

Manifold regularization based on Nyström type subsampling

Abhishake Rastogi*, Sivananthan Sampath†

Department of Mathematics

Indian Institute of Technology Delhi

New Delhi 110016, India

Abstract

In this paper, we study the Nyström type subsampling for large scale kernel methods to reduce the computational complexities of big data. We discuss the multi-penalty regularization scheme based on Nyström type subsampling which is motivated from well-studied manifold regularization schemes. We develop a theoretical analysis of multi-penalty least-square regularization scheme under the general source condition in vector-valued function setting, therefore the results can also be applied to multi-task learning problems. We achieve the optimal minimax convergence rates of multi-penalty regularization using the concept of effective dimension for the appropriate subsampling size. We discuss an aggregation approach based on linear function strategy to combine various Nyström approximants. Finally, we demonstrate the performance of multi-penalty regularization based on Nyström type subsampling on Caltech-101 data set for multi-class image classification and NSL-KDD benchmark data set for intrusion detection problem.

Keywords: Multi-task learning; Manifold learning; Multi-penalty regularization; Nyström type subsampling; Optimal rates; Linear functional strategy.

Mathematics Subject Classification 2010: 68T05, 68Q32.

1 Introduction

Multi-task learning is an approach which learns multiple tasks simultaneously. The problem has potential to learn the structure of the related tasks. The idea is that exploring task relatedness can lead to improved performance. Various learning algorithms are studied to incorporate the structure of task relations in literature [?, ?, ?, ?]. In agreement with past empirical work on multi-task learning, learning multiple related tasks simultaneously has been empirically [?, ?, ?, ?, ?, ?, ?, ?] and theoretically [?, ?, ?] shown significantly improved performance relative to learning each task independently. Multi-task learning is becoming interesting due to its applications in computer vision, image processing and many other fields such as object detection/classification [?], image denoising, inpainting, finance and economics forecasting predicting [?], marketing modeling for the preferences of many individuals [?, ?] and in bioinformatics for example to study tumor prediction from multiple microarray data sets or analyze data from multiple related diseases.

In this work, we discuss multi-task learning approach that considers a notion of relatedness based on the concept of manifold regularization. In scalar-valued function setting, Belkin et al. [?] introduced the

*Corresponding Author, Email address: abhishekrastogi2012@gmail.com

†Email address: siva@maths.iitd.ac.in

concept of manifold regularization which focus on a semi-supervised framework that incorporates labeled and unlabeled data in a general-purpose learner. Minh and Sindhwani [?] generalized the concept of manifold learning for vector-valued functions which exploits output inter-dependencies while enforcing smoothness with respect to input data geometry. Further, Minh and Sindhwani [?] present a general learning framework that encompasses learning across three different paradigms, namely vector-valued, multi-view and semi-supervised learning simultaneously. Multi-view learning approach is considered to construct the regularized solution based on different views of the input data using different hypothesis spaces [?, ?, ?, ?, ?, ?, ?]. Micchelli and Pontil [?] introduced the concept of vector-valued reproducing kernel Hilbert spaces to facilitate theory of multi-task learning. Also, the fact that every vector-valued RKHS is corresponding to some operator-valued positive definite kernel, reduces the problem of choosing appropriate RKHS (hypothesis space) to choosing appropriate kernel [?]. In paper [?, ?, ?], the authors proposed multiple kernel learning from a set of kernels. Here we consider the direct sum of reproducing kernel Hilbert spaces as the hypothesis space.

Multi-task learning is studied under the elegant and effective framework of kernel methods. The expansion of automatic data generation and acquisition bring data of huge size and complexity which raises challenges to computational capacities. In order to tackle these difficulties, various techniques are discussed in the literature such as replacing the empirical kernel matrix with a smaller matrix obtained by (column) subsampling [?, ?, ?], greedy-type algorithms [?], divide-and-conquer approach [?, ?, ?]. We are inspired from the work of Kriukova et al. [?] in which the authors discussed an approach to aggregate various regularized solutions based on Nyström subsampling in single-penalty regularization. Here we consider the so-called Nyström type subsampling in large scale kernel methods for dealing with big data which particularly can be seen as a regularized projection scheme. We achieve the optimal convergence rates for multi-penalty regularization based on the Nyström type subsampling approach, provided the subsampling size is appropriate. We adapt the aggregation approach for multi-task learning manifold regularization scheme to improve the accuracy of the results. We consider the linear combination of Nyström approximants and try to obtain a combination of the approximants which is closer to the target function. The coefficients of the linear combination are estimated by means of the linear functional strategy. The aggregation approach tries to accumulate the information hidden inside various approximants to produce the estimator of the target function [?, ?] (also see reference therein).

The paper is organized as follows. In Section 2, we describe the framework of vector-valued multi-penalized learning problem with some basic definitions and notations. In Section 3, we discuss the convergence issues of the vector-valued multi-penalty regularization scheme based on Nyström type subsampling in the norm in $\mathcal{L}_{\rho_X}^2$ and the norm in \mathcal{H} . In Section 4, we discuss the aggregation approach to accumulate various estimators based on the Nyström type subsampling. In the last section, we demonstrate the performance of multi-penalty regularization based on Nyström type subsampling on Caltech-101 data set for multi-class image classification and NSL-KDD benchmark data set for intrusion detection problem.

2 Multi-task learning via vector-valued RKHS

The problem of learning multiple tasks jointly can be modeled by the vector-valued functions $f : X \rightarrow \mathbb{R}^T$ whose components represent individual task-predictors, i.e., $f = (f_1, \dots, f_T)$ for $f_t : X \rightarrow \mathbb{R}$ ($1 \leq t \leq T$). Here we consider general framework of vector-valued functions $f : X \rightarrow Y$ developed by Micchelli and Pontil [?] to address the multi-task learning algorithm. We consider the concept of vector-valued reproducing kernel Hilbert space which is the extension of well-known scalar-valued reproducing kernel Hilbert space.

Definition 2.1. Vector-valued reproducing kernel Hilbert space (RKHS_{vv}). Let X be a non-empty set, $(Y, \langle \cdot, \cdot \rangle_Y)$ be a real separable Hilbert space. The Hilbert space 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.

Suppose $\mathcal{L}(Y)$ be the Banach space of bounded linear operators on Y . A function $K : X \times X \rightarrow \mathcal{L}(Y)$ is said to be an operator-valued positive definite kernel if for each pair $(x, z) \in X \times X$, $K(x, z)^* = K(z, x)$, and for every finite set of points $\{x_i\}_{i=1}^N \subset X$ and $\{y_i\}_{i=1}^N \subset Y$,

$$\sum_{i,j=1}^N \langle y_i, K(x_i, x_j) y_j \rangle_Y \geq 0.$$

There exists a unique Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ of functions on X satisfying the following conditions:

(i) for all $x \in X$ and $y \in Y$, the functions $K_x y = K(\cdot, x)y \in \mathcal{H}$, defined by

$$(K_x y)z = K(z, x)y \text{ for all } z \in X,$$

(ii) the span of the set $\{K_x y : x \in X, y \in Y\}$ is dense in \mathcal{H} , and

(iii) for all $f \in \mathcal{H}$, $\langle f(x), y \rangle_Y = \langle f, K_x y \rangle_{\mathcal{H}}$ (reproducing property).

Moreover, there is one to one correspondence between operator-valued positive definite kernels and vector-valued RKHS [?].

In the learning theory, we are given with the random samples $\{(x_i, y_i) : 1 \leq i \leq m\}$ drawn identically and independently from a unknown joint probability measure ρ on the sample space $X \times Y$. We assume that the input space X is a locally compact countable Hausdorff space and the output space $(Y, \langle \cdot, \cdot \rangle_Y)$ is a real separable space. The goal is to predict the output values for the inputs. Suppose we predict y for the input x based on our algorithm but the true output is y' . Then we suffer a loss $\ell(y, y')$, where the loss function $\ell : Y \times Y \rightarrow \mathbb{R}^+$. A widely used approach based on the square loss function $\ell(y, y') = \|y - y'\|_Y^2$ in regularization theory is Tikhonov type regularization:

$$\frac{1}{m} \sum_{i=1}^m \|f(x_i) - y_i\|_Y^2 + \lambda \|f\|_{\mathcal{H}}^2,$$

The regularization parameter λ controls the trade off between the error measuring the fitness of data and the complexity of the solution measured in the RKHS-norm.

We discuss the multi-task learning approach that considers a notion of task relatedness based on the concept of manifold regularization. In this approach, different RKHS_{vv} are used to estimate the target functions based on different views of input data, such as different features or modalities and a data-dependent regularization term is used to enforce consistency of output values from different views of the same input example.

We consider the following regularization scheme to analyze the multi-task manifold learning scheme corresponding to different views:

$$\arg \min_{f \in \mathcal{H}_{K_1} \oplus \dots \oplus \mathcal{H}_{K_v}} \left\{ \frac{1}{m} \sum_{i=1}^m \|f(x_i) - y_i\|_Y^2 + \lambda_A \|f\|_{\mathcal{H}}^2 + \lambda_I \langle \mathbf{f}, M\mathbf{f} \rangle_{Y^n} \right\}, \quad (1)$$

where $\{(x_i, y_i) \in X \times Y : 1 \leq i \leq m\} \cup \{x_i \in X : m < i \leq n\}$ is given set of labeled and unlabeled data, M is a symmetric, positive operator, $\lambda_A, \lambda_I \geq 0$ and $\mathbf{f} = (f(x_1), \dots, f(x_n))^T \in Y^n$.

The direct sum of reproducing kernel Hilbert spaces $\mathcal{H} = \mathcal{H}_{K_1} \oplus \dots \oplus \mathcal{H}_{K_v}$ is also a RKHS. Suppose K is the kernel corresponding to RKHS \mathcal{H} .

Throughout this paper we assume the following hypothesis:

Assumption 2.1. *Let \mathcal{H} be a reproducing kernel Hilbert space of functions $f : X \rightarrow Y$ such that*

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

Hilbert-Schmidt operator $A \in \mathcal{L}(\mathcal{H})$, $\text{Tr}(A) := \sum_{k=1}^{\infty} \langle A e_k, e_k \rangle$ for an orthonormal basis $\{e_k\}_{k=1}^{\infty}$ of \mathcal{H} .

(ii) *The real-valued function $\phi : X \times X \rightarrow \mathbb{R}$, defined by $\phi(x, t) = \langle K_t v, K_x w \rangle_{\mathcal{H}}$, is measurable $\forall v, w \in Y$.*

By the representation theorem [?], the solution of the multi-penalized regularization problem (1) will be of the form:

$$f_{\mathbf{z}, \lambda} = \sum_{i=1}^n K_{x_i} c_i, \text{ for some } \mathbf{c} = (c_1, \dots, c_n) = (\mathbb{J} \mathbb{K}_n + \lambda_A m \mathbb{I}_n + \lambda_I m L \mathbb{K}_n)^{-1} \mathbf{y}_n, \quad (2)$$

where $(\mathbb{K}_n)_{ij} = K(x_i, x_j)$ with $i, j \in \{1, \dots, n\}$, $\mathbb{J} = \text{diag}(1, \dots, 1, 0, \dots, 0)$ is $n \times n$ diagonal matrix with the first m diagonal entries as 1 and the rest 0, \mathbb{I}_n is identity of order n and $\mathbf{y}_n = [y_1, \dots, y_m, 0, \dots, 0]^T \in Y^n$.

In order to obtain the computationally efficient algorithm from the functional (1), we consider the Nyström type subsampling which uses the idea of replacing the empirical kernel matrix with a smaller matrix obtained by (column) subsampling [?, ?, ?]. This can also be seen as a restriction of the optimization functional (1) over the space:

$$\mathcal{H}^{\mathbf{x}_s} := \{f | f = \sum_{i=1}^s K_{x_i} c_i, \mathbf{c} = (c_1, \dots, c_s) \in Y^s\},$$

where $s \ll n$ and $\mathbf{x}_s = (x_1, \dots, x_s)$ is a subset of the input points in the training set.

The minimizer of the manifold regularization scheme (1) over the space $\mathcal{H}^{\mathbf{x}_s}$ will be of the form:

$$f_{\mathbf{z}, \lambda}^s = \sum_{i=1}^s K_{x_i} c_i, \text{ for } \mathbf{c} = (c_1, \dots, c_s) = (\mathbb{K}_{ms}^T \mathbb{K}_{ms} + \lambda_A m \mathbb{K}_{ss} + \lambda_I m \mathbb{K}_{ns}^T L \mathbb{K}_{ns})^\dagger \mathbb{K}_{ms}^T \mathbf{y}, \quad (3)$$

where A^\dagger denotes the Moore-Penrose pseudoinverse of a matrix A , $(\mathbb{K}_{ms})_{ij} = K(x_i, \tilde{x}_j)$, $(\mathbb{K}_{ss})_{kj} = K(\tilde{x}_k, \tilde{x}_j)$ with $i \in \{1, \dots, m\}$ and $j, k \in \{1, \dots, s\}$ and $\mathbf{y} = [y_1, \dots, y_m]^T \in Y^m$.

The computational time of the Nyström approximation (3) is of order $\mathcal{O}(sn^2)$ while the computational time complexity of standard manifold regularized solution (2) is of order $\mathcal{O}(n^3)$. Therefore, the randomized subsampling methods can break the memory barriers and consequently achieve much better time complexity compare with standard manifold regularization algorithm.

We analyze the more general vector-valued multi-penalty regularization scheme based on Nyström subsampling:

$$f_{\mathbf{z}, \lambda}^s = \arg \min_{f \in \mathcal{H}^{\mathbf{x}_s}} \left\{ \frac{1}{m} \sum_{i=1}^m \|f(x_i) - y_i\|_Y^2 + \lambda_0 \|f\|_{\mathcal{H}}^2 + \sum_{j=1}^p \lambda_j \|B_j f\|_{\mathcal{H}}^2 \right\}, \quad (4)$$

where $B_j : \mathcal{H} \rightarrow \mathcal{H}$ ($1 \leq j \leq p$) are bounded operators, $\lambda_0 > 0$, λ_j ($1 \leq j \leq p$) are non-negative real numbers and λ denotes the ordered set $(\lambda_0, \lambda_1, \dots, \lambda_p)$.

Here we introduce the sampling operator which is useful in the analysis of regularization schemes.

Definition 2.2. The **sampling operator** $S_{\mathbf{x}} : \mathcal{H} \rightarrow Y^m$ associated with a discrete subset $\mathbf{x} = (x_i)_{i=1}^m$ is defined by

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

Then its adjoint is given by

$$S_{\mathbf{x}}^* \mathbf{y} = \frac{1}{m} \sum_{i=1}^m K_{x_i} y_i, \quad \forall \mathbf{y} = (y_1, \dots, y_m) \in Y^m.$$

The sampling operator is bounded by κ .

We obtain the following explicit expression of the minimizer of the regularization scheme (4). The proof of the theorem follows the same steps as of Lemma 1 [?].

Theorem 2.1. For the positive choice of λ_0 , the functional (4) has unique minimizer:

$$f_{\mathbf{z}, \lambda}^s = \left(P_{\mathbf{x}_s} S_{\mathbf{x}}^* S_{\mathbf{x}} P_{\mathbf{x}_s} + \lambda_0 I + \sum_{j=1}^p \lambda_j P_{\mathbf{x}_s} B_j^* B_j P_{\mathbf{x}_s} \right)^{-1} P_{\mathbf{x}_s} S_{\mathbf{x}}^* \mathbf{y},$$

where $P_{\mathbf{x}_s}$ is the orthogonal projection operator with range $\mathcal{H}^{\mathbf{x}_s}$.

The data-free version of the considered regularization scheme (4) is

$$f_{\lambda}^s := \arg \min_{f \in \mathcal{H}^{\mathbf{x}_s}} \left\{ \int_Z \|f(x) - y\|_Y^2 d\rho(x, y) + \lambda_0 \|f\|_{\mathcal{H}}^2 + \sum_{j=1}^p \lambda_j \|B_j f\|_{\mathcal{H}}^2 \right\}. \quad (5)$$

Using the fact $\mathcal{E}(f) = \int_Z \|f(x) - y\|_Y^2 d\rho(x, y) = \|L_K^{1/2}(f - f_{\mathcal{H}})\|_{\mathcal{H}}^2 + \mathcal{E}(f_{\mathcal{H}})$, we get,

$$f_{\lambda}^s = \left(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} + \lambda_0 I + \sum_{j=1}^p \lambda_j P_{\mathbf{x}_s} B_j^* B_j P_{\mathbf{x}_s} \right)^{-1} P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} f_{\mathcal{H}}. \quad (6)$$

We assume

$$f_{\lambda_0}^s := \arg \min_{f \in \mathcal{H}^{\mathbf{x}_s}} \left\{ \int_Z \|f(x) - y\|_Y^2 d\rho(x, y) + \lambda_0 \|f\|_{\mathcal{H}}^2 \right\} \quad (7)$$

which implies

$$f_{\lambda_0}^s = (P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} + \lambda_0 I)^{-1} P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} f_{\mathcal{H}},$$

where the integral operator L_K is a self-adjoint, non-negative, compact operator on the Hilbert space $(\mathcal{L}_{\rho_X}^2, \langle \cdot, \cdot \rangle_{\mathcal{L}_{\rho_X}^2})$ of square-integrable functions from X to Y with respect to ρ_X , defined as

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

The integral operator is bounded by κ^2 . 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. Though it is notational abuse, for convenience 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 (For more properties see [?, ?]).

Our aim is to discuss the convergence issues of the regularized solution $f_{\mathbf{z},\lambda}^s$ based on Nyström type subsampling. We estimate the error bounds of $f_{\mathbf{z},\lambda}^s - f_{\mathcal{H}}$ by measuring the bounds of sample error $f_{\mathbf{z},\lambda}^s - f_{\lambda}^s$ and approximation error $f_{\lambda}^s - f_{\mathcal{H}}$. The approximation error is estimated with the help of the single-penalty regularized solution $f_{\lambda_0}^s$.

For any probability measure, we can always obtain a solution converging to the prescribed target function but the convergence rates may be arbitrarily slow. This phenomena is known as no free lunch theorem [?]. Therefore, we need some prior assumptions on the probability measure ρ in order to achieve the uniform convergence rates for learning algorithms. Following the notion of Bauer et al. [?], Caponnetto and De Vito [?], we consider the following assumptions on the joint probability measure ρ in terms of the complexity of the target function and a theoretical spectral parameter effective dimension:

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

$$\int_Z \|y\|_Y^2 d\rho(x, y) < \infty \quad (8)$$

- (ii) There exists the minimizer of the generalization error over the RKHS \mathcal{H} ,

$$f_{\mathcal{H}} := \arg \min_{f \in \mathcal{H}} \left\{ \int_Z \|f(x) - y\|_Y^2 d\rho(x, y) \right\}. \quad (9)$$

- (iii) There exist some constants M, Σ such that

$$\int_Y \left(e^{\|y - f_{\mathcal{H}}(x)\|_Y / M} - \frac{\|y - f_{\mathcal{H}}(x)\|_Y}{M} - 1 \right) d\rho(y|x) \leq \frac{\Sigma^2}{2M^2} \quad (10)$$

holds for almost all $x \in X$.

It is worthwhile to observe that for the real-valued functions and multi-task learning algorithms, the boundedness of output space Y can be easily ensured. So we can get the error estimates from our analysis without imposing any condition on the conditional probability measure (10).

The smoothness of the target function can be described in terms of the integral operator by the source condition:

Assumption 2.2. (Source condition) *Suppose*

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

where ϕ is operator monotone function on the interval $[0, \kappa^2]$ with the assumption $\phi(0) = 0$ and ϕ^2 is a concave function. Then the condition $f_{\mathcal{H}} \in \Omega_{\phi,R}$ is usually referred to as general source condition [?].

Assumption 2.3. (Polynomial decay condition) *For fixed positive constants α, β and $b > 1$, we assume that the eigenvalues t_n 's of the integral operator L_K follows the polynomial decay:*

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

We define the class of the probability measures \mathcal{P}_{ϕ} satisfying the conditions (i), (ii), (iii) and Assumption 2.2. We also consider the probability measure class $\mathcal{P}_{\phi,b}$ which satisfies the conditions (i), (ii), (iii) and Assumption 2.2, 2.3.

The convergence rates discussed in our analysis depend on the effective dimension. We achieve the optimal minimax convergence rates using the concept of the effective dimension. For the integral operator L_K , the effective dimension is defined as

$$\mathcal{N}(\gamma) := \text{Tr} \left((L_K + \gamma I)^{-1} L_K \right), \text{ for } \gamma > 0.$$

The fact, L_K is a trace class operator implies that the effective dimension is finite. The effective dimension is continuously decreasing function of γ from ∞ to 0. For further discussion on effective dimension we refer to the literature [?, ?, ?, ?, ?].

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

$$\mathcal{N}(\gamma) \leq \frac{\beta b}{b-1} \gamma^{-1/b}, \text{ for } b > 1 \quad (11)$$

and without the polynomial decay condition, we have

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

We define the random variable $\mathcal{N}_x(\gamma) = \langle K_x, (L_K + \gamma I)^{-1} K_x \rangle_{\mathcal{H}}$ for $x \in X$ and let

$$\mathcal{N}_{\infty}(\gamma) := \sup_{x \in X} \mathcal{N}_x(\gamma) < \infty.$$

	$\ f_{\mathbf{z}, \lambda} - f_{\mathcal{H}}\ _{\rho}$	$\ f_{\mathbf{z}, \lambda} - f_{\mathcal{H}}\ _{\mathcal{H}}$	Assumption (p qualification)	Scheme	general source condition	Optimal rates
Kriukova et al. [?]	$m^{-\frac{2r+1}{4r+4}}$	N/A	$r \leq \frac{1}{2}$	Single-penalty Tikhonov regularization	✓	
Rudi et al. [?]	$m^{-\frac{2br+b}{4br+2b+2}}$	N/A	$r \leq \frac{1}{2}$	Single-penalty Tikhonov regularization		✓
Our Results	$m^{-\frac{2br+b}{4br+2b+2}}$	$m^{-\frac{br}{2br+b+1}}$	$r \leq \frac{1}{2}$	Multi-penalty regularization	✓	✓

Table 1: Convergence rates of the regularized learning algorithms based on Nyström subsampling

Now we review the previous results on the regularization schemes based on Nyström subsampling which are directly comparable to our results: Kriukova et al. [?] and Rudi et al. [?]. For convenience, we tried to present the most essential points in the unified way in Table 1. We have shown the convergence rates under Hölder's source condition. Rudi et al. [?] obtained the minimax optimal convergence rates depending on the eigenvalues of L_K in $\|\cdot\|_{\rho}$ -norm. To obtain the optimal rates the concept of effective dimension is exploited. Kriukova et al. [?] considered the Tikhonov regularization with Nyström type subsampling under general source condition. They discussed the upper convergence rates and do not take into account the polynomial decay condition of the eigenvalues of the integral operator L_K . We used the idea of Nyström type subsampling to efficiently implement the multi-penalty regularization algorithm. We obtain optimal convergence rates of multi-penalty regularization with Nyström type subsampling under general source condition. In particular, we also get optimal rates of single-penalty Tikhonov regularization with Nyström type subsampling under general source condition as the special case.

3 Convergence issues

In this section, we present the optimal minimax convergence rates for vector-valued multi-penalty regularization based on Nyström type subsampling using the concept of effective dimension over the classes of the probability measures \mathcal{P}_ϕ and $\mathcal{P}_{\phi,b}$.

In order to prove the optimal convergence rates, we need the following inequality which is used in the papers [?, ?] and based on the results of Pinelis and Sakhanenko [?].

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

$$E \{ \|\xi - E(\xi)\|_{\mathcal{H}}^n \} \leq \frac{1}{2} n! S^2 Q^{n-2} \quad \forall n \geq 2, \quad (12)$$

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

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

In particular, the inequality (12) holds if

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

In the following proposition, we measure the effect of random sampling using noise assumption (10) in terms of the effective dimension $\mathcal{N}(\gamma)$. The quantity describes the probabilistic estimates of the perturbation measure due to random sampling.

Proposition 3.2. *Let \mathbf{z} be i.i.d. samples drawn according to the probability measure ρ satisfying the assumptions (8), (9), (10) and $\kappa = \sqrt{\sup_{x \in X} \text{Tr}(K_x^* K_x)}$. Then for all $0 < \eta < 1$, with the confidence $1 - \eta$, we have*

$$\|(L_K + \gamma I)^{-1/2} P_{\mathbf{x}_s} \{S_{\mathbf{x}}^* \mathbf{y} - S_{\mathbf{x}}^* S_{\mathbf{x}} f_{\mathcal{H}}\}\|_{\mathcal{H}} \leq 2 \left(\frac{\kappa M}{m \sqrt{\gamma}} + \sqrt{\frac{\Sigma^2 \mathcal{N}(\gamma)}{m}} \right) \log \left(\frac{4}{\eta} \right) \quad (13)$$

and

$$\|S_{\mathbf{x}}^* S_{\mathbf{x}} - L_K\|_{\mathcal{L}(\mathcal{H})} \leq 2 \left(\frac{\kappa^2}{m} + \frac{\kappa^2}{\sqrt{m}} \right) \log \left(\frac{4}{\eta} \right). \quad (14)$$

Proof. To estimate the first expression, we consider the random variable $\xi_1(z) = (L_K + \gamma I)^{-1/2} P_{\mathbf{x}_s} K_x (y - f_{\mathcal{H}}(x))$ from (Z, ρ) to reproducing kernel Hilbert space \mathcal{H} with

$$E_z(\xi_1) = \int_Z (L_K + \gamma I)^{-1/2} P_{\mathbf{x}_s} K_x (y - f_{\mathcal{H}}(x)) d\rho(x, y) = 0,$$

$$\frac{1}{m} \sum_{i=1}^m \xi_1(z_i) = (L_K + \gamma I)^{-1/2} P_{\mathbf{x}_s} (S_{\mathbf{x}}^* \mathbf{y} - S_{\mathbf{x}}^* S_{\mathbf{x}} f_{\mathcal{H}})$$

and

$$\begin{aligned} E_z(\|\xi_1 - E_z(\xi_1)\|_{\mathcal{H}}^n) &= E_z \left(\|(L_K + \gamma I)^{-1/2} P_{\mathbf{x}_s} K_x (y - f_{\mathcal{H}}(x))\|_{\mathcal{H}}^n \right) \\ &\leq E_z \left(\|K_x^* P_{\mathbf{x}_s} (L_K + \gamma I)^{-1} P_{\mathbf{x}_s} K_x\|_{\mathcal{L}(Y)}^{n/2} \|y - f_{\mathcal{H}}(x)\|_Y^n \right) \\ &\leq E_x \left(\|K_x^* P_{\mathbf{x}_s} (L_K + \gamma I)^{-1} P_{\mathbf{x}_s} K_x\|_{\mathcal{L}(Y)}^{n/2} E_y (\|y - f_{\mathcal{H}}(x)\|_Y^n) \right). \end{aligned}$$

Under the assumption (10) we get,

$$E_z(\|\xi_1 - E_z(\xi_1)\|_{\mathcal{H}}^n) \leq \frac{n!}{2} \left(\Sigma \sqrt{\mathcal{N}(\gamma)} \right)^2 \left(\frac{\kappa M}{\sqrt{\gamma}} \right)^{n-2}, \quad \forall n \geq 2.$$

On applying Proposition 3.1 we conclude that

$$\|(L_K + \gamma I)^{-1/2} P_{\mathbf{x}_s} \{S_{\mathbf{x}_s}^* \mathbf{y} - S_{\mathbf{x}_s}^* S_{\mathbf{x}_s} f_{\mathcal{H}}\}\|_{\mathcal{H}} \leq 2 \left(\frac{\kappa M}{m \sqrt{\gamma}} + \sqrt{\frac{\Sigma^2 \mathcal{N}(\gamma)}{m}} \right) \log \left(\frac{4}{\eta} \right)$$

with confidence $1 - \eta/2$.

The second expression can be estimated easily by considering the random variable $\xi_2(x) = K_x K_x^*$ from (X, ρ_X) to $\mathcal{L}(\mathcal{H})$. The proof can also be found in De Vito et al. [?]. \square

The following conditions on the sample size and sub-sample size are used to derive the convergence rates of regularized learning algorithms. In particular, we can assume the following inequality for sufficiently large sample with the confidence $1 - \eta$:

$$\frac{8\kappa^2}{\sqrt{m}} \log \left(\frac{4}{\eta} \right) \leq \lambda_0. \quad (15)$$

Following the notion of Rudi et al. [?] and Kriukova et al. [?] on subsampling, we measure the approximation power of the projection method induced by the projection operator $P_{\mathbf{x}_s}$ in terms of $\Delta_s := \|L_K^{1/2}(I - P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})}$. We make the assumption on Δ_s as considered in Theorem 2 [?]:

$$\Delta_s = \|L_K^{1/2}(I - P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})} \leq \sqrt{\Theta_{1/2}^{-1}(m^{-1/2})}, \text{ for } \Theta_{1/2}(t) = \sqrt{t} \phi(t). \quad (16)$$

Under the parameter choice $\lambda_0 = \Psi^{-1}(m^{-1/2})$ for $\Psi(t) = t^{\frac{1}{2} + \frac{1}{2a}} \phi(t)$ ($a \geq 1$), we obtain

$$\Delta_s^2 \leq \Theta_{1/2}^{-1}(m^{-1/2}) \leq \Psi^{-1}(m^{-1/2}) = \lambda_0. \quad (17)$$

Moreover, from Lemma 6 [?] under Assumption 2.3 and $s \geq \max \left\{ 67 \log \left(\frac{12\kappa^2}{\lambda_0 \delta} \right), 5\mathcal{N}_{\infty} \left(\frac{\lambda_0}{3} \right) \log \left(\frac{12\kappa^2}{\lambda_0 \delta} \right) \right\}$, $\lambda_0 > 0$, for every $\delta > 0$, the following inequality holds with the probability $1 - \delta$,

$$\Delta_s^2 \leq \left\| \left(L_K + \frac{\lambda_0}{3} I \right)^{1/2} (I - P_{\mathbf{x}_s}) \right\|_{\mathcal{L}(\mathcal{H})}^2 \leq \lambda_0.$$

Then under the condition (17) using Proposition 2, 3 [?] we get,

$$\|(I - P_{\mathbf{x}_s})\psi(L_K)\|_{\mathcal{L}(\mathcal{H})} \leq \psi \left(\|L_K^{1/2}(I - P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})}^2 \right) \leq \psi(\lambda_0)$$

and

$$\|P_{\mathbf{x}_s} \psi(L_K) P_{\mathbf{x}_s} - \psi(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})} \leq c_{\psi} \psi \left(\|L_K^{1/2}(I - P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})}^2 \right) \leq c_{\psi} \psi(\lambda_0).$$

In the following section, we discuss the error analysis of the multi-penalty regularization scheme based on Nyström type subsampling in probabilistic sense. In general, we derive the convergence rates for regularization algorithms in the norm in RKHS and the norm in $\mathcal{L}_{\rho_X}^2$ separately. In Theorem 3.1, 3.2, 3.3, we estimate error bounds for multi-penalty regularization based on Nyström type subsampling in ψ -weighted norm which consequently provides the convergence rates of the regularized solution $f_{\mathbf{z}, \lambda}^s$ in both

$\|\cdot\|_{\mathcal{H}}$ -norm and $\|\cdot\|_{\rho}$ -norm.

Theorem 3.1. *Let \mathbf{z} be i.i.d. samples drawn according to the probability measure $\rho \in \mathcal{P}_{\phi}$ with the assumption that $t^{-1/2}\psi(t)$, $t^{-1}\phi(t)$, $t^{-1}\phi(t)\psi(t)$ are nonincreasing functions. Then under the parameter choice $\lambda_0 = \Psi^{-1}(m^{-1/2})$ for $\Psi(t) = t^{\frac{1}{2} + \frac{1}{2a}}\phi(t)$ ($a \geq 1$), for sufficiently large sample according to (15) and for subsampling according to (17), the following convergence rates of $f_{\mathbf{z},\lambda}^s$ holds with the confidence $1 - \eta$ for all $0 < \eta < 1$,*

$$\|\psi(L_K)(f_{\mathbf{z},\lambda}^s - f_{\mathcal{H}})\|_{\mathcal{H}} \leq \psi(\lambda_0) \left\{ c_1 \phi(\lambda_0) + c_2 \frac{\mathcal{B}_{\lambda}}{\lambda_0^{3/2}} + c_3 \frac{1}{m\lambda_0} + c_4 \sqrt{\frac{\mathcal{N}(\lambda_0)}{m\lambda_0}} \right\} \log \left(\frac{4}{\eta} \right),$$

where $c_1 = 6R + (5 + c_{\psi})(3 + c_{\phi})R$, $c_2 = (5 + c_{\psi})\|f_{\mathcal{H}}\|_{\rho}$, $c_3 = 8\kappa M$, $c_4 = 8\Sigma$ and $\mathcal{B}_{\lambda} = \|\sum_{j=1}^p \lambda_j B_j^* B_j\|$.

Proof. We discuss the error bound for $\|\psi(L_K)(f_{\mathbf{z},\lambda}^s - f_{\mathcal{H}})\|_{\mathcal{H}}$ by estimating the expressions $\|\psi(L_K)(f_{\mathbf{z},\lambda}^s - f_{\lambda}^s)\|_{\mathcal{H}}$ and $\|\psi(L_K)(f_{\lambda}^s - f_{\mathcal{H}})\|_{\mathcal{H}}$. The first term can be expressed as

$$\begin{aligned} f_{\mathbf{z},\lambda}^s - f_{\lambda}^s &= (P_{\mathbf{x}_s} S_{\mathbf{x}}^* S_{\mathbf{x}} P_{\mathbf{x}_s} + \lambda_0 I + \sum_{j=1}^p \lambda_j P_{\mathbf{x}_s} B_j^* B_j P_{\mathbf{x}_s})^{-1} \{P_{\mathbf{x}_s} S_{\mathbf{x}}^* \mathbf{y} - P_{\mathbf{x}_s} S_{\mathbf{x}}^* S_{\mathbf{x}} P_{\mathbf{x}_s} f_{\mathcal{H}} \\ &\quad + (P_{\mathbf{x}_s} S_{\mathbf{x}}^* S_{\mathbf{x}} P_{\mathbf{x}_s} - P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s})(f_{\mathcal{H}} - f_{\lambda}^s)\} \end{aligned}$$

which implies

$$\|\psi(L_K)(f_{\mathbf{z},\lambda}^s - f_{\lambda}^s)\|_{\mathcal{H}} \leq \frac{\psi(\lambda_0)}{\sqrt{\lambda_0}} I_1 \left\{ I_2 + \frac{1}{\sqrt{\lambda_0}} (I_3 + I_4 \|f_{\mathcal{H}} - f_{\lambda}^s\|_{\mathcal{H}}) \right\},$$

where $I_1 = \|(L_K + \lambda_0 I)^{1/2} (P_{\mathbf{x}_s} S_{\mathbf{x}}^* S_{\mathbf{x}} P_{\mathbf{x}_s} + \lambda_0 I + \sum_{j=1}^p \lambda_j P_{\mathbf{x}_s} B_j^* B_j P_{\mathbf{x}_s})^{-1} (L_K + \lambda_0 I)^{1/2}\|_{\mathcal{L}(\mathcal{H})}$, $I_2 = \|(L_K + \lambda_0 I)^{-1/2} (P_{\mathbf{x}_s} S_{\mathbf{x}}^* \mathbf{y} - P_{\mathbf{x}_s} S_{\mathbf{x}}^* S_{\mathbf{x}} P_{\mathbf{x}_s} f_{\mathcal{H}})\|_{\mathcal{H}}$, $I_3 = \|P_{\mathbf{x}_s} S_{\mathbf{x}}^* S_{\mathbf{x}} (I - P_{\mathbf{x}_s}) f_{\mathcal{H}}\|_{\mathcal{H}}$ and $I_4 = \|S_{\mathbf{x}}^* S_{\mathbf{x}} - L_K\|_{\mathcal{L}(\mathcal{H})}$.

The estimates of I_2 and I_4 can be obtained from Proposition 3.2. Under the condition (15) using the second estimate of Proposition 3.2, we obtain

$$Tr((P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} + \lambda_0 I)^{-1} (P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} - P_{\mathbf{x}_s} S_{\mathbf{x}}^* S_{\mathbf{x}} P_{\mathbf{x}_s})) \leq \frac{I_4}{\lambda_0} \leq \frac{4\kappa^2}{\sqrt{m}\lambda_0} \log \left(\frac{4}{\eta} \right) \leq \frac{1}{2}$$

and under the norm inequalities $\|A\|_{\mathcal{L}(\mathcal{H})} \leq Tr(|A|)$, $Tr(AB) \leq Tr(A)\|B\|_{\mathcal{L}(\mathcal{H})}$ which implies

$$\begin{aligned} I_1 &\leq \|(L_K + \lambda_0 I)^{1/2} (P_{\mathbf{x}_s} S_{\mathbf{x}}^* S_{\mathbf{x}} P_{\mathbf{x}_s} + \lambda_0 I)^{-1} (L_K + \lambda_0 I)^{1/2}\|_{\mathcal{L}(\mathcal{H})} \\ &\leq Tr((P_{\mathbf{x}_s} S_{\mathbf{x}}^* S_{\mathbf{x}} P_{\mathbf{x}_s} + \lambda_0 I)^{-1} (L_K + \lambda_0 I)) \\ &= Tr((P_{\mathbf{x}_s} S_{\mathbf{x}}^* S_{\mathbf{x}} P_{\mathbf{x}_s} + \lambda_0 I)^{-1} ((I - P_{\mathbf{x}_s}) L_K + P_{\mathbf{x}_s} L_K (I - P_{\mathbf{x}_s}))) \\ &\quad + Tr((P_{\mathbf{x}_s} S_{\mathbf{x}}^* S_{\mathbf{x}} P_{\mathbf{x}_s} + \lambda_0 I)^{-1} (P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} + \lambda_0 I)) \\ &\leq \frac{2}{\lambda_0} \|L_K (I - P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})} + Tr(\{I - (P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} + \lambda_0 I)^{-1} (P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} - P_{\mathbf{x}_s} S_{\mathbf{x}}^* S_{\mathbf{x}} P_{\mathbf{x}_s})\}^{-1}) \leq 4. \end{aligned}$$

Under the smoothness assumption $f_{\mathcal{H}} \in \Omega_{\phi,R}$ there exists $g \in \mathcal{H}$ such that $f_{\mathcal{H}} = \phi(L_K)g$ and $\|g\|_{\mathcal{H}} \leq R$.

$$\begin{aligned} I_3 = \|P_{\mathbf{x}_s} S_{\mathbf{x}}^* S_{\mathbf{x}} (I - P_{\mathbf{x}_s}) f_{\mathcal{H}}\|_{\mathcal{H}} &\leq R \|P_{\mathbf{x}_s} S_{\mathbf{x}}^* S_{\mathbf{x}} (I - P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})} \|(I - P_{\mathbf{x}_s}) \phi(L_K)\|_{\mathcal{L}(\mathcal{H})} \\ &\leq R \phi(\|L_K^{1/2} (I - P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})}^2) (\|S_{\mathbf{x}}^* S_{\mathbf{x}} - L_K\|_{\mathcal{L}(\mathcal{H})} + \|L_K (I - P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})}) \\ &\leq R \phi(\|L_K^{1/2} (I - P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})}^2) \left(\frac{4\kappa^2}{\sqrt{m}} \log \left(\frac{4}{\eta} \right) + \|L_K^{1/2} (I - P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})}^2 \right). \end{aligned}$$

Using the conditions (15) and (17), we get

$$I_3 \leq \frac{3}{2} R \lambda_0 \phi(\lambda_0).$$

Therefore,

$$\|\psi(L_K)(f_{\mathbf{z},\lambda}^s - f_\lambda^s)\|_{\mathcal{H}} \leq 2\psi(\lambda_0) \left\{ 2I_2/\sqrt{\lambda_0} + 3R\phi(\lambda_0) + \|f_\lambda^s - f_{\mathcal{H}}\|_{\mathcal{H}} \right\}. \quad (18)$$

For the operator monotone function ψ , we consider the error term:

$$\|\psi(L_K)(f_\lambda^s - f_{\mathcal{H}})\|_{\mathcal{H}} \leq (I_5 + I_6 + I_7) \|f_\lambda^s - f_{\mathcal{H}}\|_{\mathcal{H}} + \|\psi(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s})(f_\lambda^s - f_{\mathcal{H}})\|_{\mathcal{H}},$$

where $I_5 = \|(I - P_{\mathbf{x}_s})\psi(L_K)\|_{\mathcal{L}(\mathcal{H})}$, $I_6 = \|P_{\mathbf{x}_s}\psi(L_K) - P_{\mathbf{x}_s}\psi(L_K)P_{\mathbf{x}_s}\|_{\mathcal{L}(\mathcal{H})}$ and $I_7 = \|P_{\mathbf{x}_s}\psi(L_K)P_{\mathbf{x}_s} - \psi(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})}$.

Hence,

$$\|\psi(L_K)(f_\lambda^s - f_{\mathcal{H}})\|_{\mathcal{H}} \leq (2 + c_\psi)\psi(\lambda_0) \|f_\lambda^s - f_{\mathcal{H}}\|_{\mathcal{H}} + \|\psi(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s})(f_\lambda^s - f_{\mathcal{H}})\|_{\mathcal{H}}. \quad (19)$$

We decompose the term $f_\lambda^s - f_{\mathcal{H}}$ into three parts $f_\lambda^s - f_{\lambda_0}^s$, $f_{\lambda_0}^s - P_{\mathbf{x}_s} f_{\mathcal{H}}$ and $P_{\mathbf{x}_s} f_{\mathcal{H}} - f_{\mathcal{H}}$. Then the expression

$$f_\lambda^s - f_{\lambda_0}^s = -(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} + \lambda_0 I + \sum_{j=1}^p \lambda_j P_{\mathbf{x}_s} B_j^* B_j P_{\mathbf{x}_s})^{-1} \sum_{j=1}^p \lambda_j P_{\mathbf{x}_s} B_j^* B_j P_{\mathbf{x}_s} f_{\lambda_0}^s$$

implies that

$$\|f_\lambda^s - f_{\lambda_0}^s\|_{\mathcal{H}} \leq \frac{\mathcal{B}_\lambda}{\lambda_0} \|f_{\lambda_0}^s\|_{\mathcal{H}} \leq \frac{\mathcal{B}_\lambda}{\lambda_0^{3/2}} \|f_{\mathcal{H}}\|_{\rho}$$

and

$$\|\psi(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s})(f_\lambda^s - f_{\lambda_0}^s)\|_{\mathcal{H}} \leq \frac{\mathcal{B}_\lambda}{\sqrt{\lambda_0}} I_8 I_9 \|f_{\lambda_0}^s\|_{\mathcal{H}} \leq \frac{\mathcal{B}_\lambda \psi(\lambda_0)}{\lambda_0^{3/2}} \|f_{\mathcal{H}}\|_{\rho},$$

where $\mathcal{B}_\lambda = \left\| \sum_{j=1}^p \lambda_j B_j^* B_j \right\|$, $I_8 = \|\psi(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s})(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} + \lambda_0 I)^{-1/2}\|_{\mathcal{L}(\mathcal{H})}$ and $I_9 = \|(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} + \lambda_0 I)^{1/2} (P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} + \lambda_0 I + \sum_{j=1}^p \lambda_j P_{\mathbf{x}_s} B_j^* B_j P_{\mathbf{x}_s})^{-1} (P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} + \lambda_0 I)^{1/2}\|_{\mathcal{L}(\mathcal{H})}$.

The expression

$$f_{\lambda_0}^s - P_{\mathbf{x}_s} f_{\mathcal{H}} = \{(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} + \lambda_0 I)^{-1} P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} - I\} P_{\mathbf{x}_s} \phi(L_K) g$$

gives that

$$\|f_{\lambda_0}^s - P_{\mathbf{x}_s} f_{\mathcal{H}}\|_{\mathcal{H}} \leq R(I_{10} + I_{11} + I_{12})$$

and

$$\|\psi(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s})(f_{\lambda_0}^s - P_{\mathbf{x}_s} f_{\mathcal{H}})\|_{\mathcal{H}} \leq R(I_{10} I_{13} + I_{11} I_{13} + I_{14}),$$

where $I_{10} = \|P_{\mathbf{x}_s} \phi(L_K)(I - P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})}$, $I_{11} = \|P_{\mathbf{x}_s} \phi(L_K)P_{\mathbf{x}_s} - \phi(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})}$, $I_{12} = \|\{(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} + \lambda_0 I)^{-1} P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} - I\} \phi(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})}$, $I_{13} = \|\psi(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s})\{(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} + \lambda_0 I)^{-1} P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} - I\}\|_{\mathcal{L}(\mathcal{H})}$ and $I_{14} = \|\psi(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s})\{(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} + \lambda_0 I)^{-1} P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s} - I\} \phi(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})}$.

Again using the conditions on ψ and ϕ , we get

$$\|f_{\lambda_0}^s - P_{\mathbf{x}_s} f_{\mathcal{H}}\|_{\mathcal{H}} \leq R(2 + c_\phi)\phi(\lambda_0)$$

and

$$\|\psi(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s})(f_{\lambda_0}^s - P_{\mathbf{x}_s} f_{\mathcal{H}})\|_{\mathcal{H}} \leq R(2 + c_{\phi})\psi(\lambda_0)\phi(\lambda_0).$$

We also have,

$$\|P_{\mathbf{x}_s} f_{\mathcal{H}} - f_{\mathcal{H}}\|_{\mathcal{H}} \leq R\|(I - P_{\mathbf{x}_s})\phi(L_K)\|_{\mathcal{L}(\mathcal{H})} \leq R\phi(\|L_K^{1/2}(I - P_{\mathbf{x}_s})\|_{\mathcal{L}(\mathcal{H})}^2) \leq R\phi(\lambda_0)$$

and

$$\|\psi(L_K)(P_{\mathbf{x}_s} f_{\mathcal{H}} - f_{\mathcal{H}})\|_{\mathcal{H}} \leq R\psi(\lambda_0)\phi(\lambda_0).$$

Hence we obtain,

$$\|f_{\lambda}^s - f_{\mathcal{H}}\|_{\mathcal{H}} \leq R(3 + c_{\phi})\phi(\lambda_0) + \frac{\mathcal{B}_{\lambda}}{\lambda_0^{3/2}}\|f_{\mathcal{H}}\|_{\rho} \quad (20)$$

and

$$\|\psi(P_{\mathbf{x}_s} L_K P_{\mathbf{x}_s})(f_{\lambda}^s - f_{\mathcal{H}})\|_{\mathcal{H}} \leq \psi(\lambda_0) \left\{ R(3 + c_{\phi})\phi(\lambda_0) + \frac{\mathcal{B}_{\lambda}}{\lambda_0^{3/2}}\|f_{\mathcal{H}}\|_{\rho} \right\}. \quad (21)$$

Combining the bounds (20), (21) with inequalities (18) and (19) we obtain the desired result. \square

In Theorem 3.1, the error estimates reveal the interesting fact that the error terms consist of increasing and decreasing function of α which led to propose a posteriori choice of regularization parameter α based on balancing principle. Hence the effective dimension plays the crucial role in the error analysis of regularized learning algorithms.

Here the upper convergence rates of the regularized solution $f_{\mathbf{z},\lambda}$ are derived from the estimates of Theorem 3.1 for the class of probability measure P_{ϕ} , $P_{\phi,b}$, respectively. In Theorem 3.2, 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 3.3 we obtain the optimal minimax convergence rates in terms of index function ϕ , the parameter b and the number of samples m .

Theorem 3.2. *Under the same assumptions of Theorem 3.1 with the parameter choice $\lambda_0 \in (0, 1]$, $\lambda_0 = \Theta^{-1}(m^{-1/2})$, $\lambda_j = (\Theta^{-1}(m^{-1/2}))^{3/2}\phi(\Theta^{-1}(m^{-1/2}))$ for $1 \leq j \leq p$, where $\Theta(t) = t\phi(t)$, the convergence rates of $f_{\mathbf{z},\lambda}^s$ can be described as follows:*

$$\text{Prob}_{\mathbf{z} \in Z^m} \left\{ \|\psi(L_K)(f_{\mathbf{z},\lambda}^s - f_{\mathcal{H}})\|_{\mathcal{H}} \leq C\psi(\Theta^{-1}(m^{-1/2}))\phi(\Theta^{-1}(m^{-1/2})) \log\left(\frac{4}{\eta}\right) \right\} \geq 1 - \eta.$$

Proof. Let $\Theta(t) = t\phi(t)$. Then it follows,

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

Under the parameter choice $\lambda_0 = \Theta^{-1}(m^{-1/2})$ we have,

$$\lim_{m \rightarrow \infty} m\lambda_0 = \infty.$$

Therefore for sufficiently large m , we get $m\lambda_0 \geq 1$ and

$$\frac{1}{m\lambda_0} = \frac{\lambda_0^{1/2}\phi(\lambda_0)}{\sqrt{m\lambda_0}} \leq \lambda_0^{1/2}\phi(\lambda_0).$$

Under the parameter choice $\lambda_0 \leq 1$, $\lambda_0 = \Theta^{-1}(m^{-1/2})$, $\lambda_j = (\Theta^{-1}(m^{-1/2}))^{3/2}\phi(\Theta^{-1}(m^{-1/2}))$ for $1 \leq j \leq$

p , from Theorem 3.1 follows that with the confidence $1 - \eta$,

$$\|\psi(L_K)(f_{\mathbf{z},\lambda}^s - f_{\mathcal{H}})\|_{\mathcal{H}} \leq C\psi(\Theta^{-1}(m^{-1/2}))\phi(\Theta^{-1}(m^{-1/2}))\log\left(\frac{4}{\eta}\right).$$

Hence our conclusion follows. \square

Theorem 3.3. *Under the same assumptions of Theorem 3.1 and Assumption 2.3 with 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 $\Psi(t) = t^{\frac{1}{2} + \frac{1}{2b}}\phi(t)$, the convergence rates of $f_{\mathbf{z},\lambda}^s$ can be described as follows:*

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

where $C' = R'(3R + 4\kappa M + 4\sqrt{\beta b \Sigma^2/(b-1)} + 3 \sum_{j=1}^p \|B_j^* B_j\| \|f_{\mathcal{H}}\|_{\rho})$.

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

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

Under the parameter choice $\lambda_0 = \Psi^{-1}(m^{-1/2})$ we have,

$$\lim_{m \rightarrow \infty} m\lambda_0 = \infty.$$

Therefore for sufficiently large m , we get $m\lambda_0 \geq 1$ and

$$\frac{1}{m\lambda_0} = \frac{\lambda_0^{\frac{1}{2b}}\phi(\lambda_0)}{\sqrt{m\lambda_0}} \leq \lambda_0^{\frac{1}{2b}}\phi(\lambda_0).$$

Under the parameter choice $\lambda_0 \leq 1$, $\lambda_0 = \Psi^{-1}(m^{-1/2})$, $\lambda_j = (\Psi^{-1}(m^{-1/2}))^{3/2}\phi(\Psi^{-1}(m^{-1/2}))$ for $1 \leq j \leq p$, from Theorem 3.1 and eqn. (11) follows that with the confidence $1 - \eta$,

$$\|\psi(L_K)(f_{\mathbf{z},\lambda}^s - f_{\mathcal{H}})\|_{\mathcal{H}} \leq C'\psi(\Psi^{-1}(m^{-1/2}))\phi(\Psi^{-1}(m^{-1/2}))\log\left(\frac{4}{\eta}\right). \quad (22)$$

Hence our conclusion follows. \square

In Theorem 3.4, 3.5, we present the convergence rates of the multi-penalty estimator $f_{\mathbf{z},\lambda}^s$ for the classes of probability measures \mathcal{P}_{ϕ} and $\mathcal{P}_{\phi,b}$ in both RKHS-norm and \mathcal{L}^2 -norm. For $\psi(t) = t^{\alpha}$, we can also obtain the convergence rates of the regularized solution $f_{\mathbf{z},\lambda}^s$ in the interpolation norm for the parameter $\alpha \in [0, \frac{1}{2}]$. In particular, we obtain the error estimates in $\|\cdot\|_{\mathcal{H}}$ -norm for $\alpha = 0$ and in $\|\cdot\|_{L_{\rho_X}^2}$ -norm for $\alpha = \frac{1}{2}$.

Theorem 3.4. *Let \mathbf{z} be i.i.d. samples drawn according to the probability measure $\rho \in \mathcal{P}_{\phi}$. Then for sufficiently large sample size m according to (15) and for subsampling according to (17) under the parameter choice $\lambda_0 \in (0, 1]$, $\lambda_0 = \Theta^{-1}(m^{-1/2})$, $\lambda_j = (\Theta^{-1}(m^{-1/2}))^{3/2}\phi(\Theta^{-1}(m^{-1/2}))$ for $1 \leq j \leq p$, where $\Theta(t) = t\phi(t)$, 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,*

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

and

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

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

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

and

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

Theorem 3.5. Let \mathbf{z} be i.i.d. samples drawn according to the probability measure $\rho \in \mathcal{P}_{\phi,b}$. Then for sufficiently large sample size m according to (15) and for subsampling according to (17) 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,

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

and

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

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

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

and

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

The lower convergence rates are discussed for any learning algorithm over the class of the probability measures $\mathcal{P}_{\phi,b}$ in Theorem 3.10, 3.12 [?]. Here we study the upper convergence rates for multi-penalty regularization based on Nyström type subsampling in vector-valued function setting. If the upper convergence rate for the parameter choice $\lambda = \lambda(m)$ coincides with the lower convergence rates, then the parameter choice $\lambda = \lambda(m)$ is said to be optimal. For the parameter choice $\lambda = (\lambda_0, \dots, \lambda_p)$, $\lambda_0 = \Psi^{-1}(m^{-1/2})$, $\lambda_j = (\Psi^{-1}(m^{-1/2}))^{3/2} \phi(\Psi^{-1}(m^{-1/2}))$, $1 \leq j \leq p$, Theorem 3.5 share the upper convergence rates with the lower minimax rates of Theorem 3.10, 3.12 [?]. Therefore the choice of the parameter is optimal.

Remark 3.1. Under Hölder source condition ($\phi(t) = t^r$), we get the order of convergence $\mathcal{O}(m^{-\frac{r}{2r+2}})$ (for $0 \leq r \leq 1$) and $\mathcal{O}(m^{-\frac{2r+1}{4r+4}})$ (for $0 \leq r \leq \frac{1}{2}$) for the class of probability measures \mathcal{P}_ϕ in $\|\cdot\|_{\mathcal{H}}$ -norm and $\|\cdot\|_{\mathcal{L}_{\rho_X}^2}$ -norm, respectively. Also, for the class of probability measures $\mathcal{P}_{\phi,b}$, we obtain the order of convergence $\mathcal{O}(m^{-\frac{br}{2br+b+1}})$ (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 $\|\cdot\|_{\mathcal{H}}$ -norm and $\|\cdot\|_{\mathcal{L}_{\rho_X}^2}$ -norm, respectively.

Therefore, the randomized subsampling methods can break the memory barriers of standard kernel methods, while preserving optimal learning guarantees.

4 An aggregation approach for Nyström approximants

The size of sub-sample is described in terms of the integral operator and behavior of its eigenvalues values or the regularity of the target function and the approximation power of the projection operator. In practice, it may be difficult to obtain the appropriate size of the sub-sample. In order to overcome this problem, we discuss an aggregation approach based on linear functional strategy. We construct various Nyström approximants corresponding to different subsampling size. Then the strategy tries to accumulate the information hidden inside various approximants to produce the best estimator of the target function. The approach is widely considered in ill-posed inverse problems [?, ?, ?, ?, ?] as well as learning theory framework [?, ?, ?, ?] to aggregate the various regularized solutions. In linear functional strategy, we consider the linear combination of the approximants and try to figure out the combination which is closure to the target function $f_{\mathcal{H}}$. For a finite set of Nyström approximants $\{f_{\mathbf{z},\lambda}^{s_i} \in \mathcal{H} : 1 \leq i \leq l\}$, the aggregation approach can be described as:

$$\arg \min_{(c_1, \dots, c_l) \in \mathbb{R}^l} \left\| \sum_{i=1}^l c_i f_{\mathbf{z},\lambda}^{s_i} - f_{\mathcal{H}} \right\|_{\rho}. \quad (23)$$

The problem of minimization (23) is equivalent to the problem of finding $\mathbf{c} = (c_1, \dots, c_l)$,

$$H\mathbf{c} = h,$$

where $H = (\langle f_{\mathbf{z},\lambda}^{s_i}, f_{\mathbf{z},\lambda}^{s_j} \rangle_{\rho})_{i,j=1}^l$ and $h = (\langle f_{\mathcal{H}}, f_{\mathbf{z},\lambda}^{s_i} \rangle_{\rho})_{i=1}^l$.

Due to the involvement of the unknown probability distribution of ρ we cannot determine H and h directly. Therefore we approximate H and h by the quantities $\bar{H} = \left(\frac{1}{n} \sum_{r=1}^n f_{\mathbf{z},\lambda}^{s_i}(x_r) f_{\mathbf{z},\lambda}^{s_j}(x_r) \right)_{i,j=1}^l$ and $\bar{h} = \left(\frac{1}{m} \sum_{r=1}^m y_r f_{\mathbf{z},\lambda}^{s_i}(x_r) \right)_{i=1}^l$, respectively.

Now the combination vector $\bar{\mathbf{c}}$ is given by $\bar{\mathbf{c}} = \bar{H}^{-1} \bar{h}$ and the aggregated solution $f_{\mathbf{z}} = \sum_{i=1}^l \bar{c}_i f_{\mathbf{z},\lambda}^{s_i}$, $\bar{\mathbf{c}} = (\bar{c}_i)_{i=1}^l$ based on linear functional strategy in $\mathcal{L}_{\rho_X}^2$ -norm (LFS- $\mathcal{L}_{\rho_X}^2$). The convergence rate of the constructed solution $f_{\mathbf{z}}$ can be described as

Theorem 4.1. *Let \mathbf{z} be i.i.d. samples drawn according to the probability measure ρ with the hypothesis (10). Then for sufficiently large sample size m according to (15) with the confidence $1 - \eta$, we have*

$$\|f_{\mathbf{z}} - f_{\mathcal{H}}\|_{\rho} = \min_{\mathbf{c} \in \mathbb{R}^l} \left\| \sum_{i=1}^l c_i f_{\mathbf{z},\lambda}^{s_i} - f_{\mathcal{H}} \right\|_{\rho} + \mathcal{O} \left(m^{-1/2} \log \left(\frac{4}{\eta} \right) \right)$$

holds with the probability $1 - \eta$.

The proof of the theorem follows the same steps as of Theorem 10 [?]. Here we observe that the individual Nyström approximants can be obtained for the particular values of combination vector. The error term in the Theorem 4.1 goes to 0 for the large sample size. Moreover, the order of the error in Theorem 4.1 is less than the order of error of Nyström approximants in Theorem 3.5. Therefore the error of aggregated solution in Theorem 4.1 is negligible.

5 Numerical realization

In our experiments, we demonstrate the performance of multi-penalty regularization based on Nyström type subsampling using the linear functional strategy for both scalar-valued functions and multi-task learning problems. We present the extensive empirical analysis of the multi-view manifold regularization scheme based on Nyström type subsampling for the challenging multi-class image classification and species recognition with attributes. Then we consider the NSL-KDD benchmark data set from UCI machine learning repository for the intrusion detection problem based on Nyström type subsampling.

5.1 Caltech-101 data set

We consider the multi-view manifold regularization scheme discussed in [?] for multi-class classification on Caltech-101 data set. The regularized solution is constructed based on the different views of the input data. For the given data set, the views are the different features extracted from the input examples. Let $f = (f^1, \dots, f^v) \in \mathcal{H} = \mathcal{H}_{K^1} \times \dots \times \mathcal{H}_{K^v}$ be the function associated to the v -views of the inputs. We define the combination operator $C : Y^v \rightarrow Y$ by $Cf(x) = \sum_{i=1}^v c_i f^i(x)$, for $\mathbf{c} = (c_1, \dots, c_v) \in \mathbb{R}^v$. We consider the multi-view manifold regularization for semi-supervised problem corresponding to the labeled data $\{(x_i, y_i)\}_{i=1}^m$ and unlabeled data $\{x_i\}_{i=m+1}^n$:

$$f_{\mathbf{z}, \lambda} = \arg \min_{f \in \mathcal{H}, \mathbf{c} \in S_{\alpha}^{v-1}} \frac{1}{m} \sum_{i=1}^m \|y_i - Cf(x_i)\|_Y^2 + \lambda_A \|f\|_{\mathcal{H}}^2 + \lambda_B \|(S_{\mathbf{x}'}^* M_B S_{\mathbf{x}'})^{1/2} f\|_{\mathcal{H}}^2 + \lambda_W \|(S_{\mathbf{x}'}^* M_W S_{\mathbf{x}'})^{1/2} f\|_{\mathcal{H}}^2, \quad (24)$$

where the regularization parameters $\lambda_A, \lambda_B, \lambda_W \geq 0$, $\mathbf{x}' = (x_i)_{i=1}^n$, $M_B = I_n \otimes (M_v \otimes I_Y)$ and $M_W = L \otimes I_Y$ are symmetric, positive operators. Here $M_v = vI_v - \mathbf{1}_v \mathbf{1}_v^T$, I_v is identity of size $v \times v$, $\mathbf{1}_v$ is a vector of size $v \times 1$ with ones, \otimes is the Kronecker product and L is a graph Laplacian. The graph Laplacian L is the block matrix of size $n \times n$, with block (i, j) being the $v \times v$ diagonal matrix, given by

$$L_{ij} = \text{diag}(L_{ij}^1, \dots, L_{ij}^v), \quad (25)$$

where the scalar graph Laplacian L^i is induced by the symmetric, nonnegative weight matrix W^i .

The first term controls the complexity of the function in the ambient space, the second term between-view regularization which measures the consistency of the component functions across different views and the third term within-view regularization which measures the smoothness of the component functions in their corresponding views.

Now we consider the Nyström type subsampling on the multi-view learning problem. We restrict the minimization problem (24) over the space:

$$\mathcal{H}^{\mathbf{x}_s} := \{f | f = \sum_{i=1}^s K_{x_i} y_i, \mathbf{y}_s = (y_1, \dots, y_s) \in Y^s\},$$

where $s \ll n$ and $\mathbf{x}_s = (x_1, \dots, x_s)$ is a subset of the input points in the training set and K is the kernel corresponding to the RKHS $\mathcal{H} = \mathcal{H}_{K^1} \times \dots \times \mathcal{H}_{K^v}$. We construct Nyström approximants by minimizing the regularization problem (24) over the space $\mathcal{H}^{\mathbf{x}_s}$.

We have to minimize the multi-view manifold regularization problem simultaneously for $f \in \mathcal{H}^{\mathbf{x}_s}$ and $\mathbf{c} \in \mathbb{R}^v$. So we first choose a weight vector \mathbf{c} on the sphere $S_{\alpha}^{v-1} = \{x \in \mathbb{R}^v : \|x\| = \alpha\}$ and minimize for the function f over $\mathcal{H}^{\mathbf{x}_s}$. Then we fix the estimator f and optimize for the weight vector \mathbf{c} . We continue

the iterative process until we get the desired accuracy. The regularized solution of the scheme (24) is constructed according to the Theorem 10 [?].

Algorithm 1 Semi-supervised least-square regression and classification based on Nyström subsampling using the aggregation approach

This algorithm computes the multi-view learning estimators $f_{\mathbf{z},\lambda}^{s_r}$ corresponding to the views (v_1, \dots, v_q) for the subsampling s_r ($1 \leq r \leq l$). Then the algorithm computes the aggregated solution $f_{\mathbf{z}}$ from the regularized solutions $f_{\mathbf{z},\lambda}^{s_r}$ ($1 \leq r \leq l$) based on the linear functional strategy in $\mathcal{L}_{\rho_X}^2$.

Input:

- Training data $\{(x_i, y_i)\}_{i=1}^m \cup \{x_i\}_{i=m+1}^n$, with m labeled and $n - m$ unlabeled examples.
- Testing data t_i .

Parameters:

- The regularization parameters $\lambda_A, \lambda_B, \lambda_W$.
- The weight vector \mathbf{c} .
- The number of classes P .
- Scalar-valued kernels K^i corresponding to i -th view.

Procedure:

- To calculate the set of estimators $f_{\mathbf{z},\lambda}^{s_r}$ ($1 \leq r \leq l$), compute kernel matrices $G_r[\mathbf{x}', \mathbf{x}^s] = (G_r(x_i, x_j))_{ij}$ for $1 \leq i \leq n$ and $1 \leq j \leq s$ corresponding to views (v_1^r, \dots, v_q^r) according to (26).
- Compute graph Laplacian L according to (25).
- Compute $B_r = ((J_m^n \otimes \mathbf{c}\mathbf{c}^T) + m\lambda_B(I_n \otimes M_v) + m\lambda_W L) G_r[\mathbf{x}', \mathbf{x}^s]$, where $J_m^n : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a diagonal matrix of size $n \times n$, with the first m entries on the main diagonal being 1 and the rest being 0.
- Compute $C = \mathbf{c}^T \otimes I_P$ and $\mathbf{C}^* = I_{n \times m} \otimes C^*$ for $I_{n \times m} = [I_m, 0_{m \times (n-m)}]^T$.
- Compute Y_C such that $\mathbf{C}^* \mathbf{y} = \text{vec}(Y_C^T)$, where the vectorization of an $n \times P$ matrix A , denoted $\text{vec}(A)$, is the $nP \times 1$ column vector obtained by stacking the columns of the matrix A on top of one another.
- Solve matrix equations $B_r A_r + m\lambda_A A_r = Y_C$ for A_r ($1 \leq r \leq l$).
- Compute kernel matrices $G_r[\mathbf{x}, \mathbf{x}'] = (G_r(x_i, x_j))_{ij}$ for $1 \leq i \leq m$ and $1 \leq j \leq n$.
- Compute $\bar{H}_{rq} = \frac{1}{n} \text{vec}(A_r^T G_r[\mathbf{x}']^T (I_{n \times m} \otimes \mathbf{c}))^T \text{vec}(A_q^T G_q[\mathbf{x}']^T (I_{n \times m} \otimes \mathbf{c}))$ and $\bar{h}_r = \frac{1}{m} \mathbf{y}^T \text{vec}(A_r^T G_r[\mathbf{x}, \mathbf{x}']^T (I_m \otimes \mathbf{c}))$ for $1 \leq r, q \leq l$.
- Compute $\bar{c} = (\bar{c}_1, \dots, \bar{c}_l)^T = \bar{H}^{-1} \bar{h}$ for $\bar{H} = (\bar{H}_{rq})_{r,q=1}^l$ and $\bar{h} = (\bar{h}_r)_{r=1}^l$.
- Compute kernel matrices $G_r[t_i, \mathbf{x}]$ between t_i and \mathbf{x} .
- Compute the value of estimator $f_{\mathbf{z}}$ based on aggregation approach at t_i , i.e.,

$$f_{\mathbf{z}}(t_i) = \sum_{r=1}^l \bar{c}_r f_{\mathbf{z},\lambda}^{K_r}(t_i) = \sum_{r=1}^l \bar{c}_r \text{vec}(A_r^T G_r[t_i, \mathbf{x}]^T).$$

Output: Multi-class classification: return index of $\max(C f_{\mathbf{z}}(t_i))$.

We demonstrate the performance of multi-view manifold regularization problem (24) based on Nyström type subsampling on Caltech-101 data set for the challenging task of multi-class image classification. Caltech-101 data set is used for object recognition problem which is provided in Fei-Fei et al. [?]. The data set contains $P = 102$ classes of objects and each class having 40 to 800 images. We have chosen 15 images randomly from each class. PHOW gray, PHOW color, geometric blurred and self-symmetry are the four

Estimators	Level-1 views	Level-2 views	Level-3 views
MVL-NS (s=2)	41.31%	42.55%	42.55%
MVL-NS (s=5)	57.71%	62.90%	63.99%
LFS-MVL-NS	57.80%	62.85%	64.27%
MVL-LS-SP	57.32%	62.11%	63.68%
MVL-LS	57.67%	63.09%	64.20%

Table 2: Performance of multi-view learning estimators based on Nyström subsampling and the aggregated solution based on LFS- $\mathcal{L}_{\rho_X}^2$ using 5 labeled and 5 unlabeled images per class from Caltech-101 data set.

Estimators	Level-1 views	Level-2 views	Level-3 views
MVL-NS (s=3)	53.31%	56.10%	56.82%
MVL-NS (s=10)	64.42%	69.74%	71.09%
LFS-MVL-NS	64.64%	69.87%	71.09%
MVL-LS-SP	64.51%	69.30%	70.68%
MVL-LS	64.36%	69.54%	70.98%

Table 3: Performance of multi-view learning estimators based on Nyström subsampling and the aggregated solution based on LFS- $\mathcal{L}_{\rho_X}^2$ using 10 labeled and 5 unlabeled images per class from Caltech-101 data set.

Estimators	Level-1 views	Level-2 views	Level-3 views
MVL-NS-optC (s=2)	35.90%	35.64%	34.90%
MVL-NS-optC (s=5)	61.72%	64.29%	64.77%
LFS-MVL-NS-optC	61.81%	64.40%	64.90%
MVL-LS-optC	60.33%	64.62%	65.16%

Table 4: Performance of multi-view learning estimators based on Nyström subsampling and the aggregated solution based on LFS- $\mathcal{L}_{\rho_X}^2$ with optimal combination operator using 5 labeled and 5 unlabeled images per class from Caltech-101 data set.

Estimators	Level-1 views	Level-2 views	Level-3 views
MVL-NS-optC (s=3)	45.93%	47.23%	48.54%
MVL-NS-optC (s=10)	67.06%	70.09%	71.57%
LFS-MVL-NS-optC	67.15%	70.24%	71.61%
MVL-LS-optC	65.90%	70.39%	71.20%

Table 5: Performance of multi-view learning estimators based on Nyström subsampling and the aggregated solution based on LFS- $\mathcal{L}_{\rho_X}^2$ with optimal combination operator using 10 labeled and 5 unlabeled images per class from Caltech-101 data set.

features considered by Vedaldi et al. [?] which are extracted on three levels of the spatial pyramid. We use the χ^2 -kernels corresponding to each view x^i of the input $x = (x^1, \dots, x^v)$ provided in [?]. We define the operator-valued kernel for the manifold learning scheme (24):

$$K(x, t) = G(x, t) \otimes I_P \text{ for } G(x, t) = \sum_{i=1}^v K^i(x^i, t^i) e_i e_i^T, \quad (26)$$

where K^i is scalar-valued kernel defined on i -th view and $e_i = (0, \dots, 1, \dots, 0) \in \mathbb{R}^v$ is the i -th coordinate vector. Three splits of the given data are considered and the results are reported in terms of accuracy by averaging over all three splits. The test set contains 15 objects per class in all experiments. We set $y = (-1, \dots, 1, \dots, -1) \in Y = \mathbb{R}^P$, i.e., 1 on the k -th place otherwise -1 if x belongs to the k -th class for the output values.

Here we take the number of labeled data per class $l = \{5, 10\}$ and the number of unlabeled data per class $u = 5$. In the results of Table 2, 3, 4 & 5, we use the same choice of regularization parameters $\lambda_A = 10^{-5}$, $\lambda_B = 10^{-6}$, $\lambda_W = 10^{-6}$ as considered in Minh et al. [?]. We have chosen the uniform combination operator $\mathbf{c} = \frac{1}{v}(1, \dots, 1)^T \in \mathbb{R}^v$ in the results of Table 2 & 3. MVL-NS represents the multi-view estimator based on Nyström subsampling with subsampling size s . Level-1 views represents the upper level of views, i.e., PHOW color and gray ℓ_0 , SSIM ℓ_0 , GB. Similarly, Level-2 views and Level-3 views represent the middle and lower level of feature’s spatial pyramid. LFS-MVL-NS is the aggregated estimator based on LFS- $\mathcal{L}_{\rho_X}^2$ using Nyström approximants (MVL-NS). MVL-SP is the single-penalty multi-view learning estimator corresponding to the problem (24), i.e. for $\lambda_B = 0$ & $\lambda_W = 0$.

We present the performance of multi-view estimators based on Nyström type subsampling using the feature on different levels of spatial pyramid. In Table 2, 3, we observe that on the aggregating the Nyström approximants based on linear functional strategy we are able to achieve the accuracy of multi-view learning estimator (MVL-LS) of the problem (24). We also show the accuracy of the single-penalty regularizer (MVL-LS-SP). In Table 4, 5, we go one step further by choosing the optimal combination operator \mathbf{c} . We optimize the combination operator over the sphere S_{α}^{v-1} for the fixed function f . To obtain the optimal weight vector \mathbf{c} , we created a validation set by selecting 5 examples for each class from the training set. The validation set is used to determine the best value of \mathbf{c} found over all the iterations using different initializations. We iterate 25 times in the optimization procedure of \mathbf{c} . We fixed $\|\alpha\| = 1$ in the optimization problem of combination operator. The optimal choice of combination operator is powerful with clear improvements in classification accuracies over the uniform weight approach.

The results demonstrate that the Nyström type subsampling can be effectively used in order to reduce the complexity of kernel methods. The aggregation approach automatically produces the best approximant. In all the cases shown in the tables, we obtain better results by the aggregation approach compared to the Nyström approximants.

5.2 NSL-KDD data set

We consider the NSL-KDD benchmark data set [?] from UCI machine learning repository for an empirical analysis of Nyström type subsampling. NSL-KDD data set is the refined version of KDD Cup 99 data set which is used in the 3rd International Knowledge Discovery and Data Mining Tools Competition for intrusion detection. The training set of NSL-KDD data set does not include redundant records and contains reasonable number of records to run the experiments on the complete set. The NSL-KDD data contains 41 attributes and 5 classes that are normal and 4 categories of attacks: DoS, Probe, R2L and U2R. Denial of Service Attack (DoS) is an attack in which the attacker attempts to block system or network resources and services. In Probing Attack, the attacker attempts to gain the information about potential vulnerabilities of a network of computers for the apparent purpose of circumventing its security controls. User to Root Attack (U2R) is a class of exploit in which the attacker access the system as a normal user and break the vulnerabilities to gain administrative privileges. Remote to Local Attack (R2L) occurs when an attacker who has unauthorized ability to dump data packets to remote system over network and exploits some vulnerability to gain access either as a user or root to do their unauthorized activity.

The features in NSL-KDD data set can be classified into three groups: (a) **the basic input features** encapsulate all the attributes that can be extracted from a TCP/IP connection. It includes some flags in TCP connections, duration, prototype, number of bytes from source IP addresses or from destination IP addresses and service, (b) **the content input features** use domain knowledge to assess the payload of the original TCP packets. It includes features such as the number of failed login attempts and (c) **the**

statistical input features that are determined either by a time window or a window of certain kind of connections. It examines only the connections in the past 2 seconds that have the same destination host or service as the current connection.

In our experiment, we used first 25000 patterns from “20 Percent Training Set” available on [?]. NSL-KDD data set contains numeric and nonnumeric attributes. First, we convert the nonnumeric attributes: protocol_type, service and flag as numeric attributes. We represent the corresponding values of individual strings “tcp”, “udp”, “icmp” for protocol name as 0, 1, 2, respectively. All the remaining attributes are also represented according to the above encoding system. We removed the two columns of attribute which contain only zero values. The attributes of the input data are normalized to the interval $[0, 1]$ (see [?]).

We consider a multi-penalty regularization scheme based on Nyström type subsampling which can be viewed as a special case of proposed problem (4),

$$\arg \min_{f \in \mathcal{H}^{\mathbf{x}_s}} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda_0 \|f\|_{\mathcal{H}}^2 + \lambda_1 \|(S_{\mathbf{x}}^* L S_{\mathbf{x}})^{1/2} f\|_{\mathcal{H}}^2 \right\},$$

where the Laplacian $L = D - W$ with $W = (\omega_{ij})$ is a weight matrix with non-negative entries and D is a diagonal matrix with $D_{ii} = \sum_{j=1}^m \omega_{ij}$. We apply this scheme on NSL-KDD data set for binary classification among attacks and normal situations. We demonstrate the performance of single-penalty regularization ($\lambda_1 = 0$) versus multi-penalty regularization and also describe the efficiency of proposed regularization algorithm statistically using the standard error measures. All regularized solutions appearing in this experiment are constructed in the reproducing kernel Hilbert space corresponding to the Gaussian 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 the regularization parameters for single-penalty regularization according to the balancing principle [?] and for multi-penalty regularization according to the balanced-discrepancy principle [?].

The performance of the proposed approach is evaluated using various performance measures. To measure the performance of learning classifiers we need to know the terms: True Positive TP (the number of correctly classified positive instances), False Negative FN (the number of misclassified positive instances), False Positive FP (the number of misclassified negative instances) and True Negative TN (the number of correctly classified negative instances). Standard performance measures: Classification accuracy, Precision, Sensitivity, Specificity and F-measure can be defined as:

$$\text{Classification accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

$$\text{Sensitivity (True positive rate)} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{Specificity (True negative rate)} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

and

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}}$$

We use the 10-fold cross validation which divides 25000 patterns of the considered data set into 10 sub-data sets of size 2500. We train our algorithm on first 9 sub-data sets and test on the last sub-data

set. In our experiment, we illustrate the performance of multi-penalty regularized solutions $f_{\mathbf{z},\lambda}^{10}$, $f_{\mathbf{z},\lambda}^{50}$, $f_{\mathbf{z},\lambda}^{250}$ corresponding to subsampling size $s = 10, 50, 250$ and their aggregated solution $f_{\mathbf{z}}$ for $|\mathbf{z}| = 2500$. We also compare the performance of these estimators with single-penalty regularized solution $f_{\mathbf{z},\lambda_0}$ and multi-penalty regularized solution $f_{\mathbf{z},\lambda}$. In the context of binary classification, the output y_i takes only two values, designated by 1 for attack and -1 for normal. For the regularized solutions we consider the decision rule/classifier $\{y = 1 \text{ for } f(x) \geq 0 \text{ and } y = -1 \text{ for } f(x) < 0\}$ in discriminating the elements \mathbf{x} of two classes. In the experiments, the initial parameters are $\lambda_0 = 10^{-8}$, $\lambda_1 = 1$, the kernel parameter $\gamma = 4 \times 10^{-2}$ and the weight parameter $b = 10^{-3}$.

Estimators	Fold1 (%)	Fold2 (%)	Fold3 (%)	Fold4 (%)	Fold5 (%)	Fold6 (%)	Fold7 (%)	Fold8 (%)	Fold9 (%)
$f_{\mathbf{z},\lambda_0}$	97.80	83.48	98.56	98.56	98.44	98.40	80.20	98.64	98.12
$f_{\mathbf{z},\lambda}$	98.96	98.64	98.48	98.48	98.36	98.44	98.84	98.84	98.08
$f_{\mathbf{z},\lambda}^{10}$	92.52 (0.16)	92.48 (0.18)	92.57 (0.20)	92.64 (0.20)	92.47 (0.15)	92.47 (0.20)	92.65 (0.15)	92.06 (0.23)	92.76 (0.17)
$f_{\mathbf{z},\lambda}^{50}$	95.79 (0.07)	96.45 (0.05)	96.19 (0.06)	96.48 (0.07)	95.88 (0.06)	95.92 (0.06)	96.17 (0.08)	96.15 (0.07)	95.84 (0.07)
$f_{\mathbf{z},\lambda}^{250}$	98.33 (0.02)	98.31 (0.03)	98.03 (0.02)	98.36 (0.02)	98.41 (0.03)	98.18 (0.03)	98.58 (0.02)	98.06 (0.03)	98.56 (0.03)
$f_{\mathbf{z}}$	98.33 (0.02)	98.32 (0.03)	98.03 (0.02)	98.37 (0.02)	98.42 (0.02)	98.19 (0.03)	98.60 (0.02)	98.06 (0.03)	98.57 (0.02)

Table 6: Statistical performance of various estimators on different sub-data sets (Folds) of NSL-KDD data set using random subsampling 50 times.

Estimators	Accuracy (%)	Precision	Sensitivity	Specificity	F-measure	Parameter choice
$f_{\mathbf{z},\lambda_0}$	94.32 (3.79)	0.93314 (0.04480)	0.98539 (0.00330)	0.90603 (0.07239)	0.95220 (0.02842)	$\lambda_0 = 5.52 \times 10^{-7}$
$f_{\mathbf{z},\lambda}$	98.56 (0.10)	0.98100 (0.00163)	0.98852 (0.00121)	0.98311 (0.00148)	0.98474 (0.00105)	$\lambda_0 = 5.52 \times 10^{-7}$ $\lambda_1 = 4.31 \times 10^{-3}$
$f_{\mathbf{z},\lambda}^{10}$	92.29 (0.30)	0.93164 (0.00279)	0.90179 (0.00734)	0.94156 (0.00276)	0.91628 (0.00358)	$\lambda_0 = 1.69 \times 10^{-8}$ $\lambda_1 = 5.71 \times 10^{-6}$
$f_{\mathbf{z},\lambda}^{50}$	96.29 (0.12)	0.97816 (0.00132)	0.94193 (0.00318)	0.98144 (0.00118)	0.95967 (0.00140)	$\lambda_0 = 2.25 \times 10^{-7}$ $\lambda_1 = 5.66 \times 10^{-5}$
$f_{\mathbf{z},\lambda}^{250}$	98.33 (0.09)	0.98171 (0.00085)	0.98273 (0.00135)	0.98386 (0.00075)	0.98222 (0.00098)	$\lambda_0 = 8.42 \times 10^{-8}$ $\lambda_1 = 4.72 \times 10^{-4}$
$f_{\mathbf{z}}$	98.33 (0.09)	0.98171 (0.00085)	0.98273 (0.00135)	0.98386 (0.00075)	0.98222 (0.00098)	

Table 7: Statistical performance of various estimators over all 9 sub-data sets (Folds) of NSL-KDD data set.

Note that the performance of the Nyström type subsampling depends not only on the size s of a subsampling set but also on the sub-data set $\{(x_i, y_i)\}_{i=1}^s$. To demonstrate reliability of Nyström type

Hybrid Naive based classifier using 2500 (approximately) patterns [?]	82.39%
DBN-SVM on normalized encoded data set using 5000 patterns [?]	96.90%
Cross-breed type Bayesian network on 2500 patterns [?]	97.27%
Multi-penalty (Manifold) regularization algorithm on 2500 patterns [?]	98.56%
Proposed multi-penalty kernel method based on Nyström type subsampling on 2500 patterns	98.33%

Table 8: Performance comparison of the proposed model with some existing research work on NSL-KDD data set.

subsampling we generate 50 times subsamples of size $s = 10, 50, 250$ from 9 sub-data sets (Folds) and report the mean and standard deviation of performance accuracy over 9 sub-data sets in Table 6. We observe that the performance of the single-penalty regularization varies over the different folds. On the other hand, multi-penalty regularized solution performs consistently. The accuracy of the multi-penalty regularized solutions varies over the subsampling size s but the aggregated solution based on linear functional strategy almost provides the better solution among the Nyström approximants. Here we also note that the performance of the Nyström approximants does not vary much over the different random subsampling of same size. In Table 7, we report the results in terms of mean and standard deviation of the performance measures over nine sub-data sets with the balanced-discrepancy parameter choice.

In Table 8, we compare the performance of proposed approach with existing approaches applied on NSL-KDD data set for intrusion detection. The multi-penalty regularization scheme illustrates the better performance than the benchmark result achieved in [?]. Moreover, the proposed multi-penalty kernel method based on Nyström type subsampling is able to achieve the more accurate results. The reported results of the experiments demonstrate that the aggregation approach automatically uses the best Nyström approximant and achieves the accuracy of standard multi-penalty regularization scheme. The results shows that the Nyström subsampling can yield very good performance while substantially reducing the computational requirements. So the approach can be efficiently applied as a reliable strategy when dealing with the data of big size.

References