Journal of Complexity 36 (2016) 141–165

Contents lists available at ScienceDirect

Journal of Complexity

journal homepage: www.elsevier.com/locate/jco

Multi-penalty regularization in learning theory

Abhishake, S. Sivananthan *

Department of Mathematics, Indian Institute of Technology Delhi, New Delhi 110016, India

ARTICLE INFO

Article history:
Received 2 December 2015
Accepted 2 May 2016
Available online 21 May 2016

Keywords:
Learning theory
Manifold learning
Multi-penalty regularization
Error estimate
Adaptive parameter choice

ABSTRACT

In this paper we establish the error estimates for multi-penalty regularization under the general smoothness assumption in the context of learning theory. One of the motivation for this work is to study the convergence analysis of two-parameter regularization theoretically in the manifold learning setting. In this spirit, we obtain the error bounds for the manifold learning problem using more general framework of multi-penalty regularization. We propose a new parameter choice rule "the balanced-discrepancy principle" and analyze the convergence of the scheme with the help of estimated error bounds. We show that multi-penalty regularization with the proposed parameter choice exhibits the convergence rates similar to single-penalty regularization. Finally on a series of test samples we demonstrate the superiority of multi-parameter regularization over single-penalty regularization.

1. Introduction

Suppose $\{(x_i, y_i)\}_{i=1}^m \subset X \times Y$ is a finite set of samples, where $X \subset \mathbb{R}^n$ denotes the input space and $Y \subset \mathbb{R}$ denotes the output space of a system. Then learning theory [9,15,35] aims to develop an algorithm which finds a function $f : X \to Y$ based on given samples such that $f(x)$ predicts the response variable $y \in Y$ for the input $x \in X$. The problem of construction of estimator from samples is ill-conditioned. To overcome this situation we use regularization schemes [3,14,17,34] in order to get a stable solution. In learning theory, single-penalty regularization is well-studied. Convergence issues are briefly discussed in literature [3,6,30–33] while various parameter choice approaches are proposed

* Corresponding author.
E-mail addresses: abhishekrastogi2012@gmail.com (Abhishake), siva@maths.iitd.ac.in (S. Sivananthan).

http://dx.doi.org/10.1016/j.jco.2016.05.003
0885-064X/

and justified in [7,11] (also see references therein). But sometimes to incorporate various features in the solution such as boundedness, monotonicity, smoothness we require multiple penalties.

Multi-penalty regularization has been widely accepted as a stable and robust method to construct regularized solution for ill-posed problems [21,19]. The advantage of multi-penalty methods is that one can incorporate any prior information through additional penalties. For example, in extrapolation problem, we may incorporate the prediction points as a priori information in the construction of extrapolating estimator (see Belkin et al. [4]). So it is interesting to study more sophisticated multi-penalty regularization methods in learning theory framework. In the context of inverse problems, the convergence of multi-penalty regularization and its various parameter choice rules are well-studied in the literature [2,16,18–20,27]. But it is well-known that we cannot directly apply the inverse problem framework to learn the regression function from sampled data. The fact is that unlike the inverse problems, in learning theory infinite dimensional embedding operator is discretized which is finite dimensional sampling operator acts on Euclidean space [3,12].

Unlike the regularized learning algorithms in single-penalty, the multi-penalty regularization is not well-studied in the literature. In the context of learning theory, Belkin et al. [4] proposed a multi-penalty regularization scheme which exploits the geometry of the marginal distribution. The authors successfully demonstrated performance of the multi-penalty functional under selected parameters (by inspection) with the help of various numerical experiments. Wood [36] developed a parameter choice strategy for multi-penalized likelihood methods using generalized cross-validation score which requires data-splitting. Shuai Lu et al. [22] proposed a parameter choice rule using discrepancy principle and numerically illustrated with an academic example but no theoretical justification was provided.

In this paper we address the convergence issues for a more general two-parameter regularization and present a new parameter choice rule "the balanced-discrepancy principle" in learning theory which does not require any data-splitting. The paper is organized as follows. In Section 2, we recall general framework of learning and some basic concepts. In Section 3, the error estimates are established under the general source condition for the estimator of the regression function in $\|\cdot\|_{\mathcal{H}_K}$-norm and $\|\cdot\|_{\mathscr{L}^2_{\rho_X}}$-norm. In Section 4, the balanced-discrepancy principle is proposed to choose the regularization parameters and analyzed the convergence of the scheme. In Section 5, we illustrate the superiority of the multi-penalty regularizer over the single-penalty regularizers using various examples.

2. Learning from examples

Let X be a compact metric space, Y be the set of real numbers and $\{(x_i, y_i) \in X \times Y,\ i = 1, \ldots, m\}$ are i.i.d. samples drawn according to an unknown probability measure ρ on the sample space $Z = X \times Y$. The aim of "learning from examples" [5,10,15,35] is to estimate a function f based on the finite examples $\{(x_i, y_i)\}_{i=1}^m$ such that $f(x)$ predicts the output y for any given input x. For a given loss function ℓ, choose a function f such that which minimizes the generalization error

$$\mathcal{E}(f) := \mathcal{E}_\rho(f) = \int_Z \ell(x, y, f) d\rho(x, y). \tag{1}$$

Throughout the paper we consider the square loss function $\ell(x, y, f) = (f(x) - y)^2$ in our analysis. The splitting form of the probability measure ρ can be described as

$$\rho(x, y) = \rho(y|x)\rho_X(x),$$

where $\rho(y|x)$ and $\rho_X(x)$ are conditional probability measure on Y and marginal probability measure on X respectively. Then the minimizer of the generalization error over $\mathscr{L}^2_{\rho_X}$ is given by the function

$$f_\rho(x) := \int_Y y d\rho(y|x), \tag{2}$$

which is called the regression function of ρ.

Learning algorithms minimize the generalization error over an appropriate subspace $\mathcal{H} \subset \mathcal{L}^2_{\rho_X}$, called hypothesis space. Minimization of the generalization error is equivalent to finding the estimator of the regression function f_ρ over the hypothesis space which is clear from the following proposition:

$$\mathcal{E}(f) = \int_X (f(x) - f_\rho(x))^2 d\rho_X(x) + \sigma_\rho^2,$$

where $\sigma_\rho^2 = \int_X \int_Y (y - f_\rho(x))^2 d\rho(y|x) d\rho_X(x)$.

Now the problem is reduced to construct a good estimator of the regression function over the hypothesis space $\mathcal{H}$. But it is important to note that, in general, the probability measure ρ and the regression function f_ρ are unknown. The only available information is finite samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$. The empirical estimate of the generalization error, i.e., empirical error of f (w.r.t. $\mathbf{z}$) is defined by

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2. \tag{3}$$

So instead of minimizing the generalization error, learning algorithms minimize empirical error over the hypothesis space.

Generally, the problem of finding the optimizer of (3) is ill-conditioned. Regularization techniques [3,14,17,34] stabilize the problem by incorporating a priori information [24], e.g., boundedness, monotonicity and smoothness. The standard regularization scheme [15,34] is

$$\widetilde{f} := arg \min_{f \in \mathcal{H}} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda_1 \|f\|^2_{\mathcal{H}} \right\}.$$

But sometimes to incorporate more features we require more than one penalty. In manifold regularization [4], a multi-penalty regularization scheme is discussed which controls the complexity of the function in ambient space as well as geometry of the probability space. Multi-penalty regularization has been successfully applied in various inspiring applications such as image reconstruction [23], reconstruction of the Earth gravity potential [37], option pricing and parameter estimation problem [13]. Therefore we consider a more general framework in the context of multi-penalty regularization,

$$\widetilde{f} := arg \min_{f \in \mathcal{H}} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda_1 \|f\|^2_{\mathcal{H}} + \lambda_2 \|Bf\|^2_{\mathcal{H}} \right\},$$

where B is a bounded operator on $\mathcal{H}$, λ_1 and λ_2 are non-negative real numbers.

Definition 2.1. Let $K : X \times X \to \mathbb{R}$ be continuous, symmetric and positive semidefinite, i.e., for all finite sets $\{x_1, \ldots, x_d\} \subset X$, the matrix $\{K(x_i, x_j)\}_{i,j=1}^d$ is positive semidefinite. Then K is called a Mercer kernel.

Definition 2.2. Let X be a non-empty set and $\mathcal{H}$ be a Hilbert space of real-valued functions on X. If the pointwise evaluation map $F_x : \mathcal{H} \to \mathbb{R}$, defined by

$$F_x(f) = f(x) \quad \forall f \in \mathcal{H},$$

is continuous for every $x \in X$. Then $\mathcal{H}$ is called reproducing kernel Hilbert space.

For each reproducing kernel Hilbert space $\mathcal{H}$ there exists a Mercer kernel $K : X \times X \to \mathbb{R}$ such that for $K_x : X \to \mathbb{R}$, defined as $K_x(t) = K(x, t)$, the span of the set $\{K_x : x \in X\}$ is dense in $\mathcal{H}$. Moreover, there is one to one correspondence between Mercer kernels and reproducing kernel Hilbert spaces [1,8]. So a reproducing kernel Hilbert space $\mathcal{H}$ corresponding to a Mercer kernel K can be denoted as $\mathcal{H}_K$ and norm in the space $\mathcal{H}$ can be denoted as $\|\cdot\|_{\mathcal{H}_K}$ or $\|\cdot\|_K$.

Choosing the hypothesis $\mathcal{H}$ as reproducing kernel Hilbert space $\mathcal{H}_K$, the learning scheme can be written as

$$f_{\mathbf{z},\lambda} = arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda_1 \|f\|^2_K + \lambda_2 \|Bf\|^2_K \right\}, \tag{4}$$

where λ denotes the ordered pair (λ_1, λ_2).

By the representation theorem [4], the solution of the above problem will be of the form:

$$f_{\mathbf{z},\lambda} = \sum_{i=1}^{m} c_i K_{x_i} + \int_{\mathcal{M}} c(t)K(\cdot,t)d\rho_X(t),$$

where $\mathcal{M}$ is the support of marginal probability measure ρ_X.

Define the sampling operator $S_{\mathbf{x}} : \mathcal{H}_K \to \mathbb{R}^m$ by

$$S_{\mathbf{x}}(f) = (f(x))_{x\in\mathbf{x}},$$

where $\mathbf{x} = (x_i)_{i=1}^m$. Denote $S_{\mathbf{x}}^* : \mathbb{R}^m \to \mathcal{H}_K$ as the adjoint of $S_{\mathbf{x}}$. Then for $c \in \mathbb{R}^m$,

$$\langle f, S_{\mathbf{x}}^* c\rangle_K = \langle S_{\mathbf{x}} f, c\rangle_m = \frac{1}{m}\sum_{i=1}^{m} c_i f(x_i) = \left\langle f, \frac{1}{m}\sum_{i=1}^{m} c_i K_{x_i}\right\rangle_K, \quad \forall f \in \mathcal{H}_K,$$

where $\|\cdot\|_m$ in $\mathbb{R}^m$ is $1/m$ times of Euclidean norm (see [3,8] for details). Then it follows that

$$S_{\mathbf{x}}^* c = \frac{1}{m}\sum_{i=1}^{m} c_i K_{x_i}, \quad \forall c \in \mathbb{R}^m.$$

From the following assertion we observe that $S_{\mathbf{x}}$ is a bounded operator:

$$\|S_{\mathbf{x}} f\|_m = \frac{1}{m}\left\{\sum_{i=1}^{m}\langle f, K_{x_i}\rangle_K^2\right\}^{1/2} \le \frac{1}{m}\left\{\sum_{i=1}^{m}\|f\|_K^2\|K_{x_i}\|_K^2\right\}^{1/2} \le \frac{\kappa\|f\|_K}{\sqrt{m}},$$

which implies $\|S_{\mathbf{x}}\| \le \frac{\kappa}{\sqrt{m}}$, where $\kappa := \sqrt{\sup_{x\in X} K(x,x)}$.

For each $(x_i, y_i) \in Z$, assume that $y_i = f_\rho(x_i) + \eta_{x_i}$, then the probability distribution $\rho(\cdot|x_i)$ of η_{x_i} has mean 0 and denote its variance by $\sigma_{x_i}^2$ with $\sigma^2 := \frac{1}{m}\sum_{i=1}^m \sigma_{x_i}^2$.

3. Convergence analysis

In the following section, we discuss the error analysis of the estimator in terms of expectation and probability under the general smoothness assumption of the target function.

The optimization functional (4) can be reformulated as

$$f_{\mathbf{z},\lambda} = arg\min_{f\in\mathcal{H}_K}\left\{\|S_{\mathbf{x}} f - \mathbf{y}\|_m^2 + \lambda_1\|f\|_K^2 + \lambda_2\|Bf\|_K^2\right\}, \tag{5}$$

where $\mathbf{y} = (y_i)_{i=1}^m$. By taking the functional derivative of $\|S_{\mathbf{x}} f - \mathbf{y}\|_m^2 + \lambda_1\|f\|_K^2 + \lambda_2\|Bf\|_K^2$ over $\mathcal{H}_K$, we get the minimizer $f_{\mathbf{z},\lambda}$.

Theorem 3.1. *For the positive value of λ_1, the functional (4) has unique minimizer:*

$$f_{\mathbf{z},\lambda} = (S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^* B)^{-1} S_{\mathbf{x}}^* \mathbf{y}. \tag{6}$$

Define $f_{\mathbf{x},\lambda}$ as the minimizer of the optimization problem

$$f_{\mathbf{x},\lambda} := arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{m}\sum_{i=1}^{m}(f(x_i) - f_\rho(x_i))^2 + \lambda_1\|f\|_K^2 + \lambda_2\|Bf\|_K^2\right\}, \tag{7}$$

which gives

$$f_{\mathbf{x},\lambda} = (S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^* B)^{-1} S_{\mathbf{x}}^* S_{\mathbf{x}} f_\rho \tag{8}$$

and

$$f_\lambda := arg\min_{f\in\mathcal{H}_K}\left\{\|f - f_\rho\|_\rho^2 + \lambda_1\|f\|_K^2 + \lambda_2\|Bf\|_K^2\right\}, \tag{9}$$

where $\|f\|_\rho = \|f\|_{\mathscr{L}^2_{\rho_X}} = \{\int_X |f(x)|^2 d\rho_X(x)\}^{1/2}$ denotes norm in $\mathscr{L}^2_{\rho_X}$. Then we get,

$$f_\lambda = (L_K + \lambda_1 I + \lambda_2 B^*B)^{-1} L_K f_\rho, \tag{10}$$

where the integral operator $L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{L}^2_{\rho_X}$ is a self-adjoint, non-negative, compact operator, defined as

$$L_K(f)(x) := \int_X K(x,t) f(t) d\rho_X(t), \quad x \in X.$$

The smoothness of the regression function f_ρ can be described in terms of integral operator L_K through general source condition [3]:

$$\Omega_{\phi,R} := \left\{ f \in \mathscr{L}^2_{\rho_X} : f = \phi(L_K)g \text{ and } \|g\|_\rho \le R \right\},$$

where ϕ is a continuous increasing index function defined on the interval $[0, \|L_K\|]$ with the assumption $\phi(0) = 0$.

Proposition 3.1. *For the positive choice of λ_1, we have*

$$\|(S^*_{\mathbf{x}} S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^*B)^{-1}\| \le \frac{1}{(\lambda^2_{\mathbf{x}} + \lambda_1 + \lambda_2 \mu_B^2)},$$

where $\lambda_{\mathbf{x}} := \inf_{f \in \mathscr{H}_K} \frac{\|S_{\mathbf{x}} f\|_m}{\|f\|_K}$ and $\mu_B := \inf_{f \in \mathscr{H}_K} \frac{\|Bf\|_K}{\|f\|_K}$.

Proof. Suppose $v = (S^*_{\mathbf{x}} S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^*B)u$. Then we have,

$$\langle (S^*_{\mathbf{x}} S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^*B)u, u\rangle_K = \langle v, u\rangle_K, \quad \forall u \in \mathscr{H}_K.$$

Applying Cauchy–Schwarz inequality, we get

$$\|S_{\mathbf{x}} u\|^2_m + \lambda_1 \|u\|^2_K + \lambda_2 \|Bu\|^2_K \le \|v\|_K \|u\|_K.$$

This gives

$$(\lambda^2_{\mathbf{x}} + \lambda_1 + \lambda_2 \mu_B^2)\|u\|^2_K \le \|v\|_K \|u\|_K.$$
$$\|u\|_K \le (\lambda^2_{\mathbf{x}} + \lambda_1 + \lambda_2 \mu_B^2)^{-1} \|v\|_K.$$

Consequently,

$$\|(S^*_{\mathbf{x}} S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^*B)^{-1}\| \le \frac{1}{(\lambda^2_{\mathbf{x}} + \lambda_1 + \lambda_2 \mu_B^2)}. \quad \square$$

Remark 3.1. It is important to note that we can apply the above framework to the sampling theory. In case of sampling theory $\lambda_{\mathbf{x}}$ may be positive (see [32]) while for learning theory, in general, it becomes zero.

The procedure of obtaining the error estimates of $f_{\mathbf{z},\lambda} - f_\rho$ consists of two steps, in first step we formulate the error bounds for $f_{\mathbf{z},\lambda} - f_\lambda$ while in second we establish the estimates of $f_\lambda - f_\rho$ in different norms.

Theorem 3.2. *Let $\mathbf{z}$ be i.i.d. samples with the hypothesis $|y| \le M$. Then under the positive choice of λ_1, we have*

$$E_{\mathbf{z}}(\|f_{\mathbf{z},\lambda} - f_\lambda\|_K) \le \frac{4\kappa M}{\sqrt{m}(\lambda^2_{\mathbf{x}} + \lambda_1 + \lambda_2 \mu_B^2)},$$

where $E_{\mathbf{z}}(\phi(\mathbf{z})) = \int_{Z^m} \phi(\mathbf{z}) d\rho(z_1) \dots d\rho(z_m)$.

Proof. To estimate the error bound, we decompose $f_{\mathbf{z},\lambda}-f_\lambda$ into $f_{\mathbf{z},\lambda}-f_{\mathbf{x},\lambda}+f_{\mathbf{x},\lambda}-\widetilde{f}_\lambda+\widetilde{f}_\lambda-f_\lambda$, where $\widetilde{f}_\lambda$ is defined by

$$\widetilde{f}_\lambda := (S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda_1 I+\lambda_2 B^*B)^{-1}L_K f_\rho.$$

Using the definitions of optimizing functions (6), (8) and (10), we get

$$f_{\mathbf{z},\lambda}-f_{\mathbf{x},\lambda}=(S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda_1 I+\lambda_2 B^*B)^{-1}(S_{\mathbf{x}}^*\mathbf{y}-S_{\mathbf{x}}^*S_{\mathbf{x}}f_\rho),$$
$$f_{\mathbf{x},\lambda}-\widetilde{f}_\lambda=(S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda_1 I+\lambda_2 B^*B)^{-1}(S_{\mathbf{x}}^*S_{\mathbf{x}}f_\rho-L_K f_\rho)$$

and

$$\begin{aligned}\widetilde{f}_\lambda-f_\lambda &= (S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda_1 I+\lambda_2 B^*B)^{-1}(L_K+\lambda_1 I+\lambda_2 B^*B)f_\lambda-f_\lambda\\ &= (S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda_1 I+\lambda_2 B^*B)^{-1}(L_K f_\lambda-S_{\mathbf{x}}^*S_{\mathbf{x}}f_\lambda).\end{aligned}$$

Under Proposition 3.1, we get

$$\begin{aligned}&\|f_{\mathbf{z},\lambda}-f_\lambda\|_K\\ &\le \frac{1}{(\lambda_{\mathbf{x}}^2+\lambda_1+\lambda_2\mu_B^2)}\left\{\|S_{\mathbf{x}}^*\mathbf{y}-S_{\mathbf{x}}^*S_{\mathbf{x}}f_\rho\|_K+\|S_{\mathbf{x}}^*S_{\mathbf{x}}f_\rho-L_K f_\rho\|_K+\|S_{\mathbf{x}}^*S_{\mathbf{x}}f_\lambda-L_K f_\lambda\|_K\right\}.\end{aligned}\tag{11}$$

Now,

$$S_{\mathbf{x}}^*(\mathbf{y}-S_{\mathbf{x}}f_\rho)=\frac{1}{m}\sum_{i=1}^m(y_i-f_\rho(x_i))K_{x_i},$$

which implies

$$\|S_{\mathbf{x}}^*(\mathbf{y}-S_{\mathbf{x}}f_\rho)\|_K^2=\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m(y_i-f_\rho(x_i))(y_j-f_\rho(x_j))\langle K_{x_i},K_{x_j}\rangle_K.$$

We define

$$E_{\mathbf{x}}(\phi(\mathbf{x})):=\int_{X^m}\phi(\mathbf{x})d\rho_X(x_1)\dots d\rho_X(x_m)\quad\text{and}$$
$$E_{\mathbf{y}}(\phi(\mathbf{z})):=\int_{Y^m}\phi(\mathbf{z})d\rho(y_1|x_1)\dots d\rho(y_m|x_m).$$

Then by the independence of the samples and $E_{\mathbf{y}}(y_i-f_\rho(x_i))=0,\ E_{\mathbf{y}}(y_i-f_\rho(x_i))^2=\sigma_{x_i}^2$, we get

$$E_{\mathbf{y}}(\|S_{\mathbf{x}}^*(\mathbf{y}-S_{\mathbf{x}}f_\rho)\|_K^2)=\frac{1}{m^2}\sum_{i=1}^m\sigma_{x_i}^2\|K_{x_i}\|_K^2.$$

Since $E_{\mathbf{x}}(\sigma_{x_i}^2)=\sigma^2(\rho)$, therefore we obtain

$$E_{\mathbf{z}}(\|S_{\mathbf{x}}^*(\mathbf{y}-S_{\mathbf{x}}f_\rho)\|_K^2)\le\frac{\kappa^2\sigma^2(\rho)}{m}.\tag{12}$$

In the light of Lemma 1 of the paper [32]

$$E_{\mathbf{x}}\left(\|S_{\mathbf{x}}^*S_{\mathbf{x}}f-L_K f\|_K\right)\le\frac{\kappa\|f\|_\rho}{\sqrt{m}},$$

from Eq. (11) together with (12) we conclude that

$$E_{\mathbf{z}}(\|f_{\mathbf{z},\lambda}-f_\lambda\|_K)\le\frac{\kappa}{\sqrt{m}(\lambda_{\mathbf{x}}^2+\lambda_1+\lambda_2\mu_B^2)}\left\{\sqrt{\sigma^2(\rho)}+\|f_\rho\|_\rho+\|f_\lambda\|_\rho\right\}.\tag{13}$$

Since f_λ is minimizer of (9), by taking $f = 0$ yields

$$\|f_\lambda - f_\rho\|_\rho^2 + \lambda_1\|f_\lambda\|_K^2 + \lambda_2\|Bf_\lambda\|_K^2 \leq \|f_\rho\|_\rho^2. \tag{14}$$

Consequently, we get

$$\|f_\lambda\|_\rho \leq 2\|f_\rho\|_\rho \quad \text{and} \quad \|f_\lambda\|_K \leq \frac{\|f_\rho\|_\rho}{\sqrt{\lambda_1 + \lambda_2\mu_B^2}}. \tag{15}$$

Then under the condition $|y| \leq M$, Eq. (13) in connection with (15) provides the desired estimate. □

Proposition 3.2. *For the positive choice of* λ_1,

$$\|(L_K + \lambda_1 I + \lambda_2 B^*B)^{-1}\| \leq \frac{1}{(\lambda_1 + \lambda_2\mu_B^2)}.$$

The proof of the proposition is similar to Proposition 3.1.

To establish the error estimates of $f_\lambda - f_\rho$ we define f_{λ_1} minimizer of the functional,

$$f_{\lambda_1} := arg\min_{f\in\mathcal{H}_K}\left\{\|f - f_\rho\|_\rho^2 + \lambda_1\|f\|_K^2\right\} \tag{16}$$

gives

$$f_{\lambda_1} = (L_K + \lambda_1 I)^{-1}L_K f_\rho.$$

Further we use the relation between the qualification of a regularization and index function ϕ.

Definition 3.1. The qualification of a regularization scheme g_λ is the maximal p such that

$$\sup_{0<\alpha\leq\|L_K\|}|1 - g_\lambda(\alpha)\alpha|\alpha^p \leq \gamma_p\lambda^p,$$

where γ_p does not depend on λ.

Definition 3.2. We say the qualification p covers ϕ if $\exists\, c > 0$ such that

$$c\frac{\alpha^p}{\phi(\alpha)} \leq \inf_{\alpha\leq\lambda\leq\|L_K\|}\frac{\lambda^p}{\phi(\lambda)},$$

where $0 < \alpha \leq \|L_K\|$.

The qualification of Tikhonov regularization (16) is 1 and $\gamma_p = 1$. Therefore the qualification of Tikhonov regularization covers ϕ is equivalent to the condition that $\alpha^{-1}\phi(\alpha)$ is a nonincreasing function.

Theorem 3.3. *For a bounded operator B with positive* λ_1, *if* $f_\rho \in \Omega_{\phi,R}$.

(i) *Under the assumption that* $\phi(\alpha)$ *and* $\alpha/\phi(\alpha)$ *are nondecreasing functions, we have*

$$\|f_\lambda - f_\rho\|_\rho \leq \phi(\lambda_1)R + \frac{2\lambda_2\|B^*B\|\,\|f_\rho\|_\rho}{(\lambda_1 + \lambda_2\mu_B^2)}. \tag{17}$$

(ii) *Under the assumption that* $\psi(\alpha) = \alpha^{-1/2}\phi(\alpha)$ *and* $\alpha/\psi(\alpha)$ *are nondecreasing functions, we have*

$$\|f_\lambda - f_\rho\|_K \leq \frac{1}{\sqrt{\lambda_1}}\left\{\phi(\lambda_1)R + \frac{\lambda_2\|B^*B\|\,\|f_\rho\|_\rho}{(\lambda_1 + \lambda_2\mu_B^2)}\right\}. \tag{18}$$

Proof. To realize the above error estimates, we decompose $f_\lambda - f_\rho$ into $f_\lambda - f_{\lambda_1} + f_{\lambda_1} - f_\rho$.

The first term can be expressed as

$$\begin{aligned} f_\lambda - f_{\lambda_1} &= (L_K + \lambda_1 I + \lambda_2 B^*B)^{-1} L_K f_\rho - f_{\lambda_1} \\ &= (L_K + \lambda_1 I + \lambda_2 B^*B)^{-1} (L_K + \lambda_1 I) f_{\lambda_1} - f_{\lambda_1} \\ &= -\lambda_2 (L_K + \lambda_1 I + \lambda_2 B^*B)^{-1} B^*B f_{\lambda_1}. \end{aligned}$$

Employing $\|\cdot\|_\rho$-norm, we get

$$\|f_\lambda - f_{\lambda_1}\|_\rho \le \lambda_2 \|(L_K + \lambda_1 I + \lambda_2 B^*B)^{-1}\| \, \|B^*B\| \, \|f_{\lambda_1}\|_\rho.$$

Applying Proposition 3.2,

$$\|f_\lambda - f_{\lambda_1}\|_\rho \le \frac{\lambda_2}{(\lambda_1 + \lambda_2 \mu_B^2)} \|B^*B\| \, \|f_{\lambda_1}\|_\rho.$$

Similarly for $\|\cdot\|_K$-norm, we get

$$\|f_\lambda - f_{\lambda_1}\|_K \le \frac{\lambda_2}{(\lambda_1 + \lambda_2 \mu_B^2)} \|B^*B\| \, \|f_{\lambda_1}\|_K.$$

Note that f_{λ_1} is minimizer of (16) and taking $f = 0$ yields

$$\|f_{\lambda_1} - f_\rho\|_\rho^2 + \lambda_1 \|f_{\lambda_1}\|_K^2 \le \|f_\rho\|_\rho^2,$$

which gives

$$\|f_{\lambda_1}\|_\rho \le 2\|f_\rho\|_\rho \quad \text{and} \quad \|f_{\lambda_1}\|_K \le \frac{\|f_\rho\|_\rho}{\sqrt{\lambda_1}}.$$

Therefore,

$$\|f_\lambda - f_{\lambda_1}\|_\rho \le \frac{2\lambda_2 \|B^*B\| \, \|f_\rho\|_\rho}{(\lambda_1 + \lambda_2 \mu_B^2)} \tag{19}$$

and

$$\|f_\lambda - f_{\lambda_1}\|_K \le \frac{\lambda_2 \|B^*B\| \, \|f_\rho\|_\rho}{\sqrt{\lambda_1}(\lambda_1 + \lambda_2 \mu_B^2)}. \tag{20}$$

The hypothesis $f_\rho \in \Omega_{\phi,R}$ implies $f_\rho = \phi(L_K)g$ for some $g \in \mathscr{L}^2_{\rho_X}$ with $\|g\|_\rho \le R$. To estimate the second term, we consider

$$f_{\lambda_1} - f_\rho = \{(L_K + \lambda_1 I)^{-1} L_K - I\} \phi(L_K) g.$$

Therefore,

$$\|f_{\lambda_1} - f_\rho\|_\rho \le \|\{(L_K + \lambda_1 I)^{-1} L_K - I\} \phi(L_K)\| \, \|g\|_\rho.$$

Then under the assumptions on ϕ described in (i), using functional calculus we get

$$\|f_{\lambda_1} - f_\rho\|_\rho \le R \sup_{\alpha \in [0, \|L_K\|]} |1 - (\alpha + \lambda_1)^{-1} \alpha| \phi(\alpha) \le R\phi(\lambda_1) \tag{21}$$

and in the same manner with the assumptions on ψ described in (ii), we get

$$\begin{aligned} \|f_{\lambda_1} - f_\rho\|_K &= \|\{(L_K + \lambda_1 I)^{-1} L_K - I\} \phi(L_K) L_K^{-1/2} g\|_\rho \\ &\le R \sup_{\alpha \in [0, \|L_K\|]} |1 - (\alpha + \lambda_1)^{-1} \alpha| \alpha^{-1/2} \phi(\alpha) \\ &= R \sup_{\alpha \in [0, \|L_K\|]} |1 - (\alpha + \lambda_1)^{-1} \alpha| \psi(\alpha) \\ &\le R\psi(\lambda_1) = R\lambda_1^{-1/2} \phi(\lambda_1). \end{aligned} \tag{22}$$

Combining the error bounds (21) and (22) together with (19) and (20), we achieve the required estimates. □

Corollary 3.1. *Under the same assumptions of Theorems 3.2 and 3.3, we obtain*

$$E_{\mathbf{z}}(\|f_{\mathbf{z},\lambda}-f_\rho\|_K) \leq C_{\rho,K}\left\{\frac{1}{\sqrt{m}\lambda_1}+\frac{\lambda_2}{\lambda_1^{3/2}}+\frac{\phi(\lambda_1)}{\lambda_1^{1/2}}\right\}$$

and

$$E_{\mathbf{z}}(\|f_{\mathbf{z},\lambda}-f_\rho\|_\rho) \leq C'_{\rho,K}\left\{\frac{1}{\sqrt{m}\lambda_1}+\frac{\lambda_2}{\lambda_1}+\phi(\lambda_1)\right\},$$

where $C_{\rho,K} := 4\kappa M + \|B^*B\|M + R$ *and* $C'_{\rho,K} := 4\kappa^2 M + 2\|B^*B\|M + R$ *are independent of the sample size.*

Corollary 3.2 (*Hölder's Source Condition*)**.** *Suppose* $\phi(t) = t^r$. *Then choosing* $\lambda_1 = m^{-\frac{1}{2r+1}}$, $\lambda_2 = m^{-\frac{r+1}{2r+1}}$, *we get*

$$E_{\mathbf{z}}(\|f_{\mathbf{z},\lambda}-f_\rho\|_K) \leq C_{\rho,K}\left(\frac{1}{m}\right)^{\frac{2r-1}{4r+2}}, \quad \left(\frac{1}{2}<r\leq\frac{3}{2}\right)$$

and for $\lambda_1 = m^{-\frac{1}{2r+2}}$, $\lambda_2 = m^{-\frac{1}{2}}$, *we get*

$$E_{\mathbf{z}}(\|f_{\mathbf{z},\lambda}-f_\rho\|_\rho) \leq C'_{\rho,K}\left(\frac{1}{m}\right)^{\frac{r}{2r+2}}, \quad (0<r\leq 1),$$

where $C_{\rho,K}$ *and* $C'_{\rho,K}$ *are defined above.*

Remark 3.2. It is worthwhile to note that the smoothness of the regression function f_ρ under the Hölder's source condition ($L_K^{-r}f_\rho \in \mathscr{L}^2_{\rho_X}$) can be interpreted in terms of r [3,12,14,33]. As greater will be the value of r, f_ρ will be more smooth. Observe that for $r \geq \frac{1}{2}$, the regression function f_ρ belongs to the reproducing kernel Hilbert space and no more depends on the distribution of marginal probability measure ρ_X which allows to consider the $\|\cdot\|_K$-norm error estimate of f_ρ.

In order to improve the error estimate discussed in Theorem 3.2, we consider the following inequality used in the paper [3] which is based on the results of Pinelis and Sakhanenko [29].

Proposition 3.3. *Let* $\mathcal{H}$ *be a real separable Hilbert space and* ξ *be a random variable on* (Ω,ρ) *with values in* $\mathcal{H}$. *If there exist two constants* Q *and* S *satisfying*

$$E\left\{\|\xi-E(\xi)\|^n_{\mathcal{H}}\right\} \leq \frac{1}{2}n!S^2Q^{n-2} \quad \forall n\geq 2,$$

then for any $0<\eta<1$ *and for all* $m\in\mathbb{N}$,

$$Prob\left\{(\omega_1,\ldots,\omega_m)\in\Omega^m : \left\|\frac{1}{m}\sum_{i=1}^m \xi(\omega_i)-E(\xi)\right\|_{\mathcal{H}} \leq 2\left(\frac{Q}{m}+\frac{S}{\sqrt{m}}\right)\log\left(\frac{2}{\eta}\right)\right\} \geq 1-\eta.$$

Theorem 3.4. *Let* $\mathbf{z}$ *be i.i.d. samples with the hypothesis* $|y|\leq M$. *Then under the positive choice of* λ_1, *for any* $0<\eta<1$,

$$\|f_{\mathbf{z},\lambda}-f_\lambda\|_K \leq \frac{6\kappa M\log(2/\eta)}{\sqrt{m}(\lambda_1+\lambda_2\mu_B^2)}$$

with confidence $1-\eta$.

Proof. From the definition of $f_{\mathbf{z},\lambda}$, we can express $f_{\mathbf{z},\lambda} - f_\lambda$ as

$$f_{\mathbf{z},\lambda} - f_\lambda = (S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^* B)^{-1}\{S_{\mathbf{x}}^* \mathbf{y} - S_{\mathbf{x}}^* S_{\mathbf{x}} f_\lambda - \lambda_1 f_\lambda - \lambda_2 B^* B f_\lambda\}. \tag{23}$$

Then $f_\lambda = (L_K + \lambda_1 I + \lambda_2 B^* B)^{-1} L_K f_\rho$ implies

$$L_K f_\rho = L_K f_\lambda + \lambda_1 f_\lambda + \lambda_2 B^* B f_\lambda.$$

Therefore (23) becomes

$$f_{\mathbf{z},\lambda} - f_\lambda = (S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^* B)^{-1}\{S_{\mathbf{x}}^* \mathbf{y} - S_{\mathbf{x}}^* S_{\mathbf{x}} f_\lambda - L_K(f_\rho - f_\lambda)\}.$$

In view of Proposition 3.1, we get

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_K \leq \frac{1}{(\lambda_{\mathbf{x}}^2 + \lambda_1 + \lambda_2 \mu_B^2)} \left\| \frac{1}{m} \sum_{i=1}^{m} (y_i - f_\lambda(x_i)) K_{x_i} - L_K(f_\rho - f_\lambda) \right\|_K. \tag{24}$$

Consider a random variable $\xi(z) = (y - f_\lambda(x))K_x$ from (Z, ρ) to Hilbert space $\mathcal{H}_K$ with

$$E_z(\xi) = \int_Z (y - f_\lambda(x)) K_x d\rho = L_K(f_\rho - f_\lambda)$$

and $\|\xi\|_K \leq \kappa(M + \|f_\lambda\|_\infty)$, $\|E_z(\xi)\|_K \leq \kappa(M + \|f_\lambda\|_\infty)$, $E_z(\|\xi\|_K^2) \leq \kappa^2 \int_Z (y - f_\lambda(x))^2 d\rho$.

In order to apply Proposition 3.3 we evaluate

$$\begin{aligned} E_z(\|\xi - E_z(\xi)\|_K^n) &\leq E_z\left\{(\|\xi\|_K + \|E_z(\xi)\|_K)^{n-2} \|\xi - E_z(\xi)\|_K^2\right\} \\ &\leq \{2\kappa(M + \|f_\lambda\|_\infty)\}^{n-2} \left\{\kappa^2 \int_Z (y - f_\lambda(x))^2 d\rho\right\}. \end{aligned} \tag{25}$$

The relation (14) says that

$$\|f_\lambda - f_\rho\|_\rho^2 \leq \|f_\rho\|_\rho^2. \tag{26}$$

We have the identity,

$$\int_Z (f(x) - y)^2 d\rho = \|f - f_\rho\|_\rho^2 + \int_Z (f_\rho(x) - y)^2 d\rho. \tag{27}$$

For the particular values $f = 0$ and $f = f_\lambda$, we get

$$\|f_\rho\|_\rho^2 + \int_Z (f_\rho(x) - y)^2 d\rho = \int_Z y^2 d\rho \tag{28}$$

and

$$\int_Z (f_\lambda(x) - y)^2 d\rho = \|f_\lambda - f_\rho\|_\rho^2 + \int_Z (f_\rho(x) - y)^2 d\rho.$$

Under the condition $|y| \leq M$, by (26) and (28), we obtain

$$\int_Z (f_\lambda(x) - y)^2 d\rho \leq \int_Z y^2 d\rho \leq M^2.$$

Then from inequality (25) we conclude that

$$E_z(\|\xi - E(\xi)\|_K^n) \leq \frac{1}{2} n! (\kappa M)^2 \{\kappa(M + \|f_\lambda\|_\infty)\}^{n-2} \quad \forall\ n \in \mathbb{N}.$$

Applying Proposition 3.3 to the random variable ξ with $Q = \kappa(M + \|f_\lambda\|_\infty)$ and $S = \kappa M$ leads to the following estimate

$$\left\| \frac{1}{m}\sum_{i=1}^{m}(y_i - f_\lambda(x_i))K_{x_i} - L_K(f_\rho - f_\lambda) \right\|_K \leq 2\left\{ \frac{\kappa(M + \|f_\lambda\|_\infty)}{m} + \frac{\kappa M}{\sqrt{m}} \right\} \log\left(\frac{2}{\eta}\right),$$

which holds with probability $1 - \eta$.

Using the norm inequality $\|f\|_\infty \leq \kappa\|f\|_K$ in connection with (15) implies that with confidence $1 - \eta$,

$$\|S_{\mathbf{x}}^* y - S_{\mathbf{x}}^* S_{\mathbf{x}} f_\lambda - L_K(f_\rho - f_\lambda)\|_K \leq \frac{2\kappa M}{\sqrt{m}} \left\{ \frac{1}{\sqrt{m}} + \frac{\kappa}{\sqrt{m(\lambda_1 + \lambda_2\mu_B^2)}} + 1 \right\} \log\left(\frac{2}{\eta}\right). \tag{29}$$

When $\frac{\kappa}{\sqrt{m(\lambda_1+\lambda_2\mu_B^2)}} \leq 1$, then with inequality (24) the estimate becomes

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_K \leq \frac{6\kappa M \log(2/\eta)}{\sqrt{m}(\lambda_1 + \lambda_2\mu_B^2)},$$

which holds with probability $1 - \eta$.

In case $\frac{\kappa}{\sqrt{m(\lambda_1+\lambda_2\mu_B^2)}} > 1$, we observe that

$$\frac{2M}{\sqrt{\lambda_1 + \lambda_2\mu_B^2}} < \frac{6\kappa M \log(2/\eta)}{\sqrt{m}(\lambda_1 + \lambda_2\mu_B^2)}. \tag{30}$$

Since $1 > \frac{1}{3\log\left(\frac{2}{\eta}\right)}$. With $f = 0$ from the functional (4) we find that

$$\|f_{\mathbf{z},\lambda}\|_K \leq \frac{M}{\sqrt{\lambda_1 + \lambda_2\mu_B^2}},$$

which together with (15) implies

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_K < \frac{2M}{\sqrt{\lambda_1 + \lambda_2\mu_B^2}}.$$

Therefore (30) says that in this case also

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_K \leq \frac{6\kappa M \log(2/\eta)}{\sqrt{m}(\lambda_1 + \lambda_2\mu_B^2)}.$$

So in both cases we conclude the desired error estimate. □

Theorem 3.5. *Let* $\mathbf{z}$ *be i.i.d. samples with the hypothesis* $|y| \leq M$. *Then under the positive choice of* λ_1, *for any* $0 < \eta < 1$, *with confidence* $1 - \eta$ *there holds*

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_\rho \leq \frac{12\kappa M \log(4/\eta)}{\sqrt{m(\lambda_1 + \lambda_2\mu_B^2)}}$$

provided that

$$(\lambda_1 + \lambda_2\mu_B^2) \geq \frac{8\kappa^2 \log(4/\eta)}{\sqrt{m}}. \tag{31}$$

Proof. Under the isometry $L_K^{1/2} : \mathscr{L}^2_{\rho_X} \to \mathcal{H}_K$ we have

$$\begin{aligned}\|f_{\mathbf{z},\lambda} - f_\lambda\|_\rho &= \|L_K^{1/2}(S_{\mathbf{x}}^*S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^*B)^{-1}\{S_{\mathbf{x}}^*\mathbf{y} - S_{\mathbf{x}}^*S_{\mathbf{x}}f_\lambda - L_K(f_\rho - f_\lambda)\}\|_K \\ &\leq \|L_K^{1/2}(S_{\mathbf{x}}^*S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^*B)^{-1}\|\,\|S_{\mathbf{x}}^*\mathbf{y} - S_{\mathbf{x}}^*S_{\mathbf{x}}f_\lambda - L_K(f_\rho - f_\lambda)\|_K. \end{aligned} \tag{32}$$

We can write $(S_{\mathbf{x}}^*S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^*B)$ as

$$S_{\mathbf{x}}^*S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^*B = \{I - (L_K - S_{\mathbf{x}}^*S_{\mathbf{x}})(L_K + \lambda_1 I + \lambda_2 B^*B)^{-1}\}(L_K + \lambda_1 I + \lambda_2 B^*B),$$

which follows

$$\begin{aligned}&L_K^{1/2}(S_{\mathbf{x}}^*S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^*B)^{-1} \\ &\quad = L_K^{1/2}(L_K + \lambda_1 I + \lambda_2 B^*B)^{-1}\{I - (L_K - S_{\mathbf{x}}^*S_{\mathbf{x}})(L_K + \lambda_1 I + \lambda_2 B^*B)^{-1}\}^{-1}. \end{aligned} \tag{33}$$

Consider a random variable $\xi(x) = K_x\langle\cdot, K_x\rangle_K$ with

$$E_x(\xi) = \int_X K_x\langle\cdot, K_x\rangle_K d\rho_X(x) = L_K$$

and $\|\xi\| \leq \kappa^2$, $\|E_x(\xi)\| \leq \kappa^2$, $E_x(\|\xi\|^2) \leq \kappa^4$.

In order to apply Proposition 3.3 we evaluate

$$\begin{aligned}E_x(\|\xi - E_x(\xi)\|^n) &\leq E_x\left\{(\|\xi\| + \|E_x(\xi)\|)^{n-2}\|\xi - E_x(\xi)\|^2\right\} \\ &\leq (2\kappa^2)^{n-2}(\kappa^4) \leq \frac{1}{2}n!(\kappa^2)^{n-2}(\kappa^2)^2 \quad \forall n \in \mathbb{N}.\end{aligned}$$

Applying Proposition 3.3 to the random variable ξ with $Q = S = \kappa^2$, we conclude that

$$\|S_{\mathbf{x}}^*S_{\mathbf{x}} - L_K\| \leq 2\left(\frac{\kappa^2}{m} + \frac{\kappa^2}{\sqrt{m}}\right)\log\left(\frac{2}{\eta}\right)$$

holds with probability $1 - \eta$.

Proposition 3.2 implies that with confidence $1 - \eta/2$,

$$\|(L_K - S_{\mathbf{x}}^*S_{\mathbf{x}})(L_K + \lambda_1 I + \lambda_2 B^*B)^{-1}\| \leq \frac{4\kappa^2\log(4/\eta)}{\sqrt{m}(\lambda_1 + \lambda_2\mu_B^2)}.$$

Since we assume $(\lambda_1 + \lambda_2\mu_B^2) \geq \frac{8\kappa^2\log(4/\eta)}{\sqrt{m}}$, this gives

$$\|(L_K - S_{\mathbf{x}}^*S_{\mathbf{x}})(L_K + \lambda_1 I + \lambda_2 B^*B)^{-1}\| \leq \frac{1}{2}$$

holds with probability $1 - \eta/2$.

Using (33) we conclude that

$$\|L_K^{1/2}(S_{\mathbf{x}}^*S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^*B)^{-1}\| \leq 2\|L_K^{1/2}(L_K + \lambda_1 I + \lambda_2 B^*B)^{-1}\| \leq \frac{2}{\sqrt{\lambda_1 + \lambda_2\mu_B^2}}, \tag{34}$$

with confidence $1 - \eta/2$.

Under the parameters choice (31) from (29) with confidence $1 - \eta/2$, we obtain

$$\|S_{\mathbf{x}}^*\mathbf{y} - S_{\mathbf{x}}^*S_{\mathbf{x}}f_\lambda - L_K(f_\rho - f_\lambda)\|_K \leq \frac{6\kappa M\log(4/\eta)}{\sqrt{m}}. \tag{35}$$

Eq. (32) in connection with (34) and (35) provides the desired result. □

Theorem 3.6. *Suppose $\Theta(t) = \phi(t)\sqrt{t}$. Then under the assumptions of Theorems 3.3 and 3.4 with the parameters choice, $\lambda_1 = \Theta^{-1}(m^{-1/2})$ and $\lambda_2 = (m^{-1}\Theta^{-1}(m^{-1/2}))^{1/2}$,*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_K \leq \widetilde{C}_{\rho,K}\phi(\Theta^{-1}(m^{-1/2}))(\Theta^{-1}(m^{-1/2}))^{-1/2}\log(4/\eta)$$

and under the assumptions of Theorems 3.3 and 3.5 with the above parameters choice,

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_\rho \leq \widetilde{\widetilde{C}}_{\rho,K}\phi(\Theta^{-1}(m^{-1/2}))\log(4/\eta)$$

holds with confidence $1-\eta$, where

$$\widetilde{C}_{\rho,K} := 6\kappa M + \|B^*B\|M + R$$

and

$$\widetilde{\widetilde{C}}_{\rho,K} := 12\kappa M + 2\|B^*B\|M + R$$

are constants independent of sample size.

Corollary 3.3. *Under the same assumptions of Theorem 3.6 corresponding to Hölder's source condition $\phi(t) = t^r$, we get*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_K \leq \widetilde{C}_{\rho,K}\left(m^{-\frac{2r-1}{4r+2}}\right)\log(4/\eta), \quad \left(\frac{1}{2} < r \leq \frac{3}{2}\right)$$

and

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_\rho \leq \widetilde{\widetilde{C}}_{\rho,K}\left(m^{-\frac{r}{2r+1}}\right)\log(4/\eta), \quad (0 < r \leq 1)$$

with confidence $1-\eta$, where $\widetilde{C}_{\rho,K}$ and $\widetilde{\widetilde{C}}_{\rho,K}$ are as in Theorem 3.6.

Remark 3.3. The error estimates obtained in Corollary 3.3 are order optimal. Due to the fact that multi-penalty regularization reduces to single penalty Tikhonov regularization with the operator $B = I$ but Tikhonov regularization suffers saturation effect. Hence the order of convergence cannot be improved after a certain regularity level r ($r = 1$ for $\|\cdot\|_\rho$-norm and $r = \frac{3}{2}$ for $\|\cdot\|_K$-norm). This is a well-known fact in inverse problem theory [20].

Remark 3.4. We observe from Theorems 3.2 and 3.5 that in $\|\cdot\|_\rho$-norm, the estimate for $f_{\mathbf{z},\lambda} - f_\lambda$ can be improved for special values of $\lambda = (\lambda_1, \lambda_2)$ based on (31) while in the light of Theorem 3.4 we conclude that bound of $f_{\mathbf{z},\lambda} - f_\lambda$ becomes better irrespective of choice of λ in $\|\cdot\|_K$-norm.

3.1. Manifold Learning as multi-penalty regularization

Belkin et al. [4] proposed a learning algorithm based on multi-penalty regularization to incorporate the structure of probability measure as

$$f^* = arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{m}\sum_{i=1}^{m}\ell(x_i, y_i, f) + \lambda_A\|f\|_K^2 + \lambda_I\|f\|_I^2\right\}, \tag{36}$$

where $\|f\|_I^2$ is the smoothness penalty which deals with intrinsic structure of the probability distribution ρ_X while the first penalty controls the complexity of the function in the ambient space. When the support of marginal probability measure ρ_X is a compact submanifold $\mathcal{M} \subset \mathbb{R}^n$. Then, naturally we can take $\|f\|_I^2 = \int_{x\in\mathcal{M}}\|\nabla_{\mathcal{M}} f\|^2 d\rho_X(x)$, where $\nabla_{\mathcal{M}}$ is the gradient of f along the manifold $\mathcal{M}$.

For the square loss function, the optimization problem will be

$$f^* = arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{m}\sum_{i=1}^{m}S(x_i, y_i, f) + \lambda_A\|f\|_K^2 + \lambda_I\int_{x\in\mathcal{M}}\|\nabla_{\mathcal{M}} f\|^2 d\rho_X(x)\right\}. \tag{37}$$

In general, the marginal probability measure ρ_X is unknown. So the second penalty can be approximated by the graph Laplacian associated to the labeled data $\{(x_i, y_i)\}_{i=1}^m$ and unlabeled data $\{x_j\}_{j=m+1}^n$, then the optimization problem can be expressed as

$$f^* = arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2+\lambda_A\|f\|_K^2+\lambda_I\sum_{i,j=1}^{n}(f(x_i)-f(x_j))^2\omega_{ij}\right\}$$
$$= arg\min_{f\in\mathcal{H}_K}\left\{\|S_{\mathbf{x}}f-\mathbf{y}\|_m^2+\lambda_A\|f\|_K^2+\lambda_I\|(S_{\mathbf{x}'}^*LS_{\mathbf{x}'})^{1/2}f\|_K^2\right\}, \tag{38}$$

where $\mathbf{x}' = \{x_i \in X : 1 \le i \le n\}$ and $L = D - W$ with $W = (\omega_{ij})$ is a weight matrix with non-negative entries and D is a diagonal matrix with $D_{ii} = \sum_{j=1}^n \omega_{ij}$.

We observe that the functional (38) is a particular case of (5) with $B = (S_{\mathbf{x}'}^*LS_{\mathbf{x}'})^{1/2}$. To evaluate error bound for the optimizer f^* we need bound for $\|B^*B\|$.

Proposition 3.4. *For* $B = (S_{\mathbf{x}'}^*LS_{\mathbf{x}'})^{1/2}$,

$$\|B^*B\| \le 2\omega\kappa^2.$$

Proof.

$$\|S_{\mathbf{x}'}f\|_n = \frac{1}{n}\left\{\sum_{i=1}^{n}\langle f, K_{x_i}\rangle_K^2\right\}^{1/2} \le \frac{1}{n}\left\{\sum_{i=1}^{n}\|f\|_K^2\|K_{x_i}\|_K^2\right\}^{1/2} \le \frac{\kappa\|f\|_K}{\sqrt{n}}$$

which implies

$$\|S_{\mathbf{x}'}\| \le \frac{\kappa}{\sqrt{n}}.$$
$$\|L\| \le 2\max_{1\le j\le n}\left\{\sum_{i=1,\ i\ne j}^{n}\omega_{ij}\right\} \le 2(n-1)\omega,$$

where $\omega = \max_{1\le i,j\le n}\{\omega_{ij}\}$. Therefore

$$\|B^*B\| \le \|L\|\,\|S_{\mathbf{x}'}\|^2 \le 2\omega\kappa^2. \quad \square$$

In the light of Corollary 3.3, for the bounded operator $B = (S_{\mathbf{x}'}^*LS_{\mathbf{x}'})^{1/2}$, the error bounds for the optimizer $f^* = (S_{\mathbf{x}}^*S_{\mathbf{x}} + \lambda_A I + \lambda_I S_{\mathbf{x}'}^*LS_{\mathbf{x}'})^{-1}S_{\mathbf{x}}^*\mathbf{y}$ of the functional (38) can be estimated.

It is important to observe that in Corollaries 3.1–3.3, we produced a parameter choice dependent on smoothness of the target function which is not known in practice generally. In the preceding section we are presenting a data-driven approach which requires information about the samples only in order to select the regularization parameter.

4. Parameter choice rule

One of the main issue for regularization schemes is to choose appropriate parameters. Various parameter choice strategies are studied in the context of inverse problems. The most widely used approach to select regularization parameters for ill-posed inverse problems is discrepancy principle [16,19,20,26,27]. S. Lu et al. [19] and S. Lu et al. [22] also considered the discrepancy principle to choose the regularization parameters for multi-penalty regularization in learning theory framework. Discrepancy principle produces a discrepancy curve of regularization parameters $\lambda = (\lambda_1, \lambda_2)$. Our aim is to choose appropriate point on the discrepancy curve such that it incorporates various features in the regularized estimator. In order to adaptively select the first parameter we employ the well-established data-driven parameter choice rule the balancing principle [11] which does not require any data-splitting [7] and any prior information about the regression function. After

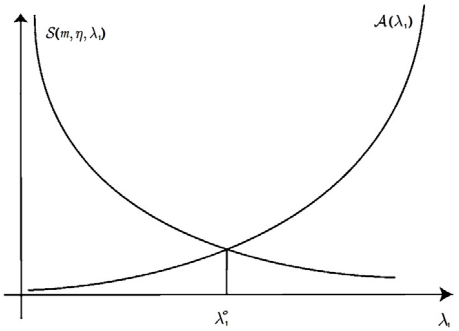

Fig. 1. Sample error vs Approximation error.

the selection of first parameter we obtain the corresponding point $\bar{\lambda} = (\bar{\lambda}_1, \bar{\lambda}_2)$ on the discrepancy curve.

We choose first parameter λ_1 according to the well-known balancing principle discussed in the paper [11]. In single-penalty regularization, the error estimates consist of the two terms, sample error and approximation error. Generally, sample error is a decreasing function of λ_1 while approximation error is increasing function of λ_1. The principle tries to provide a value of regularization parameter λ_1 for which both the errors are equal.

Assume an error estimate of the form:

$$\|f_{\mathbf{z},\lambda_1} - f_\rho\| \leq S(m,\eta,\lambda_1) + \mathcal{A}(\lambda_1) \tag{39}$$

and further assume that $S(m,\eta,\lambda_1) = \frac{\alpha(\eta)}{\omega(\lambda_1)\gamma(m)}$, where $\omega(\lambda_1)$, $\gamma(m)$ are positive and monotonically increasing function of λ_1, m respectively.

In general $\alpha(\eta) \geq 1$, therefore for sake of convenience, error bound can be expressed as

$$\|f_{\mathbf{z},\lambda_1} - f_\rho\| \leq \alpha(\eta)\left(\frac{1}{\omega(\lambda_1)\gamma(m)} + \mathcal{A}(\lambda_1)\right).$$

It can be observed from Fig. 1 that there exists an ideal parameter λ_1^o with same sample and approximation error, i.e.,

$$\|f_{\mathbf{z},\lambda_1^o} - f_\rho\| \leq 2S(m,\eta,\lambda_1^o) = 2\mathcal{A}(\lambda_1^o). \tag{40}$$

Consider one-parameter least square regularization which is a special case of multi-penalty regularization (4) with $\lambda_2 = 0$,

$$\min_{f\in\mathcal{H}_K}\left\{\frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2 + \lambda_1\|f\|_K^2\right\}. \tag{41}$$

For the regularization, the error estimate [33] is given as

$$\|f_{\mathbf{z},\lambda_1} - f_\rho\|_K \leq C\log\left(\frac{2}{\eta}\right)\left(\frac{1}{\lambda_1\sqrt{m}} + \lambda_1^{r-\frac{1}{2}}\right), \quad \frac{1}{2} < r \leq \frac{3}{2}$$

with probability $1-\eta$.

Here we observe that single-penalty Tikhonov regularization has error bounds of type (39) with $\alpha(\eta) = \log\left(\frac{2}{\eta}\right)$, $\omega(\lambda_1) = \lambda_1$, $\mathcal{A}(\lambda_1) = \lambda_1^{r-\frac{1}{2}}$ and $\gamma(m) = \sqrt{m}$.

Consider a discretized ordered sequence of parameters $(\lambda_1^i)_{i\in\mathbb{N}}$ such that λ_1^o falls within the grid. Then the estimate for λ_1^o is given by

$$\bar{\lambda}_1 = \max\left\{\lambda_1^i : \|f_{\mathbf{z},\lambda_1^j} - f_{\mathbf{z},\lambda_1^{j-1}}\|_K \leq \frac{4C\alpha(\eta)}{\omega(\lambda_1^{j-1})\gamma(m)} = \frac{4C\log(2/\eta)}{\lambda_1^{j-1}\sqrt{m}},\ j = 1,\ldots,i-1\right\}. \tag{42}$$

To get the detailed idea of such a choice for the balancing parameter λ_1^o and for further discussion we refer to the paper [11]. From (40) it is clear that $\lambda_1^o = m^{-\frac{1}{2r+1}} > m^{-\frac{1}{2}}$. Therefore choose a geometric sequence as

$$\lambda_1^i = \lambda_1^{start} q^i, \quad \text{with } q > 1 \text{ and } \lambda_1^{start} \leq \frac{1}{\sqrt{m}}. \tag{43}$$

For a discretized data (43) containing $\lambda_1^o(m)$, the best approximation for $\lambda_1^o(m)$ will be

$$\lambda_1^* = \max\left\{\lambda_1^i : \mathcal{A}(\lambda_1^i) \leq \frac{1}{\omega(\lambda_1^i)\gamma(m)}\right\}. \tag{44}$$

Here we are discussing the error estimate of single-penalty regularizer with the balancing principle parameter choice which is useful to analyze the convergence issues of multi-penalty regularizer in accordance with the balanced-discrepancy principle. In order to prove the following theorem we use the same ideas of [11], which are applied to establish a general error estimates for single-penalty regularization.

Theorem 4.1. *For a sequence of regularization parameter values according to (43), for $0 < \eta < 1$, with confidence $1 - \eta$,*

$$\|f_{\mathbf{z},\bar{\lambda}_1} - f_\rho\|_K \leq C'\left(\frac{\log(2/\eta)}{\lambda_1^o\sqrt{m}}\right),$$

where $C' = Cq\left(\frac{6q-2}{q-1}\right)$.

Proof. To determine the error bound for $f_{\mathbf{z},\bar{\lambda}_1} - f_\rho$, first we relate $f_{\mathbf{z},\bar{\lambda}_1}$ and $f_{\mathbf{z},\lambda_1^*}$. For this consider $\lambda_1^l = \lambda_1^* \leq \bar{\lambda}_1 = \lambda_1^m$, then

$$\|f_{\mathbf{z},\bar{\lambda}_1} - f_{\mathbf{z},\lambda_1^*}\|_K \leq \sum_{j=l+1}^{m} \|f_{\mathbf{z},\lambda_1^j} - f_{\mathbf{z},\lambda_1^{j-1}}\|_K.$$

By the definition of $\bar{\lambda}_1$, we get

$$\begin{aligned}\|f_{\mathbf{z},\bar{\lambda}_1} - f_{\mathbf{z},\lambda_1^*}\|_K &\leq \left(\frac{4C\log(2/\eta)}{\sqrt{m}}\right)\sum_{j=l+1}^{m}\frac{1}{\lambda_1^{j-1}} = \left(\frac{4C\log(2/\eta)}{\sqrt{m}}\right)\sum_{j=0}^{m-l-1}\frac{1}{\lambda_1^* q^j} \\ &= 4C\left(\frac{1-\frac{1}{q^{m-l}}}{1-\frac{1}{q}}\right)\frac{\log(2/\eta)}{\lambda_1^*\sqrt{m}}.\end{aligned} \tag{45}$$

Since $\lambda_1^* \leq \lambda_1^o$, therefore (44) says

$$\|f_{\mathbf{z},\lambda_1^*} - f_\rho\|_K \leq \frac{2C\log(2/\eta)}{\lambda_1^*\sqrt{m}}. \tag{46}$$

From (45) and (46), we get

$$\|f_{\mathbf{z},\bar{\lambda}_1} - f_\rho\|_K \leq C\left(\frac{6q-2}{q-1}\right)\frac{\log(2/\eta)}{\lambda_1^*\sqrt{m}}. \tag{47}$$

Assume that $\lambda_1^* = \lambda_1^l \leq \lambda_1^o \leq \lambda_1^{l+1}$. Then we conclude that

$$\frac{1}{q\lambda_1^*} \leq \frac{1}{\lambda_1^o}. \tag{48}$$

Therefore,

$$\|f_{\mathbf{z},\bar{\lambda}_1} - f_\rho\|_K \leq Cq\left(\frac{6q-2}{q-1}\right)\frac{\log(2/\eta)}{\lambda_1^o\sqrt{m}}. \quad \square$$

Now we choose second parameter λ_2 corresponding to $\bar{\lambda}_1$ using discrepancy principle,

$$\{\lambda \in \mathbb{R}^2 : \|S_{\mathbf{x}} f_{\mathbf{z},\lambda} - \mathbf{y}\|_m = c\delta\}, \tag{49}$$

where $\|S_{\mathbf{x}} f_\rho - \mathbf{y}\|_m \leq \delta$ and c is some positive constant.

As we know that

$$f_{\mathbf{z},\lambda} = (S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^* B)^{-1} S_{\mathbf{x}}^* \mathbf{y}.$$

It gives

$$S_{\mathbf{x}}^* S_{\mathbf{x}} f_{\mathbf{z},\lambda} + \lambda_1 f_{\mathbf{z},\lambda} + \lambda_2 B^* B f_{\mathbf{z},\lambda} = S_{\mathbf{x}}^* \mathbf{y}.$$

Taking inner product with $f_{\mathbf{z},\lambda}$, we get

$$\|S_{\mathbf{x}} f_{\mathbf{z},\lambda}\|_m^2 + \lambda_1 \|f_{\mathbf{z},\lambda}\|_K^2 + \lambda_2 \|B f_{\mathbf{z},\lambda}\|_K^2 = \langle \mathbf{y}, S_{\mathbf{x}} f_{\mathbf{z},\lambda}\rangle_m. \tag{50}$$

Discrepancy principle $\|S_{\mathbf{x}} f_{\mathbf{z},\lambda} - \mathbf{y}\|_m = c\delta$ implies

$$\|S_{\mathbf{x}} f_{\mathbf{z},\lambda}\|_m^2 + \|\mathbf{y}\|_m^2 - 2\langle S_{\mathbf{x}} f_{\mathbf{z},\lambda}, \mathbf{y}\rangle_m = c^2\delta^2.$$

From (50), we get

$$\|\mathbf{y}\|_m^2 - \|S_{\mathbf{x}} f_{\mathbf{z},\lambda}\|_m^2 - 2\lambda_1 \|f_{\mathbf{z},\lambda}\|_K^2 - 2\lambda_2 \|B f_{\mathbf{z},\lambda}\|_K^2 = c^2\delta^2. \tag{51}$$

This is easy to see that this equation is equivalent to the following form and for a given $\bar{\lambda}_1$, $\lambda_2 = \lambda_2^{k+1}$ can be updated as

$$\lambda_2^{k+1} = \frac{2(\lambda_2^k)^2 \|B f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)}\|_K^2}{\|\mathbf{y}\|_m^2 - \|S_{\mathbf{x}} f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)}\|_m^2 - 2\bar{\lambda}_1 \|f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)}\|_K^2 - c^2\delta^2}. \tag{52}$$

Continue the process until $|\lambda_2^{k+1} - \lambda_2^k| > \varepsilon$, for some $\varepsilon > 0$.

Algorithm. 1. For a discretized sequence according to (43), calculate $\lambda_1 = \bar{\lambda}_1$ with the criteria (42) of the balancing principle.
2. For an initial value λ_2^0, c, δ, start with $k = 0$.
3. Calculate $f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)}$ and update $\lambda_2 = \lambda_2^{k+1}$ according to (52).
4. If stopping criteria $|\lambda_2^{k+1} - \lambda_2^k| < \varepsilon$ satisfied then stop otherwise set $k = k + 1$ and GOTO (3).

Proposition 4.1. *For a non-negative parameter λ_2^k satisfying $\|S_{\mathbf{x}} f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)} - \mathbf{y}\|_m > c\delta$, $\lambda_2 = \lambda_2^{k+1}$ given by the formula (52) is a non-negative value with $\lambda_2^{k+1} \leq \lambda_2^k$.*

Proof. From Eq. (52), $\lambda_2^{k+1} - \lambda_2^k$ can be expressed as

$$\begin{aligned}\lambda_2^{k+1} - \lambda_2^k &= \lambda_2^k \left\{ \frac{2\lambda_2^k \|B f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)}\|_K^2}{\|\mathbf{y}\|_m^2 - \|S_{\mathbf{x}} f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)}\|_m^2 - 2\bar{\lambda}_1 \|f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)}\|_K^2 - c^2\delta^2} - 1 \right\} \\ &= \lambda_2^k \left\{ \frac{c^2\delta^2 - \|S_{\mathbf{x}} f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)} - \mathbf{y}\|_m^2}{\|\mathbf{y}\|_m^2 - \|S_{\mathbf{x}} f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)}\|_m^2 - 2\bar{\lambda}_1 \|f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)}\|_K^2 - c^2\delta^2} \right\}. \end{aligned} \tag{53}$$

Under the assumption $\|S_{\mathbf{x}} f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)} - \mathbf{y}\|_m^2 > c^2\delta^2$, from (50) we get

$$\{\|\mathbf{y}\|_m^2 - \|S_{\mathbf{x}} f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)}\|_m^2 - 2\bar{\lambda}_1 \|f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)}\|_K^2 - c^2\delta^2\} > 2\lambda_2^k \|B f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)}\|^2 \geq 0.$$

So using (53) we conclude that

$$\lambda_2^{k+1} \leq \lambda_2^k$$

and (52) says $\lambda_2^{k+1} \geq 0$. □

Suppose $\{\lambda_2^k\}$ is a sequence according to (52) satisfying $\|S_{\mathbf{x}} f_{\mathbf{z},(\bar{\lambda}_1,\lambda_2^k)} - \mathbf{y}\|_m > c\delta$. Then from Proposition 4.1 follows that $\{\lambda_2^k\}$ converges to some $\bar{\lambda}_2$ and under continuity of $f_{\mathbf{z},\lambda}$ as a function of λ from (52) we obtain $\|S_{\mathbf{x}} f_{\mathbf{z},\bar{\lambda}} - \mathbf{y}\|_m = c\delta$.

Remark 4.1. In order to implement the balanced-discrepancy principle we require the knowledge about the free parameter c and the perturbation δ. Proposition 4.2 and Lemma 4.1 may provide the information about the behavior of these parameters to produce appropriate parameter choice.

4.1. Convergence analysis of the balanced-discrepancy principle

In the preceding section we are analyzing, one of the main result of this paper, the convergence of multi-penalty regularizer with the balanced-discrepancy principle parameter choice. Theorem 4.2 shows that multi-penalty regularizer $f_{\mathbf{z},\lambda}$ achieves the order of convergence same as single-penalty regularizer corresponding to the balancing principle.

Proposition 4.2. *Let* $\mathbf{z}$ *be i.i.d. random samples with the assumption* $|y| \leq M$, *then for* $0 < \eta < 1$,

$$Prob_{\mathbf{z}\in Z^m}\left\{\|S_{\mathbf{x}} f_\rho - \mathbf{y}\|_m \leq 2\left(\frac{M}{m} + \sqrt{\frac{\sigma^2(\rho)}{m}}\right)\log\left(\frac{2}{\eta}\right)\right\} \geq 1-\eta$$

with probability $1-\eta$.

Proof. Let $\xi : X \times Y \to \mathbb{R}$ be a random variable which is defined as

$$\xi(z) = f_\rho(x) - y,$$

which has mean zero and variance $\sigma^2(\rho)$. Therefore,

$$E_z(|\xi(z) - E_z(\xi(z))|^n) \leq \frac{1}{2}n!\sigma^2(\rho)M^{n-2} \quad \forall n \geq 2.$$

Then from Proposition 3.3 with $Q = M$ and $S = \sqrt{\sigma^2(\rho)}$ follows the result. □

Let $\bar{\lambda}_2$ is parameter corresponding to $\bar{\lambda}_1$ which satisfies discrepancy principle,

$$\bar{\lambda}_2 := \left\{\lambda_2 \in \mathbb{R} : \|S_{\mathbf{x}} f_{\mathbf{z}}(\bar{\lambda}_1, \lambda_2) - \mathbf{y}\|_m = c\delta\right\}. \tag{54}$$

Theorem 4.2. *For the parameter choice* $\bar{\lambda} = (\bar{\lambda}_1, \bar{\lambda}_2)$ *described in* (42) *and* (54), *for* $0 < \eta < 1$ *with probability* $1-\eta$ *holds*

$$\|f_{\mathbf{z},\bar{\lambda}} - f_\rho\|_K \leq \widehat{C}\left(\frac{1}{\sqrt{m}\lambda_1^o} + \frac{\bar{\lambda}_2}{m(\lambda_1^o)^2}\right)\log\left(\frac{4}{\eta}\right),$$

where $\widehat{C} = C' + 4\kappa cM\|B^*B\|$.

Proof. To evaluate error bound for $f_{\mathbf{z},\bar{\lambda}} - f_\rho$, we decompose it into $f_{\mathbf{z},\bar{\lambda}} - f_{\mathbf{z},\bar{\lambda}_1}$ and $f_{\mathbf{z},\bar{\lambda}_1} - f_\rho$.

We estimate the error bound for first term:

$$f_{\mathbf{z},\bar{\lambda}} = (S_{\mathbf{x}}^* S_{\mathbf{x}} + \bar{\lambda}_1 I + \bar{\lambda}_2 B^*B)^{-1} S_{\mathbf{x}}^* \mathbf{y}, \tag{55}$$

which implies

$$f_{\mathbf{z},\bar{\lambda}} = -(\bar{\lambda}_1 I + \bar{\lambda}_2 B^*B)^{-1} S_{\mathbf{x}}^*(S_{\mathbf{x}} f_{\mathbf{z},\bar{\lambda}} - \mathbf{y}). \tag{56}$$

Now,

$$f_{\mathbf{z},\bar{\lambda}} - f_{\mathbf{z},\bar{\lambda}_1} = f_{\mathbf{z},\bar{\lambda}} - (S_{\mathbf{x}}^* S_{\mathbf{x}} + \bar{\lambda}_1 I)^{-1} S_{\mathbf{x}}^* \mathbf{y}.$$

Using (55), we obtain

$$f_{\mathbf{z},\bar{\lambda}} - f_{\mathbf{z},\bar{\lambda}_1} = -\bar{\lambda}_2 (S_{\mathbf{x}}^* S_{\mathbf{x}} + \bar{\lambda}_1 I)^{-1} B^* B f_{\mathbf{z},\bar{\lambda}}.$$

From (56), we get

$$\|f_{\mathbf{z},\bar{\lambda}} - f_{\mathbf{z},\bar{\lambda}_1}\|_K \leq c\delta\bar{\lambda}_2 \|(S_{\mathbf{x}}^* S_{\mathbf{x}} + \bar{\lambda}_1 I)^{-1}\| \, \|B^* B\| \, \|(\bar{\lambda}_1 I + \bar{\lambda}_2 B^* B)^{-1}\| \, \|S_{\mathbf{x}}^*\|.$$

It gives

$$\|f_{\mathbf{z},\bar{\lambda}} - f_{\mathbf{z},\bar{\lambda}_1}\|_K \leq \frac{\kappa c\delta}{\sqrt{m}} \left(\frac{\bar{\lambda}_2}{\bar{\lambda}_1^2}\right) \|B^* B\|.$$

Proposition 4.2 implies that

$$\|f_{\mathbf{z},\bar{\lambda}} - f_{\mathbf{z},\bar{\lambda}_1}\|_K \leq 2\kappa c \|B^* B\| \left(\frac{\bar{\lambda}_2}{\sqrt{m}\bar{\lambda}_1^2}\right) \left\{\frac{M}{m} + \sqrt{\frac{\sigma^2(\rho)}{m}}\right\} \log\left(\frac{2}{\eta}\right),$$

with probability $1 - \eta$.

From the definition of λ_1^o we observe that $\lambda_1^o \leq \bar{\lambda}_1$ which gives

$$Prob_{\mathbf{z}\in Z^m} \left\{ \|f_{\mathbf{z},\bar{\lambda}} - f_{\mathbf{z},\bar{\lambda}_1}\|_K \leq \left(\frac{4\kappa c M \|B^* B\| \bar{\lambda}_2}{m(\lambda_1^o)^2}\right) \log\left(\frac{2}{\eta}\right) \right\} \geq 1 - \eta. \tag{57}$$

Combining Theorem 4.1 together with (57) follows the desired estimate. □

The choice of free parameter c in discrepancy principle plays role in finding the appropriate parameters. If we have some prior estimates of $\|f_{\mathbf{z},\lambda} - f_\rho\|$, then we can find the range of constant c in which it lies.

Lemma 4.1. *For a given error bound of $\|f_{\mathbf{z},\lambda} - f_\rho\| \leq S$, we have*

$$c \leq \frac{\kappa S}{\sqrt{m}\delta} + 1.$$

Proof. We have

$$S_{\mathbf{x}}(f_{\mathbf{z},\lambda} - f_\rho) = (S_{\mathbf{x}} f_{\mathbf{z},\lambda} - \mathbf{y}) - (S_{\mathbf{x}} f_\rho - \mathbf{y}),$$

which says

$$\|S_{\mathbf{x}} f_{\mathbf{z},\lambda} - \mathbf{y}\| - \|S_{\mathbf{x}} f_\rho - \mathbf{y}\| \leq \|S_{\mathbf{x}}\| \, \|f_{\mathbf{z},\lambda} - f_\rho\|.$$

Using the bounds we get

$$c\delta - \delta \leq \frac{\kappa}{\sqrt{m}} \|f_{\mathbf{z},\lambda} - f_\rho\|, \tag{58}$$

which gives

$$c \leq \frac{\kappa \|f_{\mathbf{z},\lambda} - f_\rho\|}{\sqrt{m}\delta} + 1.$$

Then follows the required estimate. □

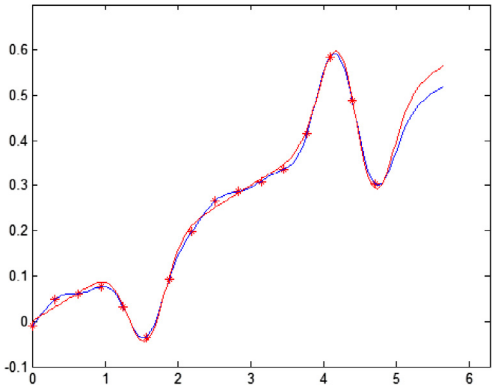

Fig. 2. The target function f_ρ (red line) and its estimator $f_{\mathbf{z},\lambda_1}$ (blue line) based on the balancing principle with $\lambda_1 = 1.4835 \times 10^{-3}$ and the empirical data $\mathbf{z}_{16}$ (red stars). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Numerical realization

In the following section, we demonstrate the performance of one-parameter regularization versus two-parameter regularization and also describe the efficiency of proposed regularization algorithm statistically using the relative error measure.

To test the multi-penalty regularization under the balanced-discrepancy parameter choice rule, we consider the well-known academic example [11,22,25],

$$f_\rho(x) = \frac{1}{10}\left\{x + 2\left(e^{-8(\frac{4\pi}{3}-x)^2} - e^{-8(\frac{\pi}{2}-x)^2} - e^{-8(\frac{3\pi}{2}-x)^2}\right)\right\}, \quad x \in [0, 2\pi], \tag{59}$$

which belongs to reproducing kernel Hilbert space $\mathcal{H}_K$ corresponding to the kernel $K(x, y) = xy + \exp(-8(x-y)^2)$.

Consider a training set $\mathbf{z}_m = \{(x_i, y_i)\}_{i=1}^m$, where corresponding to the empirical inputs $\mathbf{x} = \{x_i\}_{i=1}^m = \{\frac{\pi}{10}(i-1)\}_{i=1}^m$ outputs are generated as

$$y_i = f_\rho(x_i) + \delta\xi_i, \quad i = 1, \ldots, m \tag{60}$$

and ξ_i follows the uniform distribution over $[-1, 1]$ with $\delta = 0.02$.

In case of interpolation, single-penalty regularization (41) with the balancing principle parameter choice (42) gives a reasonably good estimator (Fig. 2).

Now consider the situation where one has prior information that except the given data there may appear some inputs $\{x_i\}_{m+1}^n \nsubseteq \overline{Co\{x_i\}_{i=1}^m}$ corresponding to which outputs are unavailable. But prediction has to be made corresponding to these points also, this problem is referred as extrapolation. This situation occurs in many practical problems, such as blood glucose prediction in medical diagnostic [28]. So development of learning algorithms to predict estimator based on labeled and unlabeled data deserves great consideration. Let us try to find an estimator for the samples $\{(x_i, y_i)\}_{i=1}^{15}$ with unlabeled point $\{x_{16}\}$ under the single-parameter regularization (41). It is clear from Fig. 3 that the single-penalty estimator does not extrapolate good enough in case of extrapolation. So the aim is to find the estimator which not only interpolates good but also extrapolates well to the points for which outputs are unknown.

We consider a multi-penalty regularization functional which can be viewed as a special case of problem proposed in [4],

$$arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda_1\|f\|_K^2 + \lambda_2\|S_{\mathbf{x}'}f\|_n^2\right\},$$

where $\mathbf{x}' = \{x_i : 1 \le i \le n\}$.

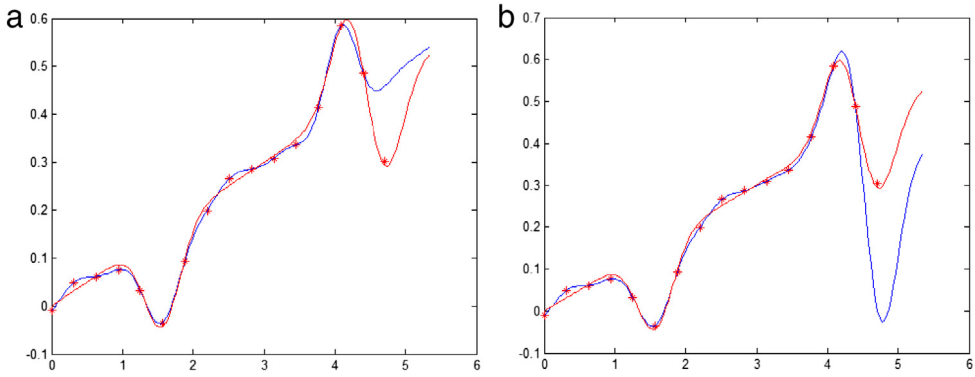

Fig. 3. (a) The estimator $f_{\mathbf{z},\lambda_1}$ (blue line) based on the balancing principle with $\lambda_1 = 5.9855 \times 10^{-4}$, (b) The estimator $f_{\mathbf{z},\lambda_2}$ (blue line) based on discrepancy principle with $\lambda_2 = 3.4994 \times 10^{-4}$. The target function f_ρ (red line) and the empirical data $\mathbf{z}_{15}$ with additional input x_{16} (red stars). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

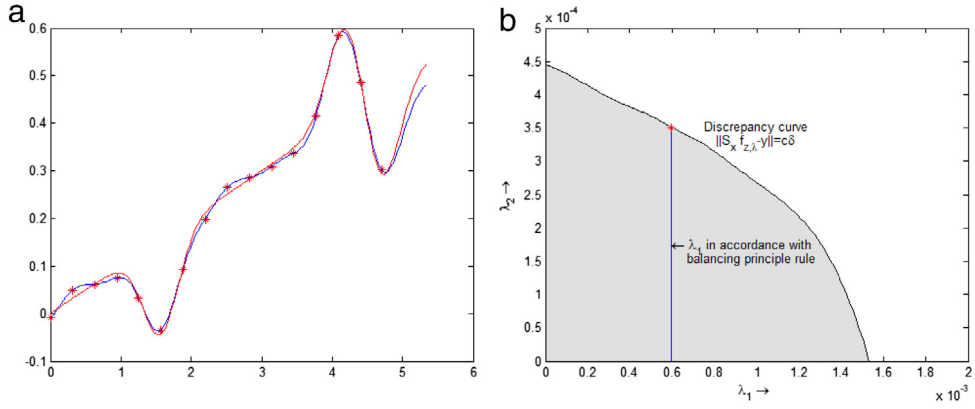

Fig. 4. (a) The target function f_ρ (red line) and its estimator $f^{(1)}_{\mathbf{z},\lambda}$ (blue line) based on the balanced-discrepancy principle with $\lambda_1 = 5.9855 \times 10^{-4}$, $\lambda_2 = 3.4518 \times 10^{-4}$ and the empirical data $\mathbf{z}_{15}$ with additional input x_{16} (red stars), (b) The discrepancy curve is shown with the balanced-discrepancy parameter choice (red star). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

When either of parameter is zero, it reduces to single-penalty regularization. Single-penalty regularizers $f_{\lambda_1}, f_{\lambda_2}$ and double-penalty regularizer $f^{(1)}_{\mathbf{z},\lambda}$ are demonstrated in Figs. 3(a), (b) and 4(a) respectively.

For single-penalty regularization, parameters are chosen according to the balancing principle and discrepancy principle while for two-parameter regularization according to the balanced-discrepancy principle. In learning theory, we do not have deterministic sharp bound for δ but on the other hand we can obtain probabilistic estimate for δ with Proposition 4.2 and using the tuning parameter c we may construct a good estimator. It is important to notice that the estimator f_{λ_1} lies above the actual function f_ρ while f_{λ_2} shows high deep in the extrapolation region. With the appropriate choice of regularization parameters according to the balanced-discrepancy principle, multi-penalty estimator $f^{(1)}_{\mathbf{z},\lambda}$ executes the trend of the function f_ρ in the extrapolation region. Hence it demonstrates the compensatory property of the multi-penalty regularization.

On the other hand we have a natural question, 'can we interchange the role of parameter choice of λ_1 and λ_2?' That is, first choose λ_2 in accordance with balancing principle [11] by assuming $\lambda_1 = 0$ and then find the corresponding parameter value of λ_1 on the discrepancy curve. Hence we construct

Table 1
Statistical performance interpretation of single-penalty regularizer $f_{\mathbf{z},\lambda_1}$.

Error measure	Mean relative error	Median relative error	Standard deviation of relative error	Regularization parameter value
Sup norm	0.3023	0.3059	0.0149	
$\|\cdot\|_m$-empirical norm	0.0916	0.0936	0.0090	$\lambda_1 = 5.9855 \times 10^{-4}$
$\|\cdot\|_K$-norm	0.0415	0.0433	0.0166	

Table 2
Statistical performance interpretation of single-penalty regularizer $f_{\mathbf{z},\lambda_2}$.

Error measure	Mean relative error	Median relative error	Standard deviation of relative error	Regularization parameter value
Sup norm	0.5324	0.5351	0.0106	
$\|\cdot\|_m$-empirical norm	0.2835	0.2865	0.0113	$\lambda_2 = 3.4994 \times 10^{-4}$
$\|\cdot\|_K$-norm	0.3027	0.3086	0.0135	

Table 3
Statistical performance interpretation of multi-penalty regularizer $f^{(1)}_{\mathbf{z},\lambda}$.

Error measure	Mean relative error	Median relative error	Standard deviation of relative error	Regularization parameter value
Sup norm	0.0648	0.0640	0.0089	
				$\lambda_1 = 5.9855 \times 10^{-4}$
$\|\cdot\|_m$-empirical norm	0.0043	0.0042	0.0012	$\lambda_2 = 3.4518 \times 10^{-4}$
$\|\cdot\|_K$-norm	0.0322	0.0325	0.0126	

Table 4
Statistical performance interpretation of multi-penalty regularizer $f^{(2)}_{\mathbf{z},\lambda}$.

Error measure	Mean relative error	Median relative error	Standard deviation of relative error	Regularization parameter value
Sup norm	0.0614	0.0603	0.0085	
				$\lambda_1 = 5.9899 \times 10^{-4}$
$\|\cdot\|_m$-empirical norm	0.0038	0.0037	0.0011	$\lambda_2 = 3.3 \times 10^{-4}$
$\|\cdot\|_K$-norm	0.0344	0.0348	0.0129	

a multi-penalty regularizer $f^{(2)}_{\mathbf{z},\lambda}$ based on this approach for numerical comparison. Further we also compare these approaches to the discrepancy parameter choice rule proposed by Shuai Lu et al. [19] which gives the estimator $f^{(3)}_{\mathbf{z},\lambda}$ through model function approach.

To demonstrate reliability of the balanced-discrepancy rule we generate samples 100 times in accordance with Eq. (60) for $\delta = 0.02$. In our experiment, we compare the performance of single-penalty regularization against the multi-penalty regularization using the relative error measure $\frac{\|f - f_\rho\|}{\|f\|}$ with $f = f_{\mathbf{z},\lambda_1}, f = f_{\mathbf{z},\lambda_2}, f = f^{(1)}_{\mathbf{z},\lambda}, f = f^{(2)}_{\mathbf{z},\lambda}$ and $f = f^{(3)}_{\mathbf{z},\lambda}$ which is listed in Tables 1–5. We illustrate the error estimates for different multi-penalty regularizers in sup norm, $\mathcal{H}_K$-norm and $\|\cdot\|_m$-empirical norm in Fig. 5(a), (b) and (c) respectively.

From the statistical analysis, we observe that the proposed multi-penalty regularizer outperforms the single-penalty regularizers. We also observe that the multi-penalty regularizers corresponding to various parameter choice rules which provides the parameters $\lambda = (\lambda_1, \lambda_2)$ on the discrepancy curve perform similar. Further, we illustrate the comparison of multi-penalty regularization over single-penalty regularization method using the well-known two moon data set (Fig. 6) in the context of manifold learning.

Two moon data consists of two classes (moons): one class (lower moon) M_1 while other class (upper moon) M_{-1} both of size $n/2$. In order to classify two classes with $m = 2k$ labeled samples (k labels from each class) we exploit the geometry of the probability distribution through $(n - m)$

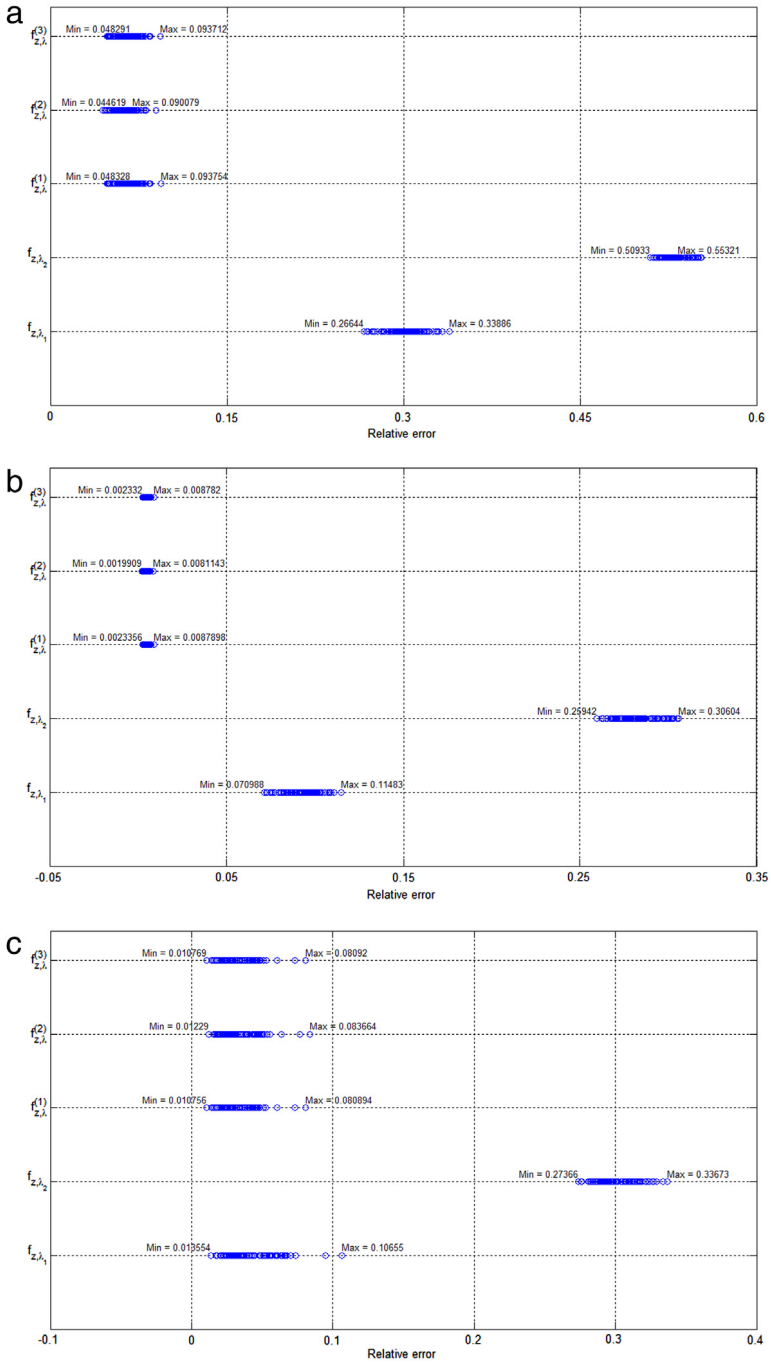

Fig. 5. Figures show relative errors in infinity norm (a), $\|\cdot\|_m$-empirical norm (b) and $\|\cdot\|_K$-norm (c) corresponding to 100 test problems with samples according to (60) with $\delta = 0.02$ for all estimators.

Table 5
Statistical performance interpretation of multi-penalty regularizer $f^{(3)}_{\mathbf{z},\lambda}$.

Error measure	Mean relative error	Median relative error	Standard deviation of relative error	Regularization parameter value
Sup norm	0.0647	0.0640	0.0089	$\lambda_1 = 5.9892 \times 10^{-4}$
$\|\cdot\|_m$-empirical norm	0.0043	0.0042	0.0012	$\lambda_2 = 3.4990 \times 10^{-4}$
$\|\cdot\|_K$-norm	0.0322	0.0326	0.0126	

Table 6
Statistical performance interpretation of single-penalty ($\lambda_I = 0$) and multi-penalty regularizers of the functional (38).

	Single-penalty regularizer		Multi-penalty Regularizer	
	(SP %)	(WC)	(SP %)	(WC)
$m = 2$	76.794	143	100	0
$m = 4$	83.057	126	100	0
$m = 8$	91.967	76	100	0
$m = 16$	96.957	049	100	0

Symbols: labeled points (m); successfully predicted (SP); maximum of wrongly classified points (WC).

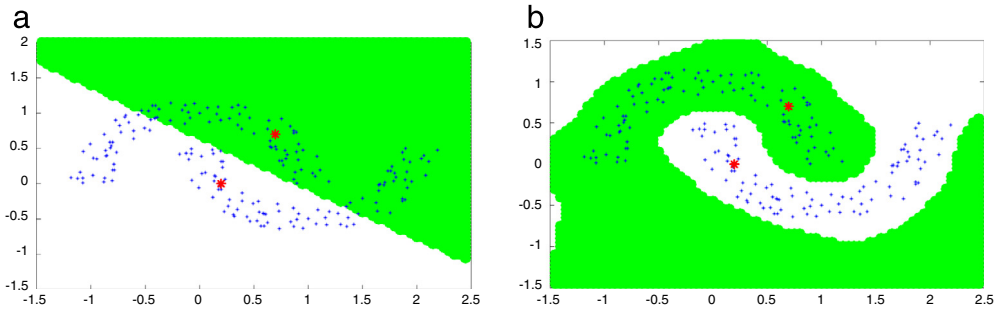

Fig. 6. The figures show the decision surfaces generated with two labeled samples (red star) by single-penalty regularizer (a) based on balancing principle ($\lambda_A = 6 \times 10^{-9}$) and manifold regularizer (b) based on balanced-discrepancy parameter choice $\lambda_A = 6 \times 10^{-9}$, $\lambda_I = 0.0077$.

unlabeled inputs. We optimize the multi-penalty functional (38) discussed in the literature [4,19] over the RKHS $\mathcal{H}_K$ corresponding to mercer kernel $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$ with the exponential weights $\omega_{ij} = \exp(-\|x_i - x_j\|^2/4b)$, for some $b, \gamma > 0$. We employ balanced-discrepancy to choose the regularization parameters λ_A, λ_I in the same manner as discussed above.

We performed experiment 500 times by randomly choosing labeled samples ($m = 2k$, $k = 1, 2, 4, 8$) over 200 points. In all experiments initial parameters are $\lambda_A = 5 \times 10^{-9}$, $\lambda_I = 1$, the kernel parameter $\gamma = 1.95$ and the weight parameter $b = 6.25 \times 10^{-3}$. The performance of single-penalty ($\lambda_I = 0$) and the proposed multi-penalty regularizer (38) is presented in Fig. 6, Table 6.

Overall, we observe that the proposed multi-penalty regularization with the balanced-discrepancy principle parameter choice outperforms single-penalty regularizers.

Acknowledgments

The authors are grateful for the valuable suggestions and comments of the anonymous referees that led to improve the quality of the paper.

References

[1] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (1950) 337–404.
[2] F. Bauer, S.V. Pereverzev, An utilization of a rough approximation of a noise covariance within the framework of multi-parameter regularization, Int. J. Tomogr. Stat. 4 (2006) 1–12.

[3] F. Bauer, S. Pereverzev, L. Rosasco, On regularization algorithms in learning theory, J. Complexity 23 (1) (2007) 52–72.
[4] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: A geometric framework for learning from labeled and unlabeled examples, J. Mach. Learn. Res. 7 (2006) 2399–2434.
[5] O. Bousquet, S. Boucheron, G. Lugosi, Introduction to Statistical Learning Theory, in: Advanced Lectures in Machine Learning, Springer, Berlin/Heidelberg, 2004, pp. 169–207.
[6] A. Caponnetto, E. De Vito, Optimal rates for the regularized least-squares algorithm, Found. Comput. Math. 7 (3) (2007) 331–368.
[7] A. Caponnetto, Y. Yao, Cross-validation based adaptation for regularization operators in learning theory, Anal. Appl. 8 (2) (2010) 161–183.
[8] C. Carmeli, E. De Vito, A. Toigo, Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem, Anal. Appl. 4 (4) (2006) 377–408.
[9] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. (N.S.) 39 (1) (2002) 1–49.
[10] F. Cucker, D.X. Zhou, Learning Theory: An Approximation Theory Viewpoint, in: Cambridge Monographs on Applied and Computational Mathematics, vol. 24, Cambridge University Press, Cambridge, UK, 2007.
[11] E. De Vito, S. Pereverzyev, L. Rosasco, Adaptive kernel methods using the balancing principle, Found. Comput. Math. 10 (4) (2010) 455–479.
[12] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, F. Odone, Learning from examples as an inverse problem, J. Mach. Learn. Res. 6 (2005) 883–904.
[13] D. Düvelmeyer, B. Hofmann, A multi-parameter regularization approach for estimating parameters in jump diffusion processes, J. Inverse Ill-Posed Probl. 14 (9) (2006) 861–880.
[14] H.W. Engl, M. Hanke, A. Neubauer, Regularization of Inverse Problems, in: Mathematics and its Applications, vol. 375, Kluwer Academic, Dordrecht, 1996.
[15] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (1) (2000) 1–50.
[16] M. Fornasier, V. Naumova, S.V. Pereverzyev, Parameter choice strategies for multipenalty regularization, SIAM J. Numer. Anal. 52 (4) (2014) 1770–1794.
[17] L.L. Gerfo, L. Rosasco, F. Odone, E. De Vito, A. Verri, Spectral algorithms for supervised learning, Neural Comput. 20 (7) (2008) 1873–1897.
[18] K. Ito, B. Jin, T. Takeuchi, Multi-parameter Tikhonov regularization—an augmented approach, Chin. Ann. Math. Ser. B 35 (3) (2014) 383–398.
[19] S. Lu, S.V. Pereverzev, Multi-parameter regularization and its numerical realization, Numer. Math. 118 (1) (2011) 1–31.
[20] S. Lu, S. Pereverzev, Regularization Theory for Ill-posed Problems: Selected Topics, Vol. 58, DeGruyter, Berlin, 2013.
[21] S. Lu, S.V. Pereverzev, U. Tautenhahn, A model function method in regularized total least squares, Appl. Anal. 89 (11) (2010) 1693–1703.
[22] S. Lu, S. Pereverzyev Jr., S. Sivananthan, Multiparameter Regularization for Construction of Extrapolating Estimators in Statistical Learning Theory, in: Multiscale Signal Analysis and Modeling, Springer, New York, 2013, pp. 347–366.
[23] Y. Lu, L. Shen, Y. Xu, Multi-parameter regularization methods for high-resolution image reconstruction with displacement errors, IEEE Trans. Circuits Syst. I 54 (8) (2007) 1788–1799.
[24] P. Mathé, S.V. Pereverzev, Geometry of linear ill-posed problems in variable Hilbert scales, Inverse Problems 19 (3) (2003) 789–803.
[25] C.A. Micchelli, M. Pontil, Learning the kernel function via regularization, J. Mach. Learn. Res. 6 (2) (2005) 1099–1125.
[26] V.A. Morozov, On the solution of functional equations by the method of regularization, Sov. Math. Dokl. 7 (1) (1966) 414–417.
[27] V. Naumova, S.V. Pereverzyev, Multi-penalty regularization with a component-wise penalization, Inverse Problems 29 (7) (2013) 075002.
[28] V. Naumova, S.V. Pereverzyev, S. Sivananthan, A meta-learning approach to the regularized learning–Case study: Blood glucose prediction, Neural Netw. 33 (2012) 181–193.
[29] I.F. Pinelis, A.I. Sakhanenko, Remarks on inequalities for the probabilities of large deviations, Theory Probab. Appl. 30 (1) (1985) 127–131.
[30] S. Smale, D.X. Zhou, Estimating the approximation error in learning theory, Anal. Appl. 1 (1) (2003) 17–41.
[31] S. Smale, D.X. Zhou, Shannon sampling and function reconstruction from point values, Bull. Amer. Math. Soc. 41 (3) (2004) 279–306.
[32] S. Smale, D.X. Zhou, Shannon sampling II: Connections to learning theory, Appl. Comput. Harmon. Anal. 19 (3) (2005) 285–302.
[33] S. Smale, D.X. Zhou, Learning theory estimates via integral operators and their approximations, Constr. Approx. 26 (2) (2007) 153–172.
[34] A.N. Tikhonov, V.Y. Arsenin, Solutions of Ill-posed Problems, W.H. Winston, Washington, DC, 1977.
[35] V.N. Vapnik, V. Vapnik, Statistical Learning Theory, Vol. 1, Wiley, New York, 1998.
[36] S.N. Wood, Modelling and smoothing parameter estimation with multiple quadratic penalties, J. R. Stat. Soc. Ser. B 62 (2000) 413–428.
[37] P. Xu, Y. Fukuda, Y. Liu, Multiple parameter regularization: Numerical solutions and applications to the determination of geopotential from precise satellite orbits, J. Geod. 80 (1) (2006) 17–27.