frontiers in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 27 March 2017
doi: 10.3389/fams.2017.00003

Optimal Rates for the Regularized Learning Algorithms under General Source Condition

Abhishake Rastogi and Sivananthan Sampath*

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

We consider the learning algorithms under general source condition with the polynomial decay of the eigenvalues of the integral operator in vector-valued function setting. We discuss the upper convergence rates of Tikhonov regularizer under general source condition corresponding to increasing monotone index function. The convergence issues are studied for general regularization schemes by using the concept of operator monotone index functions in minimax setting. Further we also address the minimum possible error for any learning algorithm.

Keywords: learning theory, general source condition, vector-valued RKHS, error estimate, optimal rates
Mathematics Subject Classification 2010: 68T05, 68Q32

OPEN ACCESS

Edited by:
Yiming Ying,
University at Albany, SUNY, USA

Reviewed by:
Xin Guo,
The Hong Kong Polytechnic University, Hong Kong
Ernesto De Vito,
University of Genoa, Italy

****Correspondence:***
Abhishake Rastogi
abhishekrastogi2012@gmail.com

Specialty section:
This article was submitted to Mathematics of Computation and Data Science, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *02 November 2016*
Accepted: *09 March 2017*
Published: *27 March 2017*

Citation:
Rastogi A and Sampath S (2017) Optimal Rates for the Regularized Learning Algorithms under General Source Condition. Front. Appl. Math. Stat. 3:3. doi: 10.3389/fams.2017.00003

1. INTRODUCTION

Learning theory [1–3] aims to learn the relation between the inputs and outputs based on finite random samples. We require some underlying space to search the relation function. From the experiences we have some idea about the underlying space which is called hypothesis space. Learning algorithms tries to infer the best estimator over the hypothesis space such that $f(x)$ gives the maximum information of the output variable y for any unseen input x. The given samples $\{x_i, y_i\}_{i=1}^m$ are not exact in the sense that for underlying relation function $f(x_i) \neq y_i$ but $f(x_i) \approx y_i$. We assume that the uncertainty follows the probability distribution ρ on the sample space $X \times Y$ and the underlying function (called the regression function) for the probability distribution ρ is given by

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

where $\rho(y|x)$ is the conditional probability measure for given x. The problem of obtaining estimator from examples is ill-posed. Therefore, we apply the regularization schemes [4–7] to stabilize the problem. Various regularization schemes are studied for inverse problems. In the context of learning theory [2, 3, 8–10], the square loss-regularization (Tikhonov regularization) is widely considered to obtain the regularized estimator [9, 11–16]. Gerfo et al. [6] introduced general regularization in the learning theory and provided the error bounds under Hölder's source condition [5]. Bauer et al. [4] discussed the convergence issues for general regularization under general source condition [17] by removing the Lipschitz condition on the regularization considered in Gerfo et al. [6]. Caponnetto and De Vito [12] discussed the square-loss regularization under the polynomial decay of the eigenvalues of the integral operator L_K with Hölder's source condition. For the inverse statistical learning problem, Blanchard and Mücke [18] analyzed the convergence rates for general regularization scheme under Hölder's source condition in scalar-valued function setting. Here we are discussing the convergence issues of general regularization schemes under general

source condition and the polynomial decay of the eigenvalues of the integral operator in vector-valued framework. We present the minimax upper convergence rates for Tikhonov regularization under general source condition $\Omega_{\phi,R}$, for a monotone increasing index function ϕ. For general regularization the minimax rates are obtained using the operator monotone index function ϕ. The concept of effective dimension [19, 20] is exploited to achieve the convergence rates. In the choice of regularization parameters, the effective dimension plays the important role. We also discuss the lower convergence rates for any learning algorithm under the smoothness conditions. We present the results in vector-valued function setting. Therefore, in particular they can be applied to multi-task learning problems.

The structure of the paper is as follows. In the second section, we introduce some basic assumptions and notations for supervised learning problems. In Section 3, we present the upper and lower convergence rates under the smoothness conditions in minimax setting.

2. LEARNING FROM EXAMPLES: NOTATIONS AND ASSUMPTIONS

In the learning theory framework [2, 3, 8–10], the sample space $Z = X \times Y$ consists of two spaces: The input space X (locally compact second countable Hausdorff space) and the output space $(Y, \langle\cdot,\cdot\rangle_Y)$ (the real separable Hilbert space). The input space X and the output space Y are related by some unknown probability distribution ρ on Z. The probability measure can be split as $\rho(x,y) = \rho(y|x)\rho_X(x)$, where $\rho(y|x)$ is the conditional probability measure of y given x and ρ_X is the marginal probability measure on X. The only available information is the random i.i.d. samples $\mathbf{z} = ((x_1,y_1),\ldots,(x_m,y_m))$ drawn according to the probability measure ρ. Given the training set $\mathbf{z}$, learning theory aims to develop an algorithm which provides an estimator $f_{\mathbf{z}} : X \to Y$ such that $f_{\mathbf{z}}(x)$ predicts the output variable y for any given input x. The goodness of the estimator can be measured by the generalization error of a function f which can be defined as

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

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. The regression function f_ρ belongs to the space of square integrable functions provided that

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

We search the minimizer of the generalization error over a hypothesis space $\mathcal{H}$,

$$f_{\mathcal{H}} := \underset{f\in\mathcal{H}}{\operatorname{argmin}} \left\{ \int_Z ||f(x)-y||_Y^2 d\rho(x,y) \right\}, \tag{4}$$

where $f_{\mathcal{H}}$ is called the target function. In case $f_\rho \in \mathcal{H}$, $f_{\mathcal{H}}$ becomes the regression function f_ρ.

Because of inaccessibility of the probability distribution ρ, we minimize the regularized empirical estimate of the generalization error over the hypothesis space $\mathcal{H}$,

$$f_{\mathbf{z},\lambda} := \underset{f\in\mathcal{H}}{\operatorname{argmin}} \left\{ \frac{1}{m}\sum_{i=1}^m ||f(x_i)-y_i||_Y^2 + \lambda||f||_{\mathcal{H}}^2 \right\}, \tag{5}$$

where λ is the positive regularization parameter. The regularization schemes [4–7, 10] are used to incorporate various features in the solution such as boundedness, monotonicity and smoothness. In order to optimize the vector-valued regularization functional, one of the main problems is to choose the appropriate hypothesis space which is assumed to be a source to provide the estimator.

2.1. Reproducing Kernel Hilbert Space as a Hypothesis Space

Definition 2.1. (Vector-valued reproducing kernel Hilbert space) *For non-empty set X and the real Hilbert space $(Y, \langle\cdot,\cdot\rangle_Y)$, the Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of functions from X to Y is called reproducing kernel Hilbert space if for any $x \in X$ and $y \in Y$ the linear functional which maps $f \in \mathcal{H}$ to $\langle y, f(x)\rangle_Y$ is continuous.*

By Riesz lemma [21], for every $x \in X$ and $y \in Y$ there exists a linear operator $K_x : Y \to \mathcal{H}$ such that

$$\langle y, f(x)\rangle_Y = \langle K_x y, f\rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}.$$

Therefore, the adjoint operator $K_x^* : \mathcal{H} \to Y$ is given by $K_x^* f = f(x)$. Through the linear operator $K_x : Y \to \mathcal{H}$ we define the linear operator $K(x,t) : Y \to Y$,

$$K(x,t)y := K_t y(x).$$

From Proposition 2.1 [22], the linear operator $K(x,t) \in \mathcal{L}(Y)$ (the set of bounded linear operators on Y), $K(x,t) = K(t,x)^*$ and $K(x,x)$ is non-negative bounded linear operator. For any $m \in \mathbb{N}, \{x_i : 1 \le i \le m\} \in X, \{y_i : 1 \le i \le m\} \in Y$, we have that $\sum_{i,j=1}^m \langle y_i, K(x_i,x_j)y_j\rangle \ge 0$. The operator valued function $K : X \times X \to \mathcal{L}(Y)$ is called the kernel.

There is one to one correspondence between the kernels and reproducing kernel Hilbert spaces [22, 23]. So a reproducing kernel Hilbert space $\mathcal{H}$ corresponding to a kernel K can be denoted as $\mathcal{H}_K$ and the norm in the space $\mathcal{H}$ can be denoted as $||\cdot||_{\mathcal{H}_K}$. In the following article, we suppress K by simply using $\mathcal{H}$ for reproducing kernel Hilbert space and $||\cdot||_{\mathcal{H}}$ for its norm.

Throughout the paper we assume the reproducing kernel Hilbert space $\mathcal{H}$ is separable such that

(i) $K_x : Y \to \mathcal{H}$ is a Hilbert-Schmidt operator for all $x \in X$ and $\kappa := \sqrt{\sup_{x\in X} Tr(K_x^* K_x)} < \infty$.

(ii) The real function from $X \times X$ to $\mathbb{R}$, defined by $(x,t) \mapsto \langle K_t v, K_x w\rangle_{\mathcal{H}}$, is measurable $\forall v, w \in Y$.

By the representation theorem [22], the solution of the penalized regularization problem (5) will be of the form:

$$f_{\mathbf{z},\lambda} = \sum_{i=1}^{m} K_{x_i} c_i, \text{ for } (c_1, \ldots, c_m) \in Y^m.$$

Definition 2.2. *let $\mathcal{H}$ be a separable Hilbert space and $\{e_k\}_{k=1}^{\infty}$ be an orthonormal basis of $\mathcal{H}$. Then for any positive operator $A \in \mathcal{L}(\mathcal{H})$ we define $Tr(A) = \sum_{k=1}^{\infty} \langle Ae_k, e_k \rangle$. It is well-known that the number $Tr(A)$ is independent of the choice of the orthonormal basis.*

Definition 2.3. *An operator $A \in \mathcal{L}(\mathcal{H})$ is called Hilbert-Schmidt operator if $Tr(A^*A) < \infty$. The family of all Hilbert-Schmidt operators is denoted by $\mathcal{L}_2(\mathcal{H})$. For $A \in \mathcal{L}_2(\mathcal{H})$, we define $Tr(A) = \sum_{k=1}^{\infty} \langle Ae_k, e_k \rangle$ for an orthonormal basis $\{e_k\}_{k=1}^{\infty}$ of $\mathcal{H}$.*

It is well-known that $\mathcal{L}_2(\mathcal{H})$ is the separable Hilbert space with the inner product,

$$\langle A, B \rangle_{\mathcal{L}_2(\mathcal{H})} = Tr(B^*A)$$

and its norm satisfies

$$||A||_{\mathcal{L}(\mathcal{H})} \leq ||A||_{\mathcal{L}_2(\mathcal{H})} \leq Tr(|A|),$$

where $|A| = \sqrt{A^*A}$ and $||\cdot||_{\mathcal{L}(\mathcal{H})}$ is the operator norm (For more details see [24]).

For the positive trace class operator $K_x K_x^*$, we have

$$||K_x K_x^*||_{\mathcal{L}(\mathcal{H})} \leq ||K_x K_x^*||_{\mathcal{L}_2(\mathcal{H})} \leq Tr(K_x K_x^*) \leq \kappa^2.$$

Given the ordered set $\mathbf{x} = (x_1, \ldots, x_m) \in X^m$, the sampling operator $S_{\mathbf{x}} : \mathcal{H} \to Y^m$ is defined by $S_{\mathbf{x}}(f) = (f(x_1), \ldots, f(x_m))$ and its adjoint $S_{\mathbf{x}}^* : Y^m \to \mathcal{H}$ is given by $S_{\mathbf{x}}^* \mathbf{y} = \frac{1}{m} \sum_{i=1}^{m} K_{x_i} y_i$, $\forall\, \mathbf{y} = (y_1, \ldots, y_m) \in Y^m$.

The regularization scheme (5) can be expressed as

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

where $||\mathbf{y}||_m^2 = \frac{1}{m} \sum_{i=1}^{m} ||y_i||_Y^2$.

We obtain the explicit expression of $f_{\mathbf{z},\lambda}$ by taking the functional derivative of above expression over RKHS $\mathcal{H}$.

Theorem 2.1. *For the positive choice of λ, the functional (6) has unique minimizer:*

$$f_{\mathbf{z},\lambda} = \left(S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda I \right)^{-1} S_{\mathbf{x}}^* \mathbf{y}. \tag{7}$$

Define f_λ as the minimizer of the optimization functional,

$$f_\lambda := \underset{f \in \mathcal{H}}{\operatorname{argmin}} \left\{ \int_Z ||f(x) - y||_Y^2 d\rho(x, y) + \lambda ||f||_{\mathcal{H}}^2 \right\}. \tag{8}$$

Using the fact $\mathcal{E}(f) = ||L_K^{1/2}(f - f_{\mathcal{H}})||_{\mathcal{H}}^2 + \mathcal{E}(f_{\mathcal{H}})$, we get the expression of f_λ,

$$f_\lambda = (L_K + \lambda I)^{-1} L_K f_{\mathcal{H}}, \tag{9}$$

where the integral operator $L_K : \mathscr{L}_{\rho_X}^2 \to \mathscr{L}_{\rho_X}^2$ 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 $\mathcal{H}$. We use the same notation L_K for both the operators defined on different domains. It is well-known that $L_K^{1/2}$ is an isometry from the space of square integrable functions to reproducing kernel Hilbert space.

In order to achieve the uniform convergence rates for learning algorithms we need some prior assumptions on the probability measure ρ. Following the notion of Bauer et al. [4] and Caponnetto and De Vito [12], we consider the class of probability measures $\mathcal{P}_\phi$ which satisfies the assumptions:

(i) For the probability measure ρ on $X \times Y$,

$$\int_Z ||y||_Y^2 \, d\rho(x, y) < \infty. \tag{10}$$

(ii) The minimizer of the generalization error $f_{\mathcal{H}}$ (4) over the hypothesis space $\mathcal{H}$ exists.

(iii) There exist some constants M, Σ such that for almost all $x \in X$,

$$\int_Y \left(e^{||y - f_{\mathcal{H}}(x)||_Y / M} - \frac{||y - f_{\mathcal{H}}(x)||_Y}{M} - 1 \right) d\rho(y|x) \leq \frac{\Sigma^2}{2M^2}. \tag{11}$$

(iv) The target function $f_{\mathcal{H}}$ belongs to the class $\Omega_{\phi,R}$ with

$$\Omega_{\phi,R} := \left\{ f \in \mathcal{H} : f = \phi(L_K) g \text{ and } ||g||_{\mathcal{H}} \leq R \right\}, \tag{12}$$

where ϕ is a continuous increasing index function defined on the interval $[0, \kappa^2]$ with the assumption $\phi(0) = 0$. This condition is usually referred to as general source condition [17].

In addition, we consider the set of probability measures $\mathcal{P}_{\phi,b}$ which satisfies the conditions (i), (ii), (iii), (iv) and the eigenvalues t_n's of the integral operator L_K follow the polynomial decay: For fixed positive constants α, β and $b > 1$,

$$\alpha n^{-b} \leq t_n \leq \beta n^{-b} \ \ \forall n \in \mathbb{N}. \tag{13}$$

Under the polynomial decay of the eigenvalues the effective dimension $\mathcal{N}(\lambda)$, to measure the complexity of RKHS, can be estimated from Proposition 3 [12] as follows,

$$\mathcal{N}(\lambda) := Tr\left((L_K + \lambda I)^{-1} L_K \right) \leq \frac{\beta b}{b - 1} \lambda^{-1/b}, \text{ for } b > 1 \tag{14}$$

and without the polynomial decay condition (13), we have

$$\mathcal{N}(\lambda) \leq ||(L_K + \lambda I)^{-1}||_{\mathcal{L}(\mathcal{H})} Tr(L_K) \leq \frac{\kappa^2}{\lambda}.$$

We discuss the convergence issues for the learning algorithms ($\mathbf{z} \to f_{\mathbf{z}} \in \mathcal{H}$) in probabilistic sense by exponential tail inequalities such that

$$Prob_{\mathbf{z}}\left\{||f_{\mathbf{z}} - f_\rho||_\rho \le \varepsilon(m)\log\left(\frac{1}{\eta}\right)\right\} \ge 1-\eta$$

for all $0 < \eta \le 1$ and $\varepsilon(m)$ is a positive decreasing function of m. Using these probabilistic estimates we can obtain error estimates in expectation by integration of tail inequalities:

$$\begin{aligned} E_{\mathbf{z}}\left(||f_{\mathbf{z}} - f_\rho||_\rho\right) &= \int_0^\infty Prob_{\mathbf{z}}\left(||f_{\mathbf{z}} - f_\rho||_\rho > t\right) dt \\ &\le \int_0^\infty \exp\left(-\frac{t}{\varepsilon(m)}\right) dt = \varepsilon(m), \end{aligned}$$

where $||f||_\rho = ||f||_{\mathcal{L}^2_{\rho_X}} = \{\int_X ||f(x)||^2_Y d\rho_X(x)\}^{1/2}$ and $E_{\mathbf{z}}(\xi) = \int_{Z^m} \xi\, d\rho(z_1)\ldots d\rho(z_m)$.

3. CONVERGENCE ANALYSIS

In this section, we analyze the convergence issues of the learning algorithms on reproducing kernel Hilbert space under the smoothness priors in the supervised learning framework. We discuss the upper and lower convergence rates for vector-valued estimators in the standard minimax setting. Therefore, the estimates can be utilized particularly for scalar-valued functions and multi-task learning algorithms.

3.1. Upper Rates for Tikhonov Regularization

In General, we consider Tikhonov regularization in learning theory. Tikhonov regularization is briefly discussed in the literature [7, 9, 10, 25]. The error estimates for Tikhonov regularization are discussed theoretically under Hölder's source condition [12, 15, 16]. We establish the error estimates for Tikhonov regularization scheme under general source condition $f_{\mathcal{H}} \in \Omega_{\phi,R}$ for some continuous increasing index function ϕ and the polynomial decay of the eigenvalues of the integral operator L_K.

In order to estimate the error bounds, we consider the following inequality used in the papers [4, 12] which is based on the results of Pinelis and Sakhanenko [26].

Proposition 3.1. *Let ξ be a random variable on the probability space $(\Omega, \mathcal{B}, P)$ with values in real separable Hilbert space $\mathcal{H}$. If there exist two constants Q and S satisfying*

$$E\left\{||\xi - E(\xi)||^n_{\mathcal{H}}\right\} \le \frac{1}{2}n!S^2Q^{n-2} \quad \forall n \ge 2, \tag{15}$$

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

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

In particular, the inequality (15) holds if

$$||\xi(\omega)||_{\mathcal{H}} \le Q \text{ and } E(||\xi(\omega)||^2_{\mathcal{H}}) \le S^2.$$

We estimate the error bounds for the regularized estimators by measuring the effect of random sampling and the complexity of $f_{\mathcal{H}}$. The quantities described in Proposition 3.2 express the probabilistic estimates of the perturbation measure due to random sampling. The expressions of Proposition 3.3 describe the complexity of the target function $f_{\mathcal{H}}$ which are usually referred to as the approximation errors. The approximation errors are independent of the samples $\mathbf{z}$.

Proposition 3.2. *Let $\mathbf{z}$ be i.i.d. samples drawn according to the probability measure ρ satisfying the assumptions (10), (11) and $\kappa = \sqrt{\sup_{x\in X} Tr(K^*_x K_x)}$. Then for all $0 < \eta < 1$, we have*

$$\begin{aligned} &||(L_K + \lambda I)^{-1/2}\{S^*_{\mathbf{x}}\mathbf{y} - S^*_{\mathbf{x}}S_{\mathbf{x}}f_{\mathcal{H}}\}||_{\mathcal{H}} \\ &\le 2\left(\frac{\kappa M}{m\sqrt{\lambda}} + \sqrt{\frac{\Sigma^2\mathcal{N}(\lambda)}{m}}\right)\log\left(\frac{4}{\eta}\right) \end{aligned} \tag{16}$$

and

$$||S^*_{\mathbf{x}}S_{\mathbf{x}} - L_K||_{\mathcal{L}_2(\mathcal{H})} \le 2\left(\frac{\kappa^2}{m} + \frac{\kappa^2}{\sqrt{m}}\right)\log\left(\frac{4}{\eta}\right). \tag{17}$$

with the confidence $1-\eta$.

The proof of the first expression is the content of the step 3.2 of Theorem 4 [12] while the proof of the second expression can be obtained from Theorem 2 in De Vito et al. [25].

Proposition 3.3. *Suppose $f_{\mathcal{H}} \in \Omega_{\phi,R}$. Then,*

(i) Under the assumption that $\phi(t)\sqrt{t}$ and $\sqrt{t}/\phi(t)$ are non-decreasing functions, we have

$$||f_\lambda - f_{\mathcal{H}}||_\rho \le R\phi(\lambda)\sqrt{\lambda}. \tag{18}$$

(ii) Under the assumption that $\phi(t)$ and $t/\phi(t)$ are non-decreasing functions, we have

$$||f_\lambda - f_{\mathcal{H}}||_\rho \le R\kappa\phi(\lambda) \tag{19}$$

and

$$||f_\lambda - f_{\mathcal{H}}||_{\mathcal{H}} \le R\phi(\lambda). \tag{20}$$

Under the source condition $f_{\mathcal{H}} \in \Omega_{\phi,R}$, the proposition can be proved using the ideas of Theorem 10 [4].

Theorem 3.1. *Let* $\mathbf{z}$ *be i.i.d. samples drawn according to the probability measure* $\rho \in \mathcal{P}_\phi$ *where* ϕ *is the index function satisfying the conditions that* $\phi(t)$, $t/\phi(t)$ *are non-decreasing functions. Then for all* $0 < \eta < 1$, *with confidence* $1-\eta$, *for the regularized estimator* $f_{\mathbf{z},\lambda}$ *(7) the following upper bound holds:*

$$||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_{\mathcal{H}} \leq 2\left\{R\phi(\lambda) + \frac{2\kappa M}{m\lambda} + \sqrt{\frac{4\Sigma^2\mathcal{N}(\lambda)}{m\lambda}}\right\}\log\left(\frac{4}{\eta}\right)$$

provided that

$$\sqrt{m}\lambda \geq 8\kappa^2\log(4/\eta). \tag{21}$$

Proof. The error of regularized solution $f_{\mathbf{z},\lambda}$ can be estimated in terms of the sample error and the approximation error as follows:

$$||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_{\mathcal{H}} \leq ||f_{\mathbf{z},\lambda} - f_\lambda||_{\mathcal{H}} + ||f_\lambda - f_{\mathcal{H}}||_{\mathcal{H}}. \tag{22}$$

Now $f_{\mathbf{z},\lambda} - f_\lambda$ can be expressed as

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

Then $f_\lambda = (L_K + \lambda I)^{-1}L_K f_{\mathcal{H}}$ implies

$$L_K f_{\mathcal{H}} = L_K f_\lambda + \lambda f_\lambda.$$

Therefore,

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

Employing RKHS-norm we get,

$$\begin{aligned} ||f_{\mathbf{z},\lambda} - f_\lambda||_{\mathcal{H}} &\leq ||(S_{\mathbf{x}}^*S_{\mathbf{x}} + \lambda I)^{-1}\{S_{\mathbf{x}}^*\mathbf{y} - S_{\mathbf{x}}^*S_{\mathbf{x}}f_{\mathcal{H}} \\ &\quad + (S_{\mathbf{x}}^*S_{\mathbf{x}} - L_K)(f_{\mathcal{H}} - f_\lambda)\}||_{\mathcal{H}} \\ &\leq I_1I_2 + I_3||f_\lambda - f_{\mathcal{H}}||_{\mathcal{H}}/\lambda, \end{aligned} \tag{23}$$

where $I_1 = ||(S_{\mathbf{x}}^*S_{\mathbf{x}} + \lambda I)^{-1}(L_K + \lambda I)^{1/2}||_{\mathcal{L}(\mathcal{H})}$, $I_2 = ||(L_K + \lambda I)^{-1/2}(S_{\mathbf{x}}^*\mathbf{y} - S_{\mathbf{x}}^*S_{\mathbf{x}}f_{\mathcal{H}})||_{\mathcal{H}}$ and $I_3 = ||S_{\mathbf{x}}^*S_{\mathbf{x}} - L_K||_{\mathcal{L}(\mathcal{H})}$.

The estimates of I_2, I_3 can be obtained from Proposition 3.2 and the only task is to bound I_1. For this we consider

$$\begin{aligned}(S_{\mathbf{x}}^*S_{\mathbf{x}} + \lambda I)^{-1}(L_K + \lambda I)^{1/2} &= \{I - (L_K + \lambda I)^{-1}(L_K - S_{\mathbf{x}}^*S_{\mathbf{x}})\}^{-1} \\ &\quad (L_K + \lambda I)^{-1/2}\end{aligned}$$

which implies

$$I_1 \leq \sum_{n=0}^{\infty}||(L_K + \lambda I)^{-1}(L_K - S_{\mathbf{x}}^*S_{\mathbf{x}})||_{\mathcal{L}(\mathcal{H})}^n||(L_K + \lambda I)^{-1/2}||_{\mathcal{L}(\mathcal{H})} \tag{24}$$

provided that $||(L_K + \lambda I)^{-1}(L_K - S_{\mathbf{x}}^*S_{\mathbf{x}})||_{\mathcal{L}(\mathcal{H})} < 1$. To verify this condition, we consider

$$||(L_K + \lambda I)^{-1}(S_{\mathbf{x}}^*S_{\mathbf{x}} - L_K)||_{\mathcal{L}(\mathcal{H})} \leq I_3/\lambda.$$

Now using Proposition 3.2 we get with confidence $1-\eta/2$,

$$||(L_K + \lambda I)^{-1}(S_{\mathbf{x}}^*S_{\mathbf{x}} - L_K)||_{\mathcal{L}(\mathcal{H})} \leq \frac{4\kappa^2}{\sqrt{m}\lambda}\log\left(\frac{4}{\eta}\right).$$

From the condition (21) we get with confidence $1-\eta/2$,

$$||(L_K + \lambda I)^{-1}(S_{\mathbf{x}}^*S_{\mathbf{x}} - L_K)||_{\mathcal{L}(\mathcal{H})} \leq \frac{1}{2}. \tag{25}$$

Consequently, using (25) in the inequality (24) we obtain with probability $1-\eta/2$,

$$\begin{aligned} I_1 &= ||(S_{\mathbf{x}}^*S_{\mathbf{x}} + \lambda I)^{-1}(L_K + \lambda I)^{1/2}||_{\mathcal{L}(\mathcal{H})} \\ &\leq 2||(L_K + \lambda I)^{-1/2}||_{\mathcal{L}(\mathcal{H})} \leq \frac{2}{\sqrt{\lambda}}. \end{aligned} \tag{26}$$

From Proposition 3.2 we have with confidence $1-\eta/2$,

$$||S_{\mathbf{x}}^*S_{\mathbf{x}} - L_K||_{\mathcal{L}(\mathcal{H})} \leq 2\left(\frac{\kappa^2}{m} + \frac{\kappa^2}{\sqrt{m}}\right)\log\left(\frac{4}{\eta}\right).$$

Again from the condition (21) we get with probability $1-\eta/2$,

$$I_3 = ||S_{\mathbf{x}}^*S_{\mathbf{x}} - L_K||_{\mathcal{L}(\mathcal{H})} \leq \frac{\lambda}{2}. \tag{27}$$

Therefore, the inequality (23) together with (16), (20), (26), (27) provides the desired bound. □

The following theorem discuss the error estimates in $\mathscr{L}^2$-norm. The proof is similar to the above theorem.

Theorem 3.2. *Let* $\mathbf{z}$ *be i.i.d. samples drawn according to the probability measure* $\rho \in \mathcal{P}_\phi$ *and* $f_{\mathbf{z},\lambda}$ *is the regularized solution (7) corresponding to Tikhonov regularization. Then for all* $0 < \eta < 1$, *with confidence* $1-\eta$, *the following upper bounds holds:*

(i) Under the assumption that $\phi(t)$, $\sqrt{t}/\phi(t)$ *are non-decreasing functions,*

$$||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_\rho \leq 2\left\{R\phi(\lambda)\sqrt{\lambda} + \frac{2\kappa M}{m\sqrt{\lambda}} + \sqrt{\frac{4\Sigma^2\mathcal{N}(\lambda)}{m}}\right\}\log\left(\frac{4}{\eta}\right)$$

(ii) Under the assumption that $\phi(t)$, $t/\phi(t)$ *are non-decreasing functions,*

$$||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_\rho \leq \left\{R(\kappa + \sqrt{\lambda})\phi(\lambda) + \frac{4\kappa M}{m\sqrt{\lambda}} + \sqrt{\frac{16\Sigma^2\mathcal{N}(\lambda)}{m}}\right\}\log\left(\frac{4}{\eta}\right)$$

provided that

$$\sqrt{m}\lambda \geq 8\kappa^2\log(4/\eta). \tag{28}$$

We derive the convergence rates of Tikhonov regularizer based on data-driven strategy of the parameter choice of λ for the class of probability measure $\mathcal{P}_{\phi,b}$.

Theorem 3.3. *Under the same assumptions of Theorem 3.2 and hypothesis (13), the convergence of the estimator* $f_{\mathbf{z},\lambda}$ *(7) to the target function* $f_{\mathcal{H}}$ *can be described as:*

(i) If $\phi(t)$ and $\sqrt{t}/\phi(t)$ are non-decreasing functions. Then under the parameter choice $\lambda \in (0,1]$, $\lambda = \Psi^{-1}(m^{-1/2})$ where $\Psi(t) = t^{\frac{1}{2}+\frac{1}{2b}}\phi(t)$, we have

$$Prob_{\mathbf{z}}\left\{\begin{array}{c} ||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_\rho \leq C(\Psi^{-1}(m^{-1/2}))^{1/2}\phi \\ (\Psi^{-1}(m^{-1/2}))\log\left(\frac{4}{\eta}\right) \end{array}\right\} \geq 1-\eta$$

and

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

(ii) If $\phi(t)$ and $t/\phi(t)$ are non-decreasing functions. Then under the parameter choice $\lambda \in (0,1]$, $\lambda = \Theta^{-1}(m^{-1/2})$ where $\Theta(t) = t^{\frac{1}{2b}}\phi(t)$, we have

$$Prob_{\mathbf{z}}\left\{||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_\rho \leq C'\phi(\Theta^{-1}(m^{-1/2}))\log\left(\frac{4}{\eta}\right)\right\} \geq 1-\eta$$

and

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

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

$$\lim_{t\to 0}\frac{\Psi(t)}{\sqrt{t}} = \lim_{t\to 0}\frac{t^2}{\Psi^{-1}(t)} = 0.$$

Under the parameter choice $\lambda = \Psi^{-1}(m^{-1/2})$ we have,

$$\lim_{m\to\infty} m\lambda = \infty.$$

Therefore, for sufficiently large m,

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

Under the fact $\lambda \leq 1$ from Theorem 3.2 and Equation (14) follows that with confidence $1-\eta$,

$$||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_\rho \leq C(\Psi^{-1}(m^{-1/2}))^{1/2}\phi(\Psi^{-1}(m^{-1/2}))\log\left(\frac{4}{\eta}\right), \tag{29}$$

where $C = 2R + 4\kappa M + 4\sqrt{\beta b\Sigma^2/(b-1)}$.

Now defining $\tau := C\log\left(\frac{4}{\eta}\right)$ gives

$$\eta = \eta_\tau = 4e^{-\tau/C}.$$

The estimate (29) can be reexpressed as

$$Prob_{\mathbf{z}}\{||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_\rho > \tau(\Psi^{-1}(m^{-1/2}))^{1/2}\phi(\Psi^{-1}(m^{-1/2}))\} \leq \eta_\tau. \tag{30}$$

(ii) Suppose $\Theta(t) = t^{\frac{1}{2b}}\phi(t)$. Then the condition (28) follows that

$$\sqrt{m\lambda} \geq \frac{8\kappa^2\log(4/\eta)}{\sqrt{\lambda}} \geq \frac{8\kappa^2}{\sqrt{\lambda}}.$$

Hence under the parameter choice $\lambda \in (0,1]$, $\lambda = \Theta^{-1}(m^{-1/2})$ we have

$$\frac{1}{m\sqrt{\lambda}} \leq \frac{\sqrt{\lambda}}{8\kappa^2\sqrt{m}} \leq \frac{\lambda^{\frac{1}{2}+\frac{1}{2b}}\phi(\lambda)}{8\kappa^2} \leq \frac{\phi(\lambda)}{8\kappa^2}.$$

From Theorem 3.2 and Equation (14), it follows that with confidence $1-\eta$,

$$||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_\rho \leq C'\phi(\Theta^{-1}(m^{-1/2}))\log\left(\frac{4}{\eta}\right), \tag{31}$$

where $C' := R(\kappa+1) + M/2\kappa + 4\sqrt{\beta b\Sigma^2/(b-1)}$.

Now defining $\tau := C'\log\left(\frac{4}{\eta}\right)$ gives

$$\eta = \eta_\tau = 4e^{-\tau/C'}.$$

The estimate (31) can be reexpressed as

$$Prob_{\mathbf{z}}\left\{||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_\rho > \tau\phi(\Theta^{-1}(m^{-1/2}))\right\} \leq \eta_\tau. \tag{32}$$

Then from Equations (30) and (32) our conclusions follow. □

Theorem 3.4. *Under the same assumptions of Theorem 3.1 and hypothesis (13) with the parameter choice $\lambda \in (0,1]$, $\lambda = \Psi^{-1}(m^{-1/2})$ where $\Psi(t) = t^{\frac{1}{2}+\frac{1}{2b}}\phi(t)$, the convergence of the estimator $f_{\mathbf{z},\lambda}$ (7) to the target function $f_{\mathcal{H}}$ can be described as*

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

and

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

The proof of the theorem follows the same steps as of Theorem 3.3 (i). We obtain the following corollary as a consequence of Theorem 3.3, 3.4.

Corollary 3.1. *Under the same assumptions of Theorem 3.3, 3.4 for Tikhonov regularization with Hölder's source condition $f_{\mathcal{H}} \in \Omega_{\phi,R}$, $\phi(t) = t^r$, for all $0 < \eta < 1$, with confidence $1-\eta$, for the parameter choice $\lambda = m^{-\frac{b}{2br+b+1}}$, we have*

$$||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_{\mathcal{H}} \leq Cm^{-\frac{br}{2br+b+1}}\log\left(\frac{4}{\eta}\right) \text{ for } 0 \leq r \leq 1,$$

$$||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_\rho \leq Cm^{-\frac{2br+b}{4br+2b+2}}\log\left(\frac{4}{\eta}\right) \text{ for } 0 \leq r \leq \frac{1}{2}$$

and for the parameter choice $\lambda = m^{-\frac{b}{2br+1}}$, we have

$$||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_\rho \leq C'm^{-\frac{br}{2br+1}}\log\left(\frac{4}{\eta}\right) \text{ for } 0 \leq r \leq 1.$$

3.2. Upper Rates for General Regularization Schemes

Bauer et al. [4] discussed the error estimates for general regularization schemes under general source condition. Here we study the convergence issues for general regularization schemes under general source condition and the polynomial decay of the eigenvalues of the integral operator L_K. We define the regularization in learning theory framework similar to considered for ill-posed inverse problems (See Section 3.1 [4]).

Definition 3.1. *A family of functions $g_\lambda : [0,\kappa^2] \to \mathbb{R}$, $0 < \lambda \leq \kappa^2$, is said to be the regularization if it satisfies the following conditions:*

- $\exists D : \sup_{\sigma\in(0,\kappa^2]} |\sigma g_\lambda(\sigma)| \leq D.$
- $\exists B : \sup_{\sigma\in(0,\kappa^2]} |g_\lambda(\sigma)| \leq \frac{B}{\lambda}.$
- $\exists \gamma : \sup_{\sigma\in(0,\kappa^2]} |1 - g_\lambda(\sigma)\sigma| \leq \gamma.$
- *The maximal p satisfying the condition:*

$$\sup_{\sigma\in(0,\kappa^2]} |1 - g_\lambda(\sigma)\sigma|\sigma^p \leq \gamma_p\lambda^p$$

is called the qualification of the regularization g_λ, where γ_p does not depend on λ.

The properties of general regularization are satisfied by the large class of learning algorithms which are essentially all the linear regularization schemes. We refer to Section 2.2 [10] for brief discussion of the regularization schemes. Here we consider general regularized solution corresponding to the above regularization:

$$f_{\mathbf{z},\lambda} = g_\lambda(S_{\mathbf{x}}^* S_{\mathbf{x}})S_{\mathbf{x}}^*\mathbf{y}. \tag{33}$$

Here we are discussing the connection between the qualification of the regularization and general source condition [17].

Definition 3.2. *The qualification p covers the index function ϕ if the function $t \to \frac{t^p}{\phi(t)}$ on $t \in (0,\kappa^2]$ is non-decreasing.*

The following result is a restatement of Proposition 3 [17].

Proposition 3.4. *Suppose ϕ is a non-decreasing index function and the qualification of the regularization g_λ covers ϕ. Then*

$$\sup_{\sigma\in(0,\kappa^2]} |1 - g_\lambda(\sigma)\sigma|\phi(\sigma) \leq c_g\phi(\lambda), \;\; c_g = \max(\gamma, \gamma_p).$$

Generally, the index function ϕ is not stable with respect to perturbation in the integral operator L_K. In practice, we are only accessible to the perturbed empirical operator $S_{\mathbf{x}}^* S_{\mathbf{x}}$ but the source condition can be expressed in terms of L_K only. So we want to control the difference $\phi(L_K) - \phi(S_{\mathbf{x}}^* S_{\mathbf{x}})$. In order to obtain the error estimates for general regularization, we further restrict the index functions to operator monotone functions which is defined as

Definition 3.3. *A function $\phi_1 : [0,d] \to [0,\infty)$ is said to be operator monotone index function if $\phi_1(0) = 0$ and for every non-negative pair of self-adjoint operators A, B such that $||A||, ||B|| \leq d$ and $A \leq B$ we have $\phi_1(A) \leq \phi_1(B)$.*

We consider the class of operator monotone index functions:

$$\begin{aligned}\mathcal{F}_\mu = \{&\phi_1 : [0,\kappa^2] \to [0,\infty) \text{ operator monotone,}\\ &\phi_1(0) = 0, \phi_1(\kappa^2) \leq \mu\}.\end{aligned}$$

For the above class of operator monotone functions from Theorem 1 [4], given $\phi_1 \in F_\mu$ there exists c_{ϕ_1} such that

$$||\phi_1(S_{\mathbf{x}}^* S_{\mathbf{x}}) - \phi_1(L_K)||_{\mathcal{L}(\boldsymbol{\mathcal{H}})} \leq c_{\phi_1}\phi_1(||S_{\mathbf{x}}^* S_{\mathbf{x}} - L_K||_{\mathcal{L}(\boldsymbol{\mathcal{H}})}).$$

Here we observe that the rate of convergence of $\phi_1(S_{\mathbf{x}}^* S_{\mathbf{x}})$ to $\phi_1(L_K)$ is slower than the convergence rate of $S_{\mathbf{x}}^* S_{\mathbf{x}}$ to L_K. Therefore, we consider the following class of index functions:

$$\begin{aligned}\mathcal{F} = \{&\phi = \phi_2\phi_1 : \phi_1 \in \mathcal{F}_\mu, \phi_2 : [0,\kappa^2]\\ &\to [0,\infty) \text{ non-decreasing Lipschitz}, \phi_2(0) = 0\}.\end{aligned}$$

The splitting of $\phi = \phi_2\phi_1$ is not unique. So we can take ϕ_2 as a Lipschitz function with Lipschitz constant 1. Now using Corollary 1.2.2 [27] we get

$$||\phi_2(S_{\mathbf{x}}^* S_{\mathbf{x}}) - \phi_2(L_K)||_{\mathcal{L}_2(\boldsymbol{\mathcal{H}})} \leq ||S_{\mathbf{x}}^* S_{\mathbf{x}} - L_K||_{\mathcal{L}_2(\boldsymbol{\mathcal{H}})}.$$

General source condition $f_{\boldsymbol{\mathcal{H}}} \in \Omega_{\phi,R}$ corresponding to index class functions $\mathcal{F}$ covers wide range of source conditions as Hölder's source condition $\phi(t) = t^r$, logarithm source condition $\phi(t) = t^p \log^{-\nu}\left(\frac{1}{t}\right)$. Following the analysis of Bauer et al. [4] we develop the error estimates of general regularization for the index class function $\mathcal{F}$ under the suitable priors on the probability measure ρ.

Theorem 3.5. *Let $\mathbf{z}$ be i.i.d. samples drawn according to the probability measure $\rho \in \mathcal{P}_\phi$. Suppose $f_{\mathbf{z},\lambda}$ is the regularized solution (33) corresponding to general regularization and the qualification of the regularization covers ϕ. Then for all $0 < \eta < 1$, with confidence $1 - \eta$, the following upper bound holds:*

$$||f_{\mathbf{z},\lambda} - f_{\boldsymbol{\mathcal{H}}}||_{\boldsymbol{\mathcal{H}}} \leq \left\{ \begin{array}{l} Rc_g(1 + c_{\phi_1})\phi(\lambda) + \frac{4R\mu\gamma\kappa^2}{\sqrt{m}} + \frac{2\sqrt{2}\nu_1\kappa M}{m\lambda} \\ \quad + \sqrt{\frac{8\nu_1^2\Sigma^2\mathcal{N}(\lambda)}{m\lambda}} \end{array} \right\} \log\left(\frac{4}{\eta}\right)$$

provided that

$$\sqrt{m}\lambda \geq 8\kappa^2 \log(4/\eta). \tag{34}$$

Proof. We consider the error expression for general regularized solution (33),

$$f_{\mathbf{z},\lambda} - f_{\boldsymbol{\mathcal{H}}} = g_\lambda(S_{\mathbf{x}}^* S_{\mathbf{x}})(S_{\mathbf{x}}^*\mathbf{y} - S_{\mathbf{x}}^* S_{\mathbf{x}} f_{\boldsymbol{\mathcal{H}}}) - r_\lambda(S_{\mathbf{x}}^* S_{\mathbf{x}})f_{\boldsymbol{\mathcal{H}}}, \tag{35}$$

where $r_\lambda(\sigma) = 1 - g_\lambda(\sigma)\sigma$.

Now the first term can be expressed as

$$\begin{aligned} g_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*\mathbf{y} - S_{\mathbf{x}}^*S_{\mathbf{x}}f_{\mathcal{H}}) &= g_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda I)^{1/2} \\ &\quad (S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda I)^{-1/2}(L_K+\lambda I)^{1/2} \\ &\quad (L_K+\lambda I)^{-1/2}(S_{\mathbf{x}}^*\mathbf{y} - S_{\mathbf{x}}^*S_{\mathbf{x}}f_{\mathcal{H}}). \end{aligned}$$

On applying RKHS-norm we get,

$$\begin{aligned} ||g_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*\mathbf{y} - S_{\mathbf{x}}^*S_{\mathbf{x}}f_{\mathcal{H}})||_{\mathcal{H}} &\leq I_2 I_5 ||g_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}}) \\ &\quad (S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda I)^{1/2}||_{\mathcal{L}(\mathcal{H})}, \quad (36) \end{aligned}$$

where $I_2 = ||(L_K+\lambda I)^{-1/2}(S_{\mathbf{x}}^*\mathbf{y} - S_{\mathbf{x}}^*S_{\mathbf{x}}f_{\mathcal{H}})||_{\mathcal{H}}$ and $I_5 = ||(S_{\mathbf{x}}^*S_{\mathbf{x}} + \lambda I)^{-1/2}(L_K+\lambda I)^{1/2}||_{\mathcal{L}(\mathcal{H})}$.

The estimate of I_2 can be obtained from the first estimate of Proposition 3.2 and from the second estimate of Proposition 3.2 with the condition (34) we obtain with probability $1-\eta/2$,

$$\begin{aligned} &||(L_K+\lambda I)^{-1/2}(L_K - S_{\mathbf{x}}^*S_{\mathbf{x}})(L_K+\lambda I)^{-1/2}||_{\mathcal{L}(\mathcal{H})} \\ &\leq \frac{1}{\lambda}||S_{\mathbf{x}}^*S_{\mathbf{x}} - L_K||_{\mathcal{L}(\mathcal{H})} \leq \frac{4\kappa^2}{\sqrt{m}\lambda}\log\left(\frac{4}{\eta}\right) \leq \frac{1}{2}. \end{aligned}$$

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

$$\begin{aligned} I_5 &= ||(S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda I)^{-1/2}(L_K+\lambda I)^{1/2}||_{\mathcal{L}(\mathcal{H})} \\ &= ||(L_K+\lambda I)^{1/2}(S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda I)^{-1}(L_K+\lambda I)^{1/2}||_{\mathcal{L}(\mathcal{H})}^{1/2} \\ &= ||\{I-(L_K+\lambda I)^{-1/2}(L_K - S_{\mathbf{x}}^*S_{\mathbf{x}}) \\ &\quad (L_K+\lambda I)^{-1/2}\}^{-1}||_{\mathcal{L}(\mathcal{H})}^{1/2} \\ &\leq \sqrt{2}. \quad (37) \end{aligned}$$

From the properties of the regularization we have,

$$\begin{aligned} ||g_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*S_{\mathbf{x}})^{1/2}||_{\mathcal{L}(\mathcal{H})} &\leq \sup_{0<\sigma\leq\kappa^2}|g_\lambda(\sigma)\sqrt{\sigma}| \\ = \left(\sup_{0<\sigma\leq\kappa^2}|g_\lambda(\sigma)\sigma| \sup_{0<\sigma\leq\kappa^2}|g_\lambda(\sigma)|\right)^{1/2} &\leq \sqrt{\frac{BD}{\lambda}}. \quad (38) \end{aligned}$$

Hence it follows,

$$\begin{aligned} ||g_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda I)^{1/2}||_{\mathcal{L}(\mathcal{H})} &\leq \sup_{0<\sigma\leq\kappa^2}|g_\lambda(\sigma)(\sigma+\lambda)^{1/2}| \\ \leq \sup_{0<\sigma\leq\kappa^2}|g_\lambda(\sigma)\sqrt{\sigma}| + \sqrt{\lambda}\sup_{0<\sigma\leq\kappa^2}|g_\lambda(\sigma)| &\leq \frac{\nu_1}{\sqrt{\lambda}}, \quad (39) \end{aligned}$$

where $\nu_1 := B + \sqrt{BD}$.

Therefore, using (16), (37) and (39) in Equation (36) we conclude that with probability $1-\eta$,

$$\begin{aligned} ||g_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*\mathbf{y} - S_{\mathbf{x}}^*S_{\mathbf{x}}f_{\mathcal{H}})||_{\mathcal{H}} &\leq 2\sqrt{2}\nu_1\left\{\frac{\kappa M}{m\lambda} + \sqrt{\frac{\Sigma^2\mathcal{N}(\lambda)}{m\lambda}}\right\} \\ &\quad \log\left(\frac{4}{\eta}\right). \quad (40) \end{aligned}$$

Now we consider the second term,

$$\begin{aligned} r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})f_{\mathcal{H}} &= r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})\phi(L_K)v = r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})\phi(S_{\mathbf{x}}^*S_{\mathbf{x}})v \\ &\quad + r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})\phi_2(S_{\mathbf{x}}^*S_{\mathbf{x}})(\phi_1(L_K) - \phi_1(S_{\mathbf{x}}^*S_{\mathbf{x}}))v \\ &\quad + r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(\phi_2(L_K) - \phi_2(S_{\mathbf{x}}^*S_{\mathbf{x}}))\phi_1(L_K)v. \end{aligned}$$

Employing RKHS-norm we get

$$\begin{aligned} ||r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})f_{\mathcal{H}}||_{\mathcal{H}} &\leq Rc_g\phi(\lambda) + Rc_g c_{\phi_1}\phi_2(\lambda)\phi_1 \\ &(||L_K - S_{\mathbf{x}}^*S_{\mathbf{x}}||_{\mathcal{L}(\mathcal{H})}) + R\mu\gamma||L_K - S_{\mathbf{x}}^*S_{\mathbf{x}}||_{\mathcal{L}_2(\mathcal{H})}. \end{aligned}$$

Here we used the fact that if the qualification of the regularization covers $\phi = \phi_1\phi_2$, then the qualification also covers ϕ_1 and ϕ_2 both separately.

From Equations (17) and (34) we have with probability $1-\eta/2$,

$$||S_{\mathbf{x}}^*S_{\mathbf{x}} - L_K||_{\mathcal{L}(\mathcal{H})} \leq \frac{4\kappa^2}{\sqrt{m}}\log\left(\frac{4}{\eta}\right) \leq \lambda/2. \quad (41)$$

Therefore, with probability $1-\eta/2$,

$$||r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})f_{\mathcal{H}}||_{\mathcal{H}} \leq Rc_g(1+c_{\phi_1})\phi(\lambda) + \frac{4R\mu\gamma\kappa^2}{\sqrt{m}}\log\left(\frac{4}{\eta}\right). \quad (42)$$

Combining the bounds (40) and (42) we get the desired result. □

Theorem 3.6. *Let* $\mathbf{z}$ *be i.i.d. samples drawn according to the probability measure* $\rho \in \mathcal{P}_\phi$ *and* $f_{\mathbf{z},\lambda}$ *is the regularized solution (33) corresponding to general regularization. Then for all* $0 < \eta < 1$, *with confidence* $1-\eta$, *the following upper bounds holds:*

(i) If the qualification of the regularization covers ϕ,

$$||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_\rho \leq \left\{\begin{array}{l} Rc_g(1+c_{\phi_1})(\kappa+\sqrt{\lambda})\phi(\lambda) \\ +\frac{4R\mu\gamma\kappa^2(\kappa+\sqrt{\lambda})}{\sqrt{m}} + \frac{2\sqrt{2}\nu_2\kappa M}{m\sqrt{\lambda}} \\ +\sqrt{\frac{8\nu_2^2\Sigma^2\mathcal{N}(\lambda)}{m}} \end{array}\right\}\log\left(\frac{4}{\eta}\right),$$

(ii) If the qualification of the regularization covers $\phi(t)\sqrt{t}$,

$$||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_\rho \leq \left\{\begin{array}{l} 2Rc_g(1+c_{\phi_1})\phi(\lambda)\sqrt{\lambda} + \frac{4R\mu(\gamma+c_g)\kappa^2\sqrt{\lambda}}{\sqrt{m}} \\ +\frac{2\sqrt{2}\nu_2\kappa M}{m\sqrt{\lambda}} + \sqrt{\frac{8\nu_2^2\Sigma^2\mathcal{N}(\lambda)}{m}} \end{array}\right\} \log\left(\frac{4}{\eta}\right)$$

provided that

$$\sqrt{m}\lambda \geq 8\kappa^2\log(4/\eta). \quad (43)$$

Proof. Here we establish $\mathscr{L}^2$-norm estimate for the error expression:

$$f_{\mathbf{z},\lambda} - f_{\mathcal{H}} = g_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*\mathbf{y} - S_{\mathbf{x}}^*S_{\mathbf{x}}f_{\mathcal{H}}) - r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})f_{\mathcal{H}}.$$

On applying $\mathscr{L}^2$-norm in the first term we get,

$$||g_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*\mathbf{y}-S_{\mathbf{x}}^*S_{\mathbf{x}}f_{\mathcal{H}})||_\rho \le I_2I_5||L_K^{1/2}g_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda I)^{1/2}||_{\mathcal{L}(\mathcal{H})}, \quad (44)$$

where $I_2 = ||(L_K+\lambda I)^{-1/2}(S_{\mathbf{x}}^*\mathbf{y}-S_{\mathbf{x}}^*S_{\mathbf{x}}f_{\mathcal{H}})||_{\mathcal{H}}$ and $I_5 = ||(S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda I)^{-1/2}(L_K+\lambda I)^{1/2}||_{\mathcal{L}(\mathcal{H})}$.

The estimates of I_2 and I_5 can be obtained from Proposition 3.2 and Theorem 3.5 respectively. Now we consider

$$\begin{aligned}||L_K^{1/2}g_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda I)^{1/2}||_{\mathcal{L}(\mathcal{H})} \le\ & ||L_K^{1/2}\\ & -(S_{\mathbf{x}}^*S_{\mathbf{x}})^{1/2}||_{\mathcal{L}(\mathcal{H})}||g_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda I)^{1/2}\\ & ||_{\mathcal{L}(\mathcal{H})}+||(S_{\mathbf{x}}^*S_{\mathbf{x}})^{1/2}g_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})\\ & (S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda I)^{1/2}||_{\mathcal{L}(\mathcal{H})}.\end{aligned}$$

Since $\phi(t)=\sqrt{t}$ is operator monotone function. Therefore, from Equation (41) with probability $1-\eta/2$, we get

$$||L_K^{1/2}-(S_{\mathbf{x}}^*S_{\mathbf{x}})^{1/2}||_{\mathcal{L}(\mathcal{H})} \le (||L_K-S_{\mathbf{x}}^*S_{\mathbf{x}}||_{\mathcal{L}(\mathcal{H})})^{1/2} \le \sqrt{\lambda}.$$

Then using the properties of the regularization and Equation (38) we conclude that with probability $1-\eta/2$,

$$\begin{aligned}&||\ L_K^{1/2}g_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*S_{\mathbf{x}}+\lambda I)^{1/2}||_{\mathcal{L}(\mathcal{H})}\\ &\le \sqrt{\lambda}\sup_{0<\sigma\le\kappa^2}|g_\lambda(\sigma)(\sigma+\lambda)^{1/2}|+\sup_{0<\sigma\le\kappa^2}|g_\lambda(\sigma)(\sigma^2+\lambda\sigma)^{1/2}|\\ &\le \sup_{0<\sigma\le\kappa^2}|g_\lambda(\sigma)\sigma|+\lambda\sup_{0<\sigma\le\kappa^2}|g_\lambda(\sigma)|+2\sqrt{\lambda}\sup_{0<\sigma\le\kappa^2}|g_\lambda(\sigma)\sqrt{\sigma}|\\ &\le B+D+2\sqrt{BD}=\nu_2(\text{let}). \quad (45)\end{aligned}$$

From Equations (44) with Equations (16), (37), and (45) we obtain with probability $1-\eta$,

$$||g_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*\mathbf{y}-S_{\mathbf{x}}^*S_{\mathbf{x}}f_{\mathcal{H}})||_\rho \le 2\sqrt{2}\nu_2\left\{\frac{\kappa M}{m\sqrt{\lambda}}+\sqrt{\frac{\Sigma^2\mathcal{N}(\lambda)}{m}}\right\}\log\left(\frac{4}{\eta}\right). \quad (46)$$

The second term can be expressed as

$$\begin{aligned}&||\ r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})f_{\mathcal{H}}||_\rho \le ||L_K^{1/2}-(S_{\mathbf{x}}^*S_{\mathbf{x}})^{1/2}||_{\mathcal{L}(\mathcal{H})}||r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})f_{\mathcal{H}}||_{\mathcal{H}}\\ &\quad +||(S_{\mathbf{x}}^*S_{\mathbf{x}})^{1/2}r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})f_{\mathcal{H}}||_{\mathcal{H}}\\ &\le ||L_K-S_{\mathbf{x}}^*S_{\mathbf{x}}||_{\mathcal{L}(\mathcal{H})}^{1/2}||r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})f_{\mathcal{H}}||_{\mathcal{H}}\\ &\quad +||r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*S_{\mathbf{x}})^{1/2}\phi(S_{\mathbf{x}}^*S_{\mathbf{x}})\nu||_{\mathcal{H}}\\ &\quad +||r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*S_{\mathbf{x}})^{1/2}\phi_2(S_{\mathbf{x}}^*S_{\mathbf{x}})(\phi_1(S_{\mathbf{x}}^*S_{\mathbf{x}})-\phi_1(L_K))\nu||_{\mathcal{H}}\\ &\quad +||r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})(S_{\mathbf{x}}^*S_{\mathbf{x}})^{1/2}(\phi_2(S_{\mathbf{x}}^*S_{\mathbf{x}})-\phi_2(L_K))\phi_1(L_K)\nu||_{\mathcal{H}}.\end{aligned}$$

Here two cases arises:

Case 1. If the qualification of the regularization covers ϕ. Then we get with confidence $1-\eta/2$,

$$||r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})f_{\mathcal{H}}||_\rho \le (\kappa+\sqrt{\lambda})\left(Rc_g(1+c_{\phi_1})\phi(\lambda)+R\mu\gamma||S_{\mathbf{x}}^*S_{\mathbf{x}}-L_K||_{\mathcal{L}_2(\mathcal{H})}\right).$$

Therefore, using Equation (17) we obtain with probability $1-\eta/2$,

$$||r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})f_{\mathcal{H}}||_\rho \le (\kappa+\sqrt{\lambda})\left(Rc_g(1+c_{\phi_1})\phi(\lambda)+\frac{4R\mu\gamma\kappa^2}{\sqrt{m}}\log\left(\frac{4}{\eta}\right)\right). \quad (47)$$

Case 2. If the qualification of the regularization covers $\phi(t)\sqrt{t}$, we get with probability $1-\eta/2$,

$$\begin{aligned}||r_\lambda(S_{\mathbf{x}}^*S_{\mathbf{x}})f_{\mathcal{H}}||_\rho \ \le\ & 2Rc_g(1+c_{\phi_1})\phi(\lambda)\sqrt{\lambda}\\ & +4R\mu(\gamma+c_g)\kappa^2\sqrt{\frac{\lambda}{m}}\log\left(\frac{4}{\eta}\right). \quad (48)\end{aligned}$$

Combining the error estimates (46), (47) and (48) we get the desired results. □

We discuss the convergence rates of general regularizer based on data-driven strategy of the parameter choice of λ for the class of probability measure $\mathcal{P}_{\phi,b}$. The proof of Theorem 3.7, 3.8 are similar to Theorem 3.3.

Theorem 3.7. *Under the same assumptions of Theorem 3.5 and hypothesis (13) with the parameter choice* $\lambda \in (0,1]$, $\lambda = \Psi^{-1}(m^{-1/2})$ *where* $\Psi(t) = t^{\frac{1}{2}+\frac{1}{2b}}\phi(t)$, *the convergence of the estimator* $f_{\mathbf{z},\lambda}$ *(33) to the target function* $f_{\mathcal{H}}$ *can be described as*

$$Prob_{\mathbf{z}}\left\{||f_{\mathbf{z},\lambda}-f_{\mathcal{H}}||_{\mathcal{H}} \le \widetilde{C}\phi(\Psi^{-1}(m^{-1/2}))\log\left(\frac{4}{\eta}\right)\right\} \ge 1-\eta,$$

where $\widetilde{C} = Rc_g(1+c_{\phi_1})+4R\mu\gamma\kappa^2+2\sqrt{2}\nu_1\kappa M+\sqrt{8\beta b\nu_1^2\Sigma^2/(b-1)}$ *and*

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

Theorem 3.8. *Under the same assumptions of Theorem 3.6 and hypothesis (13), the convergence of the estimator* $f_{\mathbf{z},\lambda}$ *(33) to the target function* $f_{\mathcal{H}}$ *can be described as*

(i) If the qualification of the regularization covers ϕ. *Then under the parameter choice* $\lambda \in (0,1]$, $\lambda = \Theta^{-1}(m^{-1/2})$ *where* $\Theta(t) = t^{\frac{1}{2b}}\phi(t)$, *we have*

$$Prob_{\mathbf{z}}\left\{||f_{\mathbf{z},\lambda}-f_{\mathcal{H}}||_\rho \le \widetilde{C}_1\phi(\Theta^{-1}(m^{-1/2}))\log\left(\frac{4}{\eta}\right)\right\} \ge 1-\eta,$$

where $\widetilde{C}_1 = Rc_g(1+c_{\phi_1})(\kappa+1)+4R\mu\gamma\kappa^2(\kappa+1)+\nu_2M/2\sqrt{2}\kappa+\sqrt{8\beta b\nu_2^2\Sigma^2/(b-1)}$ *and*

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

(ii) If the qualification of the regularization covers $\phi(t)\sqrt{t}$. Then under the parameter choice $\lambda \in (0,1]$, $\lambda = \Psi^{-1}(m^{-1/2})$ where $\Psi(t) = t^{\frac{1}{2}+\frac{1}{2b}}\phi(t)$, we have

$$Prob_{\mathbf{z}}\left\{||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_\rho \leq \widetilde{C}_2(\Psi^{-1}(m^{-1/2}))^{1/2}\phi(\Psi^{-1}(m^{-1/2})) \log\left(\frac{4}{\eta}\right)\right\} \geq 1-\eta,$$

where $\widetilde{C}_2 = 2Rc_g(1+c_{\phi_1}) + 4R\mu(\gamma+c_g)\kappa^2 + 2\sqrt{2}v_2\kappa M + \sqrt{8\beta b v_2^2\Sigma^2/(b-1)}$ *and*

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

We obtain the following corollary as a consequence of Theorem 3.7, 3.8.

Corollary 3.2. *Under the same assumptions of Theorem 3.7, 3.8 for general regularization of the qualification p with Hölder's source condition $f_{\mathcal{H}} \in \Omega_{\phi,R}$, $\phi(t) = t^r$, for all $0 < \eta < 1$, with confidence $1-\eta$, for the parameter choice $\lambda = m^{-\frac{b}{2br+b+1}}$, we have*

$$||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_{\mathcal{H}} \leq \widetilde{C}m^{-\frac{br}{2br+b+1}}\log\left(\frac{4}{\eta}\right) \text{ for } 0 \leq r \leq p,$$

$$||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_\rho \leq \widetilde{C}_2 m^{-\frac{2br+b}{4br+2b+2}}\log\left(\frac{4}{\eta}\right) \text{ for } 0 \leq r \leq p-\frac{1}{2}$$

and for the parameter choice $\lambda = m^{-\frac{b}{2br+1}}$, we have

$$||f_{\mathbf{z},\lambda} - f_{\mathcal{H}}||_\rho \leq \widetilde{C}_1 m^{-\frac{br}{2br+1}}\log\left(\frac{4}{\eta}\right) \text{ for } 0 \leq r \leq p.$$

Remark 3.1. *It is important to observe from Corollary 3.1, 3.2 that using the concept of operator monotonicity of index function we are able to achieve the same error estimates for general regularization as of Tikhonov regularization up to a constant multiple.*

Remark 3.2. *(Related work) Corollary 3.1 provides the order of convergence same as of Theorem 1 [12] for Tikhonov regularization under the Hölder's source condition $f_{\mathcal{H}} \in \Omega_{\phi,R}$ for $\phi(t) = t^r$ $\left(\frac{1}{2} \leq r \leq 1\right)$ and the polynomial decay of the eigenvalues (13). Blanchard and Mücke [18] addressed the convergence rates for inverse statistical learning problem for general regularization under the Hölder's source condition with the assumption $f_\rho \in \mathcal{H}$. In particular, the upper convergence rates discussed in Blanchard and Mücke [18] agree with Corollary 3.2 for considered learning problem which is referred as direct learning problem in Blanchard and Mücke[18]. Under the fact $\mathcal{N}(\lambda) \leq \frac{\kappa^2}{\lambda}$ from Theorem 3.5, 3.6 we obtain the similar estimates as of Theorem 10 [4] for general regularization schemes without the polynomial decay condition of the eigenvalues (13).*

Remark 3.3. *For the real valued functions and multi-task algorithms (the output space $Y \subset \mathbb{R}^m$ for some $m \in \mathbb{N}$) we can obtain the error estimates from our analysis without imposing any condition on the conditional probability measure (11) for the bounded output space Y.*

Remark 3.4. *We can address the convergence issues of binary classification problem [28] using our error estimates as similar to discussed in Section 3.3 [4] and Section 5 [16].*

3.3. Lower Rates for General Learning Algorithms

In this section, we discuss the estimates of minimum possible error over a subclass of the probability measures $\mathcal{P}_{\phi,b}$ parameterized by suitable functions $f \in \mathcal{H}$. Throughout this section we assume that Y is finite-dimensional.

Let $\{v_j\}_{j=1}^d$ be a basis of Y and $f \in \Omega_{\phi,R}$. Then we parameterize the probability measure based on the function f,

$$\rho_f(x,y) := \frac{1}{2dL}\sum_{j=1}^{d}\left(a_j(x)\delta_{y+dLv_j} + b_j(x)\delta_{y-dLv_j}\right)\nu(x), \quad (49)$$

where $a_j(x) = L - \langle f, K_x v_j\rangle_{\mathcal{H}}$, $b_j(x) = L + \langle f, K_x v_j\rangle_{\mathcal{H}}$, $L = 4\kappa\phi(\kappa^2)R$ and δ_ξ denotes the Dirac measure with unit mass at ξ. It is easy to observe that the marginal distribution of ρ_f over X is ν and the regression function for the probability measure ρ_f is f (see Proposition 4 [12]). In addition to this, for the conditional probability measure $\rho_f(y|x)$ we have,

$$\int_Y\left(e^{||y-f(x)||_Y/M} - \frac{||y-f(x)||_Y}{M} - 1\right)d\rho_f(y|x) \leq \left(d^2L^2 - ||f(x)||_Y^2\right)\sum_{i=2}^{\infty}\frac{(dL+||f(x)||_Y)^{i-2}}{M^i i!} \leq \frac{\Sigma^2}{2M^2}$$

provided that

$$dL + L/4 \leq M \text{ and } \sqrt{2}dL \leq \Sigma.$$

We assume that the eigenvalues of the integral operator L_K follow the polynomial decay (13) for the marginal probability measure ν. Then we conclude that the probability measure ρ_f parameterized by f belongs to the class $\mathcal{P}_{\phi,b}$.

The concept of information theory such as the Kullback-Leibler information and Fano inequalities (Lemma 3.3 [29]) are the main ingredients in the analysis of lower bounds. In the literature [12, 29], the closeness of probability measures is described through Kullback-Leibler information: Given two probability measures ρ_1 and ρ_2, it is defined as

$$\mathcal{K}(\rho_1,\rho_2) := \int_Z \log(g(z))d\rho_1(z),$$

where g is the density of ρ_1 with respect to ρ_2, that is, $\rho_1(E) = \int_E g(z)d\rho_2(z)$ for all measurable sets E.

Following the analysis of Caponnetto and De Vito [12] and DeVore et al. [29] we establish the lower rates of accuracy that can be attained by any learning algorithm.

To estimate the lower rates of learning algorithms, we generate N_ε-functions belonging to $\Omega_{\phi,R}$ for given $\varepsilon > 0$ such that (53), (54) holds. Then we construct the probability measures $\rho_i \in \mathcal{P}_{\phi,b}$ from Equation (49), parameterized by these functions f_i's ($1 \leq i \leq N_\varepsilon$). On applying Lemma 3.3 [29], we obtain the lower convergence rates using Kullback-Leibler information.

Theorem 3.9. *Let* $\mathbf{z}$ *be i.i.d. samples drawn according to the probability measure* $\rho \in \mathcal{P}_{\phi,b}$ *under the hypothesis* $dim(Y) = d < \infty$. *Then for any learning algorithm* ($\mathbf{z} \to f_\mathbf{z} \in \boldsymbol{\mathcal{H}}$) *there exists a probability measure* $\rho_* \in \mathcal{P}_{\phi,b}$ *and* $f_{\rho_*} \in \boldsymbol{\mathcal{H}}$ *such that for all* $0 < \varepsilon < \varepsilon_o$, $f_\mathbf{z}$ *can be approximated as*

$$Prob_\mathbf{z}\left\{||f_\mathbf{z} - f_{\rho_*}||_{\boldsymbol{\mathcal{H}}} > \varepsilon/2\right\} \geq \min\left\{\frac{1}{1+e^{-\ell_\varepsilon/24}}, \vartheta e^{\left(\frac{\ell_\varepsilon}{48} - \frac{cm\varepsilon^2}{\ell_\varepsilon^b}\right)}\right\}$$

where $\vartheta = e^{-3/e}$, $c = \frac{64\beta}{15(b-1)dL^2}\left(1 - \frac{1}{2^{b-1}}\right)$ *and* $\ell_\varepsilon = \left\lfloor \frac{1}{2}\left(\frac{\alpha}{\phi^{-1}(\varepsilon/R)}\right)^{1/b}\right\rfloor$.

Proof. For given $\varepsilon > 0$, we define

$$g = \sum_{n=\ell+1}^{2\ell} \frac{\varepsilon \sigma^{n-\ell} e_n}{\sqrt{\ell}\phi(t_n)},$$

where $\sigma = (\sigma^1, \ldots, \sigma^\ell) \in \{-1,+1\}^\ell$, t_n's are the eigenvalues of the integral operator L_K, e_n's are the eigenvectors of the integral operator L_K and the orthonormal basis of RKHS $\boldsymbol{\mathcal{H}}$. Under the decay condition on the eigenvalues $\alpha \leq n^b t_n$, we get

$$||g||^2_{\boldsymbol{\mathcal{H}}} = \sum_{n=\ell+1}^{2\ell} \frac{\varepsilon^2}{\ell\phi^2(t_n)} \leq \sum_{n=\ell+1}^{2\ell} \frac{\varepsilon^2}{\ell\phi^2\left(\frac{\alpha}{n^b}\right)} \leq \frac{\varepsilon^2}{\phi^2\left(\frac{\alpha}{2^b\ell^b}\right)}.$$

Hence $f = \phi(L_K)g \in \Omega_{\phi,R}$ provided that $||g||_{\boldsymbol{\mathcal{H}}} \leq R$ or equivalently,

$$\ell \leq \frac{1}{2}\left(\frac{\alpha}{\phi^{-1}(\varepsilon/R)}\right)^{1/b}. \tag{50}$$

For $\ell = \ell_\varepsilon = \left\lfloor \frac{1}{2}\left(\frac{\alpha}{\phi^{-1}(\varepsilon/R)}\right)^{1/b}\right\rfloor$, choose ε_o such that $\ell_{\varepsilon_o} > 16$. Then according to Proposition 6 [12], for every positive $\varepsilon < \varepsilon_o$ ($\ell_\varepsilon > \ell_{\varepsilon_o}$) there exists $N_\varepsilon \in \mathbb{N}$ and $\sigma_1, \ldots, \sigma_{N_\varepsilon} \in \{-1,+1\}^{\ell_\varepsilon}$ such that

$$\sum_{n=1}^{\ell_\varepsilon} (\sigma_i^n - \sigma_j^n)^2 \geq \ell_\varepsilon, \text{ for all } 1 \leq i,j \leq N_\varepsilon,\ i \neq j \tag{51}$$

and

$$N_\varepsilon \geq e^{\ell_\varepsilon/24}. \tag{52}$$

Now we suppose $f_i = \phi(L_K)g_i$ and for $\varepsilon > 0$,

$$g_i = \sum_{n=\ell_\varepsilon+1}^{2\ell_\varepsilon} \frac{\varepsilon \sigma_i^{n-\ell_\varepsilon} e_n}{\sqrt{\ell_\varepsilon}\phi(t_n)}, \text{ for } i = 1, \ldots, N_\varepsilon,$$

where $\sigma_i = (\sigma_i^1, \ldots, \sigma_i^{\ell_\varepsilon}) \in \{-1,+1\}^{\ell_\varepsilon}$. Then from Equation (51) we get,

$$\varepsilon \leq ||f_i - f_j||_{\boldsymbol{\mathcal{H}}}, \text{ for all } 1 \leq i,j \leq N_\varepsilon,\ i \neq j. \tag{53}$$

For $1 \leq i,j \leq N_\varepsilon$, we have

$$||f_i - f_j||^2_{\mathscr{L}_\nu(X)} \leq \sum_{n=\ell_\varepsilon+1}^{2\ell_\varepsilon} \frac{\beta\varepsilon^2\left(\sigma_i^{n-\ell_\varepsilon} - \sigma_j^{n-\ell_\varepsilon}\right)^2}{\ell_\varepsilon n^b} \leq \sum_{n=\ell_\varepsilon+1}^{2\ell_\varepsilon} \frac{4\beta\varepsilon^2}{\ell_\varepsilon n^b}$$
$$\leq \frac{4\beta\varepsilon^2}{\ell_\varepsilon}\int_{\ell_\varepsilon}^{2\ell_\varepsilon} \frac{1}{x^b}dx = c'\frac{\varepsilon^2}{\ell_\varepsilon^b}, \tag{54}$$

where $c' = \frac{4\beta}{(b-1)}\left(1 - \frac{1}{2^{b-1}}\right)$.

We define the sets,

$$A_i = \left\{\mathbf{z} : ||f_\mathbf{z} - f_i||_{\boldsymbol{\mathcal{H}}} < \frac{\varepsilon}{2}\right\}, \text{ for } 1 \leq i \leq N_\varepsilon.$$

It is clear from Equation (53) that A_i's are disjoint sets. On applying Lemma 3.3 [29] with probability measures $\rho_{f_i}^m$, $1 \leq i \leq N_\varepsilon$, we obtain that either

$$p := \max_{1\leq i\leq N_\varepsilon} \rho_{f_i}^m(A_i^c) \geq \frac{N_\varepsilon}{N_\varepsilon + 1} \tag{55}$$

or

$$\min_{1\leq j\leq N_\varepsilon} \frac{1}{N_\varepsilon}\sum_{i=1,i\neq j}^{N_\varepsilon} \mathcal{K}(\rho_{f_i}^m, \rho_{f_j}^m) \geq \Psi_{N_\varepsilon}(p), \tag{56}$$

where $\Psi_{N_\varepsilon}(p) = \log(N_\varepsilon) + (1-p)\log\left(\frac{1-p}{p}\right) - p\log\left(\frac{N_\varepsilon - p}{p}\right)$. Further,

$$\Psi_{N_\varepsilon}(p) \geq (1-p)\log(N_\varepsilon) + (1-p)\log(1-p) - \log(p)$$
$$+2p\log(p) \geq -\log(p) + \log(\sqrt{N_\varepsilon}) - 3/e. \tag{57}$$

Since minimum value of $x\log(x)$ is $-1/e$ on $[0,1]$.

For the joint probability measures $\rho_{f_i}^m$, $\rho_{f_j}^m$ ($\rho_{f_i}, \rho_{f_j} \in \mathcal{P}_{\phi,b}$, $1 \leq i,j \leq N_\varepsilon$) from Proposition 4 [12] and the Equation (54) we get,

$$\mathcal{K}(\rho_{f_i}^m, \rho_{f_j}^m) = m\mathcal{K}(\rho_{f_i}, \rho_{f_j}) \leq \frac{16m}{15dL^2}||f_i - f_j||^2_{\mathscr{L}_\nu(X)} \leq \frac{cm\varepsilon^2}{\ell_\varepsilon^b}, \tag{58}$$

where $c = 16c'/15dL^2$.

Therefore, Equations (55), (56), together with Equations (57) and (58) implies

$$p := \max_{1\leq i\leq N_\varepsilon}\left(Prob\left\{\mathbf{z} : ||f_\mathbf{z} - f_i||_{\boldsymbol{\mathcal{H}}} > \frac{\varepsilon}{2}\right\}\right)$$
$$\geq \min\left\{\frac{N_\varepsilon}{N_\varepsilon+1}, \sqrt{N_\varepsilon}e^{-\frac{3}{e} - \frac{cm\varepsilon^2}{\ell_\varepsilon^b}}\right\}.$$

From Equation (52) for the probability measure ρ_* such that $p = \rho_*^m(A_i^c)$ follows the result. □

The lower estimates in $\mathscr{L}^2$-norm can be obtained similar to above theorem.

Theorem 3.10. *Let* $\mathbf{z}$ *be i.i.d. samples drawn according to the probability measure* $\rho \in \mathcal{P}_{\phi,b}$ *under the hypothesis* $dim(Y) = d < \infty$. *Then for any learning algorithm* ($\mathbf{z} \to f_{\mathbf{z}} \in \mathcal{H}$) *there exists a probability measure* $\rho_* \in \mathcal{P}_{\phi,b}$ *and* $f_{\rho_*} \in \mathcal{H}$ *such that for all* $0 < \varepsilon < \varepsilon_o$, $f_{\mathbf{z}}$ *can be approximated as*

$$Prob_{\mathbf{z}}\left\{||f_{\mathbf{z}} - f_{\rho_*}||_{\mathscr{L}^2_\nu(X)} > \varepsilon/2\right\} \geq \min\left\{\frac{1}{1+e^{-\ell_\varepsilon/24}}, \vartheta e^{\left(\frac{\ell_\varepsilon}{48} - \frac{64m\varepsilon^2}{15dL^2}\right)}\right\}$$

where $\vartheta = e^{-3/e}$, $\ell_\varepsilon = \left\lfloor \left(\frac{\alpha}{\psi^{-1}(\varepsilon/R)}\right)^{1/b} \right\rfloor$ *and* $\psi(t) = \sqrt{t}\phi(t)$.

Theorem 3.11. *Under the same assumptions of Theorem 3.10 for* $\psi(t) = t^{1/2}\phi(t)$ *and* $\Psi(t) = t^{\frac{1}{2}+\frac{1}{2b}}\phi(t)$, *the estimator* $f_{\mathbf{z}}$ *corresponding to any learning algorithm converges to the regression function* f_ρ *with the following lower rate:*

$$\lim_{\tau\to 0}\liminf_{m\to\infty}\inf_{l\in\mathcal{A}}\sup_{\rho\in\mathcal{P}_{\phi,b}} Prob_{\mathbf{z}}\left\{||f^l_{\mathbf{z}} - f_\rho||_{\mathscr{L}^2_\nu(X)} > \tau\psi\left(\Psi^{-1}(m^{-1/2})\right)\right\} = 1,$$

where $\mathcal{A}$ *denotes the set of all learning algorithms* $l : \mathbf{z} \to f^l_{\mathbf{z}}$.

Proof. Under the condition $\ell_\varepsilon = \left\lfloor \left(\frac{\alpha}{\psi^{-1}(\varepsilon/R)}\right)^{1/b} \right\rfloor$ from Theorem 3.10 we get,

$$Prob_{\mathbf{z}}\left\{||f_{\mathbf{z}} - f_{\rho_*}||_{\mathscr{L}^2_\nu(X)} > \tfrac{\varepsilon}{2}\right\} \geq \min\left\{\frac{1}{1+e^{-\ell_\varepsilon/24}}, \vartheta e^{-\frac{1}{48}} e^{\left\{\frac{1}{48}\left(\frac{\alpha}{\psi^{-1}(\varepsilon/R)}\right)^{1/b} - \frac{64m\varepsilon^2}{15dL^2}\right\}}\right\}.$$

Choosing $\varepsilon_m = \tau R\psi(\Psi^{-1}(m^{-1/2}))$, we obtain

$$Prob_{\mathbf{z}}\left\{||f_{\mathbf{z}} - f_{\rho_*}||_{\mathscr{L}^2_\nu(X)} > \tau\frac{R}{2}\psi(\Psi^{-1}(m^{-1/2}))\right\} \geq \min\left\{\frac{1}{1+e^{-\ell_\varepsilon/24}}, \vartheta e^{-\frac{1}{48}} e^{c(\Psi^{-1}(m^{-1/2}))^{-1/b}}\right\},$$

where $c = \left(\frac{\alpha^{1/b}}{48} - \frac{64R^2\tau^2}{15dL^2}\right) > 0$ for $0 < \tau < \min\left(\frac{\sqrt{5d}L\alpha^{\frac{1}{2b}}}{32R}, 1\right)$.

Now as m goes to ∞, $\varepsilon \to 0$ and $\ell_\varepsilon \to \infty$. Therefore, for $c > 0$ we conclude that

$$\liminf_{m\to\infty}\inf_{l\in\mathcal{A}}\sup_{\rho\in\mathcal{P}_{\phi,b}} Prob_{\mathbf{z}}\left\{||f^l_{\mathbf{z}} - f_\rho||_{\mathscr{L}^2_\nu(X)} > \tau\frac{R}{2}\psi(\Psi^{-1}(m^{-1/2}))\right\} = 1.$$

□

Choosing $\varepsilon_m = \tau R\phi(\Psi^{-1}(m^{-1/2}))$ we get the following convergence rate from Theorem 3.9.

Theorem 3.12. *Under the same assumptions of Theorem 3.9 for* $\Psi(t) = t^{\frac{1}{2}+\frac{1}{2b}}\phi(t)$, *the estimator* $f_{\mathbf{z}}$ *corresponding to any learning algorithm converges to the regression function* f_ρ *with the following lower rate:*

$$\lim_{\tau\to 0}\liminf_{m\to\infty}\inf_{l\in\mathcal{A}}\sup_{\rho\in\mathcal{P}_{\phi,b}} Prob_{\mathbf{z}}\left\{||f^l_{\mathbf{z}} - f_\rho||_{\mathcal{H}} > \tau\phi\left(\Psi^{-1}(m^{-1/2})\right)\right\} = 1.$$

We obtain the following corollary as a consequence of Theorem 3.11, 3.12.

Corollary 3.3. *For any learning algorithm under Hölder's source condition* $f_\rho \in \Omega_{\phi,R}$, $\phi(t) = t^r$ *and the polynomial decay condition (13) for* $b > 1$, *the lower convergence rates can be described as*

$$\lim_{\tau\to 0}\liminf_{m\to\infty}\inf_{l\in\mathcal{A}}\sup_{\rho\in\mathcal{P}_{\phi,b}} Prob_{\mathbf{z}}\left\{||f^l_{\mathbf{z}} - f_\rho||_{\mathscr{L}^2_\nu(X)} > \tau m^{-\frac{2br+b}{4br+2b+2}}\right\} = 1$$

and

$$\lim_{\tau\to 0}\liminf_{m\to\infty}\inf_{l\in\mathcal{A}}\sup_{\rho\in\mathcal{P}_{\phi,b}} Prob_{\mathbf{z}}\left\{||f^l_{\mathbf{z}} - f_\rho||_{\mathcal{H}} > \tau m^{-\frac{br}{2br+b+1}}\right\} = 1.$$

If the minimax lower rate coincides with the upper convergence rate for $\lambda = \lambda_m$. Then the choice of parameter is said to be optimal. For the parameter choice $\lambda = \Psi^{-1}(m^{-1/2})$, Theorem 3.3 and Theorem 3.8 share the upper convergence rate with the lower convergence rate of Theorem 3.11 in $\mathscr{L}^2$-norm. For the same parameter choice, Theorem 3.4 and Theorem 3.7 share the upper convergence rate with the lower convergence rate of Theorem 3.12 in RKHS-norm. Therefore, the choice of the parameter is optimal.

It is important to observe that we get the same convergence rates for $b = 1$.

3.4. Individual Lower Rates

In this section, we discuss the individual minimax lower rates that describe the behavior of the error for the class of probability measure $\mathcal{P}_{\phi,b}$ as the sample size m grows.

Definition 3.4. *A sequence of positive numbers* a_n ($n \in \mathbb{N}$) *is called the individual lower rate of convergence for the class of probability measure* $\mathcal{P}$, *if*

$$\inf_{l\in\mathcal{A}}\sup_{\rho\in\mathcal{P}}\limsup_{m\to\infty}\left(\frac{E_{\mathbf{z}}\left(||f^l_{\mathbf{z}} - f_{\mathcal{H}}||^2\right)}{a_m}\right) > 0,$$

where $\mathcal{A}$ *denotes the set of all learning algorithms* $l : \mathbf{z} \mapsto f^l_{\mathbf{z}}$.

Theorem 3.13. *Let* $\mathbf{z}$ *be i.i.d. samples drawn according to the probability measure* $\mathcal{P}_{\phi,b}$ *where* ϕ *is the index function satisfying the conditions that* $\phi(t)/t^{r_1}$, $t^{r_2}/\phi(t)$ *are non-decreasing functions*

and dim$(Y) = d < \infty$. *Then for every* $\varepsilon > 0$, *the following lower bound holds:*

$$\inf_{l\in\mathcal{A}} \sup_{\rho\in\mathcal{P}_{\phi,b}} \limsup_{m\to\infty} \left(\frac{E_{\mathbf{z}}\left(||f_{\mathbf{z}}^{l} - f_{\mathcal{H}}||^2_{\mathscr{L}_\nu(X)} \right)}{m^{-(bc_2+\varepsilon)/(bc_1+\varepsilon+1)}} \right) > 0,$$

where $c_1 = 2r_1 + 1$ *and* $c_2 = 2r_2 + 1$.

We consider the class of probability measures such that the target function $f_{\mathcal{H}}$ is parameterized by $\mathbf{s} = (s_n)_{n=1}^{\infty} \in \{-1, +1\}^{\infty}$. Suppose for $\varepsilon > 0$,

$$g = \sum_{n=1}^{\infty} s_n R \sqrt{\frac{\varepsilon}{\varepsilon+1}\frac{\alpha}{n^b t_n}} \left(\frac{\phi(\alpha/n^b)}{\phi(t_n)} \right) n^{-(\varepsilon+1)/2} e_n,$$

where $\mathbf{s} = (s_n)_{n=1}^{\infty} \in \{-1, +1\}^{\infty}$, t_n's are the eigenvalues of the integral operator L_K, e_n's are the eigenvectors of the integral operator L_K and the orthonormal basis of RKHS $\mathcal{H}$. Then the target function $f_{\mathcal{H}} = \phi(L_K)g$ satisfies the general source condition. We assume that the conditional probability measure $\rho(y|x)$ follows the normal distribution centered at $f_{\mathcal{H}}$ and the marginal probability measure $\rho_X = \nu$. Now we can derive the individual lower rates over the considered class of probability measures from the ideas of the literature [12, 30].

Theorem 3.14. *Let* $\mathbf{z}$ *be i.i.d. samples drawn according to the probability measure* $\mathcal{P}_{\phi,b}$ *where* ϕ *is the index function satisfying the conditions that* $\phi(t)/t^{r_1}$, $t^{r_2}/\phi(t)$ *are non-decreasing functions and* $\dim(Y) = d < \infty$. *Then for every* $\varepsilon > 0$, *the following lower bound holds:*

$$\inf_{l\in\mathcal{A}} \sup_{\rho\in\mathcal{P}_{\phi,b}} \limsup_{m\to\infty} \left(\frac{E_{\mathbf{z}}\left(||f_{\mathbf{z}}^{l} - f_{\mathcal{H}}||^2_{\mathcal{H}} \right)}{m^{-(bc_2-b+\varepsilon)/(bc_1+\varepsilon+1)}} \right) > 0.$$

4. CONCLUSION

In our analysis we derive the upper and lower convergence rates over the wide class of probability measures considering general source condition in vector-valued setting. In particular, our minimax rates can be used for the scalar-valued functions and multi-task learning problems. The lower convergence rates coincide with the upper convergence rates for the optimal parameter choice based on smoothness parameters b, ϕ. We can also develop various parameter choice rules such as balancing principle [31], quasi-optimality principle [32], discrepancy principle [33] for the regularized solutions provided in our analysis.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

The authors are grateful to the reviewers for their helpful comments and pointing out a subtle error that led to improve the quality of the paper.

REFERENCES

1. Cucker F, Smale S. On the mathematical foundations of learning. *Bull Am Math Soc.* (2002) **39**:1–49. doi: 10.1090/S0273-0979-01-00923-5
2. Evgeniou T, Pontil M, Poggio T. Regularization networks and support vector machines. *Adv Comput Math.* (2000) **13**:1–50. doi: 10.1023/A:1018946025316
3. Vapnik VN, Vapnik V. *Statistical Learning Theory.* New York, NY: Wiley (1998).
4. Bauer F, Pereverzev S, Rosasco L. On regularization algorithms in learning theory. *J Complex.* (2007) **23**:52–72. doi: 10.1016/j.jco.2006.07.001
5. Engl HW, Hanke M, Neubauer A. *Regularization of Inverse Problems.* Dordrecht: Kluwer Academic Publishers Group (1996).
6. Gerfo LL, Rosasco L, Odone F, De Vito E, Verri A. Spectral algorithms for supervised learning. *Neural Comput.* (2008) **20**:1873–97. doi: 10.1162/neco.2008.05-07-517
7. Tikhonov AN, Arsenin VY. *Solutions of Ill-Posed Problems.* Washington, DC: W. H. Winston (1977).
8. Bousquet O, Boucheron S, Lugosi G. Introduction to statistical learning theory. In: Bousquet O, von Luxburg U, Ratsch G editors. *Advanced Lectures on Machine Learning, Volume 3176 of Lecture Notes in Computer Science.* Berlin; Heidelberg: Springer (2004). pp. 169–207.
9. Cucker F, Zhou DX. *Learning Theory: An Approximation Theory Viewpoint.* Cambridge, UK: Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press (2007).
10. Lu S, Pereverzev S. *Regularization Theory for Ill-posed Problems: Selected Topics,* Berlin: DeGruyter (2013).
11. Abhishake Sivananthan S. Multi-penalty regularization in learning theory. *J Complex.* (2016) **36**:141–65. doi: 10.1016/j.jco.2016.05.003
12. Caponnetto A, De Vito E. Optimal rates for the regularized least-squares algorithm. *Found Comput Math.* (2007) **7**:331–68. doi: 10.1007/s10208-006-0196-8
13. Smale S, Zhou DX. Estimating the approximation error in learning theory. *Anal Appl.* (2003) **1**:17–41. doi: 10.1142/S0219530503000089
14. Smale S, Zhou DX. Shannon sampling and function reconstruction from point values. *Bull Am Math Soc.* (2004) **41**:279–306. doi: 10.1090/S0273-0979-04-01025-0
15. Smale S, Zhou DX. Shannon sampling II: connections to learning theory. *Appl Comput Harmon Anal.* (2005) **19**:285–302. doi: 10.1016/j.acha.2005.03.001
16. Smale S, Zhou DX. Learning theory estimates via integral operators and their approximations. *Constr Approx.* (2007) **26**:153–72. doi: 10.1007/s00365-006-0659-y
17. Mathé P, Pereverzev SV. Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse Probl.* (2003) **19**:789–803. doi: 10.1088/0266-5611/19/3/319
18. Blanchard G, Mücke N. Optimal rates for regularization of statistical inverse learning problems. *arXiv:1604.04054* (2016).
19. Mendelson S. On the performance of kernel classes. *J Mach Learn Res.* (2003) **4**:759–71.
20. Zhang T. Effective dimension and generalization of kernel learning. In: Thrun S, Becker S, Obermayer K. editors. *Advances in Neural Information Processing Systems.* Cambridge, MA: MIT Press, (2003). pp. 454–61.
21. Akhiezer NI, Glazman IM. *Theory of Linear Operators in Hilbert Space,* Translated from the Russian and with a preface by Merlynd Nestell. New York, NY: Dover Publications Inc (1993).

22. Micchelli CA, Pontil M. On learning vector-valued functions. *Neural Comput.* (2005) **17**:177–204. doi: 10.1162/0899766052530802
23. Aronszajn N. Theory of reproducing kernels. *Trans Am Math Soc.* (1950) **68**:337–404. doi: 10.1090/S0002-9947-1950-0051437-7
24. Reed M, Simon B. *Functional Analysis, Vol. 1,* San Diego, CA: Academic Press (1980).
25. De Vito E, Rosasco L, Caponnetto A, De Giovannini U, Odone F. Learning from examples as an inverse problem. *J Mach Learn Res.* (2005) **6**:883–904.
26. Pinelis IF, Sakhanenko AI. Remarks on inequalities for the probabilities of large deviations. *Theory Prob Appl.* (1985) **30**:127–31. doi: 10.1137/1130013
27. Peller VV. Multiple operator integrals in perturbation theory. *Bull Math Sci.* (2016) **6**:15–88. doi: 10.1007/s13373-015-0073-y
28. Boucheron S, Bousquet O, Lugosi G. Theory of classification: a survey of some recent advances. *ESAIM: Prob Stat.* (2005) **9**:323–75. doi: 10.1051/ps:2005018
29. DeVore R, Kerkyacharian G, Picard D, Temlyakov V. Approximation methods for supervised learning. *Found Comput Math.* (2006) **6**:3–58. doi: 10.1007/s10208-004-0158-6
30. Györfi L, Kohler M, Krzyzak A, Walk H. *A Distribution-Free Theory of Nonparametric Regression.* New York, NY: Springer Series in Statistics, Springer-Verlag (2002).
31. De Vito E, Pereverzyev S, Rosasco L. Adaptive kernel methods using the balancing principle. *Found Comput Math.* (2010) **10**:455–79. doi: 10.1007/s10208-010-9064-2
32. Bauer F, Reiss M. Regularization independent of the noise level: an analysis of quasi-optimality. *Inverse Prob.* (2008) **24**:055009. doi: 10.1088/0266-5611/24/5/055009
33. Lu S, Pereverzev SV, Tautenhahn U. A model function method in regularized total least squares. *Appl Anal.* (2010) **89**:1693–703. doi: 10.1080/00036811.2010.492502

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.