are simple eigenvalues of* T_K;
(iv) *the second largest eigenvalue* λ_2 *of* S_K *is positive.*

Let $\gamma = n^{-\theta}$ *with* $\theta = \frac{1}{2(1+s)(1+\alpha)}$. *Then for every* $0 < \delta < 1$ *and* $n > N_1(\delta)$, *with confidence at least* $1 - 9\delta$, *there exist signs* $\beta_{\mathbf{x},\gamma} \in \{1,-1\}$ *such that*

$$\left\|\beta_{\mathbf{x},\gamma} f_{\mathbf{x},\gamma} - f^*\right\|_2^2 \leqslant C\log^4(2/\delta)n^{-\theta\alpha}.$$

Here C *is a constant depending on* s, C_0, μ_2, μ_3 κ, l, α *and* $\|S_K^{-\alpha/2} f^*\|_2$, *and we take*

$$N_1(\delta) = \max\left\{M_\delta, (4M_1)^{\frac{1}{\alpha\theta}}, \left(4D_1\log(2/\delta)\right)^{\frac{1}{2\alpha\theta}}, \left(D_2\log(2/\delta)+C_0\right)^{\frac{1}{\alpha\theta}}\right\},$$

with

$$\begin{aligned}
&D_1 = 16\kappa^4\left\|S_K^{-\alpha/2} f^*\right\|_2^2\left(24\kappa^6 l^{-3}+1\right)\left(\mu_2\kappa^2\left(\kappa^2 l^{-1}+1\right)^4 + 4\left(1+\kappa^3 l^{-2}\right)^2\right),\\
&M_\delta = \max\left\{64 l^{-2}\kappa^4\log^2(2/\delta), 64\lambda_2^{-2}\kappa^4\log^2(2/\delta)\right\},\\
&M_1 = \kappa^2\left(1+\kappa^2 l^{-1}\right)^2\left\|S_K^{-\alpha/2} f^*\right\|_2^2 \quad \textit{and} \quad D_2 = 64\kappa^4\left(1+2l^{-2}\kappa^4 + l^{-4}\kappa^4\right).
\end{aligned}$$

It was known that the assumption (i) holds if $K \in C^{2d/s}(X)$ with $X \subset \mathbb{R}^d$, and then we say the RKHS $\mathcal{H}_K$ has polynomial complexity exponent $s > 0$ (see [20]). In particular, if $K \in C^\infty(X)$, (i) is valid for any $s > 0$. Gaussian kernel is such an example. The assumption (ii) is common in consistency analysis of learning theory, see, e.g., [4, Theorem 3], [5] and [13]. (iii) is required by Proposition 3.1 and we need (iv) to bound the norm of $f_{\mathbf{x},\gamma}$ which will be discussed in Section 4.4. The selection of signs $\beta_{\mathbf{x},\gamma}$ will be given in Section 5.

If $K \in C^\infty(X)$ and f^* is in the range of L_K, then, as a consequence of Theorem 3.3, for any $0 < \delta < 1$, $\varepsilon > 0$, $\|\beta_{\mathbf{x},\gamma} f_{\mathbf{x},\gamma} - f^*\|_2^2 \leqslant C\log^4(2/\delta)(1/n)^{\frac{1}{4}-\varepsilon}$, with confidence $1 - 9\delta$. The learning rate is not as good as that in von Luxburg et al. [17].

4. An upper bound of $\varepsilon(f_{\mathbf{x},\gamma}) - \varepsilon(f^*)$

Due to the presence of the regularization term in our setting, the methods provided in von Luxburg et al. [17] and Rosasco et al. [10] do not work well in this situation. Our approach to establish the consistency of the regularized spectral clustering is mainly an elaborate analysis of the clustering objective $\varepsilon(f)$.

In this section, we study the convergence of the sequence of $\varepsilon(f_{\mathbf{x},\gamma})$ to $\varepsilon(f^*)$ when $n \to \infty$. An upper bound of $\varepsilon(f_{\mathbf{x},\gamma}) - \varepsilon(f^*)$ is given in the following theorem. The proof is given in the last subsection of this section.

Theorem 4.1. *Under the General assumption assume K and ρ satisfy*

(i) $\log\mathcal{N}(B_1,\eta) \leqslant C_0(1/\eta)^s$ *for some* $s > 0$;
(ii) $S_K^{-\alpha/2} f^* \in \mathcal{L}^2_\rho$ *for some* $\alpha \in (0,1]$;
(iii) 1 *is a simple eigenvalue of* T_K;
(iv) *the second largest eigenvalue* λ_2 *of* S_K *is positive.*

Let $\gamma = n^{-\theta}$ *with* $\theta = \frac{1}{2(1+s)(1+\alpha)}$. *Then for every* $0 < \delta < 1$ *and* $n > N_2(\delta)$, *with confidence at least* $1 - 7\delta$, *there holds*

$$\varepsilon(f_{\mathbf{x},\gamma}) - \varepsilon\left(f^*\right) \leqslant C'\log^4(2/\delta)n^{-\theta\alpha}.$$

Here C' *is a constant depending on* s, C_0, μ_2, κ, l, α *and* $\|S_K^{-\alpha/2} f^*\|_2$, *and we take*

$$N_2(\delta) = \max\left\{M_\delta, (4M_1)^{\frac{1}{\alpha\theta}}, \left(4D_1\log^2(2/\delta)\right)^{\frac{1}{2\alpha\theta}}\right\},$$

where M_δ, M_1 *and* D_1 *are given as in Theorem* 3.3.

Unlike (iii) in Theorem 3.3, Theorem 4.1 only requires $\mu_1 = 1$ be a simple eigenvalue, since just the first result (3.4) in Proposition 3.1 is required.

The technical difficulties for proving Theorems 3.3 and 4.1 are that $f_{\mathbf{x},\gamma}$ and f^* lie in different spaces, $\mathcal{A} \cap \mathcal{H}_K$ and $\mathcal{B} \cap \mathcal{H}_K$ respectively, and $\mathcal{A}$ is data dependent. To cooperate the constrains on $\mathcal{A}$ and $\mathcal{B}$, we introduce the following construction method denoted by **PN**.

PN. For a function $f \in \mathcal{L}^2_\rho$, define two functions in $\mathcal{A}$ and $\mathcal{B}$ respectively as follows.

1. Projection:

$$g(x) = f(x) - \frac{\int fp\,d\rho}{\int p^2\,d\rho} p(x), \qquad g_{\rho_n}(x) = f(x) - \frac{\int fp_{\mathbf{x}}\,d\rho_n}{\int p_{\mathbf{x}}^2\,d\rho_n} p_{\mathbf{x}}(x),$$

where $\rho_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is the empirical distribution corresponding to the sample set $\mathbf{x} = \{x_i\}_{i=1}^n$.

2. Normalization:

$$h(x) = \omega_g^{-1/2} g(x) \in \mathcal{B}, \qquad h_{\rho_n}(x) = \omega_{g_{\rho_n}}^{-1/2} g_{\rho_n}(x) \in \mathcal{A},$$

with $\omega_g = \int g^2 p\,d\rho$, $\hat{\omega}_{g_{\rho_n},\mathbf{x}} = \int g_{\rho_n}^2 p_{\mathbf{x}}\,d\rho_n$ provided $\omega_g \neq 0$, $\hat{\omega}_{g_{\rho_n},\mathbf{x}} \neq 0$.

When **PN** is applied, checking the conditions $\omega_g \neq 0$, $\hat{\omega}_{g_{\rho_n},\mathbf{x}} \neq 0$ is an issue we have to face.

In order to bound the term $\mathcal{E}(f_{\mathbf{x},\gamma}) - \mathcal{E}(f^*)$, the idea of the error analysis in Cucker and Zhou [5], Wu et al. [18] will be used. We first introduce some definitions and notations.

Given a set $\mathbf{x} = \{x_i\}_{i=1}^n$ of samples independently drawn according to ρ, the *empirical clustering objective* of a function f, denoted by $\mathcal{E}_{\mathbf{x}}(f)$, is defined as

$$\mathcal{E}_{\mathbf{x}}(f) = \frac{1}{n^2}\hat{f}^T L_n \hat{f} = \frac{1}{2n^2}\sum_{i,j=1}^n \big(f(x_i) - f(x_j)\big)^2 K_{ij}, \tag{4.1}$$

where $\hat{f}$ is defined as above. The functional $\mathcal{E}_{\mathbf{x},\gamma}(f)$ is given by

$$\mathcal{E}_{\mathbf{x},\gamma}(f) := \mathcal{E}_{\mathbf{x}}(f) + \gamma\|f\|_K^2. \tag{4.2}$$

It is easy to see that the target function $f_{\mathbf{x},\gamma}$ is the minimizer of $\mathcal{E}_{\mathbf{x},\gamma}(f)$ over $\mathcal{A} \cap \mathcal{H}_K$.

In Section 4.1, an auxiliary function $f_\gamma \in \mathcal{B} \cap \mathcal{H}_K$ is given. Note that $\mathcal{E}(f_{\mathbf{x},\gamma}) - \mathcal{E}(f^*) + \gamma\|f_{\mathbf{x},\gamma}\|_K^2$ is an upper bound of $\mathcal{E}(f_{\mathbf{x},\gamma}) - \mathcal{E}(f^*)$. It can be decomposed as

$$\mathcal{E}(f_\gamma) - \mathcal{E}\big(f^*\big) + \gamma\|f_\gamma\|_K^2 + \big\{\mathcal{E}_{\mathbf{x},\gamma}(f_{\mathbf{x},\gamma}) - \mathcal{E}_{\mathbf{x},\gamma}(f_\gamma)\big\} + \big\{\mathcal{E}(f_{\mathbf{x},\gamma}) - \mathcal{E}_{\mathbf{x}}(f_{\mathbf{x},\gamma}) + \mathcal{E}_{\mathbf{x}}(f_\gamma) - \mathcal{E}(f_\gamma)\big\}. \tag{4.3}$$

The quantity

$$\mathcal{D}(\gamma) := \mathcal{E}(f_\gamma) - \mathcal{E}\big(f^*\big) + \gamma\|f_\gamma\|_K^2$$

is called the *regularization error* and its bound will be given in Section 4.1. The second term in (4.3), $\mathcal{E}_{\mathbf{x},\gamma}(f_{\mathbf{x},\gamma}) - \mathcal{E}_{\mathbf{x},\gamma}(f_\gamma)$ is called the *space error*. It is discussed in Section 4.2. A probability inequality in a Hilbert space is used to deal with the space error. The last term in (4.3) is the *sample error*. In Section 4.3, an upper bound of the sample error is derived, which depends on the capacity of the RKHS measured by covering numbers. Finally, making use of the bound of $\|f_{\mathbf{x},\gamma}\|_K$ given in Section 4.4, we complete the proof of Theorem 4.1.

4.1. The regularization error

In this section, the estimation of the regularization error $\mathcal{D}(\gamma)$ will be provided. We first recall some approximation properties discussed in [5] and [18]. Define

$$\tilde{\mathcal{D}}(\gamma) = \min_{f\in\mathcal{H}_K} \big\|f - f^*\big\|_2^2 + \gamma\|f\|_K^2.$$

If $X \subset \mathbb{R}^d$ is a compact domain and $S_K^{-\alpha/2} f^* \in \mathcal{L}^2_\rho$ for some $0 < \alpha \leqslant 1$, there holds [5, Proposition 8.5]

$$\tilde{\mathcal{D}}(\gamma) \leqslant \gamma^\alpha \big\|S_K^{-\alpha/2} f^*\big\|_2^2, \tag{4.4}$$

and the minimizer F_γ of $\tilde{\mathcal{D}}(\gamma)$ exists. Moreover, (4.4) implies

$$\big\|F_\gamma - f^*\big\|_2^2 \leqslant \gamma^\alpha \big\|S_K^{-\alpha/2} f^*\big\|_2^2. \tag{4.5}$$

Apply the method **PN** to F_γ, and denote the two functions g and h by G_γ and H_γ respectively. When $\gamma^\alpha \leqslant 1/(4M_1)$, we claim $\omega_{G_\gamma} > 0$, for otherwise $F_\gamma = 0$ which in connection with (4.5) yields $\gamma^\alpha \geqslant 1/(\kappa^2\|S_K^{-\alpha/2} f^*\|_2^2) \geqslant 1/(4M_1)$, a contradiction. That is to say, H_γ exists when $\gamma^\alpha \leqslant 1/(4M_1)$. Hereinafter, let $f_\gamma = H_\gamma$. Then $\mathcal{D}(\gamma) = \mathcal{E}(H_\gamma) - \mathcal{E}(f^*) + \gamma\|H_\gamma\|_K^2$. In the following, an estimation of $\mathcal{D}(\gamma)$ is obtained.

Theorem 4.2. *Suppose that K is a Mercer kernel such that $S_K^{-\alpha/2} f^* \in \mathcal{L}^2_\rho$ for some $0 < \alpha \leqslant 1$, and $\mu_1 = 1$ is a simple eigenvalue of T_K. Then for every $0 < \gamma \leqslant \min\{1, (1/(4M_1))^{1/\alpha}\}$, there holds*

$$\mathcal{D}(\gamma) \leqslant \tilde{c}\gamma^\alpha,$$

where $\tilde{c} = (\mu_2\kappa^2(\kappa^2 l^{-1} + 1)^4 + 4(1 + \kappa^3 l^{-2})^2)\|S_K^{-\alpha/2} f^*\|_2^2$, and M_1 is defined as in Theorem 3.3.

Proof. Under the assumption, a function $F_\gamma \in \mathcal{H}_K$ exists minimizing $\tilde{\mathcal{D}}(\gamma)$ and satisfying (4.5). Two functions G_γ and H_γ are given by **PN** with f replaced by F_γ.

Since $f_\gamma = H_\gamma \in \mathcal{H}_K \cap \mathcal{B}$, Proposition 3.1 ensures

$$\mathcal{D}(\gamma) = \mathcal{E}(H_\gamma) - \mathcal{E}(f^*) + \gamma\|H_\gamma\|_K^2 \leqslant \mu_2\kappa^2\|H_\gamma - f^*\|_2^2 + \gamma\|H_\gamma\|_K^2. \tag{4.6}$$

It remains to estimate each of terms $\|H_\gamma - f^*\|_2$ and $\|H_\gamma\|_K$.

Clearly,

$$\|p\|_2^2 \geqslant l^2 \quad \text{and} \quad \|p\|_K^2 = \iint K(x, y)\, d\rho(x)\, d\rho(y) \leqslant \kappa^2. \tag{4.7}$$

Since $\int f^* p\, d\rho = 0$, by (3.3), (4.5) and (4.7),

$$\|F_\gamma - G_\gamma\|_2^2 = \|p\|_2^{-2}\left(\int (F_\gamma - f^*)p\, d\rho\right)^2 \leqslant \kappa^4 l^{-2}\gamma^\alpha\|S_K^{-\alpha/2} f^*\|_2^2. \tag{4.8}$$

On the other hand, $\|G_\gamma\|_2^2 = \int G_\gamma^2\, d\rho \leqslant l^{-1}\omega_{G_\gamma}$. Note that $\omega_{G_\gamma}^{1/2} = \|G_\gamma\sqrt{p}\|_2$ and $\|f^*\sqrt{p}\|_2 = 1$. Hence

$$\|G_\gamma - H_\gamma\|_2 = \left|1 - \omega_{G_\gamma}^{-1/2}\right|\|G_\gamma\|_2 \leqslant l^{-1/2}\left|\omega_{G_\gamma}^{1/2} - 1\right| \leqslant l^{-1/2}\left\|(G_\gamma - f^*)\sqrt{p}\right\|_2 \leqslant \kappa l^{-1/2}\|G_\gamma - f^*\|_2.$$

This, together with (4.5) and (4.8), verifies

$$\|H_\gamma - f^*\|_2 \leqslant (1 + \kappa l^{-1/2})(1 + \kappa^2 l^{-1})\gamma^{\alpha/2}\|S_K^{-\alpha/2} f^*\|_2. \tag{4.9}$$

We now turn to $\|H_\gamma\|_K$. Clearly,

$$\|F_\gamma - G_\gamma\|_K^2 = \frac{\|p\|_K^2(\int F_\gamma p\, d\rho)^2}{\|p\|_2^4} = \frac{\|p\|_K^2}{\|p\|_2^2}\|F_\gamma - G_\gamma\|_2^2.$$

Then it follows from (4.7) and (4.8)

$$\|F_\gamma - G_\gamma\|_K^2 \leqslant \kappa^6 l^{-4}\gamma^\alpha\|S_K^{-\alpha/2} f^*\|_2^2.$$

Observing $\gamma\|F_\gamma\|_K^2 \leqslant \tilde{\mathcal{D}}(\gamma) \leqslant \gamma^\alpha\|S_K^{-\alpha/2} f^*\|_2^2$, we find

$$\|G_\gamma\|_K^2 \leqslant (\|F_\gamma - G_\gamma\|_K + \|F_\gamma\|_K)^2 \leqslant (\kappa^3 l^{-2} + \gamma^{-1/2})^2\gamma^\alpha\|S_K^{-\alpha/2} f^*\|_2^2. \tag{4.10}$$

When $\gamma^\alpha \leqslant 1/(4M_1)$ with $M_1 = \kappa^2(1 + \kappa^2 l^{-1})^2\|S_K^{-\alpha/2} f^*\|_2^2$,

$$\omega_{G_\gamma}^{1/2} \geqslant 1 - \kappa\|G_\gamma - f^*\|_2 \geqslant 1/2.$$

Consequently, (4.10) implies

$$\|H_\gamma\|_K^2 = \omega_{G_\gamma}^{-1}\|G_\gamma\|_K^2 \leqslant 4(\kappa^3 l^{-2} + \gamma^{-1/2})^2\gamma^\alpha\|S_K^{-\alpha/2} f^*\|_2^2. \tag{4.11}$$

Note $l^{-1/2}\kappa \geqslant 1$. When $\gamma < 1$ our statement follows from (4.6), (4.9) and (4.11). □

Theorem 4.2 shows a polynomial decay of the regularization error, $\mathcal{D}(\gamma) = O(\gamma^\alpha)$ with some $0 < \alpha \leqslant 1$. The rate is not only important for bounding the first term in (4.3), but also crucial for bounding the second and third terms. Moreover, it contributes to the understanding of choice of the parameter γ.

4.2. The space error

Usually in learning theory, least-square regularized regression for example, $\varepsilon_{\mathbf{x},\gamma}(f_{\mathbf{x},\gamma}) - \varepsilon_{\mathbf{x},\gamma}(f_\gamma) \leqslant 0$, since $f_{\mathbf{x},\gamma}$ is the minimizer of $\varepsilon_{\mathbf{x},\gamma}(f)$ over $\mathcal{H}_K$. Unfortunately, in our setting this conclusion is no longer correct because $f_{\mathbf{x},\gamma}$ and f_γ live in different spaces, $\mathcal{A} \cap \mathcal{H}_K$ and $\mathcal{B} \cap \mathcal{H}_K$ respectively, and this is why $\varepsilon_{\mathbf{x},\gamma}(f_{\mathbf{x},\gamma}) - \varepsilon_{\mathbf{x},\gamma}(f_\gamma)$ is named as the space error.

We recall a probability inequality for random variables with values in a Hilbert space [8,11,12]. In [10], a similar concentrate inequality is used to obtain results in spectral approximation. Our methods of estimating the space error and the sample error are based on the following lemma.

Lemma 4.3. *Let* $(H, \|\cdot\|)$ *be a Hilbert space and* $\xi(x)$ *be a random variable on* (X, ρ) *with values in* H. *Suppose that* $\|\xi(x)\| \leqslant M < \infty$ *almost surely. Then for any* $0 < \delta < 1$, *with confidence at least* $1 - \delta$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \xi(x_i) - E\xi \right\| \leqslant \frac{4M \log(2/\delta)}{\sqrt{n}}.$$

Applying the above lemma to $H = \mathcal{H}_K$ and $\xi(x) = K_x \in \mathcal{H}_K$, with confidence at least $1 - \delta$,

$$\|p_{\mathbf{x}} - p\|_K \leqslant \frac{4\kappa \log(2/\delta)}{\sqrt{n}}, \tag{4.12}$$

which in connection with (2.4) implies

$$\left|p_{\mathbf{x}}(x_i) - p(x_i)\right| \leqslant \frac{4\kappa^2 \log(2/\delta)}{\sqrt{n}}, \quad i = 1, \ldots, n. \tag{4.13}$$

Moreover, two particular results are derived from Lemma 4.3 with $\xi(x)$ replaced by $f(x)K_x$ and $f^2(x)K_x$ respectively.

Lemma 4.4. *For a function* f *satisfying* $\|f\|_\infty \leqslant R$, *let*

$$I_j(f) = \left| \int f^j(x) p(x)\, d\rho - \frac{1}{n} \sum_{i=1}^{n} f^j(x_i) p_{\mathbf{x}}(x_i) \right|, \quad j = 1, 2.$$

Then for every $0 < \delta \leqslant 1$, *with confidence at least* $1 - 3\delta$,

$$I_j(f) \leqslant 8\kappa^2 R^j \log(2/\delta)/\sqrt{n}, \quad j = 1, 2.$$

Proof. Clearly,

$$I_1(f) \leqslant \left| \int f(x)\big(p(x) - p_{\mathbf{x}}(x)\big)\, d\rho \right| + \left| \int f(x) p_{\mathbf{x}}(x)\, d\rho - \frac{1}{n} \sum_{i=1}^{n} f(x_i) p_{\mathbf{x}}(x_i) \right|.$$

Denote the first and second terms by P_1 and P_2 respectively. For $\|f\|_\infty \leqslant R$, by (2.4) and (4.12), with confidence at least $1 - \delta$,

$$P_1 \leqslant R \cdot \kappa \|p_{\mathbf{x}} - p\|_K \leqslant \frac{4\kappa^2 R \log(2/\delta)}{\sqrt{n}}. \tag{4.14}$$

Let $\xi(x) = f(x)K_x \in \mathcal{H}_K$. Observing $\|\xi\|_K \leqslant \kappa R$, with confidence at least $1 - \delta$,

$$\begin{aligned} P_2 &= \left| \frac{1}{n} \sum_{j=1}^{n} \left(\frac{1}{n} \sum_{i=1}^{n} f(x_i) K(x_i, x_j) - \int f(x) K(x, x_j)\, d\rho \right) \right| \\ &\leqslant \frac{1}{n} \sum_{j=1}^{n} \left| \left(\frac{1}{n} \sum_{i=1}^{n} \xi(x_i) \right)(x_j) - (E\xi)(x_j) \right| \leqslant \kappa \left\| \frac{1}{n} \sum_{i=1}^{n} \xi(x_i) - E\xi \right\|_K \leqslant \frac{4\kappa^2 R \log(2/\delta)}{\sqrt{n}}, \end{aligned} \tag{4.15}$$

where the last equality holds due to Lemma 4.3.

Consequently, an upper bound of $I_1(f)$ is given by (4.14) and (4.15) with confidence at least $1 - 2\delta$.

Similarly, by applying Lemma 4.3 to $\eta(x) = f^2(x)K_x$, the estimation of $I_2(f)$ is obtained. □

Recall that $f_\gamma = H_\gamma$ in Section 4.1. As f_γ and $f_{\mathbf{x},\gamma}$ are in different spaces, the technique **PN** is required once again. Replace f in **PN** by f_γ and denote the resulting functions g_{ρ_n} and h_{ρ_n} by $g_{\rho_n,\gamma}$ and $h_{\rho_n,\gamma}$ respectively. Certainly, we need the condition $\hat{\omega}_{g_{\rho_n,\gamma}} \neq 0$ to ensure the existence of $h_{\rho_n,\gamma}$.

Proposition 4.5. *For every $0 < \delta < 1$, when $n > 64l^{-2}\kappa^4 \log^2(2/\delta)$, with confidence at least $1 - 3\delta$, there holds*

$$|\hat{\omega}_{g_{\rho_n,\gamma}} - 1| \leqslant D_1(n, \delta, \gamma) := \frac{\mathcal{D}(\gamma)}{\gamma}\left\{\frac{8\kappa^4(1 + 8l^{-2}\kappa^4)\log(2/\delta)}{\sqrt{n}} + \frac{1024\kappa^{12}\log^2(2/\delta)}{l^4 n}\right\}.$$

Assume in addition that $n \geqslant 256\kappa^8(1 + 24l^{-3}\kappa^6)^2(\mathcal{D}(\gamma)/\gamma)^2\log^2(2/\delta)$. We have $\hat{\omega}_{g_{\rho_n,\gamma}} \geqslant 1/2 > 0$ with the same confidence $1 - 3\delta$.

Proof. Note that

$$\hat{\omega}_{g_{\rho_n,\gamma}} - 1 = \frac{1}{n}\sum_{i=1}^{n} g^2_{\rho_n,\gamma}(x_i)p_{\mathbf{x}}(x_i) - 1.$$

Since $\int f_\gamma p\, d\rho = 0$ and $\int f^2_\gamma p\, d\rho = 1$, Lemma 4.4 implies, with confidence at least $1 - 3\delta$,

$$I_1(f_\gamma) = \left|\frac{1}{n}\sum_{i=1}^{n} f_\gamma(x_i)p_{\mathbf{x}}(x_i)\right| \leqslant \frac{8\kappa^2\|f_\gamma\|_\infty \log(2/\delta)}{\sqrt{n}}, \tag{4.16}$$

$$I_2(f_\gamma) = \left|\frac{1}{n}\sum_{i=1}^{n} f^2_\gamma(x_i)p_{\mathbf{x}}(x_i) - 1\right| \leqslant \frac{8\kappa^2\|f_\gamma\|^2_\infty \log(2/\delta)}{\sqrt{n}}. \tag{4.17}$$

Clearly, $|p_{\mathbf{x}}(x)| \leqslant \frac{1}{n}\sum_{i=1}^{n}|K(x_i, x)| \leqslant \kappa^2$. Therefore,

$$\begin{aligned}|\hat{\omega}_{g_{\rho_n,\gamma}} - 1| &\leqslant I_2(f_\gamma) + 2I^{-1}_{\mathbf{x}}I_1(f_\gamma)\frac{1}{n}\sum_{i=1}^{n}\left|f_\gamma(x_i)p^2_{\mathbf{x}}(x_i)\right| + I^{-2}_{\mathbf{x}}I^2_1(f_\gamma)\frac{1}{n}\sum_{i=1}^{n}\left|p^3_{\mathbf{x}}(x_i)\right| \\ &\leqslant I_2(f_\gamma) + 2\kappa^4\|f_\gamma\|_\infty I^{-1}_{\mathbf{x}}I_1(f_\gamma) + \kappa^6 I^{-2}_{\mathbf{x}}I^2_1(f_\gamma),\end{aligned} \tag{4.18}$$

where $I_{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^{n} p^2_{\mathbf{x}}(x_i)$.

Recall that f^* is the minimizer of $\varepsilon(f)$ over $\mathcal{B}$. Hence

$$\gamma\|f_\gamma\|^2_K \leqslant \varepsilon(f_\gamma) - \varepsilon\big(f^*\big) + \gamma\|f_\gamma\|^2_K = \mathcal{D}(\gamma).$$

This, in connection with (2.4), implies

$$\|f_\gamma\|_\infty \leqslant \kappa\|f_\gamma\|_K \leqslant \kappa\sqrt{\mathcal{D}(\gamma)/\gamma}. \tag{4.19}$$

Moreover, when $n > 64l^{-2}\kappa^4(\log(2/\delta))^2$, (4.13) tells us that

$$p_{\mathbf{x}}(x_i) \geqslant p(x_i) - \frac{4\kappa^2\log(2/\delta)}{\sqrt{n}} \geqslant l/2, \quad i = 1, \ldots, n. \tag{4.20}$$

Consequently,

$$I_{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} p^2_{\mathbf{x}}(x_i) \geqslant l^2/4. \tag{4.21}$$

By (4.16), (4.17), (4.18), (4.19) and (4.21), we complete the proof. □

When $n > N_\gamma$, Proposition 4.5 ensures the existence of $h_{\rho_n,\gamma}$ with confidence at least $1 - 3\delta$. Here N_γ is given by

$$N_\gamma := \max\big\{64l^{-2}\kappa^4\log^2(2/\delta), 256\kappa^8\big(1 + 24l^{-3}\kappa^6\big)^2\big(\mathcal{D}(\gamma)/\gamma\big)^2\log^2(2/\delta)\big\}. \tag{4.22}$$

By the expression (2.5), $\varepsilon_{\mathbf{x},\gamma}(f_{\mathbf{x},\gamma}) - \varepsilon_{\mathbf{x},\gamma}(h_{\rho_n,\gamma}) \leqslant 0$. Therefore,

$$\varepsilon_{\mathbf{x},\gamma}(f_{\mathbf{x},\gamma}) - \varepsilon_{\mathbf{x},\gamma}(f_\gamma) \leqslant \varepsilon_{\mathbf{x},\gamma}(h_{\rho_n,\gamma}) - \varepsilon_{\mathbf{x},\gamma}(g_{\rho_n,\gamma}) + \varepsilon_{\mathbf{x},\gamma}(g_{\rho_n,\gamma}) - \varepsilon_{\mathbf{x},\gamma}(f_\gamma). \tag{4.23}$$

Now an upper bound of the space error is provided as follows.

Theorem 4.6. *Given $0 < \delta < 1$, when $n > N_\gamma$ given by* (4.22), *with confidence at least $1 - 3\delta$, there holds*

$$\varepsilon_{\mathbf{x},\gamma}(f_{\mathbf{x},\gamma}) - \varepsilon_{\mathbf{x},\gamma}(f_\gamma) \leqslant 2D_1(n, \delta, \gamma)\big(\varepsilon_{\mathbf{x},\gamma}(f_\gamma) + D_2(n, \delta, \gamma)\big) + D_2(n, \delta, \gamma), \tag{4.24}$$

where

$$D_2(n, \delta, \gamma) := \frac{32(\kappa^4 + l\kappa^2/2)(2\kappa^4 + \gamma)\mathcal{D}(\gamma)}{l^2\gamma}\left(\frac{32(\kappa^4 + l\kappa^2/2)\log^2(2/\delta)}{l^2 n} + \frac{2\log(2/\delta)}{\sqrt{n}}\right).$$

Proof. Clearly,

$$\mathcal{E}_{\mathbf{x},\gamma}(g_{\rho_n,\gamma}) - \mathcal{E}_{\mathbf{x},\gamma}(f_\gamma) = \mathcal{E}_{\mathbf{x}}(g_{\rho_n,\gamma}) - \mathcal{E}_{\mathbf{x}}(f_\gamma) + \gamma\big(\|g_{\rho_n,\gamma}\|_K^2 - \|f_\gamma\|_K^2\big).$$

The definition of $\mathcal{E}_{\mathbf{x}}(f)$ (4.1), in connection with (2.4), verifies

$$\mathcal{E}_{\mathbf{x}}(g_{\rho_n,\gamma}) - \mathcal{E}_{\mathbf{x}}(f_\gamma) \leqslant 2\kappa^4\|g_{\rho_n,\gamma} - f_\gamma\|_K\big(\|g_{\rho_n,\gamma} - f_\gamma\|_K + 2\|f_\gamma\|_K\big).$$

Moreover, $\|g_{\rho_n,\gamma}\|_K^2 - \|f_\gamma\|_K^2 \leqslant \|g_{\rho_n,\gamma} - f_\gamma\|_K(\|g_{\rho_n,\gamma} - f_\gamma\|_K + 2\|f_\gamma\|_K)$. Hence

$$\mathcal{E}_{\mathbf{x},\gamma}(g_{\rho_n,\gamma}) - \mathcal{E}_{\mathbf{x},\gamma}(f_\gamma) \leqslant \big(2\kappa^4 + \gamma\big)\|g_{\rho_n,\gamma} - f_\gamma\|_K\big(\|g_{\rho_n,\gamma} - f_\gamma\|_K + 2\|f_\gamma\|_K\big). \tag{4.25}$$

When $n > 64l^{-2}\kappa^4\log^2(2/\delta)$, by (4.7) and (4.12), with confidence at least $1-\delta$,

$$\|p_{\mathbf{x}}\|_K \leqslant \|p\|_K + \frac{4\kappa\log(2/\delta)}{\sqrt{n}} \leqslant \kappa + \frac{l}{2\kappa}.$$

Combining the estimates in (4.16) and (4.21), this yields

$$\|g_{\rho_n,\gamma} - f_\gamma\|_K = I_{\mathbf{x}}^{-1}I_1(f_\gamma)\|p_{\mathbf{x}}\|_K \leqslant \frac{32(\kappa^3 + l\kappa/2)\|f_\gamma\|_\infty\log(2/\delta)}{l^2\sqrt{n}}, \tag{4.26}$$

with confidence at least $1-2\delta$. Then it follows from (4.19), (4.25) and (4.26)

$$\mathcal{E}_{\mathbf{x},\gamma}(g_{\rho_n,\gamma}) - \mathcal{E}_{\mathbf{x},\gamma}(f_\gamma) \leqslant D_2(n,\delta,\gamma). \tag{4.27}$$

On the other hand, the definition of $h_{\rho_n,\gamma}$ tells us

$$\mathcal{E}_{\mathbf{x},\gamma}(h_{\rho_n,\gamma}) - \mathcal{E}_{\mathbf{x},\gamma}(g_{\rho_n,\gamma}) = \big(\hat{\omega}_{g_{\rho_n,\gamma}}^{-1} - 1\big)\mathcal{E}_{\mathbf{x},\gamma}(g_{\rho_n,\gamma}).$$

Appealing to Proposition 4.5 and (4.27), when $n > N_\gamma$, there holds

$$\mathcal{E}_{\mathbf{x},\gamma}(h_{\rho_n,\gamma}) - \mathcal{E}_{\mathbf{x},\gamma}(g_{\rho_n,\gamma}) \leqslant 2D_1(n,\delta,\gamma)\big(\mathcal{E}_{\mathbf{x},\gamma}(f_\gamma) + D_2(n,\delta,\gamma)\big), \tag{4.28}$$

with confidence at least $1-3\delta$. By (4.23), (4.27) and (4.28), the proof is complete. □

The bound (4.24) contains the term $\mathcal{E}_{\mathbf{x},\gamma}(f_\gamma) = \mathcal{E}_{\mathbf{x}}(f_\gamma) - \mathcal{E}(f_\gamma) + \gamma\|f_\gamma\|_K^2$. Note that $\gamma\|f_\gamma\|_K^2 \leqslant \mathcal{D}(\gamma)$ and the bound of $\mathcal{E}_{\mathbf{x}}(f_\gamma) - \mathcal{E}(f_\gamma)$ is given in Section 4.3. Hence Theorem 4.2 and Proposition 4.7 would be applied to estimate (4.24).

4.3. The sample error

Recall that the sample error is $(\mathcal{E}(f_{\mathbf{x},\gamma}) - \mathcal{E}_{\mathbf{x}}(f_{\mathbf{x},\gamma})) + (\mathcal{E}_{\mathbf{x}}(f_\gamma) - \mathcal{E}(f_\gamma))$. We begin with the second term. It can be estimated by a probability inequality for random variables taking values in a Hilbert space.

To simplify computations we split $\mathcal{E}_{\mathbf{x}}(f_\gamma) - \mathcal{E}(f_\gamma)$ in four summands

$$\mathcal{E}_{\mathbf{x}}(f_\gamma) - \mathcal{E}(f_\gamma) = A_\gamma + B_\gamma + C_\gamma + D_\gamma,$$

where

$$A_\gamma = \frac{1}{n}\sum_{j=1}^n\left(\frac{1}{n}\sum_{i=1}^n f_\gamma^2(x_i)K(x_i,x_j) - \int f_\gamma^2(x)K(x,x_j)\,d\rho(x)\right),$$

$$B_\gamma = \int f_\gamma^2(x)\left(\frac{1}{n}\sum_{j=1}^n K(x,x_j) - \int K(x,y)\,d\rho(y)\right)d\rho(x),$$

$$C_\gamma = \frac{1}{n}\sum_{j=1}^n f_\gamma(x_j)\left(\int f_\gamma(x)K(x,x_j)\,d\rho(x) - \frac{1}{n}\sum_{i=1}^n f_\gamma(x_i)K(x_i,x_j)\right),$$

$$D_\gamma = \int f_\gamma(x)\left(\int f_\gamma(y)K(x,y)\,d\rho(y) - \frac{1}{n}\sum_{j=1}^n f_\gamma(x_j)K(x,x_j)\right)d\rho(x).$$

It remains to bound each of the terms A_γ, B_γ, C_γ and D_γ, which we do in the following proposition. Then an upper bound of $\mathcal{E}_{\mathbf{x}}(f_\gamma) - \mathcal{E}(f_\gamma)$ is obtained.

Proposition 4.7. *For all $0<\delta<1$, with confidence at least $1-3\delta$, there holds*

$$\varepsilon_{\mathbf{x}}(f_\gamma)-\varepsilon(f_\gamma)\leqslant\frac{16\kappa^4\mathcal{D}(\gamma)\log(2/\delta)}{\gamma\sqrt{n}}.$$

Proof. First note that $|A_\gamma+B_\gamma|=I_2(f_\gamma)$, given in Lemma 4.4. It follows from (4.17)

$$|A_\gamma+B_\gamma|\leqslant\frac{8\kappa^2\|f_\gamma\|_\infty^2\log(2/\delta)}{\sqrt{n}},\tag{4.29}$$

with confidence at least $1-2\delta$.

Next consider C_γ and D_γ. Let $\xi(x)=f_\gamma(x)K(x,\cdot)$. It is easy to see

$$|C_\gamma|\leqslant\frac{1}{n}\sum_{j=1}^n\big|f_\gamma(x_j)\big|\cdot\left|(E\xi)(x_j)-\left(\frac{1}{n}\sum_{i=1}^n\xi(x_i)\right)(x_j)\right|,$$

$$|D_\gamma|\leqslant\int\big|f_\gamma(x)\big|\cdot\left|(E\xi)(x)-\left(\frac{1}{n}\sum_{j=1}^n\xi(x_j)\right)(x)\right|d\rho(x).$$

Consequently, by (2.4) with confidence at least $1-\delta$,

$$|C_\gamma+D_\gamma|\leqslant2\|f_\gamma\|_\infty\cdot\kappa\left\|\frac{1}{n}\sum_{i=1}^n\xi_{x_i}-\mathbf{E}_x\xi_x\right\|_K\leqslant\frac{8\kappa^2\|f_\gamma\|_\infty^2\log(2/\delta)}{\sqrt{n}},\tag{4.30}$$

where the last equality holds due to Lemma 4.3.

Our desired estimate follows from (4.19), (4.29) and (4.30). □

The estimation of $\varepsilon(f_{\mathbf{x},\gamma})-\varepsilon_{\mathbf{x}}(f_{\mathbf{x},\gamma})$ is more involved, since $f_{\mathbf{x},\gamma}$ changes with the sample $\mathbf{x}$. To bound this term, we shall provide the following lemma. It can be easily derived from (2.4) and Lemma 4.3 by the method of proving [5, Theorem 3.10].

Lemma 4.8. *Given $f\in B_R$, let $\xi^f(x)=f(x)K(x,\cdot)$ and $\eta^f(x)=f^2(x)K(x,\cdot)$ be random variables on (X,ρ) with values in $\mathcal{H}_K$. Then for all $\varepsilon>0$,*

$$\operatorname*{Prob}_{\mathbf{x}\in X^n}\left\{\sup_{f\in B_R}\left\|\frac{1}{n}\sum_{i=1}^n\xi^f(x_i)-E\xi^f\right\|_K\geqslant\varepsilon\right\}\leqslant2\mathcal{N}\left(B_R,\frac{\varepsilon}{4\kappa}\right)\exp\left\{-\frac{\sqrt{n}\,\varepsilon}{8\kappa^2R}\right\},$$

$$\operatorname*{Prob}_{\mathbf{x}\in X^n}\left\{\sup_{f\in B_R}\left\|\frac{1}{n}\sum_{i=1}^n\eta^f(x_i)-E\eta^f\right\|_K\geqslant\varepsilon\right\}\leqslant2\mathcal{N}\left(B_R,\frac{\varepsilon}{8\kappa^2R}\right)\exp\left\{-\frac{\sqrt{n}\,\varepsilon}{8\kappa^3R^2}\right\}.$$

The essential difference between the above lemma and Lemma 4.3 is the inclusion of the covering number, which extends the results from a single random variable to a family of variables.

Now we bound the first term of the sample error by Lemma 4.8 with f replaced by $f_{\mathbf{x},\gamma}$ as follows.

Proposition 4.9. *For any $\varepsilon>0$ and $R>0$, there holds $\varepsilon(f_{\mathbf{x},\gamma})-\varepsilon_{\mathbf{x}}(f_{\mathbf{x},\gamma})\leqslant\kappa(\kappa R+1)^2\varepsilon$, with confidence at least*

$$1-\operatorname*{Prob}_{\mathbf{x}\in X^n}\{f_{\mathbf{x},\gamma}\notin B_R\}-2\left(\mathcal{N}\left(\frac{\varepsilon}{8\kappa^2R^2}\right)e^{-\sqrt{n}\varepsilon/8\kappa^3R^2}+\mathcal{N}\left(\frac{\varepsilon}{4\kappa R}\right)e^{-\sqrt{n}\varepsilon/8\kappa^2R}+e^{-\sqrt{n}\varepsilon/4\kappa}\right).$$

Proof. Note that $\varepsilon(f_{\mathbf{x},\gamma})-\varepsilon_{\mathbf{x}}(f_{\mathbf{x},\gamma})$ can be decomposed as

$$\varepsilon(f_{\mathbf{x},\gamma})-\varepsilon_{\mathbf{x}}(f_{\mathbf{x},\gamma})=-A_{\mathbf{x}}-B_{\mathbf{x}}-C_{\mathbf{x}}-D_{\mathbf{x}},\tag{4.31}$$

where $A_{\mathbf{x}},B_{\mathbf{x}},C_{\mathbf{x}}$ and $D_{\mathbf{x}}$ are defined as $A_\gamma,B_\gamma,C_\gamma,D_\gamma$ with f_γ replaced by $f_{\mathbf{x},\gamma}$.

By (2.4) and Lemma 4.8 with f replaced by $f_{\mathbf{x},\gamma}$,

$$|A_{\mathbf{x}}|\leqslant\kappa\left\|\frac{1}{n}\sum_{i=1}^n\eta^{f_{\mathbf{x},\gamma}}(x_i)-E\eta^{f_{\mathbf{x},\gamma}}\right\|_K\leqslant\kappa\sup_{f\in B_R}\left\|\frac{1}{n}\sum_{i=1}^n\eta^f(x_i)-E\eta^f\right\|_K\leqslant\kappa\varepsilon,\tag{4.32}$$

with confidence at least $1-\text{Prob}_{\mathbf{x}\in X^n}\{f_{\mathbf{x},\gamma}\notin B_R\}-2\mathcal{N}(B_R,\varepsilon/8\kappa^2R)\exp\{-\sqrt{n}\varepsilon/8\kappa^3R^2\}$. Similarly, since $\|f_{\mathbf{x},\gamma}\|_\infty\leqslant\kappa\|f_{\mathbf{x},\gamma}\|_K\leqslant\kappa R$, there holds

$$|C_{\mathbf{x}}+D_{\mathbf{x}}| \leqslant 2\|f_{\mathbf{x},\gamma}\|_\infty \cdot \kappa \sup_{f\in B_R}\left\| \frac{1}{n}\sum_{i=1}^{n} \xi^f(x_i) - E\xi^f \right\|_K \leqslant 2\kappa^2 R\varepsilon, \tag{4.33}$$

with confidence at least $1-\mathrm{Prob}_{\mathbf{x}\in X^n}\{f_{\mathbf{x},\gamma}\notin B_R\} - 2\mathcal{N}(B_R,\varepsilon/4\kappa)\exp\{-\sqrt{n}\,\varepsilon/8\kappa^2 R\}$.

Let $\varepsilon = 4\kappa\log(2/\delta)/\sqrt{n}$. It follows from (4.12)

$$|B_{\mathbf{x}}| \leqslant \kappa\|f_{\mathbf{x},\gamma}\|_\infty^2 \|p_{\mathbf{x}}-p\|_K \leqslant \kappa^3 R^2\varepsilon, \tag{4.34}$$

with confidence $1-\mathrm{Prob}_{\mathbf{x}\in X^n}\{f_{\mathbf{x},\gamma}\notin B_R\} - 2\exp\{-\sqrt{n}\,\varepsilon/4\kappa\}$. This, in connection with (4.32) and (4.33), verifies

$$\mathcal{E}(f_{\mathbf{x},\gamma}) - \mathcal{E}_{\mathbf{x}}(f_{\mathbf{x},\gamma}) \leqslant |A_{\mathbf{x}}| + |B_{\mathbf{x}}| + |C_{\mathbf{x}}+D_{\mathbf{x}}| \leqslant \kappa(\kappa R+1)^2\varepsilon,$$

with confidence at least

$$1-\underset{\mathbf{x}\in X^n}{\mathrm{Prob}}\{f_{\mathbf{x},\gamma}\notin B_R\} - 2\left\{\mathcal{N}\left(B_R,\frac{\varepsilon}{8\kappa^2 R}\right)\exp\left\{-\frac{\sqrt{n}\,\varepsilon}{8\kappa^3R^2}\right\} + \mathcal{N}\left(B_R,\frac{\varepsilon}{4\kappa}\right)\exp\left\{-\frac{\sqrt{n}\,\varepsilon}{8\kappa^2 R}\right\} + \exp\left\{-\frac{\sqrt{n}\,\varepsilon}{4\kappa}\right\}\right\}.$$

Note that an $\frac{\alpha}{R}-$covering of B_1 yields an $\alpha-$covering of B_R and viceversa. So the proof is complete. □

The following conclusion [18,19] concerning the covering number is required to realize the confidence in Proposition 4.9.

Lemma 4.10. *For $0<\delta<1$, there exists a unique $\zeta(n,\delta)$ such that*

$$\log\mathcal{N}\big(\zeta(n,\delta)\big) - \frac{\sqrt{n}\,\zeta(n,\delta)}{\beta} = \log\delta.$$

Moreover, if the Mercer kernel K satisfies $\log\mathcal{N}(\zeta)\leqslant C_0(1/\zeta)^s$ for some $s>0$, then

$$\zeta(n,\delta) \leqslant \max\left\{\frac{2\beta\log(1/\delta)}{\sqrt{n}}, (2\beta C_0/\sqrt{n})^{1/(s+1)}\right\}.$$

Here β and C_0 are positive constants.

For the estimation of sample error, we shall require the confidence $\mathcal{N}(\frac{\varepsilon}{8\kappa^2R^2})\exp\{-\frac{\sqrt{n}\,\varepsilon}{8\kappa^3R^2}\}$ to be at most $\delta>0$. Appealing to (4.32) and Lemma 4.10 with $\beta=\kappa$, with confidence at least $1-2\delta-\mathrm{Prob}_{\mathbf{x}\in X^n}\{f_{\mathbf{x},\gamma}\notin B_R\}$,

$$|A_{\mathbf{x}}| \leqslant 8\kappa^3R^2\zeta_1(n,\delta) \leqslant \max\left\{\frac{16\kappa^4R^2\log(1/\delta)}{\sqrt{n}}, 8\kappa^3R^2\left(\frac{2\kappa C_0}{\sqrt{n}}\right)^{1/(s+1)}\right\}. \tag{4.35}$$

Similarly, by (4.33) and Lemma 4.10 with $\beta=2\kappa$, there holds

$$|C_{\mathbf{x}}+D_{\mathbf{x}}| \leqslant 8\kappa^3R^2\zeta_2(n,\delta) \leqslant \max\left\{\frac{32\kappa^4R^2\log(1/\delta)}{\sqrt{n}}, 8\kappa^3R^2\left(\frac{4\kappa C_0}{\sqrt{n}}\right)^{1/(s+1)}\right\}, \tag{4.36}$$

with confidence at least $1-2\delta-\mathrm{Prob}_{\mathbf{x}\in X^n}\{f_{\mathbf{x},\gamma}\notin B_R\}$.

Provided with these results, we derive an upper bound of the sample error in the following theorem.

Theorem 4.11. *Assume that the Mercer kernel K satisfies $\log\mathcal{N}(\zeta)\leqslant C_0(1/\zeta)^s$ for some $s>0$. Then for $R>0$ and $0<\delta<1$, with confidence at least $1-5\delta-\mathrm{Prob}_{\mathbf{x}\in X^n}\{f_{\mathbf{x},\gamma}\notin B_R\}$, there holds*

$$\mathcal{E}(f_{\mathbf{x},\gamma}) - \mathcal{E}_{\mathbf{x}}(f_{\mathbf{x},\gamma}) + \mathcal{E}_{\mathbf{x}}(f_\gamma) - \mathcal{E}(f_\gamma) \leqslant \frac{16\kappa^4\mathcal{D}(\gamma)\log(2/\delta)}{\gamma\sqrt{n}} + 3R^2\mu(n,\delta),$$

where

$$\mu(n,\delta) = \max\left\{\frac{32\kappa^4\log(2/\delta)}{\sqrt{n}}, 16\kappa^3\left(\frac{2\kappa C_0}{\sqrt{n}}\right)^{1/(s+1)}\right\}.$$

Proof. Taking $\varepsilon = 4\kappa\log(2/\delta)/\sqrt{n}$, (4.34) tells us

$$|B_{\mathbf{x}}| \leqslant \frac{4\kappa^4R^2\log(2/\delta)}{\sqrt{n}} \leqslant R^2\mu(n,\delta),$$

with confidence $1-\delta-\mathrm{Prob}_{\mathbf{x}\in X^n}\{f_{\mathbf{x},\gamma}\notin B_R\}$. Then appealing to (4.35) and (4.36) after replacing δ by $\delta/2$, it is easy to see that $\mathcal{E}(f_{\mathbf{x},\gamma}) - \mathcal{E}_{\mathbf{x}}(f_{\mathbf{x},\gamma}) \leqslant 3R^2\mu(n,\delta)$. This, in connection with Proposition 4.7, proves out statement. □

Remark 4.12. Note that for $n \geqslant 4\kappa^2 C_0^{-2/s}(\log(2/\delta))^{(2+2/s)}$, the quantity $\mu(n,\delta)$ is dominated by the term $16\kappa^3(2\kappa C_0/\sqrt{n})^{1/(s+1)}$.

4.4. Bounding function $f_{\mathbf{x},\gamma}$

In this section, our task is to choose a suitable $R = R_\gamma > 0$ such that $\mathrm{Prob}_{\mathbf{x}\in X^n}\{f_{\mathbf{x},\gamma} \notin B_R\}$ is small enough for a sufficiently large n. In this regard, the results of the perturbation theory on the eigenvalues of the integral operator S_K (given by (3.9)) play an important role.

We now provide the following estimates for the approximation of eigenvalues of S_K by those of the operator $S_{\mathbf{x}}$, which is defined as

$$S_{\mathbf{x}} f(x) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)K(x_i,x), \quad f \in \mathcal{H}_K.$$

Clearly, it is positive semi-definite by $\langle g, S_{\mathbf{x}} f\rangle_K = \frac{1}{n}\sum_{i=1}^{n} f(x_i)g(x_i)$, $f,g\in\mathcal{H}_K$. Let $r_1^{\mathbf{x}} \geqslant r_2^{\mathbf{x}} \geqslant \cdots \geqslant r_n^{\mathbf{x}} \geqslant 0$ be its eigenvalues. The following results are derived from [12, Propositions 1 and 2].

Lemma 4.13. *Let $\{\lambda_i\}_{i=1}^{\infty}$ with $0 \leqslant \cdots \leqslant \lambda_2 \leqslant \lambda_1$ be the set of eigenvalues of S_K. Then for any $\delta\in(0,1)$, with confidence at least $1-\delta$*

$$\left|\lambda_i - r_i^{\mathbf{x}}\right| \leqslant \frac{4\kappa^2\log(2/\delta)}{\sqrt{n}}, \quad i=1,\ldots,n. \tag{4.37}$$

Denote the eigenfunctions of $S_{\mathbf{x}}$ by $f_1^{\mathbf{x}},\ldots,f_n^{\mathbf{x}}$ with corresponding eigenvalues $r_1^{\mathbf{x}},\ldots,r_n^{\mathbf{x}}$, i.e., $S_{\mathbf{x}} f_i^{\mathbf{x}} = r_i^{\mathbf{x}} f_i^{\mathbf{x}}$. It yields

$$\frac{1}{n}\mathbf{K}\hat{f}_i^{\mathbf{x}} = r_i^{\mathbf{x}}\hat{f}_i^{\mathbf{x}}, \quad l=1,\ldots,n.$$

If $r_i^{\mathbf{x}} > 0$, $i\in\{1,2,\ldots,n\}$, we claim $\hat{f}_i^{\mathbf{x}}$ is not zero, otherwise $f_i^{\mathbf{x}}$ is zero, a contradiction. Therefore, $nr_i^{\mathbf{x}}$ is the eigenvalues of the matrix $\mathbf{K}$, and $\lambda_i^{\mathbf{x}} = nr_i^{\mathbf{x}}$, if the eigenvalues of $\mathbf{K}$ by $\{\lambda_i^{\mathbf{x}}\}_{i=1}^{n}$ is denoted by $\lambda_1^{\mathbf{x}} \geqslant \lambda_2^{\mathbf{x}} \geqslant \cdots \geqslant \lambda_n^{\mathbf{x}} \geqslant 0$.

Recall that $f_{\mathbf{x},\gamma}$ can be represented as

$$f_{\mathbf{x},\gamma} = \sum_{i=1}^{n} \alpha_i^{\mathbf{x}} K(x_i,\cdot),$$

where $\alpha^{\mathbf{x}}$ is given by (2.7). To bound $\|f_{\mathbf{x},\gamma}\|_K$, a vector $\alpha\in\mathbb{R}^n$ is firstly constructed with the eigenvectors $e_1^{\mathbf{x}} = (e_{11}^{\mathbf{x}},\ldots,e_{1n}^{\mathbf{x}})^T$, $e_2^{\mathbf{x}} = (e_{21}^{\mathbf{x}},\ldots,e_{2n}^{\mathbf{x}})^T \in\mathbb{R}^n$ of $\mathbf{K}$, corresponding to the largest two eigenvalues $\lambda_1^{\mathbf{x}}$ and $\lambda_2^{\mathbf{x}}$.

Proposition 4.14. *Assume $\lambda_2 > 0$. For any $0<\delta<1$, when $n > M_\delta$, with confidence at least $1-2\delta$, there exists a vector $\alpha = \alpha(\mathbf{x})\in\mathbb{R}^n$ satisfying*

$$\frac{1}{n^2}\alpha^T\mathbf{K}D\mathbf{K}\alpha = 1 \quad \text{and} \quad \alpha^T\mathbf{K}D\mathbf{1} = 0, \tag{4.38}$$

and for some constants c_1 and c_2 depending on $\mathbf{x}$

$$\alpha = \frac{c_1}{\lambda_1^{\mathbf{x}}}e_1^{\mathbf{x}} + \frac{c_2}{\lambda_2^{\mathbf{x}}}e_2^{\mathbf{x}}. \tag{4.39}$$

Here, as in Theorem 3.3, $M_\delta = \max\{64l^{-2}\kappa^4\log^2(2/\delta), 64\lambda_2^{-2}\kappa^4\log^2(2/\delta)\}$.

Proof. Since $\lambda_2 > 0$, when $n > 64\lambda_2^{-2}\kappa^4\log^2(2/\delta)$, (4.37) implies that $r_1^{\mathbf{x}} \geqslant r_2^{\mathbf{x}} \geqslant \lambda_2/2 > 0$, with confidence at least $1-\delta$. Consequently,

$$\lambda_i^{\mathbf{x}}/n = r_i^{\mathbf{x}} \geqslant \lambda_2/2 > 0, \quad i=1,2. \tag{4.40}$$

Moreover, recall that $D = \mathrm{diag}\{D_{11},\ldots,D_{nn}\}$ with $D_{ii} = \sum_{j=1}^{n} K(x_i,x_j)$. When $n > 64l^{-2}\kappa^4\log^2(2/\delta)$, it follows from (4.20) that

$$D_{ii} = np_{\mathbf{x}}(x_i) \geqslant nl/2. \tag{4.41}$$

It yields $e^T De \geqslant (nl/2)e^T e > 0$, for any $0 \neq e\in\mathbb{R}^n$, with confidence at least $1-\delta$. Therefore, if ${e_2^{\mathbf{x}}}^T D\mathbf{1} \neq 0$,

$$I_\theta := {e_1^{\mathbf{x}}}^T De_1 + 2\theta {e_1^{\mathbf{x}}}^T De_2^{\mathbf{x}} - \theta^2 {e_2^{\mathbf{x}}}^T De_2^{\mathbf{x}} = \left(e_1^{\mathbf{x}} - \theta e_2^{\mathbf{x}}\right)^T D\left(e_1^{\mathbf{x}} - \theta e_2^{\mathbf{x}}\right) > 0,$$

where $\theta = \frac{e_1^{\mathbf{x}T} D\mathbf{1}}{e_2^{\mathbf{x}T} D\mathbf{1}}$. Then with confidence at least $1-2\delta$, there exists a vector $\alpha \in \mathbb{R}^n$ defined as (4.39) with $c_1 = I_\theta^{-1/2} n$ and $c_2 = -\theta I_\theta^{-1/2} n$. Furthermore,

$$\alpha^T \mathbf{K} D\mathbf{1} = c_1 e_1^{\mathbf{x}T} D\mathbf{1} + c_2 e_2^{\mathbf{x}T} D\mathbf{1} = 0,$$

$$\frac{1}{n^2}\alpha^T \mathbf{K} D \mathbf{K}\alpha = I_\theta^{-1}\big(e_1^{\mathbf{x}T} D e_1^{\mathbf{x}} + 2\theta e_1^{\mathbf{x}T} D e_2^{\mathbf{x}} + \theta^2 e_2^{\mathbf{x}T} D e_2^{\mathbf{x}}\big) = 1.$$

If $e_2^{\mathbf{x}T} D\mathbf{1} = 0$, such a vector α also exists, satisfying (4.38) and (4.39) where $c_1 = 0$ and $c_2 = n(e_2^{\mathbf{x}T} D e_2^{\mathbf{x}})^{-1}$. The proof is complete. □

Now a bound of $\|f_{\mathbf{x},\gamma}\|_K$ can be derived from Proposition 4.14.

Proposition 4.15. *Assume* $\lambda_2 > 0$. *Then for any* $0 < \delta < 1$, *when* $n > M_\delta$, *with confidence at least* $1 - 2\delta$,

$$\|f_{\mathbf{x},\gamma}\|_K \leqslant \sqrt{\gamma^{-1} + l^{-1}\kappa^2}.$$

Proof. Appealing to Proposition 4.14, there exits a vector α satisfying (4.38) and (4.39), with confidence at least $1-2\delta$. It follows from (2.7)

$$\frac{1}{n^2}\alpha^{\mathbf{x}T}\mathbf{K} L_n \mathbf{K}\alpha^{\mathbf{x}} + \gamma \alpha^{\mathbf{x}T}\mathbf{K}\alpha^{\mathbf{x}} \leqslant \frac{1}{n^2}\alpha^T \mathbf{K} L_n \mathbf{K}\alpha + \gamma\alpha^T\mathbf{K}\alpha. \tag{4.42}$$

Recall that $L_n = D - \mathbf{K}$. (4.39) verifies

$$\frac{1}{n^2}\alpha^T\mathbf{K} L_n \mathbf{K}\alpha = 1 - \frac{1}{n^2}\big(c_1^2\lambda_1^{\mathbf{x}} + c_2^2\lambda_2^{\mathbf{x}}\big) \leqslant 1 - \big(c_1^2 + c_2^2\big)\lambda_2^{\mathbf{x}}/n^2. \tag{4.43}$$

If $e_2^{\mathbf{x}T} D\mathbf{1} \neq 0$, by the proof of Proposition 4.14,

$$c_1^2 + c_2^2 = I_\theta^{-1} n^2\big(1+\theta^2\big) \geqslant \frac{n}{\kappa^2},$$

since $I_\theta \leqslant n\kappa^2 (e_1^{\mathbf{x}} - \theta e_2^{\mathbf{x}})^T (e_1^{\mathbf{x}} - \theta e_2^{\mathbf{x}}) = n\kappa^2(1+\theta^2)$. Then it follows from (4.43)

$$\frac{1}{n^2}\alpha^T\mathbf{K} L_n\mathbf{K}\alpha \leqslant 1 - \frac{\lambda_2^{\mathbf{x}}}{n\kappa^2}. \tag{4.44}$$

On the other hand, when $n > 64 l^{-2}\kappa^4 \log^2(2/\delta)$, (4.41) tells us with confidence at least $1-\delta$

$$I_\theta = \sum_{i=1}^n D_{ii}(e_{1i} - \theta e_{2i})^2 \geqslant \frac{nl}{2}\big(1+\theta^2\big).$$

It implies $c_1^2 + c_2^2 \leqslant 2 l^{-1} n$, and thus

$$\alpha^T\mathbf{K}\alpha = \left(\frac{c_1}{\lambda_1^{\mathbf{x}}} e_1^{\mathbf{x}} + \frac{c_2}{\lambda_2^{\mathbf{x}}} e_2^{\mathbf{x}}\right)^T \big(c_1 e_1^{\mathbf{x}} + c_2 e_2^{\mathbf{x}}\big) = \frac{c_1^2}{\lambda_1^{\mathbf{x}}} + \frac{c_2^2}{\lambda_2^{\mathbf{x}}} \leqslant \frac{2n}{l\lambda_2^{\mathbf{x}}}. \tag{4.45}$$

Since $\varepsilon_{\mathbf{x}}(f) \geqslant 0$, by (4.40), (4.42), (4.44), and (4.45),

$$\gamma\|f_{\mathbf{x},\gamma}\|_K^2 = \gamma\alpha^{\mathbf{x}T}\mathbf{K}\alpha^{\mathbf{x}} \leqslant 1 - \frac{\lambda_2^{\mathbf{x}}}{n\kappa^2} + \frac{2\gamma n}{l\lambda_2^{\mathbf{x}}} \leqslant 1 - \frac{\lambda_2}{2\kappa^2} + \frac{4\gamma}{l\lambda_2}.$$

Note that $0 < \lambda_2 \leqslant \|S_K\| \leqslant \kappa^2$ [5, Proposition 4.5, Theorem 4.7]. It yields

$$1 + \frac{4\gamma}{l\lambda_2} \geqslant 1 - \frac{\lambda_2}{2\kappa^2} + \frac{4\gamma}{l\lambda_2} \geqslant \frac{1}{2} + \frac{4\gamma}{l\lambda_2} > 0.$$

Consequently,

$$\gamma\|f_{\mathbf{x},\gamma}\|_K^2 \leqslant 1 + \frac{4\gamma}{l\lambda_2}.$$

If $e_2^{\mathbf{x}T} D\mathbf{1} = 0$, by $c_1 = 0$ and $c_2 = n(e_2^{\mathbf{x}T} D e_2^{\mathbf{x}})^{-1}$, the same bound is also derived. □

Remark 4.16. Taking $R = (\gamma^{-1} + 4 l^{-1}\lambda_2^{-1})^{1/2}$, Proposition 4.15 yields that $\mathrm{Prob}_{\mathbf{x}\in X^n}\{f_{\mathbf{x},\gamma} \notin B_R\} \leqslant 2\delta$.

4.5. Proof of Theorem 4.1

Proof. Take $\gamma = n^{-\theta}$ with $\theta = 1/(2(1+s)(1+\alpha))$. Our statement follows directly from Theorems 4.2, 4.6, 4.11 with $R = (\gamma^{-1} + 4l^{-1}\lambda_2^{-1})^{1/2}$ and Proposition 4.15. □

5. Proof of Theorem 3.3

In this section, we would prove the main theorem, Theorem 3.3, with the help of (3.5) and Theorem 4.1. To this end, the technique **PN** should be applied to $f_{\mathbf{x},\gamma}$, since (3.5) holds only for functions in $\mathcal{B}$ but $f_{\mathbf{x},\gamma}$ may not. Of course, the condition $\omega_{g_{\mathbf{x},\gamma}} > 0$ is required, which will be given by the following corollary.

Corollary 5.1. *Assume* $\lambda_2 > 0$. *Then for every* $0 < \delta \leqslant 1$, *with confidence at least* $1 - 5\delta$, *there holds*

$$I_1(f_{\mathbf{x},\gamma}) \leqslant \kappa^{-1} R\mu(n,\delta), \qquad I_2(f_{\mathbf{x},\gamma}) \leqslant R^2\mu(n,\delta),$$

where $R = (\gamma^{-1} + 4l^{-1}\lambda_2^{-1})^{1/2}$, $I_j(f)$, $j = 1, 2$, *are defined as in Lemma* 4.4 *and* $\mu(n,\delta)$ *is given in Theorem* 4.11.

Proof. It follows directly from Lemmas 4.4, 4.8 and 4.10 that

$$I_1(f_{\mathbf{x},\gamma}) \leqslant \kappa^2 R\|p - p_{\mathbf{x}}\|_K + \kappa \sup_{f\in B_R}\left\|\frac{1}{n}\sum_{i=1}^{n}\xi^f(x_i) - \mathbf{E}\xi^f\right\|_K \leqslant \kappa^{-1}R\mu(n,\delta),$$

$$I_2(f_{\mathbf{x},\gamma}) \leqslant \kappa^3 R^2\|p - p_{\mathbf{x}}\|_K + \kappa \sup_{f\in B_R}\left\|\frac{1}{n}\sum_{i=1}^{n}\eta^f(x_i) - \mathbf{E}\eta^f\right\|_K \leqslant R^2\mu(n,\delta),$$

we have with confidence at least $1 - 3\delta - \mathrm{Prob}_{\mathbf{x}\in X^n}\{f_{\mathbf{x},\gamma} \notin B_R\}$. The proof is complete by Remark 4.16. □

After replacing f with $f_{\mathbf{x},\gamma}$, denote the functions g and h in **PN** by $g_{\mathbf{x},\gamma}$ and $h_{\mathbf{x},\gamma}$ respectively. The existence condition of $h_{\mathbf{x},\gamma}$, $\omega_{g_{\mathbf{x},\gamma}} > 0$, is provided in the following.

Proposition 5.2. *Assume* $\lambda_2 > 0$. *Take* $R = (\gamma^{-1} + 4l^{-1}\lambda_2^{-1})^{1/2}$. *For* $0 < \delta \leqslant 1$, *with confidence at least* $1 - 5\delta$,

$$|\omega_{g_{\mathbf{x},\gamma}} - 1| \leqslant \varepsilon_1(n,\delta) := R^2\big(\big(1 + 2l^{-2}\kappa^4\big)\mu(n,\delta) + l^{-4}\kappa^4\mu^2(n,\delta)\big). \tag{5.1}$$

Assume in addition that $n > N_R$, *where* N_R *is an integer satisfying* $\varepsilon_1(N_R,\delta) \leqslant 1/2$. *We have* $\omega_{g_{\mathbf{x},\gamma}} \geqslant 1/2$ *with the same confidence at least* $1 - 5\delta$.

Proof. Note $\frac{1}{n}\sum_{i=1}^{n} f_{\mathbf{x},\gamma}^2(x_i)p_{\mathbf{x}}(x_i) = 1$ and $\sum_{i=1}^{n} f_{\mathbf{x},\gamma}(x_i)p_{\mathbf{x}}(x_i) = 0$. Since $f_{\mathbf{x},\gamma} \in B_R$ with confidence at least $1 - 2\delta$, the definition of $\omega_{g_{\mathbf{x},\gamma}}$, in connection with (3.3), tells us

$$|\omega_{g_{\mathbf{x},\gamma}} - 1| \leqslant I_2(f_{\mathbf{x},\gamma}) + 2l^{-2}\kappa^5 R I_1(f_{\mathbf{x},\gamma}) + l^{-4}\kappa^6 I_1^2(f_{\mathbf{x},\gamma}).$$

Then (5.1) holds with confidence at least $1 - 5\delta$, due to Corollary 5.1.

Moreover, if $n > N_R$ satisfying $\varepsilon_1(N_R,\delta) \leqslant 1/2$, (5.1) yields $\omega_{g_{\mathbf{x},\gamma}} \geqslant 1/2$. □

It follows from Remark 4.12 that $\mu(n,\delta) \leqslant 16\kappa^3(2\kappa C_0/\sqrt{n})^{1/(s+1)}$, when $n \geqslant 4\kappa^2 C_0^{-2/s}(\log(2/\delta))^{(2+2/s)}$. Hence there exists a sufficient large N_R, only depending on l, κ, s, C_0, and R, such that $\varepsilon_1(N_R,\delta) \leqslant 1/2$. Then Proposition 5.2 ensures the existence of $h_{\mathbf{x},\gamma}$ with confidence at least $1 - 5\delta$.

Now the term $\|f_{\mathbf{x},\gamma} - h_{\mathbf{x},\gamma}\|_2$ is bounded as follows.

Proposition 5.3. *Assume* $\lambda_2 > 0$. *Take* $R = (\gamma^{-1} + 4l^{-1}\lambda_2^{-1})^{1/2}$. *For every* $0 < \delta \leqslant 1$, *with confidence at least* $1 - 5\delta$,

$$\|f_{\mathbf{x},\gamma} - h_{\mathbf{x},\gamma}\|_2 \leqslant l^{-1}\kappa^{-1}R\mu(n,\delta) + l^{-1/2}\varepsilon_1(n,\delta).$$

Proof. Since $\sum_{i=1}^{n} f_{\mathbf{x},\gamma}(x_i)p_{\mathbf{x}}(x_i) = 0$, it follows from (4.7) and Corollary 5.1 that

$$\|f_{\mathbf{x},\gamma} - g_{\mathbf{x},\gamma}\|_2 = \|p\|_2^{-1}\left|\int f_{\mathbf{x},\gamma}\,p\,d\rho\right| \leqslant l^{-1}I_1(f_{\mathbf{x},\gamma}) \leqslant l^{-1}\kappa^{-1}R\mu(n,\delta), \tag{5.2}$$

with confidence at least $1 - 5\delta$.

Moreover, $\omega_{g_{\mathbf{x},\gamma}} = \int g_{\mathbf{x},\gamma}^2 p\,d\rho \geqslant l \cdot \|g_{\mathbf{x},\gamma}\|_2^2$. Hence with confidence at east $1-5\delta$,

$$\|g_{\mathbf{x},\gamma} - h_{\mathbf{x},\gamma}\|_2 = \left|1-\omega_{g_{\mathbf{x},\gamma}}^{-1/2}\right| \cdot \|g_{\mathbf{x},\gamma}\|_2 \leqslant l^{-1/2}\left|\omega_{g_{\mathbf{x},\gamma}}^{1/2}-1\right| \leqslant l^{-1/2}|\omega_{g_{\mathbf{x},\gamma}}-1| \leqslant l^{-1/2}\varepsilon_1(n,\delta), \tag{5.3}$$

where the last inequality holds due to Proposition 5.2. Our desired estimate follows from (5.2) and (5.3). □

Let $\beta_{\mathbf{x},\gamma} = \operatorname{sign}\langle h_{\mathbf{x},\gamma}\sqrt{p}, f^*\sqrt{p}\rangle$. Proposition 3.1 implies

$$\left\|\beta_{\mathbf{x},\gamma}h_{\mathbf{x},\gamma} - f^*\right\|_2^2 \leqslant \frac{2(\varepsilon(h_{\mathbf{x},\gamma}) - \varepsilon(f^*))}{(\mu_2-\mu_3)l}. \tag{5.4}$$

It remains to bound the quantity $\varepsilon(h_{\mathbf{x},\gamma}) - \varepsilon(f^*)$.

Proposition 5.4. *Assume* $\lambda_2 > 0$. *Take* $R = (\gamma^{-1} + 4l^{-1}\lambda_2^{-1})^{1/2}$. *When* $n > N_R$, *with confidence at least* $1-5\delta$,

$$\varepsilon(h_{\mathbf{x},\gamma}) - \varepsilon(f^*) \leqslant \big(\varepsilon_1(n,\delta)+1\big)\big(\varepsilon_2(n,\delta) + \varepsilon(f_{\mathbf{x},\gamma}) - \varepsilon(f^*)\big) + \varepsilon_1(n,\delta)\varepsilon(f^*),$$

where $\varepsilon_2(n,\delta) = 2l^{-1}\kappa R^2(l^{-1}\kappa^{-1}\mu^2(n,\delta) + 2\kappa\mu(n,\delta))$.

Proof. Note that

$$\varepsilon(h_{\mathbf{x},\gamma}) - \varepsilon(f^*) = \big(\omega_{g_{\mathbf{x},\gamma}}^{-1} - 1\big)\varepsilon(g_{\mathbf{x},\gamma}) + \varepsilon(g_{\mathbf{x},\gamma}) - \varepsilon(f_{\mathbf{x},\gamma}) + \varepsilon(f_{\mathbf{x},\gamma}) - \varepsilon(f^*). \tag{5.5}$$

By (2.4) and the definition of $\varepsilon(f)$ in (3.1), it is easy to see

$$\varepsilon(g_{\mathbf{x},\gamma}) - \varepsilon(f_{\mathbf{x},\gamma}) \leqslant 2\kappa^2\|g_{\mathbf{x},\gamma} - f_{\mathbf{x},\gamma}\|_2\big(\|g_{\mathbf{x},\gamma} - f_{\mathbf{x},\gamma}\|_2 + 2\kappa R\big).$$

Then it follows from (5.2), with confidence at least $1-5\delta$,

$$\varepsilon(g_{\mathbf{x},\gamma}) - \varepsilon(f_{\mathbf{x},\gamma}) \leqslant \varepsilon_2(n,\delta). \tag{5.6}$$

When $n \geqslant N_R$, Proposition 5.2 ensures

$$\big(\omega_{g_{\mathbf{x},\gamma}}^{-1} - 1\big)\varepsilon(g_{\mathbf{x},\gamma}) \leqslant \varepsilon_1(n,\delta)\big(\varepsilon(f_{\mathbf{x},\gamma}) + \varepsilon_2(n,\delta)\big). \tag{5.7}$$

Note that (5.5), (5.6) and (5.7) yield our statement. □

Proof of Theorem 3.3. Appealing to Proposition 5.4 and (5.4),

$$\left\|\beta_{\mathbf{x},\gamma}h_{\mathbf{x},\gamma} - f^*\right\|_2^2 \leqslant \frac{(2\varepsilon_1(n,\delta)+2)(\varepsilon_2(n,\delta) + \varepsilon(f_{\mathbf{x},\gamma}) - \varepsilon(f^*)) + 2\varepsilon_1(n,\delta)\varepsilon(f^*)}{(\mu_2-\mu_3)l}, \tag{5.8}$$

with confidence at least $1-5\delta$.

Now take $\gamma = n^{-\theta}$ with $\theta = 1/(2(1+s)(1+\alpha))$ and $R = (\gamma^{-1} + 4l^{-1}\lambda_2^{-1})^{1/2}$. By Theorem 4.1, Proposition 5.3 and (5.8), the proof is complete. □

In this paper we have established the consistency of the regularized spectral clustering algorithm. Only the simplest form, partitioning the whole space to two clusters, is considered. Moreover, we choose a fixed kernel function when proving the convergence results. To consider the case of a sample-dependent kernel function, such as the Gaussian kernel with the bandwidth depending on the sample size, will be our future work.

References

[1] N. Aronszajn, Theory of reproducing kernels, Transactions of the American Mathematical Society 68 (1950) 337–404.

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

[3] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: A geometric framework for learning from labeled and unlabeled examples, Journal of Machine Learning Research 7 (2006) 2399–2434.

[4] F. Cucker, S. Smale, On the mathematical foundations of learning, Bulletin American Mathematical Society 39 (2002) 1–50.

[5] F. Cucker, D. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge Univ. Press, 2007.

[6] J. Hartigan, Consistency of single linkage for high-density clusters, Journal of the American Statistical Association 76 (1981) 388–394.

[7] U.V. Luxburg, O. Bousquet, M. Belkin, Limits of spectral clustering, in: Advances in Neural Information Processing Systems (NIPS), vol. 17, MIT Press, Cambridge, MA, 2005, pp. 857–864.

[8] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, The Annals of Probability 22 (1994) 1679–1706.

[9] D. Pollard, Strong consistency of k-means clustering, The Annals of Statistics 9 (1981) 135–140.

[10] L. Rosasco, M. Belkin, E. De Vito, On learning with integral operators, Journal of Machine Learning Research 11 (2010) 905–934.

[11] S. Smale, D. Zhou, Learning theory estimates via integral operators and their approximations, Constructive Approximation 26 (2007) 153–172.

[12] S. Smale, D. Zhou, Geometry on probability spaces, Constructive Approximation 30 (2009) 311–323.

[13] S. Smale, D.-X. Zhou, Shannon sampling II: Connections to learning theory, in: Computational Harmonic Analysis. Part 1, Applied and Computational Harmonic Analysis 19 (2005) 285–302.

[14] D. Spielman, S. Teng, Spectral partitioning works: Planar graphs and finite element meshes, Linear Algebra and its Applications 421 (2007) 284–305.

[15] J. Vert, Y. Yamanishi, Supervised graph inference, Advances in Neural Information Processing Systems 17 (2005) 1433–1440.

[16] U. Von Luxburg, A tutorial on spectral clustering, Statistics and Computing 17 (2007) 395–416.

[17] U. Von Luxburg, M. Belkin, O. Bousquet, Consistency of spectral clustering, Annals of Statistics 36 (2008) 555–586.

[18] Q. Wu, Y. Ying, D. Zhou, Learning rates of least-square regularized regression, Foundations of Computational Mathematics 6 (2006) 171–192.

[19] D. Zhou, The covering number in learning theory, Journal of Complexity 18 (2002) 739–767.

[20] D. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Transactions on Information Theory 49 (2003) 1743–1752.