Manifold learning via Multi-Penalty Regularization

Abhishake Rastogi
Department of Mathematics
Indian Institute of Technology Delhi
New Delhi 110016, India
abhishekrastogi2012@gmail.com

Abstract

Manifold regularization is an approach which exploits the geometry of the marginal distribution. The main goal of this paper is to analyze the convergence issues of such regularization algorithms in learning theory. We propose a more general multi-penalty framework and establish the optimal convergence rates under the general smoothness assumption. We study a theoretical analysis of the performance of the multi-penalty regularization over the reproducing kernel Hilbert space. We discuss the error estimates of the regularization schemes under some prior assumptions for the joint probability measure on the sample space. We analyze the convergence rates of learning algorithms measured in the norm in reproducing kernel Hilbert space and in the norm in Hilbert space of square-integrable functions. The convergence issues for the learning algorithms are discussed in probabilistic sense by exponential tail inequalities. In order to optimize the regularization functional, one of the crucial issue is to select regularization parameters to ensure good performance of the solution. We propose a new parameter choice rule "the penalty balancing principle" based on augmented Tikhonov regularization for the choice of regularization parameters. The superiority of multi-penalty regularization over single-penalty regularization is shown using the academic example and moon data set.

Keywords: Learning theory, Multi-penalty regularization, General source condition, Optimal rates, Penalty balancing principle.
Mathematics Subject Classification 2010: 68T05, 68Q32.

1 Introduction

Let X be a compact metric space and $Y \subset \mathbb{R}$ with the joint probability measure ρ on $Z = X \times Y$. Suppose $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ be a observation set drawn from the unknown probability measure ρ. The learning problem [1, 2, 3, 4] aims to approximate a function $f_{\mathbf{z}}$ based on $\mathbf{z}$ such that $f_{\mathbf{z}}(x) \approx y$. The goodness of the estimator can be measured by the generalization error of a function f which can be defined as

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

DOI : 10.5121/ijaia.2017.8506

where $V : Y \times Y \to \mathbb{R}$ is the loss function. The minimizer of $\mathcal{E}(f)$ for the square loss function $V(f(x), y) = (f(x) - y)_Y^2$ is given by

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

where f_ρ is called the regression function. It 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 $\sigma_\rho^2 = \int_X \int_Y (y - f_\rho(x))^2 d\rho(y|x) d\rho_X(x)$. The regression function f_ρ belongs to the space of square integrable functions provided that

$$\int_Z y^2 \ d\rho(x, y) < \infty. \tag{3}$$

Therefore our objective becomes to estimate the regression function f_ρ.

Single-penalty regularization is widely considered to infer the estimator from given set of random samples [5, 6, 7, 8, 9, 10]. Smale et al. [9, 11, 12] provided the foundations of theoretical analysis of square-loss regularization scheme under Hölder's source condition. Caponnetto et al. [6] improved the error estimates to optimal convergence rates for regularized least-square algorithm using the polynomial decay condition of eigenvalues of the integral operator. But sometimes, one may require to add more penalties to incorporate more features in the regularized solution. Multi-penalty regularization is studied by various authors for both inverse problems and learning algorithms [13, 14, 15, 16, 17, 18, 19, 20]. Belkin et al. [13] discussed the problem of manifold regularization which controls the complexity of the function in ambient space as well as geometry of the probability space:

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \lambda_A ||f||_{\mathcal{H}_K}^2 + \lambda_I \sum_{i,j=1}^{n} (f(x_i) - f(x_j))^2 \omega_{ij} \right\}, \tag{4}$$

where $\{(x_i, y_i) \in X \times Y : 1 \leq i \leq m\} \bigcup \{x_i \in X : m < i \leq n\}$ is given set of labeled and unlabeled data, λ_A and λ_I are non-negative regularization parameters, ω_{ij}'s are non-negative weights, $\mathcal{H}_K$ is reproducing kernel Hilbert space and $||\cdot||_{\mathcal{H}_K}$ is its norm.

Further, the manifold regularization algorithm is developed and widely considered in the vector-valued framework to analyze the multi-task learning problem [21, 22, 23, 24] (Also see references therein). So it motivates us to theoretically analyze this problem. The convergence issues of the multi-penalty regularizer are discussed under general source condition in [25] but the convergence rates are not optimal. Here we are able to achieve the optimal minimax convergence rates using the polynomial decay condition of eigenvalues of the integral operator.

In order to optimize regularization functional, one of the crucial problem is the parameter choice strategy. Various prior and posterior parameter choice rules are proposed for single-penalty regularization [26, 27, 28, 29, 30] (also see references therein). Many regularization parameter selection approaches are discussed for multi-penalized ill-posed inverse problems such as discrepancy principle [15, 31], quasi-optimality principle [18, 32], balanced-discrepancy principle [33], heuristic L-curve [34], noise structure based parameter choice rules [35, 36, 37], some approaches which require reduction to single-penalty regularization [38]. Due to growing interest in multi-penalty

regularization in learning, multi-parameter choice rules are discussed in learning theory framework such as discrepancy principle [15, 16], balanced-discrepancy principle [25], parameter choice strategy based on generalized cross validation score [19]. Here we discuss the penalty balancing principle (PB-principle) to choose the regularization parameters in our learning theory framework which is considered for multi-penalty regularization in ill-posed problems [33].

1.1 Mathematical Preliminaries and Notations

Definition 1.1. 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 $K_x(y) = K(x, y)$, 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 [39]. So we denote the reproducing kernel Hilbert space $\mathcal{H}$ by $\mathcal{H}_K$ corresponding to a mercer kernel K and its norm by $||\cdot||_K$.

Definition 1.2. The sampling operator $S_{\mathbf{x}} : \mathcal{H}_K \to \mathbb{R}^m$ associated with a discrete subset $\mathbf{x} = \{x_i\}_{i=1}^m$ is defined by

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

Denote $S_{\mathbf{x}}^* : \mathbb{R}^m \to \mathcal{H}_K$ as the adjoint of $S_{\mathbf{x}}$. Then for $c \in \mathbb{R}^m$,

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

Then its adjoint is given by

$$S_{\mathbf{x}}^* \mathbf{c} = \frac{1}{m}\sum_{i=1}^m c_i K_{x_i}, \qquad \forall \mathbf{c} = (c_1, \cdots, c_m) \in \mathbb{R}^m.$$

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

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

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

For each $(x_i, y_i) \in Z$, $y_i = f_\rho(x_i) + \eta_{x_i}$, where the probability distribution of η_{x_i} has mean 0 and variance $\sigma_{x_i}^2$. Denote $\sigma^2 := \frac{1}{m}\sum_{i=1}^m \sigma_{x_i}^2 < \infty$.

Learning Scheme. The optimization functional (4) can be expressed as

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \left\{ ||S_{\mathbf{x}} f - \mathbf{y}||_m^2 + \lambda_A ||f||_K^2 + \lambda_I ||(S_{\mathbf{x}'}^* L S_{\mathbf{x}'})^{1/2} f||_K^2 \right\}, \tag{5}$$

where $\mathbf{x}' = \{x_i \in X : 1 \leq i \leq n\}$, $||\mathbf{y}||_m^2 = \frac{1}{m}\sum_{i=1}^m y_i^2$, $L = D - W$ with $W = (\omega_{ij})$ is a weight matrix with non-negative entries and D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n} \omega_{ij}$.

Here we consider a more general regularized learning scheme based on two penalties:

$$f_{\mathbf{z},\lambda} := \underset{f\in\mathcal{H}_K}{\operatorname{argmin}} \left\{ ||S_{\mathbf{x}} f - \mathbf{y}||_m^2 + \lambda_1 ||f||_K^2 + \lambda_2 ||Bf||_K^2 \right\}, \tag{6}$$

where $B : \mathcal{H}_K \to \mathcal{H}_K$ is a bounded operator and λ_1, λ_2 are non-negative parameters.

Theorem 1.1. *If $S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^* B$ is invertible, then the optimization functional (6) has unique minimizer:*

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

Proof. we know that for $f \in \mathcal{H}_K$,

$$||S_{\mathbf{x}} f - \mathbf{y}||_m^2 + \lambda_1 ||f||_K^2 + \lambda_2 ||Bf||_K^2 = \langle (S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda_1 I + \lambda_2 B^* B) f, f\rangle_K - 2\langle S_{\mathbf{x}}^* \mathbf{y}, f\rangle_K + ||\mathbf{y}||_m^2.$$

Taking the functional derivative for $f \in \mathcal{H}_K$, we see that any minimizer $f_{\mathbf{z},\lambda}$ of (6) satisfies

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

This proves Theorem 1.1. □

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

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

which gives

$$f_{\mathbf{x},\lambda} = \Delta_S S_{\mathbf{x}}^* S_{\mathbf{x}} f_\rho. \tag{8}$$

The data-free version of the considered regularization scheme (6) is

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

where the norm $||\cdot||_\rho := ||\cdot||_{\mathscr{L}^2_{\rho_X}}$. Then we get the expression of f_λ,

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

and

$$f_{\lambda_1} := \underset{f\in\mathcal{H}_K}{\operatorname{argmin}} \left\{ ||f - f_\rho||_\rho^2 + \lambda_1 ||f||_K^2 \right\}. \tag{11}$$

which implies

$$f_{\lambda_1} = (L_K + \lambda_1 I)^{-1} L_K f_\rho, \tag{12}$$

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 integral operator L_K can also be defined as a self-adjoint operator on $\mathcal{H}_K$. We use the same notation L_K for both the operators.

Using the singular value decomposition $L_K = \sum_{i=1}^{\infty} t_i \langle \cdot, e_i \rangle_K e_i$ for orthonormal system $\{e_i\}$ in $\mathcal{H}_K$ and sequence of singular numbers $\kappa^2 \geq t_1 \geq t_2 \geq \ldots \geq 0$, we define

$$\phi(L_K) = \sum_{i=1}^{\infty} \phi(t_i) \langle \cdot, e_i \rangle_K e_i,$$

where ϕ is a continuous increasing index function defined on the interval $[0, \kappa^2]$ with the assumption $\phi(0) = 0$.

We require some prior assumptions on the probability measure ρ to achieve the uniform convergence rates for learning algorithms.

Assumption 1. (Source condition) *Suppose*

$$\Omega_{\phi,R} := \{f \in \mathcal{H}_K : f = \phi(L_K)g \text{ and } ||g||_K \leq R\},$$

Then the condition $f_\rho \in \Omega_{\phi,R}$ is usually referred as general source condition [40].

Assumption 2. (Polynomial decay condition) *We assume the eigenvalues t_n's of the integral operator L_K follows the polynomial decay: For fixed positive constants α, β and $b > 1$,*

$$\alpha n^{-b} \leq t_n \leq \beta n^{-b} \quad \forall n \in \mathbb{N}.$$

Following the notion of Bauer et al. [5] and Caponnetto et al. [6], we consider the class of probability measures $\mathcal{P}_\phi$ which satisfies the source condition and the probability measure class $\mathcal{P}_{\phi,b}$ satisfying the source condition and polynomial decay condition.

The effective dimension $\mathcal{N}(\lambda_1)$ can be estimated from Proposition 3 [6] under the polynomial decay condition as follows,

$$\mathcal{N}(\lambda_1) := Tr\left((L_K + \lambda_1 I)^{-1} L_K\right) \leq \frac{\beta b}{b-1} \lambda_1^{-1/b}, \text{ for } b > 1. \tag{13}$$

where $Tr(A) := \sum_{k=1}^{\infty} \langle Ae_k, e_k \rangle$ for some orthonormal basis $\{e_k\}_{k=1}^{\infty}$.

Shuai Lu et al. [41] and Blanchard et al. [42] considered the logarithm decay condition of the effective dimension $\mathcal{N}(\lambda_1)$,

Assumption 3. (logarithmic decay) *Assume that there exists some positive constant $c > 0$ such that*

$$\mathcal{N}(\lambda_1) \leq c \log\left(\frac{1}{\lambda_1}\right), \forall \lambda_1 > 0. \tag{14}$$

2 Convergence Analysis

In this section, we discuss the convergence issues of multi-penalty regularization scheme on reproducing kernel Hilbert space under the considered smoothness priors in learning theory framework. We address the convergence rates of the multi-penalty regularizer by estimating the sample error $f_{\mathbf{z},\lambda} - f_\lambda$ and approximation error $f_\lambda - f_\rho$ in interpolation norm. In Theorem 2.1, the upper convergence rates of multi-penalty regularized solution $f_{\mathbf{z},\lambda}$ are derived from the estimates of

Proposition 2.2, 2.3 for the class of probability measure $P_{\phi,b}$, respectively. We discuss the error estimates under the general source condition and the parameter choice rule based on the index function ϕ and sample size m. Under the polynomial decay condition, in Theorem 2.2 we obtain the optimal minimax convergence rates in terms of index function ϕ, the parameter b and the number of samples m. In particular under Hölder's source condition, we present the convergence rates under the logarithm decay condition on effective dimension in Corollary 2.2.

Proposition 2.1. *Let* $\mathbf{z}$ *be i.i.d. samples drawn according to the probability measure* ρ *with the hypothesis* $|y_i| \leq M$ *for each* $(x_i, y_i) \in Z$. *Then for* $0 \leq s \leq \frac{1}{2}$ *and for every* $0 < \delta < 1$ *with prob.* $1-\delta$,

$$||L_K^s(f_{\mathbf{z},\lambda} - f_{\mathbf{x},\lambda})||_K \leq 2\lambda_1^{s-\frac{1}{2}} \left\{ \Xi \left(1 + 2\sqrt{\log\left(\frac{2}{\delta}\right)} \right) + \frac{4\kappa M}{3m\sqrt{\lambda_1}} \log\left(\frac{2}{\delta}\right) \right\},$$

where $\mathcal{N}_{x_i}(\lambda_1) = Tr\left((L_K + \lambda_1 I)^{-1} K_{x_i} K_{x_i}^*\right)$ *and* $\Xi = \frac{1}{m}\sqrt{\sum_{i=1}^m \sigma_{x_i}^2 \mathcal{N}_{x_i}(\lambda_1)}$ *for the variance* $\sigma_{x_i}^2$ *of the probability distribution of* $\eta_{x_i} = y_i - f_\rho(x_i)$.

Proof. The expression $f_{\mathbf{z},\lambda} - f_{\mathbf{x},\lambda}$ can be written as $\Delta_S S_{\mathbf{x}}^*(\mathbf{y} - S_{\mathbf{x}} f_\rho)$. Then we find that

$$\begin{aligned} ||L_K^s(f_{\mathbf{z},\lambda} - f_{\mathbf{x},\lambda})||_K &\leq I_1 ||L_K^s (L_K + \lambda_1 I)^{-1/2}||\ ||(L_K + \lambda_1 I)^{1/2} \Delta_S (L_K + \lambda_1 I)^{1/2}|| \\ &\leq I_1 I_2 ||L_K^s (L_K + \lambda_1 I)^{-1/2}||, \end{aligned} \tag{15}$$

where $I_1 = ||(L_K + \lambda_1 I)^{-1/2} S_{\mathbf{x}}^*(\mathbf{y} - S_{\mathbf{x}} f_\rho)||_K$ and $I_2 = ||(L_K + \lambda_1 I)^{1/2}(S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda_1 I)^{-1}(L_K + \lambda_1 I)^{1/2}||$.

For sufficiently large sample size m, the following inequality holds:

$$\frac{8\kappa^2}{\sqrt{m}} \log\left(\frac{2}{\delta}\right) \leq \lambda_1 \tag{16}$$

Then from Theorem 2 [43] we have with confidence $1-\delta$,

$$\begin{aligned} I_3 = ||(L_K + \lambda_1 I)^{-1/2}(L_K - S_{\mathbf{x}}^* S_{\mathbf{x}})(L_K + \lambda_1 I)^{-1/2}|| &\leq \frac{||S_{\mathbf{x}}^* S_{\mathbf{x}} - L_K||}{\lambda_1} \\ &\leq \frac{4\kappa^2}{\sqrt{m}\lambda_1} \log\left(\frac{2}{\delta}\right) \leq \frac{1}{2}. \end{aligned}$$

Then the Neumann series gives

$$\begin{aligned} I_2 &= ||\{I - (L_K + \lambda_1 I)^{-1/2}(L_K - S_{\mathbf{x}}^* S_{\mathbf{x}})(L_K + \lambda_1 I)^{-1/2}\}^{-1}|| \\ &= ||\sum_{i=0}^{\infty} \{(L_K + \lambda_1 I)^{-1/2}(L_K - S_{\mathbf{x}}^* S_{\mathbf{x}})(L_K + \lambda_1 I)^{-1/2}\}^i|| \leq \sum_{i=0}^{\infty} I_3^i = \frac{1}{1 - I_3} \leq 2. \end{aligned} \tag{17}$$

Now we have,

$$||L_K^s (L_K + \lambda_1 I)^{-1/2}|| \leq \sup_{0 < t \leq \kappa^2} \frac{t^s}{(t + \lambda_1)^{1/2}} \leq \lambda_1^{s-1/2} \text{ for } 0 \leq s \leq \frac{1}{2}. \tag{18}$$

To estimate the error bound for $||(L_K + \lambda_1 I)^{-1/2} S_{\mathbf{x}}^*(\mathbf{y} - S_{\mathbf{x}} f_\rho)||_K$ using the McDiarmid inequality (Lemma 2 [12]), define the function $\mathcal{F} : \mathbb{R}^m \to \mathbb{R}$ as

$$\mathcal{F}(\mathbf{y}) = ||(L_K + \lambda_1 I)^{-1/2} S_{\mathbf{x}}^*(\mathbf{y} - S_{\mathbf{x}} f_\rho)||_K$$

$$= \frac{1}{m}\left|\left|(L_K+\lambda_1 I)^{-1/2}\sum_{i=1}^{m}(y_i-f_\rho(x_i))K_{x_i}\right|\right|_K.$$

So $\mathcal{F}^2(\mathbf{y}) = \frac{1}{m^2}\sum_{i,j=1}^{m}(y_i-f_\rho(x_i))(y_j-f_\rho(x_j))\langle(L_K+\lambda_1 I)^{-1}K_{x_i},K_{x_j}\rangle_K$.

The independence of the samples together with $E_{\mathbf{y}}(y_i-f_\rho(x_i))=0,\ E_{\mathbf{y}}(y_i-f_\rho(x_i))^2=\sigma_{x_i}^2$ implies

$$E_{\mathbf{y}}(\mathcal{F}^2)=\frac{1}{m^2}\sum_{i=1}^{m}\sigma_{x_i}^2\mathcal{N}_{x_i}(\lambda_1)\le\Xi^2,$$

where $\mathcal{N}_{x_i}(\lambda_1) = Tr\left((L_K+\lambda_1 I)^{-1}K_{x_i}K_{x_i}^*\right)$ and $\Xi = \frac{1}{m}\sqrt{\sum_{i=1}^{m}\sigma_{x_i}^2\mathcal{N}_{x_i}(\lambda_1)}$. Since $E_{\mathbf{y}}(\mathcal{F}) \le \sqrt{E_{\mathbf{y}}(\mathcal{F}^2)}$. It implies $E_{\mathbf{y}}(\mathcal{F})\le\Xi$.

Let $\mathbf{y}^i=(y_1,\ldots,y_{i-1},y_i',y_{i+1},\ldots,y_m)$, where y_i' is another sample at x_i. We have

$$\begin{aligned}|\mathcal{F}(\mathbf{y})-\mathcal{F}(\mathbf{y}^i)| &\le ||(L_K+\lambda_1 I)^{-1/2}S_{\mathbf{x}}^*(\mathbf{y}-\mathbf{y}^i)||_K\\ &= \frac{1}{m}||(y_i-y_i')(L_K+\lambda_1 I)^{-1/2}K_{x_i}||_K \le \frac{2\kappa M}{m\sqrt{\lambda_1}}.\end{aligned}$$

This can be taken as B in Lemma 2(2) [12]. Now

$$\begin{aligned}&E_{y_i}\left(|\mathcal{F}(\mathbf{y})-E_{y_i}(\mathcal{F}(\mathbf{y}))|^2\right)\\ \le\ &\frac{1}{m^2}\int_Y\left(\int_Y|y_i-y_i'|\ ||(L_K+\lambda_1 I)^{-1/2}K_{x_i}||_K d\rho(y_i'|x_i)\right)^2 d\rho(y_i|x_i)\\ \le\ &\frac{1}{m^2}\int_Y\int_Y(y_i-y_i')^2\mathcal{N}_{x_i}(\lambda_1)d\rho(y_i'|x_i)d\rho(y_i|x_i)\\ \le\ &\frac{2}{m^2}\sigma_{x_i}^2\mathcal{N}_{x_i}(\lambda_1)\end{aligned}$$

which implies

$$\sum_{i=1}^{m}\sigma_i^2(\mathcal{F})\le 2\Xi^2.$$

In view of Lemma 2(2) [12] for every $\varepsilon>0$,

$$\underset{\mathbf{y}\in Y^m}{Prob}\left\{\mathcal{F}(\mathbf{y})-E_{\mathbf{y}}(\mathcal{F}(\mathbf{y}))\ge\varepsilon\right\}\le\exp\left\{-\frac{\varepsilon^2}{4(\Xi^2+\varepsilon\kappa M/3m\sqrt{\lambda_1})}\right\}=\delta.\ \text{(let)}$$

In terms of δ, probability inequality becomes

$$\underset{\mathbf{y}\in Y^m}{Prob}\left\{\mathcal{F}(\mathbf{y})\le\Xi\left(1+2\sqrt{\log\left(\frac{1}{\delta}\right)}\right)+\frac{4\kappa M}{3m\sqrt{\lambda_1}}\log\left(\frac{1}{\delta}\right)\right\}\le 1-\delta.$$

Incorporating this inequality with (17), (18) in (15), we get the desired result. □

Proposition 2.2. *Let* $\mathbf{z}$ *be i.i.d. samples drawn according to the probability measure* ρ *with the hypothesis* $|y_i|\le M$ *for each* $(x_i,y_i)\in Z$. *Suppose* $f_\rho\in\Omega_{\phi,R}$. *Then for* $0\le s\le\frac{1}{2}$ *and for every* $0<\delta<1$ *with prob.* $1-\delta$,

$$\begin{aligned}||L_K^s(f_{\mathbf{z},\lambda}-f_\lambda)||_K &\le \frac{2\lambda_1^{s-\frac{1}{2}}}{\sqrt{m}}\left\{3M\sqrt{\mathcal{N}(\lambda_1)}+\frac{4\kappa}{\sqrt{\lambda_1}}||f_\lambda-f_\rho||_\rho+\frac{\sqrt{\lambda_1}}{6}||f_\lambda-f_\rho||_K\right.\\ &\left.+\frac{7\kappa M}{\sqrt{m\lambda_1}}\right\}\log\left(\frac{4}{\delta}\right).\end{aligned}$$

Proof. We can express $f_{\mathbf{x},\lambda} - f_\lambda = \Delta_S(S_{\mathbf{x}}^* S_{\mathbf{x}} - L_K)(f_\rho - f_\lambda)$, which implies

$$||L_K^s(f_{\mathbf{x},\lambda} - f_\lambda)||_K \leq I_4 \left|\left| \frac{1}{m}\sum_{i=1}^{m}(f_\rho(x_i) - f_\lambda(x_i))K_{x_i} - L_K(f_\rho - f_\lambda) \right|\right|_K .$$

where $I_4 = ||L_K^s \Delta_S||$. Using Lemma 3 [12] for the function $f_\rho - f_\lambda$, we get with confidence $1-\delta$,

$$||L_K^s(f_{\mathbf{x},\lambda} - f_\lambda)||_K \leq I_4 \left(\frac{4\kappa||f_\lambda - f_\rho||_\infty}{3m} log\left(\frac{1}{\delta}\right) + \frac{\kappa||f_\lambda - f_\rho||_\rho}{\sqrt{m}} \left(1 + \sqrt{8log\left(\frac{1}{\delta}\right)}\right)\right). \quad (19)$$

For sufficiently large sample (16), from Theorem 2 [43] we get

$$||(L_K - S_{\mathbf{x}}^* S_{\mathbf{x}})(L_K + \lambda_1 I)^{-1}|| \leq \frac{||S_{\mathbf{x}}^* S_{\mathbf{x}} - L_K||}{\lambda_1} \leq \frac{4\kappa^2}{\sqrt{m}\lambda_1} \log\left(\frac{2}{\delta}\right) \leq \frac{1}{2}$$

with confidence $1-\delta$, which implies

$$||(L_K + \lambda_1 I)(S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda_1 I)^{-1}|| = ||\{I - (L_K - S_{\mathbf{x}}^* S_{\mathbf{x}})(L_K + \lambda_1 I)^{-1}\}^{-1}|| \leq 2. \quad (20)$$

$$\text{We have, } ||L_K^s(L_K + \lambda_1 I)^{-1}|| \leq \sup_{0<t\leq\kappa^2} \frac{t^s}{(t+\lambda_1)} \leq \lambda_1^{s-1} \text{ for } 0 \leq s \leq 1. \quad (21)$$

Now equation (20) and (21) implies the following inequality,

$$I_4 \leq ||L_K^s(S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda_1 I)^{-1}|| \leq ||L_K^s(L_K + \lambda_1 I)^{-1}||\ ||(L_K + \lambda_1 I)(S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda_1 I)^{-1}|| \leq 2\lambda_1^{s-1}. \quad (22)$$

Let $\xi(x) = \sigma_x^2 \mathcal{N}_x(\lambda_1)$ be the random variable. Then it satisfies $|\xi| \leq 4\kappa^2 M^2/\lambda_1$, $E_x(\xi) \leq M^2\mathcal{N}(\lambda_1)$ and $\sigma^2(\xi) \leq 4\kappa^2 M^4 \mathcal{N}(\lambda_1)/\lambda_1$. Using the Bernstein inequality we get

$$\underset{\mathbf{x}\in X^m}{Prob}\left\{ \sum_{i=1}^{m} \left(\sigma_{x_i}^2 \mathcal{N}_{x_i}(\lambda_1) - M^2\mathcal{N}(\lambda_1)\right) > t \right\} \leq exp\left(- \frac{t^2/2}{\frac{4m\kappa^2 M^4 \mathcal{N}(\lambda_1)}{\lambda_1} + \frac{4\kappa^2 M^2 t}{3\lambda_1}} \right)$$

which implies

$$\underset{\mathbf{x}\in X^m}{Prob}\left\{ \Xi \leq \sqrt{\frac{M^2\mathcal{N}(\lambda_1)}{m}} + \sqrt{\frac{8\kappa^2 M^2}{3m^2\lambda_1} \log\left(\frac{1}{\delta}\right)} \right\} \geq 1 - \delta. \quad (23)$$

We get the required error estimate by combining the estimates of Proposition 2.1 with inequalities (19), (22), (23). □

Proposition 2.3. *Suppose $f_\rho \in \Omega_{\phi,R}$. Then under the assumption that $\phi(t)$ and $t^{1-s}/\phi(t)$ are nondecreasing functions, we have*

$$||L_K^s(f_\lambda - f_\rho)||_K \leq \lambda_1^s \left(R\phi(\lambda_1) + \lambda_2 \lambda_1^{-3/2} M||B^* B|| \right). \quad (24)$$

Proof. To realize the above error estimates, we decomposes $f_\lambda - f_\rho$ into $f_\lambda - f_{\lambda_1} + f_{\lambda_1} - f_\rho$. The first term can be expressed as

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

Then we get

$$\begin{aligned} ||L_K^s(f_\lambda - f_{\lambda_1})||_K &\leq \lambda_2 ||L_K^s(L_K+\lambda_1 I)^{-1}||\ ||B^*B||\ ||f_{\lambda_1}||_K \\ &\leq \lambda_2 \lambda_1^{s-1} ||B^*B||\ ||f_{\lambda_1}||_K \leq \lambda_2 \lambda_1^{s-3/2} M ||B^*B||. \end{aligned} \tag{25}$$

$$||L_K^s(f_{\lambda_1} - f_\rho)|| \leq R||r_{\lambda_1}(L_K) L_K^s \phi(L_K)|| \leq R\lambda_1^s \phi(\lambda_1), \tag{26}$$

where $r_{\lambda_1}(t) = 1 - (t+\lambda_1)^{-1}t$.

Combining these error bounds, we achieve the required estimate. □

Theorem 2.1. *Let* $\mathbf{z}$ *be i.i.d. samples drawn according to probability measure* $\mathcal{P}_{\phi,b}$. *Suppose* $\phi(t)$ *and* $t^{1-s}/\phi(t)$ *are nondecreasing functions. Then under parameter choice* $\lambda_1 \in (0,1]$, $\lambda_1 = \Psi^{-1}(m^{-1/2})$, $\lambda_2 = (\Psi^{-1}(m^{-1/2}))^{3/2}\phi(\Psi^{-1}(m^{-1/2}))$ *where* $\Psi(t) = t^{\frac{1}{2}+\frac{1}{2b}}\phi(t)$, *for* $0 \leq s \leq \frac{1}{2}$ *and for all* $0 < \delta < 1$, *the following error estimates holds with confidence* $1-\delta$,

$$\underset{\mathbf{z}\in Z^m}{Prob}\left\{ ||L_K^s(f_{\mathbf{z},\lambda} - f_\rho)||_K \leq C(\Psi^{-1}(m^{-1/2}))^s \phi(\Psi^{-1}(m^{-1/2})) \log\left(\frac{4}{\delta}\right) \right\} \geq 1-\delta,$$

where $C = 14\kappa M + (2+8\kappa)(R + M||B^*B||) + 6M\sqrt{\beta b/(b-1)}$ *and*

$$\lim_{\tau\to\infty} \limsup_{m\to\infty} \sup_{\rho\in\mathcal{P}_{\phi,b}} \underset{\mathbf{z}\in Z^m}{Prob}\Big\{ ||L_K^s(f_{\mathbf{z},\lambda} - f_\rho)||_K > \tau(\Psi^{-1}(m^{-1/2}))^s \phi(\Psi^{-1}(m^{-1/2})) \Big\} = 0.$$

Proof. Let $\Psi(t) = t^{\frac{1}{2}+\frac{1}{2b}}\phi(t)$. Then $\Psi(t) = y$ follows,

$$\lim_{t\to 0} \frac{\Psi(t)}{\sqrt{t}} = \lim_{y\to 0} \frac{y}{\sqrt{\Psi^{-1}(y)}} = 0.$$

Under the parameter choice $\lambda_1 = \Psi^{-1}(m^{-1/2})$ we have $\lim_{m\to\infty} m\lambda_1 = \infty$. Therefore for sufficiently large m,

$$\frac{1}{m\lambda_1} = \frac{\lambda_1^{\frac{1}{2b}}\phi(\lambda_1)}{\sqrt{m\lambda_1}} \leq \lambda_1^{\frac{1}{2b}}\phi(\lambda_1).$$

Under the fact $\lambda_1 \leq 1$ from Proposition 2.2, 2.3 and eqn. (13) follows that with confidence $1-\delta$,

$$||L_K^s(f_{\mathbf{z},\lambda} - f_\rho)||_K \leq C(\Psi^{-1}(m^{-1/2}))^s \phi(\Psi^{-1}(m^{-1/2})) \log\left(\frac{4}{\delta}\right), \tag{27}$$

where $C = 14\kappa M + (2+8\kappa)(R + M||B^*B||) + 6M\sqrt{\beta b/(b-1)}$.

Now defining $\tau := C\log\left(\frac{4}{\delta}\right)$ gives $\delta = \delta_\tau = 4e^{-\tau/C}$. The estimate (27) can be reexpressed as

$$\underset{\mathbf{z}\in Z^m}{Prob}\{ ||L_K^s(f_{\mathbf{z},\lambda} - f_\rho)||_K > \tau(\Psi^{-1}(m^{-1/2}))^s \phi(\Psi^{-1}(m^{-1/2})) \} \leq \delta_\tau. \tag{28}$$

□

Theorem 2.2. *Let* $\mathbf{z}$ *be i.i.d. samples drawn according to the probability measure* $\rho \in \mathcal{P}_{\phi,b}$. *Then under the parameter choice* $\lambda_0 \in (0,1]$, $\lambda_0 = \Psi^{-1}(m^{-1/2})$, $\lambda_j = (\Psi^{-1}(m^{-1/2}))^{3/2}\phi(\Psi^{-1}(m^{-1/2}))$ *for* $1 \leq j \leq p$, *where* $\Psi(t) = t^{\frac{1}{2}+\frac{1}{2b}}\phi(t)$, *for all* $0 < \eta < 1$, *the following error estimates hold with confidence* $1-\eta$,

(i) If $\phi(t)$ and $t/\phi(t)$ are nondecreasing functions. Then we have,

$$\underset{\mathbf{z}\in Z^m}{Prob}\left\{||f_{\mathbf{z},\lambda}-f_{\mathcal{H}}||_{\mathcal{H}}\leq \bar{C}\phi(\Psi^{-1}(m^{-1/2}))\log\left(\frac{4}{\eta}\right)\right\}\geq 1-\eta$$

and

$$\lim_{\tau\to\infty}\limsup_{m\to\infty}\sup_{\rho\in\mathcal{P}_{\phi,b}}\underset{\mathbf{z}\in Z^m}{Prob}\left\{||f_{\mathbf{z},\lambda}-f_{\mathcal{H}}||_{\mathcal{H}}>\tau\phi(\Psi^{-1}(m^{-1/2}))\right\}=0,$$

where $\bar{C}=2R+4\kappa M+2M||B^*B||_{\mathcal{L}(\mathcal{H})}+4\kappa\Sigma$.

(ii) If $\phi(t)$ and $\sqrt{t}/\phi(t)$ are nondecreasing functions. Then we have,

$$\underset{\mathbf{z}\in Z^m}{Prob}\left\{||f_{\mathbf{z},\lambda}-f_{\mathcal{H}}||_{\rho}\leq \bar{C}(\Psi^{-1}(m^{-1/2}))^{1/2}\phi(\Psi^{-1}(m^{-1/2}))\log\left(\frac{4}{\eta}\right)\right\}\geq 1-\eta$$

and

$$\lim_{\tau\to\infty}\limsup_{m\to\infty}\sup_{\rho\in\mathcal{P}_{\phi,b}}\underset{\mathbf{z}\in Z^m}{Prob}\left\{||f_{\mathbf{z},\lambda}-f_{\mathcal{H}}||_{\rho}>\tau(\Psi^{-1}(m^{-1/2}))^{1/2}\phi(\Psi^{-1}(m^{-1/2}))\right\}=0.$$

Corollary 2.1. *Under the same assumptions of Theorem 2.1 for Hölder's source condition $f_\rho\in\Omega_{\phi,R}$, $\phi(t)=t^r$, for $0\leq s\leq\frac{1}{2}$ and for all $0<\delta<1$, with confidence $1-\delta$, for the parameter choice $\lambda_1=m^{-\frac{b}{2br+b+1}}$ and $\lambda_2=m^{-\frac{2br+3b}{4br+2b+2}}$ we have the following convergence rates:*

$$||L_K^s(f_{\mathbf{z},\lambda}-f_\rho)||_K\leq Cm^{-\frac{b(r+s)}{2br+b+1}}\log\left(\frac{4}{\delta}\right)\ \text{for } 0\leq r\leq 1-s.$$

Remark 2.1. For Hölder source condition $f_{\mathcal{H}}\in\Omega_{\phi,R}$, $\phi(t)=t^r$ with the parameter choice $\lambda_0\in(0,1]$, $\lambda_0=m^{-\frac{b}{2br+b+1}}$ and for $1\leq j\leq p$, $\lambda_j=m^{-\frac{2br+3b}{4br+2b+2}}$, we obtain the optimal minimax convergence rates $\mathcal{O}(m^{-\frac{br}{2br+b+1}})$ for $0\leq r\leq 1$ and $\mathcal{O}(m^{-\frac{2br+b}{4br+2b+2}})$ for $0\leq r\leq\frac{1}{2}$ in RKHS-norm and $\mathscr{L}^2_{\rho_X}$-norm, respectively.

Corollary 2.2. *Under the logarithm decay condition of effective dimension $\mathcal{N}(\lambda_1)$, for Hölder's source condition $f_\rho\in\Omega_{\phi,R}$, $\phi(t)=t^r$, for $0\leq s\leq\frac{1}{2}$ and for all $0<\delta<1$, with confidence $1-\delta$, for the parameter choice $\lambda_1=\left(\frac{\log m}{m}\right)^{\frac{1}{2r+1}}$ and $\lambda_2=\left(\frac{\log m}{m}\right)^{\frac{2r+3}{4r+2}}$ we have the following convergence rates:*

$$||L_K^s(f_{\mathbf{z},\lambda}-f_\rho)||_K\leq C\left(\frac{\log m}{m}\right)^{\frac{s+r}{2r+1}}\log\left(\frac{4}{\delta}\right)\ \text{for } 0\leq r\leq 1-s.$$

Remark 2.2. The upper convergence rates of the regularized solution is estimated in the interpolation norm for the parameter $s\in[0,\frac{1}{2}]$. In particular, we obtain the error estimates in $||\cdot||_{\mathcal{H}_K}$-norm for $s=0$ and in $||\cdot||_{\mathscr{L}^2_{\rho_X}}$-norm for $s=\frac{1}{2}$. We present the error estimates of multi-penalty regularizer over the regularity class $\mathcal{P}_{\phi,b}$ in Theorem 2.1 and Corollary 2.1. We can also obtain the convergence rates of the estimator $f_{\mathbf{z},\lambda}$ under the source condition without the polynomial decay of the eigenvalues of the integral operator L_K by substituting $\mathcal{N}(\lambda_1)\leq\frac{\kappa^2}{\lambda_1}$. In addition, for $B=(S^*_{\mathbf{x}'}LS_{\mathbf{x}'})^{1/2}$ we obtain the error estimates of the manifold regularization scheme (30) considered in [13].

Remark 2.3. The parameter choice is said to be optimal, if the minimax lower rates coincide with the upper convergence rates for some $\lambda=\lambda(m)$. For the parameter choice $\lambda_1=\Psi^{-1}(m^{-1/2})$ and

$\lambda_2 = (\Psi^{-1}(m^{-1/2}))^{3/2}\phi(\Psi^{-1}(m^{-1/2}))$, Theorem 2.2 share the upper convergence rates with the lower convergence rates of Theorem 3.11, 3.12 [44]. Therefore the choice of parameters is optimal.

Remark 2.4. The results can be easily generalized to n-penalty regularization in vector-valued framework. For simplicity, we discuss two-parameter regularization scheme in scalar-valued function setting.

Remark 2.5. We can also address the convergence issues of binary classification problem [45] using our error estimates as similar to discussed in Section 3.3 [5] and Section 5 [9].

The proposed choice of parameters in Theorem 2.1 is based on the regularity parameters which are generally not known in practice. In the proceeding section, we discuss the parameter choice rules based on samples.

3 Parameter Choice Rules

Most regularized learning algorithms depend on the tuning parameter, whose appropriate choice is crucial to ensure good performance of the regularized solution. Many parameter choice strategies are discussed for single-penalty regularization schemes for both ill-posed problems and the learning algorithms [27, 28] (also see references therein). Various parameter choice rules are studied for multi-penalty regularization schemes [15, 18, 19, 25, 31, 32, 33, 36, 46]. Ito el al. [33] studied a balancing principle for choosing regularization parameters based on the augmented Tikhonov regularization approach for ill posed inverse problems. In learning theory framework, we are discussing the fixed point algorithm based on the penalty balancing principle considered in [33].

The Bayesian inference approach provides a mechanism for selecting the regularization parameters through hierarchical modeling. Various authors successfully applied this approach in different problems. Thompson et al. [47] applied this for selecting parameters for image restoration. Jin et al. [48] considered the approach for ill-posed Cauchy problem of steady-state heat conduction.

The posterior probability density function (PPDF) for the functional (5) is given by

$$\begin{aligned} P(f,\sigma^2,\mu,\mathbf{z}) \;\propto\; & \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}||S_{\mathbf{x}}f-\mathbf{y}||_m^2\right)\mu_1^{n_1/2}\exp\left(-\frac{\mu_1}{2}||f||_K^2\right)\mu_2^{n_2/2} \\ & \cdot\; \exp\left(-\frac{\mu_2}{2}||Bf||_K^2\right)\mu_1^{\alpha'-1}e^{-\beta'\mu_1}\mu_2^{\alpha'-1}e^{-\beta'\mu_2}\left(\frac{1}{\sigma^2}\right)^{\alpha'_o-1} e^{-\beta'_o(\frac{1}{\sigma^2})}. \end{aligned}$$

where (α',β') are parameter pairs for $\mu=(\mu_1,\mu_2)$, (α'_o,β'_o) are parameter pair for inverse variance $\frac{1}{\sigma^2}$. In the Bayesian inference approach, we select parameter set (f,σ^2,μ) which maximizes the PPDF. By taking the negative logarithm and simplifying, the problem can be reformulated as

$$\begin{aligned} \mathcal{J}(f,\tau,\mu) \;=\; & \tau||S_{\mathbf{x}}f-\mathbf{y}||_m^2+\mu_1||f||_K^2+\mu_2||Bf||_K^2 \\ & +\beta(\mu_1+\mu_2)-\alpha(log\mu_1+log\mu_2)+\beta_o\tau-\alpha_o log\tau, \end{aligned}$$

where $\tau = 1/\sigma^2$, $\beta = 2\beta'$, $\alpha = n_1+2\alpha'-2$, $\beta_o = 2\beta'_o$, $\alpha_o = n_2+2\alpha'_o-2$. We assume that the scalars τ and μ_i's have Gamma distributions with known parameter pairs. The functional is pronounced as augmented Tikhonov regularization.

For non-informative prior $\beta_o = \beta = 0$, the optimality of a-Tikhonov functional can be reduced to

$$\begin{cases} f_{\mathbf{z},\lambda} = arg \min\limits_{f\in\mathcal{H}_K} \left\{||S_{\mathbf{x}}f-\mathbf{y}||_m^2 + \lambda_1||f||_K^2 + \lambda_2||Bf||_K^2\right\} \\ \mu_1 = \frac{\alpha}{||f_{\mathbf{z},\lambda}||_K^2}, \quad \mu_2 = \frac{\alpha}{||Bf_{\mathbf{z},\lambda}||_K^2} \\ \tau = \frac{\alpha_o}{||S_{\mathbf{x}}f_{\mathbf{z},\lambda}-\mathbf{y}||_m^2} \end{cases}$$

where $\lambda_1 = \frac{\mu_1}{\tau},\ \lambda_2 = \frac{\mu_2}{\tau},\ \gamma = \frac{\alpha_o}{\alpha}$, this can be reformulated as

$$\begin{cases} f_{\mathbf{z},\lambda} = arg \min\limits_{f\in\mathcal{H}_K} \left\{||S_{\mathbf{x}}f-\mathbf{y}||_m^2 + \lambda_1||f||_K^2 + \lambda_2||Bf||_K^2\right\} \\ \lambda_1 = \frac{1}{\gamma}\frac{||S_{\mathbf{x}}f_{\mathbf{z},\lambda}-\mathbf{y}||_m^2}{||f_{\mathbf{z},\lambda}||_K^2},\ \lambda_2 = \frac{1}{\gamma}\frac{||S_{\mathbf{x}}f_{\mathbf{z},\lambda}-\mathbf{y}||_m^2}{||Bf_{\mathbf{z},\lambda}||_K^2} \end{cases}$$

which implies

$$\lambda_1||f_{\mathbf{z},\lambda}||_K^2 = \lambda_2||Bf_{\mathbf{z},\lambda}||_K^2.$$

It selects the regularization parameter λ in the functional (6) by balancing the penalty with the fidelity. Therefore the term "Penalty balancing principle" follows. Now we describe the fixed point algorithm based on PB-principle.

Algorithm 1 Parameter choice rule "Penalty-balancing Principle"

1. For an initial value $\lambda = (\lambda_1^0, \lambda_2^0)$, start with $k = 0$.
2. Calculate $f_{\mathbf{z},\lambda^k}$ and update λ by

$$\lambda_1^{k+1} = \frac{||S_{\mathbf{x}}f_{\mathbf{z},\lambda^k}-\mathbf{y}||_m^2 + \lambda_2^k||Bf_{\mathbf{z},\lambda^k}||_K^2}{(1+\gamma)||f_{\mathbf{z},\lambda^k}||_K^2},$$

$$\lambda_2^{k+1} = \frac{||S_{\mathbf{x}}f_{\mathbf{z},\lambda^k}-\mathbf{y}||_m^2 + \lambda_1^k||f_{\mathbf{z},\lambda^k}||_K^2}{(1+\gamma)||Bf_{\mathbf{z},\lambda^k}||_K^2}.$$

3. If stopping criteria $||\lambda^{k+1}-\lambda^k|| < \varepsilon$ satisfied then stop otherwise set $k = k+1$ and GOTO (2).

4 Numerical Realization

In this section, the performance of single-penalty regularization versus multi-penalty regularization is demonstrated using the academic example and two moon data set. For single-penalty regularization, parameters are chosen according to the quasi-optimality principle while for two-parameter regularization according to PB-principle.

We consider the well-known academic example [28, 16, 49] to test the multi-penalty regularization under PB-principle parameter choice rule,

$$f_\rho(x) = \frac{1}{10}\left\{x + 2\left(e^{-8(\frac{4\pi}{3}-x)^2} - e^{-8(\frac{\pi}{2}-x)^2} - e^{-8(\frac{3\pi}{2}-x)^2}\right)\right\}, \quad x\in[0,2\pi], \tag{29}$$

which belongs to reproducing kernel Hilbert space $\mathcal{H}_K$ corresponding to the kernel $K(x,y) = xy + \exp(-8(x-y)^2)$. We generate noisy data 100 times in the form $y = f_\rho(x) + \delta\xi$ corresponding to the inputs $\mathbf{x} = \{x_i\}_{i=1}^m = \{\frac{\pi}{10}(i-1)\}_{i=1}^m$, where ξ follows the uniform distribution over $[-1,1]$ with $\delta = 0.02$.

We consider the following multi-penalty functional proposed in the manifold regularization [13, 15],

$$\underset{f\in\mathcal{H}_K}{\operatorname{argmin}}\left\{\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2+\lambda_1||f||_K^2+\lambda_2||(S_{\mathbf{x}'}^{*}LS_{\mathbf{x}'})^{1/2}f||_K^2\right\}, \tag{30}$$

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

In our experiment, we illustrate the error estimates of single-penalty regularizers $f=f_{\mathbf{z},\lambda_1}$, $f=f_{\mathbf{z},\lambda_2}$ and multi-penalty regularizer $f=f_{\mathbf{z},\lambda}$ using the relative error measure $\frac{||f-f_\rho||}{||f||}$ for the academic example in sup norm, $\mathcal{H}_K$-norm and $||\cdot||_m$-empirical norm in Fig. 1 (a), (b) & (c) respectively.

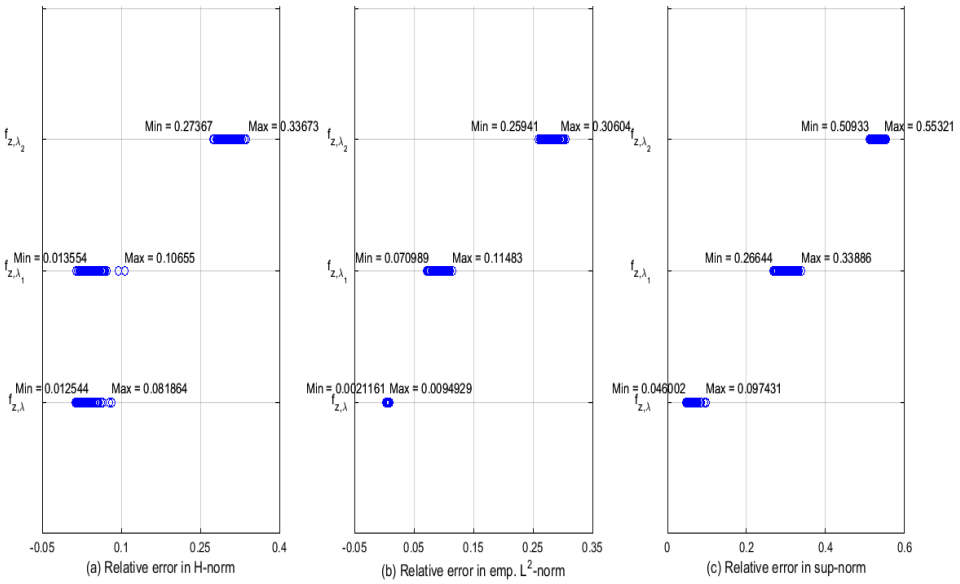

Figure 1: Figures show the relative errors of different estimators for the academic example in $||\cdot||_K$-norm (a), $||\cdot||_m$-empirical norm (b) and infinity norm (c) corresponding to 100 test problems with the noise $\delta=0.02$ for all estimators.

Now we compare the performance of multi-penalty regularization over single-penalty regularization method using the well-known two moon data set (Fig. 2) in the context of manifold learning. The data set contains 200 examples with k labeled example for each class. We perform experiments 500 times by taking $l=2k=2,6,10,20$ labeled points randomly. We solve the manifold regularization problem (30) for the 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 choose initial parameters $\lambda_1=1\times10^{-14}$, $\lambda_2=4.5\times10^{-3}$, the kernel parameter $\gamma=3.5$ and the weight parameter $b=3.125\times10^{-3}$ in all experiments. The performance of single-penalty ($\lambda_2=0$) and the proposed multi-penalty regularizer (30) is presented in Fig. 2, Table 1.

Based on the considered examples, we observe that the proposed multi-penalty regularization with the penalty balancing principle parameter choice outperforms the single-penalty regularizers.

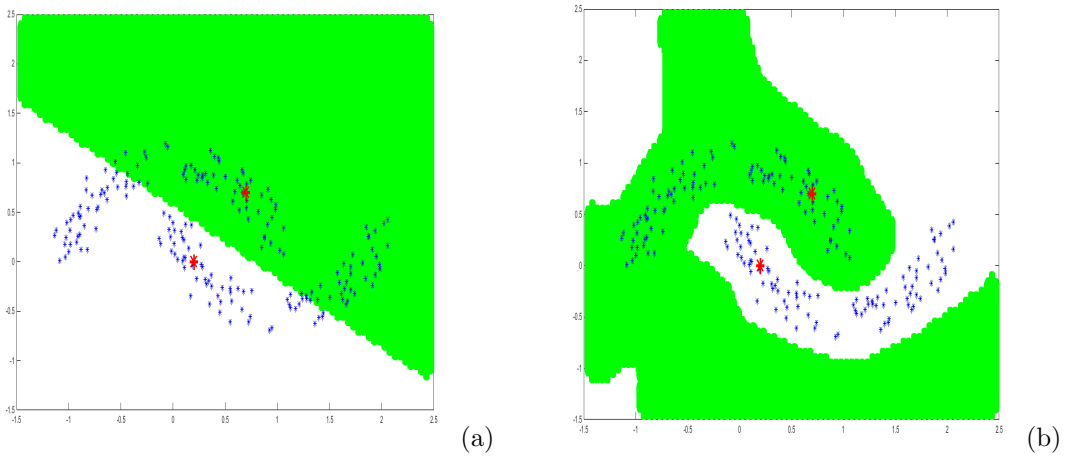

Figure 2: The figures show the decision surfaces generated with two labeled samples (red star) by single-penalty regularizer (a) based on the quasi-optimality principle and manifold regularizer (b) based on PB-principle.

	Single-penalty Regularizer			Multi-penalty Regularizer		
	(SP %)	(WC)	Parameters	(SP %)	(WC)	Parameters
$m = 2$	76.984	89	$\lambda_1 = 1.2 \times 10^{-14}$	100	0	$\lambda_1 = 1.1103 \times 10^{-14}$ $\lambda_2 = 5.9874 \times 10^{-4}$
$m = 6$	88.249	112	$\lambda_1 = 1.2 \times 10^{-14}$	100	0	$\lambda_1 = 9.8784 \times 10^{-15}$ $\lambda_2 = 5.7020 \times 10^{-4}$
$m = 10$	93.725	77	$\lambda_1 = 1.2 \times 10^{-14}$	100	0	$\lambda_1 = 1.0504 \times 10^{-14}$ $\lambda_2 = 7.3798 \times 10^{-4}$
$m = 20$	98.100	40	$\lambda_1 = 1.2 \times 10^{-14}$	100	0	$\lambda_1 = 1.0782 \times 10^{-14}$ $\lambda_2 = 7.0076 \times 10^{-4}$

Table 1: Statistical performance interpretation of single-penalty ($\lambda_2 = 0$) and multi-penalty regularizers of the functional

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

5 Conclusion

In summary, we achieved the optimal minimax rates of multi-penalized regression problem under the general source condition with the decay conditions of effective dimension. In particular, the convergence analysis of multi-penalty regularization provide the error estimates of manifold regularization problem. We can also address the convergence issues of binary classification problem using our error estimates. Here we discussed the penalty balancing principle based on augmented Tikhonov regularization for the choice of regularization parameters. Many other parameter choice rules are proposed to obtain the regularized solution of multi-parameter regularization schemes. The next problem of interest can be the rigorous analysis of different parameter choice rules of multi-penalty regularization schemes. Finally, the superiority of multi-penalty regularization over single-penalty regularization is shown using the academic example and moon data set.

Acknowledgements: 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] O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to statistical learning theory," in *Advanced lectures on machine learning*, pp. 169–207, Berlin/Heidelberg: Springer, 2004.

[2] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bull. Amer. Math. Soc. (NS)*, vol. 39, no. 1, pp. 1–49, 2002.

[3] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Adv. Comput. Math.*, vol. 13, no. 1, pp. 1–50, 2000.

[4] V. N. Vapnik and V. Vapnik, *Statistical Learning Theory*, vol. 1. New York: Wiley, 1998.

[5] F. Bauer, S. Pereverzev, and L. Rosasco, "On regularization algorithms in learning theory," *J. Complexity*, vol. 23, no. 1, pp. 52–72, 2007.

[6] A. Caponnetto and E. De Vito, "Optimal rates for the regularized least-squares algorithm," *Found. Comput. Math.*, vol. 7, no. 3, pp. 331–368, 2007.

[7] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, vol. 375. Dordrecht, The Netherlands: Math. Appl., Kluwer Academic Publishers Group, 1996.

[8] L. L. Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri, "Spectral algorithms for supervised learning," *Neural Computation*, vol. 20, no. 7, pp. 1873–1897, 2008.

[9] S. Smale and D. X. Zhou, "Learning theory estimates via integral operators and their approximations," *Constr. Approx.*, vol. 26, no. 2, pp. 153–172, 2007.

[10] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-posed Problems*, vol. 14. Washington, DC: W. H. Winston, 1977.

[11] S. Smale and D. X. Zhou, "Shannon sampling and function reconstruction from point values," *Bull. Amer. Math. Soc.*, vol. 41, no. 3, pp. 279–306, 2004.

[12] S. Smale and D. X. Zhou, "Shannon sampling II: Connections to learning theory," *Appl. Comput. Harmonic Anal.*, vol. 19, no. 3, pp. 285–302, 2005.

[13] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.

[14] D. Düvelmeyer and B. Hofmann, "A multi-parameter regularization approach for estimating parameters in jump diffusion processes," *J. Inverse Ill-Posed Probl.*, vol. 14, no. 9, pp. 861–880, 2006.

[15] S. Lu and S. V. Pereverzev, "Multi-parameter regularization and its numerical realization," *Numer. Math.*, vol. 118, no. 1, pp. 1–31, 2011.

[16] S. Lu, S. Pereverzyev Jr., and S. Sivananthan, "Multiparameter regularization for construction of extrapolating estimators in statistical learning theory," in *Multiscale Signal Analysis and Modeling*, pp. 347–366, New York: Springer, 2013.

[17] Y. Lu, L. Shen, and Y. Xu, "Multi-parameter regularization methods for high-resolution image reconstruction with displacement errors," *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 54, no. 8, pp. 1788–1799, 2007.

[18] V. Naumova and S. V. Pereverzyev, "Multi-penalty regularization with a component-wise penalization," *Inverse Problems*, vol. 29, no. 7, p. 075002, 2013.

[19] S. N. Wood, "Modelling and smoothing parameter estimation with multiple quadratic penalties," *J. R. Statist. Soc.*, vol. 62, pp. 413–428, 2000.

[20] P. Xu, Y. Fukuda, and Y. Liu, "Multiple parameter regularization: Numerical solutions and applications to the determination of geopotential from precise satellite orbits," *J. Geodesy*, vol. 80, no. 1, pp. 17–27, 2006.

[21] Y. Luo, D. Tao, C. Xu, D. Li, and C. Xu, "Vector-valued multi-view semi-supervsed learning for multi-label image classification.," in *AAAI*, pp. 647–653, 2013.

[22] Y. Luo, D. Tao, C. Xu, C. Xu, H. Liu, and Y. Wen, "Multiview vector-valued manifold regularization for multilabel image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 709–722, 2013.

[23] H. Q. Minh, L. Bazzani, and V. Murino, "A unifying framework in vector-valued reproducing kernel Hilbert spaces for manifold regularization and co-regularized multi-view learning," *J. Mach. Learn. Res.*, vol. 17, no. 25, pp. 1–72, 2016.

[24] H. Q. Minh and V. Sindhwani, "Vector-valued manifold regularization," in *International Conference on Machine Learning*, 2011.

[25] Abhishake and S. Sivananthan, "Multi-penalty regularization in learning theory," *J. Complexity*, vol. 36, pp. 141–165, 2016.

[26] F. Bauer and S. Kindermann, "The quasi-optimality criterion for classical inverse problems," *Inverse Problems*, vol. 24, p. 035002, 2008.

[27] A. Caponnetto and Y. Yao, "Cross-validation based adaptation for regularization operators in learning theory," *Anal. Appl.*, vol. 8, no. 2, pp. 161–183, 2010.

[28] E. De Vito, S. Pereverzyev, and L. Rosasco, "Adaptive kernel methods using the balancing principle," *Found. Comput. Math.*, vol. 10, no. 4, pp. 455–479, 2010.

[29] V. A. Morozov, "On the solution of functional equations by the method of regularization," *Soviet Math. Dokl*, vol. 7, no. 1, pp. 414–417, 1966.

[30] J. Xie and J. Zou, "An improved model function method for choosing regularization parameters in linear inverse problems," *Inverse Problems*, vol. 18, no. 3, pp. 631–643, 2002.

[31] S. Lu, S. V. Pereverzev, and U. Tautenhahn, "A model function method in regularized total least squares," *Appl. Anal.*, vol. 89, no. 11, pp. 1693–1703, 2010.

[32] M. Fornasier, V. Naumova, and S. V. Pereverzyev, "Parameter choice strategies for multi-penalty regularization," *SIAM J. Numer. Anal.*, vol. 52, no. 4, pp. 1770–1794, 2014.

[33] K. Ito, B. Jin, and T. Takeuchi, "Multi-parameter Tikhonov regularization–An augmented approach," *Chinese Ann. Math.*, vol. 35, no. 3, pp. 383–398, 2014.

[34] M. Belge, M. E. Kilmer, and E. L. Miller, "Efficient determination of multiple regularization parameters in a generalized L-curve framework," *Inverse Problems*, vol. 18, pp. 1161–1183, 2002.

[35] F. Bauer and O. Ivanyshyn, "Optimal regularization with two interdependent regularization parameters," *Inverse problems*, vol. 23, no. 1, pp. 331–342, 2007.

[36] F. Bauer and S. V. Pereverzev, "An utilization of a rough approximation of a noise covariance within the framework of multi-parameter regularization," *Int. J. Tomogr. Stat*, vol. 4, pp. 1–12, 2006.

[37] Z. Chen, Y. Lu, Y. Xu, and H. Yang, "Multi-parameter Tikhonov regularization for linear ill-posed operator equations," *J. Comp. Math.*, vol. 26, pp. 37–55, 2008.

[38] C. Brezinski, M. Redivo-Zaglia, G. Rodriguez, and S. Seatzu, "Multi-parameter regularization techniques for ill-conditioned linear systems," *Numer. Math.*, vol. 94, no. 2, pp. 203–228, 2003.

[39] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.

[40] P. Mathé and S. V. Pereverzev, "Geometry of linear ill-posed problems in variable Hilbert scales," *Inverse problems*, vol. 19, no. 3, pp. 789–803, 2003.

[41] S. Lu, P. Mathé, and S. Pereverzyev, "Balancing principle in supervised learning for a general regularization scheme," *RICAM-Report*, vol. 38, 2016.

[42] G. Blanchard and P. Mathé, "Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration," *Inverse problems*, vol. 28, no. 11, p. 115011, 2012.

[43] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone, "Learning from examples as an inverse problem," *J. Mach. Learn. Res.*, vol. 6, pp. 883–904, 2005.

[44] A. Rastogi and S. Sivananthan, "Optimal rates for the regularized learning algorithms under general source condition," *Front. Appl. Math. Stat.*, vol. 3, p. 3, 2017.

[45] S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of classification: A survey of some recent advances," *ESAIM: probability and statistics*, vol. 9, pp. 323–375, 2005.

[46] S. Lu and S. Pereverzev, *Regularization Theory for Ill-posed Problems: Selected Topics*, vol. 58. Berlin: Walter de Gruyter, 2013.

[47] A. M. Thompson and J. Kay, "On some choices of regularization parameter in image restoration," *Inverse Problems*, vol. 9, pp. 749–761, 1993.

[48] B. Jin and J. Zou, "Augmented Tikhonov regularization," *Inverse Problems*, vol. 25, no. 2, p. 025001, 2008.

[49] C. A. Micchelli and M. Pontil, "Learning the kernel function via regularization," *J. Mach. Learn. Res.*, vol. 6, no. 2, pp. 1099–1125, 2005.