

Reproducing Kernel Banach Spaces with the ℓ^1 Norm II: Error Analysis for Regularized Least Square Regression*

Guohui Song[†] and Haizhang Zhang[‡]

Abstract

A typical approach in estimating the learning rate of a regularized learning scheme is to bound the approximation error by the sum of the sampling error, the hypothesis error and the regularization error. Using a reproducing kernel space that satisfies the linear representer theorem brings the advantage of discarding the hypothesis error from the sum automatically. Following this direction, we illustrate how reproducing kernel Banach spaces with the ℓ^1 norm can be applied to improve the learning rate estimate of ℓ^1 -regularization in machine learning.

Keywords: reproducing kernel Banach spaces, sparse learning, regularization, least square regression, learning rate, the representer theorem

1 Introduction

A class of reproducing kernel Banach spaces (RKBS) with the ℓ^1 norm that satisfies the linear representer theorem was recently constructed in [14]. The purpose of this note is to illustrate how the obtained spaces can be applied to estimate the learning rate of the ℓ^1 -regularized least square regression in machine learning.

A general coefficient-based regularization of the least square regression has the form

$$\min_{\mathbf{c} \in \mathbb{R}^m} \frac{1}{m} \sum_{j=1}^m |K^{\mathbf{x}}(x_j)\mathbf{c} - y_j|^2 + \lambda \phi(\mathbf{c}), \quad (1.1)$$

where $\mathbf{x} := \{x_j : j \in \mathbb{N}_m\}$ with $\mathbb{N}_m := \{1, 2, \dots, m\}$ is the sequence of sampling points from an input space X , $y_j \in Y \subseteq \mathbb{R}$ is the observed data on x_j , λ is a positive regularization parameter, ϕ is a nonnegative regularization function on the coefficient column vector \mathbf{c} , and with a chosen function $K : X \times X \rightarrow \mathbb{R}$, $K^{\mathbf{x}}(x)$ is the $1 \times m$ row vector $(K(x_j, x) : j \in \mathbb{N}_m)$.

When K is a positive-definite reproducing kernel on X and

$$\phi(\mathbf{c}) := \mathbf{c}^T K[\mathbf{x}]\mathbf{c}, \quad (1.2)$$

*Supported by Guangdong Provincial Government of China through the ‘‘Computational Science Innovative Research Team’’ program.

[†]School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ 85287. E-mail address: gsong9@asu.edu.

[‡]School of Mathematics and Computational Science and Guangdong Province Key Laboratory of Computational Science, Sun Yat-sen University, Guangzhou 510275, P. R. China. E-mail address: zhhaizh2@sysu.edu.cn.

where $K[\mathbf{x}]$ is the $m \times m$ matrix defined by

$$(K[\mathbf{x}])_{j,k} := K(x_k, x_j), \quad j, k \in \mathbb{N}_m,$$

it follows from the celebrated representer theorem [7] that (1.1) is the classical regularization network and has been extensively studied in the literature [6, 9, 10, 13, 19]. Estimates for the learning rate of the regularization network can be found, for example, in [4, 5, 12, 15, 23]. Learning rates for (1.1) when $\phi(\mathbf{c}) = \sum_{j=1}^m |c_j|^p$ for $1 < p \leq 2$ and $p = 2$ were respectively obtained in [18] and [16]. The linear programming regularization where $\phi(\mathbf{c})$ is the ℓ^1 norm $\|\mathbf{c}\|_1$ of \mathbf{c} has recently attracted much attention. The increasing interest is mainly brought by the progress of the lasso in statistics [17] and compressive sensing [2, 3] in which ℓ^1 -regularization is able to yield sparse representation of the resulting minimizer, a desirable feature in model selection. Moreover, the ℓ^1 -regularization is particularly robust to non-Gaussian additive noise such as impulsive noise [1, 8].

Without making use of a reproducing kernel space, the recent references [11, 20] established estimates of the learning rate for the ℓ^1 -regularized least square regression

$$\min_{\mathbf{c} \in \mathbb{R}^m} \frac{1}{m} \sum_{j=1}^m |K^{\mathbf{x}}(x_j)\mathbf{c} - y_j|^2 + \lambda \|\mathbf{c}\|_1. \quad (1.3)$$

We attempt to show that improvement on the estimates could be made if an RKBS with the ℓ^1 norm is used. To explain how this could be done, we first introduce the popular approach [5] for learning rate estimates in machine learning.

A fundamental assumption in machine learning is that the sample data $\mathbf{z} := \{(x_j, y_j) : j \in \mathbb{N}_m\} \in X \times Y$ consists of independent and identically distributed instances of a random variable $(x, y) \in X \times Y$ subject to an unknown probability measure ρ on $X \times Y$. The performance of a predictor $f : X \rightarrow Y$ is hence measured by

$$\mathcal{E}(f) := \int_{X \times Y} |f(x) - y|^2 d\rho.$$

The predictor that minimizes the above error is the regression function

$$f_\rho(x) := \int_Y y d\rho(y|x), \quad x \in X, \quad (1.4)$$

where $\rho(y|x)$ denotes the conditional probability measure of y with respect to x . In fact, we have for every predictor f that

$$\mathcal{E}(f) = \mathcal{E}(f_\rho) + \|f - f_\rho\|_{L_{\rho_X}^2}^2, \quad (1.5)$$

where ρ_X is the marginal probability measure of ρ on X and for $p \in [1, +\infty)$, $L_{\rho_X}^p$ denotes the Banach space of measurable functions f on X with respect to ρ_X such that

$$\|f\|_{L_{\rho_X}^p} := \left(\int_X |f(x)|^p d\rho_X(x) \right)^{1/p} < +\infty.$$

The formula (1.4), though attractive, is only of theoretical value as ρ is unknown. A practical way is to find a minimizer $\mathbf{c}_{\mathbf{z}, \lambda}$ of (1.1) and hope that

$$f_{\mathbf{z}, \lambda}(x) := K^{\mathbf{x}}(x)\mathbf{c}_{\mathbf{z}, \lambda}, \quad x \in X \quad (1.6)$$

will be competitive with f_ρ in the sense that the approximation error $\mathcal{E}(f_{\mathbf{z}, \lambda}) - \mathcal{E}(f_\rho)$ would be small. To be more precise, for the learning scheme (1.1) to be useful in practice, this error should converge to zero fast in probability as the number of sampling points increases.

The approach in [5] works by introducing intermediate functions between $f_{\mathbf{z},\lambda}$ and f_ρ that are from a Banach space \mathcal{B} of functions on X with the properties that $K(x, \cdot) \in \mathcal{B}$ for all $x \in X$ and for all pairwise distinct $x_j \in X$, $j \in \mathbb{N}^m$ and $\mathbf{c} \in \mathbb{R}^m$

$$\psi(\|K^{\mathbf{x}}(\cdot)\mathbf{c}\|_{\mathcal{B}}) = \phi(\mathbf{c}),$$

for some nonnegative function ψ . Here $\|\cdot\|_{\mathcal{B}}$ is the norm on \mathcal{B} . Let g be an arbitrary function from such a space \mathcal{B} and set for each function $f : X \rightarrow \mathbb{R}$

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{j=1}^m (f(x_j) - y_j)^2.$$

The approximation error $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)$ can then be decomposed into the sum of four quantities

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) = \mathcal{S}(\mathbf{z}, \lambda, g) + \mathcal{P}(\mathbf{z}, \lambda, g) + \mathcal{D}(\lambda, g) - \lambda\psi(\|f_{\mathbf{z},\lambda}\|_{\mathcal{B}}), \quad (1.7)$$

where

$$\begin{aligned} \mathcal{S}(\mathbf{z}, \lambda, g) &:= \mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \mathcal{E}_{\mathbf{z}}(g) - \mathcal{E}(g), \\ \mathcal{P}(\mathbf{z}, \lambda, g) &:= (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda\psi(\|f_{\mathbf{z},\lambda}\|_{\mathcal{B}})) - (\mathcal{E}_{\mathbf{z}}(g) + \lambda\psi(\|g\|_{\mathcal{B}})), \\ \mathcal{D}(\lambda, g) &:= \mathcal{E}(g) - \mathcal{E}(f_\rho) + \lambda\psi(\|g\|_{\mathcal{B}}). \end{aligned}$$

The above three quantities are called the *sampling error*, the *hypothesis error* and the *regularization error*, respectively. The strategy is to choose \mathcal{B} and g carefully so that these three errors can be well bounded from above. When \mathcal{B} is the reproducing kernel Hilbert space of a positive-definite reproducing kernel K on X and the regularizer ϕ is given by (1.2), we have $\psi(t) = t^2$, $t \in \mathbb{R}$ and by the representer theorem and the definition of $f_{\mathbf{z},\lambda}$ in (1.6) that

$$f_{\mathbf{z},\lambda} = \arg \min_{f \in \mathcal{B}} (\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_{\mathcal{B}}^2). \quad (1.8)$$

In this case, one immediately has that $\mathcal{P}(\mathbf{z}, \lambda, g) \leq 0$ and thus, by (1.7) that

$$\mathcal{E}(f_{\mathbf{z},\lambda}) \leq \mathcal{S}(\mathbf{z}, \lambda, g) + \mathcal{D}(\lambda, g). \quad (1.9)$$

For the ℓ^1 -regularization where $\phi(\mathbf{c}) = \|\mathbf{c}\|_1$, the space \mathcal{B} chosen in [11, 20] does not satisfy the linear representer theorem. Consequently, the hypothesis error needed to be dealt with there.

A class of RKBS with the ℓ^1 norm that satisfies the linear representer theorem was recently constructed in [14]. In Section 2, we shall follow a similar idea to construct a slightly larger RKBS with the same desirable properties. By using the constructed space, we enjoy the same advantage as that for the RKHS case of discarding the hypothesis error automatically. Moreover, the space also leads to a better estimate of the regularization error than that in [20]. Combining these two improvements and directly using the estimates of the sampling error established in [20] or [11], one immediately has a superior learning rate. As our focus is on the advantages brought by the constructed RKBS, we shall only improve the learning rate estimate of [20] in Section 3. Interested readers may follow our strategy to engage the more sophisticated sampling error estimate given in [11] to improve the learning rate therein.

2 RKBS by Borel Measures

In this section, we construct RKBS applicable to the error analysis of the ℓ^1 -regularized least square regression. The constructed spaces are expected to have the ℓ^1 norm and satisfy the linear representer

theorem. The approach is different from the one by semi-inner products in [21, 22] as an infinite-dimensional ℓ^1 space is neither reflexive nor strictly convex.

Suppose that the input space X is a locally convex topological space and denote by $C_0(X)$ the space of continuous functions $f : X \rightarrow \mathbb{R}$ such that for all $\varepsilon > 0$, the set $\{x \in X : |f(x)| > \varepsilon\}$ is compact. We also impose the requirement that for all pairwise distinct $x_j \in X$, $j \in \mathbb{N}_m$, $m \in \mathbb{N}$, the kernel matrix $K[\mathbf{x}]$ is nonsingular. With the maximum norm $\|f\|_{C_0(X)} := \max_{x \in X} |f(x)|$, the space $C_0(X)$ is a Banach space. Its dual space is isometrically isomorphic to the space $\mathcal{M}(X)$ of all the signed Borel measures on X with bounded total variation. In other words, for each continuous linear functional T on $C_0(X)$, there exists a unique measure $\mu \in \mathcal{M}(X)$ such that

$$T(f) = \int_X f(x) d\mu(x) \text{ and } \sup_{f \in C_0(X), f \neq 0} \frac{|Tf|}{\|f\|_{C_0(X)}} = \|\mu\|, \quad (2.1)$$

where $\|\mu\|$ denotes the total variation of μ .

Let K be a real-valued function on $X \times X$ such that $K(\cdot, x) \in C_0(X)$ for all $x \in X$ and

$$\overline{\text{span}}\{K(\cdot, x) : x \in X\} = C_0(X). \quad (2.2)$$

With such a function, we introduce the following space

$$\mathcal{B} := \left\{ f_\mu := \int_X K(t, \cdot) d\mu(t) : \mu \in \mathcal{M}(X) \right\} \quad (2.3)$$

with the norm

$$\|f_\mu\|_{\mathcal{B}} := \|\mu\|. \quad (2.4)$$

Recall that a vector space V is called a *pre-RKBS* [14] on X if it is a Banach space consisting of functions on X such that point evaluation functionals are continuous on V and such that for all $f \in V$, $\|f\|_V = 0$ if and only if f vanishes everywhere on X .

Proposition 2.1 *Suppose that $K(\cdot, x) \in C_0(X)$ for all $x \in X$ and (2.2) is satisfied. Then \mathcal{B} defined by (2.3) is a pre-RKBS on X .*

Proof: We first show that the norm (2.4) is well-defined. Let μ, ν be two measures in $\mathcal{M}(X)$ such that $f_\mu(x) = f_\nu(x)$ for all $x \in X$. Then we get that

$$\int_X K(t, x) d(\mu - \nu)(t) = 0 \text{ for all } x \in X.$$

By the denseness condition (2.2), the above equation implies that $\mu - \nu = 0$. Thus, the measure μ associated with a function $f_\mu \in \mathcal{B}$ is unique. This proves that (2.4) is well-defined and that $\|f_\mu\|_{\mathcal{B}} = 0$ if and only if $f_\mu(x) = 0$ for all $x \in X$. Another consequence is that \mathcal{B} is isometrically isomorphic to $\mathcal{M}(X)$ and is hence a Banach space. Finally, we observe for all $x_0 \in X$ and $\mu \in \mathcal{M}(X)$ that

$$|f_\mu(x_0)| = \left| \int_X K(t, x_0) d\mu(t) \right| \leq \|K(\cdot, x_0)\|_{C_0(X)} \|\mu\| = \|K(\cdot, x_0)\|_{C_0(X)} \|f_\mu\|_{\mathcal{B}}.$$

Therefore, point evaluations are continuous linear functionals on \mathcal{B} . We conclude that \mathcal{B} is a pre-RKBS on X . The proof is complete. \square

Let the sampling points in \mathbf{x} be pairwise distinct. By definition, $K^{\mathbf{x}}(\cdot)\mathbf{c} \in \mathcal{B}$ for all $\mathbf{c} \in \mathbb{R}^m$. The denseness condition (2.2) implies that $K(x_j, \cdot)$, $j \in \mathbb{N}_m$ are linearly independent. As a result,

$$\|K^{\mathbf{x}}(\cdot)\mathbf{c}\|_{\mathcal{B}} = \|\mathbf{c}\|_1. \quad (2.5)$$

It is in the above sense that \mathcal{B} is said to possess the ℓ^1 norm.

We next turn to the crucial linear representer theorem in \mathcal{B} . We say that \mathcal{B} satisfies the linear representer theorem if for all continuous nonnegative loss function Q and regularizer ψ with $\lim_{t \rightarrow \infty} \psi(t) = +\infty$, the regularized learning scheme

$$\inf_{f \in \mathcal{B}} Q(f(\mathbf{x})) + \lambda \psi(\|f\|_{\mathcal{B}})$$

has a minimizer f_0 of the form $f_0 = K^{\mathbf{x}}(\cdot)\mathbf{c}$ for some $\mathbf{c} \in \mathbb{R}^m$. Here, $f(\mathbf{x}) = (f(x_j) : j \in \mathbb{N}_m)^T$.

The following lemma can be proved by arguments similar to those in [14].

Lemma 2.2 *The space \mathcal{B} satisfies the linear representer theorem if and only if for all \mathbf{x} of pairwise distinct sampling points and $\mathbf{y} \in \mathbb{R}^m$, the minimal norm interpolation*

$$\inf\{\|f\|_{\mathcal{B}} : f \in \mathcal{B}, f(\mathbf{x}) = \mathbf{y}\} \quad (2.6)$$

has a minimizer f_0 of the form $f_0 = K^{\mathbf{x}}(\cdot)\mathbf{c}$ for some $\mathbf{c} \in \mathbb{R}^m$.

A subspace of \mathcal{B} was constructed in [14] and conditions for it to satisfy the linear representer theorem were studied. In order to make use of the results obtained there, we first introduce the subspace. Denote by $\ell^1(X)$ the subset of $\mathcal{M}(X)$ of those Borel measures that are supported on a countable subset of X . Thus, for each $\nu \in \ell^1(X)$, there exist some pairwise distinct points $x_j \in X$, $j \in \mathbb{I}$ where \mathbb{I} is a countable index set, such that

$$\nu(A) = \sum_{x_j \in A} \nu(x_j) \text{ for every Borel subset } A \subseteq X.$$

Denote by $\text{supp } \nu$ the countable set of points where ν is nonzero. The space \mathcal{B}_1 considered in [14] is

$$\mathcal{B}_1 := \left\{ \sum_{x \in \text{supp } \nu} \nu(x) K(x, \cdot) : \nu \in \ell^1(X) \right\}$$

with the norm inherited from that of \mathcal{B} .

Put for all $x \in X$, $K_{\mathbf{x}}(x) := (K(x, x_j) : j \in \mathbb{N}_m)^T$, which is an $m \times 1$ vector in \mathbb{R}^m . One should not confuse $K_{\mathbf{x}}(x)$ with $K^{\mathbf{x}}(x)$. The latter is $1 \times m$ and might even not be the transpose of the former as K is not required to be symmetric. The following result about \mathcal{B}_1 is from [14].

Lemma 2.3 *For all $\mathbf{y} \in \mathbb{R}^m$, the minimal norm interpolation*

$$\inf\{\|f\|_{\mathcal{B}_1} : f \in \mathcal{B}_1, f(\mathbf{x}) = \mathbf{y}\} \quad (2.7)$$

has a minimizer f_0 of the form $f_0 = K^{\mathbf{x}}(\cdot)\mathbf{c}$ for some $\mathbf{c} \in \mathbb{R}^m$ if and only if

$$\|K[\mathbf{x}]^{-1} K_{\mathbf{x}}(x)\|_1 \leq 1 \text{ for all } x \in X. \quad (2.8)$$

Moreover, under condition (2.8), there holds for all $\mathbf{c} \in \mathbb{R}^m$ that

$$\|\mathbf{c}^T K_{\mathbf{x}}(\cdot)\|_{C_0(X)} = \|\mathbf{c}^T K[\mathbf{x}]\|_{\infty}, \quad (2.9)$$

where $\|\cdot\|_{\infty}$ is the maximum norm on \mathbb{R}^m .

We are ready to present the main result of this section.

Theorem 2.4 *The space \mathcal{B} satisfies the linear representer theorem if and only if (2.8) holds true.*

Proof: Suppose that (2.8) holds true. By Lemma 2.2, to show that \mathcal{B} satisfies the linear representer theorem, it suffices to show that $f_0 = K^{\mathbf{x}}(\cdot)K[\mathbf{x}]^{-1}\mathbf{y}$ is a minimizer of (2.6). Clearly, $f_0(\mathbf{x}) = \mathbf{y}$. Let f_μ , $\mu \in \mathcal{M}(X)$, be an arbitrary function in \mathcal{B} that satisfies the interpolation condition $f_\mu(\mathbf{x}) = \mathbf{y}$. We then have for all $\mathbf{c} \in \mathbb{R}^m$ that

$$\int_X \mathbf{c}^T K_{\mathbf{x}}(t) d\mu(t) = \int_X \sum_{j=1}^m c_j K(t, x_j) d\mu(t) = \sum_{j=1}^m c_j f_\mu(x_j) = \mathbf{c}^T \mathbf{y}.$$

It follows from (2.1) that for all $\mathbf{c} \in \mathbb{R}^m$

$$|\mathbf{c}^T \mathbf{y}| \leq \|\mathbf{c}^T K_{\mathbf{x}}(\cdot)\|_{C_0(X)} \|\mu\|.$$

This together with (2.9) implies that

$$\|\mu\| \geq \sup_{\mathbf{c} \in \mathbb{R}^m, \mathbf{c} \neq 0} \frac{|\mathbf{c}^T \mathbf{y}|}{\|\mathbf{c}^T K_{\mathbf{x}}(\cdot)\|_{C_0(X)}} = \sup_{\mathbf{c} \in \mathbb{R}^m, \mathbf{c} \neq 0} \frac{|\mathbf{c}^T \mathbf{y}|}{\|\mathbf{c}^T K[\mathbf{x}]\|_\infty} = \sup_{\mathbf{a} \in \mathbb{R}^m, \mathbf{a} \neq 0} \frac{|\mathbf{a}^T K[\mathbf{x}]^{-1} \mathbf{y}|}{\|\mathbf{a}\|_\infty} = \|K[\mathbf{x}]^{-1} \mathbf{y}\|_1.$$

Now, recall by (2.5) that $\|f_0\|_{\mathcal{B}} = \|K[\mathbf{x}]^{-1} \mathbf{y}\|_1$ and by definition of $\|\cdot\|_{\mathcal{B}}$ that $\|f_\mu\|_{\mathcal{B}} = \|\mu\|$. These two facts combined with the above inequality imply that $\|f_\mu\|_{\mathcal{B}} \geq \|f_0\|_{\mathcal{B}}$. Thus, f_0 is indeed a minimizer of (2.6).

On the other hand, suppose that \mathcal{B} satisfies the linear representer theorem and we want to prove (2.8). Let $\mathbf{y} \in \mathbb{R}^m$. By Lemma 2.2, the minimal norm interpolation (2.6) has a minimizer f_0 of the form $f_0 = K^{\mathbf{x}}(\cdot)\mathbf{c}$ for some $\mathbf{c} \in \mathbb{R}^m$. Clearly, f_0 is also a minimizer of (2.7) because $f_0 \in \mathcal{B}_1$ and

$$\|f_0\|_{\mathcal{B}_1} = \|f_0\|_{\mathcal{B}} = \inf\{\|f\|_{\mathcal{B}} : f \in \mathcal{B}, f(\mathbf{x}) = \mathbf{y}\} \leq \inf\{\|f\|_{\mathcal{B}_1} : f \in \mathcal{B}_1, f(\mathbf{x}) = \mathbf{y}\}.$$

By Lemma 2.3, (2.8) holds true. The proof is complete. \square

It will become clear in the next section that the above theorem makes \mathcal{B} a useful space for error analysis of the ℓ^1 -regularized least square regression.

We present two examples of K that satisfy all the assumptions, especially (2.8), in this section:

– the exponential kernel

$$K(s, t) := e^{-|s-t|}, \quad s, t \in \mathbb{R},$$

– the Brownian bridge kernel

$$K(s, t) := \min\{s, t\} - st, \quad s, t \in (0, 1).$$

That these two kernels satisfy (2.8) has been proved in [14]. It remains to verify the denseness requirement (2.2). The exponential kernel is a particular case of the following result.

Proposition 2.5 *If ϕ is Lebesgue integrable on \mathbb{R}^d that is nonzero almost everywhere then the function*

$$K(\mathbf{s}, \mathbf{t}) := \int_{\mathbb{R}^d} e^{-i(\mathbf{s}-\mathbf{t}) \cdot \boldsymbol{\xi}} \phi(\boldsymbol{\xi}) d\boldsymbol{\xi}, \quad \mathbf{s}, \mathbf{t} \in \mathbb{R}^d \quad (2.10)$$

satisfies that $K(\cdot, \mathbf{t}) \in C_0(\mathbb{R}^d)$ for all $\mathbf{t} \in \mathbb{R}^d$ and the denseness condition (2.2). So does $K(\mathbf{s}, \mathbf{t}) := \psi(\mathbf{s} - \mathbf{t})$, $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d$ where ψ is a nontrivial continuous function on \mathbb{R}^d of compact support.

Proof: That the function given by (2.10) belongs to $C_0(\mathbb{R}^d)$ for all $t \in \mathbb{R}^d$ follows from the Riemann-Lebesgue lemma. The denseness condition (2.2) for the two kernels can be proved by arguments similar to those in [14]. \square

The Brownian bridge kernel is handled with a manner different from that in [14].

Proposition 2.6 *The Brownian bridge kernel satisfies (2.2).*

Proof: Clearly, for the Brownian bridge kernel, $K(\cdot, t)$ is continuous for all $t \in (0, 1)$. Let ν be a Borel measure on $X := (0, 1)$ such that

$$\int_X K(s, t) d\nu(s) = 0 \quad \text{for all } t \in (0, 1). \quad (2.11)$$

Note that K has the representation

$$K(s, t) = \int_X \Gamma_s(z) \Gamma_t(z) dz, \quad s, t \in (0, 1),$$

where $\Gamma_s := \chi_{(0,s)} - s$ with $\chi_{(0,s)}$ denoting the characteristic function of $(0, s)$. Arguments similar to those in [14] yield that there exists a constant C such that

$$\int_0^s d\nu(s) = C \quad \text{for all } s \in (0, 1).$$

It follows that $\nu((s_1, s_2)) = 0$ for all $0 < s_1 < s_2 < 1$. Consequently, ν is the zero Borel measure on $(0, 1)$. Thus, the Brownian bridge kernel satisfies (2.2). \square

Finally, we remark that the function K can be regarded as the reproducing kernel for \mathcal{B} constructed by (2.3). To see this, we introduce a bilinear form on $\mathcal{B} \times C_0(X)$ by setting

$$\langle f_\mu, g \rangle := \int_X g(x) d\mu(x) \quad \text{for all } \mu \in \mathcal{M}(X) \text{ and } g \in C_0(X).$$

We observe by (2.1) that

$$|\langle f_\mu, g \rangle| \leq \|\mu\| \|g\|_{C_0(X)} = \|f_\mu\|_{\mathcal{B}} \|g\|_{C_0(X)}$$

and that for all $x \in X$,

$$f_\mu(x) = \langle f_\mu, K(\cdot, x) \rangle, \quad g(x) = \langle K(x, \cdot), g \rangle.$$

In the above senses, K is said to be the reproducing kernel for both \mathcal{B} and $C_0(X)$.

3 Error Analysis of the ℓ^1 -Regularization

We apply the constructed space \mathcal{B} to estimate the learning rate of the ℓ^1 -regularized least square regression (1.3) in this section. To this end, we first introduce some standard assumptions in the literature imposed on the regression function f_ρ , the input space X and the function K .

Let X be compact metric space with the distance d and assume that ρ_X is a Borel probability measure on X . In this note, we suppose that K is a positive-definite reproducing kernel on X with the Lipschitz condition

$$|K(x, t) - K(x, t')| \leq C_\alpha (d(t, t'))^\alpha \quad \text{for some positive constants } \alpha, C_\alpha \text{ and for all } x, t, t' \in X. \quad (3.1)$$

Denote for all $r > 0$ by $\mathcal{N}(X, r)$ the least number of open balls with radius r that cover X . Assume that this covering number satisfies for some positive constants η, C_η that

$$\mathcal{N}(X, r) \leq \frac{C_\eta}{r^\eta} \quad \text{for all } 0 < r \leq 1. \quad (3.2)$$

The requirement on f_ρ is that it is contained in the range $\text{ran}(L_K^s)$ of L_K^s for some $s > 0$. Here, L_K is the compact positive operator on $L_{\rho_X}^2$ defined by

$$L_K f := \int_X K(t, \cdot) f(t) d\rho_X(t), \quad f \in L_{\rho_X}^2.$$

Let $\phi_j, j \in \mathbb{N}$ be an orthonormal basis for $L_{\rho_X}^2$ consisting of eigenfunctions of L_K with the corresponding eigenvalues $\lambda_j \geq \lambda_{j+1}, j \in \mathbb{N}$. The assumption $f_\rho \in \text{ran}(L_K^s)$ implies that

$$f_\rho = \sum_{j=1}^{\infty} \lambda_j^s a_j \phi_j$$

for some $h = \sum_{j=1}^{\infty} a_j \phi_j$ in $L_{\rho_X}^2$. In order to make use of the space constructed in the last section, our last requirement is that K satisfies that $\overline{\text{span}}\{K(\cdot, x) : x \in X\} = C(X)$ and condition (2.8).

Let $\mathbf{c}_{\mathbf{z}, \lambda}$ be a minimizer of (1.3) and let $f_{\mathbf{z}, \lambda}$ be given by (1.6). For the minimization problem (1.3), the hypothesis error and regularization error have the specific forms

$$\begin{aligned} \mathcal{P}(\mathbf{z}, \lambda, g) &:= (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}, \lambda}) + \lambda \|f_{\mathbf{z}, \lambda}\|_{\mathcal{B}}) - (\mathcal{E}_{\mathbf{z}}(g) + \lambda \|g\|_{\mathcal{B}}), \\ \mathcal{D}(\lambda, g) &:= \mathcal{E}(g) - \mathcal{E}(f_\rho) + \lambda \|g\|_{\mathcal{B}}, \end{aligned}$$

where g is a function in \mathcal{B} to be carefully chosen.

The use of the space \mathcal{B} enables us to discard the hypothesis error immediately.

Lemma 3.1 *Under the above assumptions on K , there holds $\mathcal{E}(f_{\mathbf{z}, \lambda}) - \mathcal{E}(f_\rho) \leq \mathcal{S}(\mathbf{z}, \lambda, g) + \mathcal{D}(\lambda, g)$ for all $g \in \mathcal{B}$.*

Proof: By Theorem 2.4,

$$f_{\mathbf{z}, \lambda} = \arg \min_{f \in \mathcal{B}} \mathcal{E}_{\mathbf{z}}(f) + \lambda \|f\|_{\mathcal{B}}.$$

As a consequence, $\mathcal{P}(\mathbf{z}, \lambda, g) \leq 0$, which together with inequality (1.7) completes the proof. \square

We next estimate the regularization error.

Lemma 3.2 *If $0 < s < 1$ then*

$$\inf_{g \in \mathcal{B}} \mathcal{D}(\lambda, g) \leq (\|h\|_{L_{\rho_X}^2} + \|h\|_{L_{\rho_X}^2}^2) \lambda^{\frac{2s}{1+s}}. \quad (3.3)$$

If $s \geq 1$ then $f_\rho \in \mathcal{B}$ and

$$\mathcal{D}(\lambda, f_\rho) \leq (\lambda_1^{s-1} \|h\|_{L_{\rho_X}^2}) \lambda. \quad (3.4)$$

Proof: Firstly, we have for each $\varphi \in L_{\rho_X}^2$ that $L_K \varphi \in \mathcal{B}$ and by the Cauchy-Schwartz inequality that

$$\|L_K \varphi\|_{\mathcal{B}} = \|\varphi\|_{L_{\rho_X}^1} \leq \|\varphi\|_{L_{\rho_X}^2}. \quad (3.5)$$

If $s \geq 1$ then $f_\rho = L_K \varphi$ where

$$\varphi = \sum_{j=1}^{\infty} \lambda_j^{s-1} a_j \phi_j.$$

As λ_j is non-increasing,

$$\|\varphi\|_{L_{\rho_X}^2} \leq \lambda_1^{s-1} \left(\sum_{j=1}^{\infty} |a_j|^2 \right)^{1/2} = \lambda_1^{s-1} \|h\|_{L_{\rho_X}^2}.$$

We then get by the above equation and (3.5) that

$$\mathcal{D}(\lambda, f_\rho) = \lambda \|f_\rho\|_{\mathcal{B}} \leq \lambda \|\varphi\|_{L_{\rho_X}^2} \leq \lambda \lambda_1^{s-1} \|h\|_{L_{\rho_X}^2},$$

which is (3.4).

Suppose now that $0 < s < 1$. If $\lambda_1 \leq \lambda^{\frac{1}{1+s}}$ then by (1.5),

$$\mathcal{D}(\lambda, 0) = \mathcal{E}(0) - \mathcal{E}(f_\rho) = \|f_\rho\|_{L_{\rho_X}^2}^2 = \sum_{j=1}^{\infty} \lambda_j^{2s} a_j^2 \leq \lambda^{\frac{2s}{1+s}} \|h\|_{L_{\rho_X}^2}^2,$$

which implies (3.3). If $\lambda_1 > \lambda^{\frac{1}{1+s}}$ then since λ_j decreases to zero as j tends to infinity, there exists some $N \in \mathbb{N}$ such that $\lambda_{N+1} < \lambda^{\frac{1}{1+s}} \leq \lambda_N$. Put

$$\varphi := \sum_{j=1}^N \lambda_j^{s-1} a_j \phi_j.$$

It follows from (1.5) and (3.5) that

$$\mathcal{D}(\lambda, L_K \varphi) \leq \|L_K \varphi - f_\rho\|_{L_{\rho_X}^2}^2 + \lambda \|\varphi\|_{L_{\rho_X}^2}.$$

We estimate that

$$\lambda \|\varphi\|_{L_{\rho_X}^2} = \lambda \left(\sum_{j=1}^N a_j^2 \lambda_j^{2s-2} \right)^{1/2} \leq \lambda \lambda^{\frac{s-1}{1+s}} \left(\sum_{j=1}^N a_j^2 \right)^{1/2} \leq \lambda^{\frac{2s}{1+s}} \|h\|_{L_{\rho_X}^2}$$

and that

$$\|L_K \varphi - f_\rho\|_{L_{\rho_X}^2}^2 = \sum_{j=N+1}^{\infty} \lambda_j^{2s} a_j^2 \leq \lambda^{\frac{2s}{1+s}} \sum_{j=N+1}^{\infty} a_j^2 \leq \lambda^{\frac{2s}{1+s}} \|h\|_{L_{\rho_X}^2}^2.$$

Combing the above two inequalities leads to (3.3). The proof is complete. \square

We remark that the estimated regularization error in [20] was of the order $O(\lambda^{\frac{2s}{2+s}})$ for $0 < s \leq 2$.

Turning to the sampling error, we follow the approach in [20] to decompose it into the sum $\mathcal{S}(\mathbf{z}, \lambda, g) = \mathcal{S}_1(\mathbf{z}, \lambda, g) + \mathcal{S}_2(\mathbf{z}, \lambda)$ where

$$\mathcal{S}_1(\mathbf{z}, \lambda, g) = (\mathcal{E}_{\mathbf{z}}(g) - \mathcal{E}_{\mathbf{z}}(f_\rho)) - (\mathcal{E}(g) - \mathcal{E}(f_\rho)), \quad \mathcal{S}_2(\mathbf{z}, \lambda) = (\mathcal{E}(f_{\mathbf{z}, \lambda}) - \mathcal{E}(f_\rho)) - (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}, \lambda}) - \mathcal{E}_{\mathbf{z}}(f_\rho)).$$

The first summand $\mathcal{S}_1(\mathbf{z}, \lambda, g)$ can be bounded by using the law of large numbers. By the same arguments as those in [5, 20], we use the estimate in Lemma 3.2 to obtain an improved bound.

Lemma 3.3 Suppose that the output of sample data is bounded by a positive constant almost surely. If $0 < s < 1$ then for each $\varepsilon > 0$ there exists some $g \in \mathcal{B}$ such that for all $0 < \delta < 1$, we have with confidence $1 - \frac{\delta}{2}$ that

$$\mathcal{S}_1(\mathbf{z}, \lambda, g) \leq C_1 \left(\frac{\lambda^{\frac{2(s-1)}{1+s}}}{m} + \frac{\lambda^{\frac{2s-1}{1+s}}}{\sqrt{m}} \right) \log \frac{2}{\delta}$$

for some positive constant C_1 . If $s \geq 1$ then $\mathcal{S}_1(\mathbf{z}, \lambda, f_\rho) = 0$.

For $\mathcal{S}_2(\mathbf{z}, \lambda)$, we cite the following result from [20].

Lemma 3.4 Suppose that (3.1) and (3.2) hold true. If $\lambda \leq 1$ then we have with confidence $1 - \frac{\delta}{2}$ that

$$\mathcal{S}_2(\mathbf{z}, \lambda) \leq \frac{1}{2}(\mathcal{E}(f_{\mathbf{z}, \lambda}) - \mathcal{E}(f_\rho)) + C_2 \frac{\log \frac{2}{\delta} + \log(1+m)}{\lambda^2} m^{-\frac{1}{1+\eta/\alpha}}$$

for some positive constant C_2 .

Combining Lemmas 3.1, 3.2, 3.3, and 3.4, we reach a new learning rate estimate of the ℓ^1 -regularized least square regression.

Theorem 3.5 Suppose that X satisfy (3.2), the output is bounded by a positive constant almost surely, and $f_\rho \in \text{ran}(L_K^s)$ for some $s > 0$. Let K be a positive-definite reproducing kernel satisfying $\overline{\text{span}}\{K(\cdot, x) : x \in X\} = C(X)$, the condition (2.8) and the Lipschitz condition (3.1). Then there exists some constant $C > 0$ such that with the choice $\lambda = m^{-\frac{1}{2} \frac{1}{1+\eta/\alpha} \frac{1+s}{1+2s}}$, we have for all $0 < \delta < 1$ with confidence $1 - \delta$ that

$$\mathcal{E}(f_{\mathbf{z}, \lambda}) - \mathcal{E}(f_\rho) \leq C m^{-\frac{s}{1+2s} \frac{1}{1+\eta/\alpha}} \log \frac{2+2m}{\delta}, \quad \text{when } 0 < s < 1 \quad (3.6)$$

and

$$\mathcal{E}(f_{\mathbf{z}, \lambda}) - \mathcal{E}(f_\rho) \leq C m^{-\frac{1}{3(1+\eta/\alpha)}} \log \frac{1+m}{\delta}, \quad \text{when } s \geq 1.$$

Proof: We only discuss the case when $0 < s < 1$ as the other situation is easier and can be shown in a similar way. We choose $\lambda = m^{-\theta}$, $\theta > 0$ and get by Lemmas 3.1, 3.2, 3.3, and 3.4 that there exists some constant $C > 0$ such that with confidence $1 - \delta$

$$\mathcal{E}(f_{\mathbf{z}, \lambda}) - \mathcal{E}(f_\rho) \leq C m^{-\gamma} \log \frac{2+2m}{\delta}, \quad (3.7)$$

where

$$\gamma = \min \left\{ \frac{1}{1+\eta/\alpha} - 2\theta, 1 - \frac{2\theta(1-s)}{1+s}, \frac{1}{2} - \frac{1-2s}{1+s}\theta, \frac{2\theta s}{1+s} \right\}.$$

The maximum of γ is achieved when

$$\theta = \frac{1}{2} \frac{1}{1+\eta/\alpha} \frac{1+s}{1+2s}.$$

Substituting the above choice into (3.7) yields (3.6). \square

Improvements of the learning rate can be achieved if higher regularity is imposed on the kernel K [24] or better estimates of the sampling error are engaged [11]. Another remark is that the assumption of positive-definiteness and symmetry on K might be abandoned by using the strategy in [20].

References

- [1] S. Alliney. A property of the minimum vectors of a regularizing functional defined by means of the absolute norm. *IEEE Transactions on Signal Processing*, 45:913–917, 1997.
- [2] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.
- [3] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.
- [4] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.
- [5] F. Cucker and D.-X. Zhou. *Learning theory: an approximation theory viewpoint*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 2007. With a foreword by Stephen Smale.
- [6] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Adv. Comput. Math.*, 13(1):1–50, 2000.
- [7] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.*, 33:82–95, 1971.
- [8] M. Nikolova. A variational approach to remove outliers and impulse noise. *J. Math. Imaging Vision*, 20(1-2):99–120, 2004. Special issue on mathematics and image analysis.
- [9] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press, Cambridge, December 2001.
- [10] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, 2004.
- [11] L. Shi, Y.-L. Feng, and D.-X. Zhou. Concentration estimates for learning with ℓ^1 -regularizer and data dependent hypothesis spaces. *Appl. Comput. Harmon. Anal.* to appear.
- [12] S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26(2):153–172, 2007.
- [13] G. Song and Y. Xu. Approximation of high-dimensional kernel matrices by multilevel circulant matrices. *J. Complexity*, 26(4):375–405, 2010.
- [14] G. Song, H. Zhang, and F. J. Hickernell. Reproducing kernel banach spaces with the ℓ^1 norm. *preprint, arXiv:1101.4388v1*, 2011.
- [15] H. Sun and Q. Wu. Regularized least square regression with dependent samples. *Adv. Comput. Math.*, 32(2):175–189, 2010.
- [16] H. Sun and Q. Wu. Coefficient regularization in least square kernel regression. *preprint*, 2011.
- [17] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.
- [18] H. Tong, D.-R. Chen, and F. Yang. Least square regression with ℓ^p -coefficient regularization. *Neural Comput.*, 22:3221–3235, 2010.
- [19] V. N. Vapnik. *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.
- [20] Q.-W. Xiao and D.-X. Zhou. Learning by nonsymmetric kernels with data dependent spaces and ℓ^1 -regularizer. *Taiwanese J. Math.*, 14(5):1821–1836, 2010.

- [21] H. Zhang, Y. Xu, and J. Zhang. Reproducing kernel Banach spaces for machine learning. *J. Mach. Learn. Res.*, 10:2741–2775, 2009.
- [22] H. Zhang and J. Zhang. Frames, Riesz bases, and sampling expansions in Banach spaces via semi-inner products. *Appl. Comput. Harmon. Anal.* to appear.
- [23] T. Zhang. Leave-one-out bounds for kernel methods. *Neural Comput.*, 15:1397–1437, 2003.
- [24] D.-X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Trans. Inform. Theory*, 49(7):1743–1752, 2003.