Journal of Applied and Numerical Analysis
2023, Vol. 1, pp. 58–71
doi: 10.30970/ana.2023.1.58

On learning rates for regularized Nyström subsampling in unsupervised domain adaptation

H. Myleiko, S. Solodky

Institute of Mathematics NAS of Ukraine,
3, Tereschenkivska str., 01024, Kyiv, Ukraine

Анотація. Регуляризована підвибірка Нистрьома є популярним підходом до проблем навчання, які мають справу з надвеликим обсягом вхідних даних. Такий алгоритм ми використовуємо в контексті задач доменної адаптації без учителя за умови коваріаційного зсуву. В рамках концепції гільбертового простору з породжуючим ядром побудовано алгоритм, який є комбінацією підвибірки Нистрьома та двокрокової ітерованої тихоновської регуляризації. Запропонований підхід дозволяє не тільки суттєво скоротити обсяг задіяних обчислювальних ресурсів, але до того ж зберегти такі самі швидкості навчання як у стандартному машинному навчанні.

Abstract. The regularized Nyström subsampling is a popular approach for learning problems that deals with big data. We employ such technique in the context of the unsupervised domain adaptation problems with covariate shift assumption. Within the framework of the Reproducing Kernel Hilbert Space concept, an algorithm is constructed that is a combination of the Nyström subsampling and the two-steps iterated Tikhonov regularization. This approach allows significantly reduce the amount of computing resources involved and at the same time maintains the same learning rates as for the standard machine learning algorithms.

1 Introduction

In statistical learning theory, regularized kernel methods are the most theoretically studied algorithms provided acceptable results for the problems when the number of data is not too large. But most these methods require computing a kernel matrix which leads to at least quadratic computational cost in the sample size, which means that larger data sets are typically out of reach. Nyström subsampling is an effective approach to analyze big data, which serves as standard tool for reducing computational complexity in machine learning problems where massive data sets are involved. The present study is focused on the use of the regularized Nyström subsampling in the context of unsupervised domain adaptation problems dealing with big data.

Recall, in the supervised learning, it is commonly assumed that the training data comes from the same distribution as that of the test data. However, many real world applications, for example, in natural language processing or computer vision, do not meet this assumption. This obstacle can be overcome by embedding domain adaptation. Domain adaptation is sub-discipline of machine learning which aims to improve the performance of a learning model on the target domain by borrowing knowledge from a well- established source domain and also by reducing the difference between domain distributions or the domain shift. To be more precise, domain adaptation scenario arises when one studies relationship between the explanatory (input) variable $x \in X \subset \mathbb{R}^d$ and the response (output) variable under the assumption that they are governed by different probability distributions with respect to measures $\rho(x, y)$ and $q(x, y)$ on $X \times Y$. This, generally, means that an input $x \in X$ does not determine uniquely the output $y \in Y$, but rather some conditional probability $\rho(y|x)$ of y given x, which is assumed to be unknown. The inputs $x \in X$ is also assume to be random and governed by marginal probabilities $\rho_S(x)$ in the source domain (S) and

2020 *Mathematics Subject Classification*: 62R07, 65C20, 65J20, 68Q32, 68T09.
Key words: unsupervised domain adaptation, Big Data, Nyström subsampling, regularization; source condition, Radon-Nikodym derivative, computational complexity.
© Myleiko H., Solodky S., 2023

$\rho_T(x)$ in the target domain (T). Thus, the learning task can be seen as a minimization of the expected risk of the prediction y from x with respect to the one measure, say, $q(x,y)$ by using a training data sample $\mathbf{z} = \{(x_i, y_i),\ x_i \in X,\ y_i \in Y,\ i = 1, 2, \ldots, n\},\ |z| = n$, drawn independently and identically (i.i.d.) over the another measure $\rho(x,y)$. In domain adaptation settings, $\rho(x,y)$ and $q(x,y)$ are usually called the source and the target probabilities, respectively. Generally, the problem of domain adaptation with two different distributions is unsolvable, since $\rho(x,y)$ and $q(x,y)$ can be arbitrarily far apart. To guarantee a solvability of the problem, the covariate shift assumption should be imposed (see [5, 20]). Namely, we assume that $\rho_S(x) \neq \rho_T(x)$, while the conditional probability $\rho(y|x)$ remains unchanged for both source and target probabilities. This means that the joint probabilities $\rho(x,y)$ and $q(x,y)$ can be factorized as follows

$$\rho(x,y) = \rho(y|x)\rho_S(x), \quad q(x,y) = \rho(y|x)\rho_T(x).$$

It should be noted that sample selection bias and missing data are two causes for the covariate shift. Most of the knowing domain adaptation techniques aim to solve this class of domain gap, which typically appear in many applications such as classification , handwriting recognition , segmentation and regression for multimedia data, for example if the background, shape deformation, or quality are different across domains. Domain adaptation aims to mitigate this and has successfully been applied for object recognition, AI planning, reinforcement learning and natural language processing (e.g., the adaptation from document on the one language to another language).

In the present study, we restrict ourselves to learning with the least square loss where the expected risk of the prediction y from x by means of a function $f\colon X \to Y$ is defined in the target domain as follows

$$\mathcal{R}_q(f) := \int_{X\times Y} (f(x) - y)^2 dq(x,y),$$

which is minimized by so-called regression function

$$f(x) = f_q(x) = \int_Y y d\rho(y|x). \tag{1.1}$$

But in the unsupervised domain adaptation settings neither $\mathcal{R}_q(f)$, nor $f_q(x)$ can be computed, since the information about underlying probability $q(x,y)$ is only given as a set $X^{'} = (x_1^{'}, x_2^{'}, \ldots, x_m^{'}),\ \ |X^{'}| = m$, of unlabeled examples $x_i^{'}$ of inputs drawn i.i.d. from the target marginal probability measure $\rho_T(x)$. Thus, the goal is to use this information together with a training data set $\mathbf{z}$ to approximate the ideal minimizer f_q by an empirical estimator $f_{\mathbf{z}}$ in the sense of excess risk

$$\mathcal{R}_q(f_{\mathbf{z}}) - \mathcal{R}_q(f_q) := \|f_{\mathbf{z}} - f_q\|^2_{L_{2,\rho_T}},$$

where L_{2,ρ_T} is the space of square-integrable functions $f\colon X \to \mathbb{R}$ with respect to the marginal probability measure ρ_T. Following [6], we employ the idea that the unsupervised domain adaptation problems approximate the same regression function given by (1.1) as in the standard supervised learning. Therefore, the supervised learning algorithms based on regularization techniques in a reproducing kernel Hilbert space (RKHS) can be profitably used in the context of unsupervised domain adaptation. Here we refer to [5, 6, 17, 20] and to references therein.

The paper is organized as follows. In the next section, we give the strict problem settings and define the Nyström subsampling method. In Section 3, we obtain error estimates for the regularized Nyström subsampling under the assumption that the values of the Radon-Nikodym derivate at the unlabeled target inputs are known. In Section 4, we prove a theorem which demonstrate how the Radon-Nikodym derivate can be approximately reconstructed from unlabeled examples of inputs drawn according to source and target probabilities. For this end, we employ the combination of Nyström subsampling and the two-steps iterated Tikhonov regularization in RKHS. In the last section, we estimate learning rates of regularized Nyström subsampling in the case of the unknown values of the Radon-Nikodym derivate and provide analysis related to computational cost of the proposed method in the context of unsupervised domain adaptation problems.

2 Problem setting

From now on, we assume that the regression function $f^* = f_q$, minimizing the expected risk $\mathcal{R}_q(f)$, belongs to a specified Hilbert space with reproducing kernel $\mathcal{H}_\mathsf{K}$. Let $J_T : \mathcal{H}_\mathsf{K} \hookrightarrow L_{2,\rho_T}$ and $J_S : \mathcal{H}_\mathsf{K} \hookrightarrow L_{2,\rho_S}$ be the inclusion operators. Recall that the information about the source and the target marginal measures are only provided in the form of samples $X_S = \{x_1, x_2, \dots, x_n\}$ and $X_T = \{x_1^{'}, x_2^{'}, \dots, x_m^{'}\}$, drawn independently and identically (i.i.d.) from ρ_S and ρ_T, respectively. In the sequel, we define two sample operators

$$S_{X_T} f = (f(x_1^{'}), f(x_2^{'}), \cdots, f(x_m^{'})) \in \mathbb{R}^m,$$

$$S_{X_S} f = (f(x_1), f(x_2), \cdots, f(x_n)) \in \mathbb{R}^n,$$

acting from $\mathcal{H}_\mathsf{K}$ to $\mathbb{R}^m$ and $\mathbb{R}^n$, where the norms in later spaces are m^{-1}-times and n^{-1}-times the standard Euclidian norms, such that the adjoint operators $S^*_{X_T} \colon \mathbb{R}^m \to \mathcal{H}_\mathsf{K}$ and $S^*_{X_S} \colon \mathbb{R}^n \to \mathcal{H}_\mathsf{K}$ are given as

$$S^*_{X_T} u(\cdot) = \frac{1}{m} \sum_{j=1}^{m} \mathsf{K}(\cdot, x_j^{'}) u_j, \quad u = (u_1, u_2, \dots, u_m) \in \mathbb{R}^m,$$

$$S^*_{X_S} v(\cdot) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{K}(\cdot, x_i) v_i, \quad v = (v_1, v_2, \dots, v_n) \in \mathbb{R}^n.$$

Since we have no direct access to both the target probability measure ρ_T and the space L_{2,ρ_T} in which we are going to approximate the regression function $f^* = f_q$, then an assumption should be put on the relation between the source probability ρ_S and the target probability ρ_T. As in [5], we assume that there is a function $\beta \colon X \to \mathbb{R}_+$ such that

$$d\rho_T(x) = \beta(x) d\rho_S(x).$$

Then $\beta(x)$ is considered as the Radon-Nikodym derivative $\frac{d\rho_T}{d\rho_S}$ of the target measure with respect to the source one. We also assume that we only have access to the values $\beta(x_i)$ of the Radon-Nikodym derivative $\beta(x) = \frac{d\rho_T(x)}{d\rho_S(x)}$ at the points x_i, $i = 1, 2, \dots, n$, drawn i.i.d. from $\rho_S(x)$ and we consider a diagonal $n \times n$ matrix $B = \operatorname{diag}(\beta(x_1), \beta(x_2), \dots, \beta(x_n))$. Moreover, we assume that $\beta(x)$ is uniformly bounded on X, such that $0 \le \beta(x) \le b_0$ for some $b_0 > 0$ and any $x \in X$.

The subsequent analysis is based on two additional assumptions which are common and not restrictive. We assume that $\mathsf{K} \colon X \times X \to \mathbb{R}$ is a continuous and bounded kernel that for any $x \in X$ it holds

$$\|\mathsf{K}(\cdot, x)\|_{\mathcal{H}_\mathsf{K}} = \langle \mathsf{K}(\cdot, x) \mathsf{K}(\cdot, x) \rangle^{1/2}_{\mathcal{H}_\mathsf{K}} = [\mathsf{K}(x, x)]^{1/2} < \kappa_0 < \infty.$$

In addition, we assume that for any input $x \in X$ corresponding output $y \in Y \subset \mathbb{R}$ is bounded $|y| \le y_0$ with $y_0 > 0$.

Further, we are going to approximate a solution of the equation arising from the minimization of the excess risk

$$\mathcal{R}_q(f) - \mathcal{R}_q(f_q) = \|f - f_q\|^2_{L_{2,\rho_T}}. \tag{2.1}$$

In RKHS $\mathcal{H}_\mathsf{K}$ the above mentioned minimization problem (2.1) can be rewritten by means of the inclusion operator $J_T \colon \mathcal{H}_\mathsf{K} \hookrightarrow L_{2,\rho_T}$ as a variational problem

$$\|J_T f - f_q\|_{L_{2,\rho_T}} \to \min,$$

and it leads to the finite-dimensional normal equation

$$L_T f = J^*_T f_q, \tag{2.2}$$

where $L_T = J_T^* J_T$. Note that Eq. (2.2) is ill-posed because the involved operator L_T is compact and its inverse can not be bounded in $\mathcal{H}_\mathsf{K}$. Thus, this equation should be analyzed by the methods of the Regularization theory.

In the Regularization theory, there is a continuous strictly increasing function $\varphi\colon [0,l] \to \mathbb{R}, \quad l \geq \|L_T\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}$ such that $\varphi(0) = 0$, and allows to present f_q by the means of the so-called source conditions

$$f_q = \varphi(L_T)\mu_q, \quad \mu_q \in \mathcal{H}_\mathsf{K}. \tag{2.3}$$

The function φ is usually called the index function of the source condition. This function specifies the smoothness properties of f_q and characterizes the convergence rate of the regularization method.

Let's consider the class $\mathcal{F}_{1/2}$ of operator monotone index functions $\varphi\colon [0,l] \to \mathbb{R}_{+\infty}$ such that

$$\varphi(t) \leq c_1\sqrt{t}, \quad t \in [0,T]. \tag{2.4}$$

Note that $\mathcal{F}_{1/2}$ contains functions $\varphi\colon [0,l] \to \mathbb{R}_{+\infty}$ such that $c_2 t \leq \varphi(t) \leq c_1\sqrt{t}, \quad t \in [0,T]$.

Recall, that a function φ is operator monotone if for any non-negative self-adjoint operators $A, B\colon \mathcal{H}_\mathsf{K} \to \mathcal{H}_\mathsf{K}$ with spectra in $[0,l]$ it holds

$$\|\varphi(A) - \varphi(B)\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \leq c_3\varphi\left(\|A-B\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}\right).$$

It is known (see [1]) that the most regularization schemes can also be indexed by parameterized function $g_\lambda\colon [0,l] \to \mathbb{R}$, $\lambda > 0$. The only requirements are that there are positive constants $\gamma_0, \overline{\gamma}, \tilde{\gamma}$ such that

$$\sup_{0<t\leq l} |1 - t g_\lambda(t)| \leq \gamma_0, \quad \sup_{0<t\leq l} \sqrt{t}|g_\lambda(t)| \leq \frac{\overline{\gamma}}{\sqrt{\alpha}}, \quad \sup_{0<t\leq l} |g_\lambda(t)| \leq \frac{\tilde{\gamma}}{\lambda}. \tag{2.5}$$

Further important property of the regularization method indexed by g_λ is its qualification that is the maximum positive number p for which

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

where γ_p does not depend on λ. For example, the standard Tikhonov method has the qualification $p = 1$.

In our research, we apply the two-times iterated Tikhonov regularization, the index function of which has a form

$$g_\lambda(t) = \sum_{i=1}^{2} \lambda^{i-1}(\lambda + t)^{-i} = \frac{1}{t}\Big(1 - \frac{\lambda^2}{(\lambda+t)^2}\Big), \quad \lambda \neq 0. \tag{2.6}$$

It should be noted that the standard Tikhonov method provides optimal order of accuracy for the index functions of the form $\varphi(t) = t^\alpha$, $0 \leq \alpha \leq \frac{1}{2}$. For domain adaptation problems the solutions with indicated above smoothness were employed in [6, 7, 22]. Regarding the considered functions $\varphi(t)$ (2.3), (2.4) for the standard Tikhonov method the saturation effect will be observed. To achieve optimal accuracy for (2.3), (2.4) one needs to employ the regularization with qualification $p \geq 2$. In such situation, the two-times iterated Tikhonov regularization with $p = 2$ is the most applicable. In addition, the index functions $\varphi(t) = t^\alpha$, $\alpha > 1$, is out of our study for now. The solutions with such high smoothness call for the implementation of regularization with higher qualification ($p \geq 3$). The implementation of such method requires the modified techniques and will be considered in our further research.

The following definition [11, 13] shows a relation between the qualification and the source condition.

Definition 2.1. *We say that the qualification p covers the index function φ if the function $t \to t^p/\varphi(t)$ is non-decreasing for $t \in (0, l]$.*

Proposition 2.1. *[11, Proposition 2.7] Let the regularization method is indexed by $g_\lambda(t)$ and has the qualification p. If this qualification covers the index function φ, then*

$$\sup_{0<t\leq l} |1 - tg_\lambda(t)|\varphi(t) \leq \gamma_*\varphi(\lambda),$$

where $\gamma_* = \max\{\gamma_0, \gamma_p\}$.

It is known (see, e.g., [2,8]) that for full use of the smoothness of the unknown function f_q it is necessary that the qualification of the method g_λ implemented in the learning task covers not only the index function $\varphi(t)$, but also the product $\varphi(t)\sqrt{t}$.

Remark 2.1. *It is well-known that various regularization scheme can be profitably used in the standard supervised learning context. For instance, in [21], the Tikhonov regularization was analyzed as a supervised learning algorithm in RKHS, and the best-known learning rates were obtained for this scheme. Then in [2] it has been shown that the same type of results are true for a large class of supervised learning algorithms which are essentially all the linear regularization schemes. In [6] the authors extend the analysis of [2] to the setting of domain adaptation with covariate shift. Below we show how the technique from [6] based on iterated Tikhonov regularization can be extended to the domain adaptation with covariate shift under the big data settings.*

One of the most studied approaches to the approximation of the minimizer $f^* = f_q$ of the target expected risk $\mathcal{R}_q(f)$ by using the data $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n$, sampled from the source measure $\rho(x, y)$, is importance weighted regularized least squares (IWRLS) (see, e.g., [5,7,22]). Usually, the standard Tikhonov regularization known as Kernel Ridge Regression (KRR) in machine learning is employed in the context of this method. It is known that the low qualification is a disadvantage of such regularization. In other words, as we mentioned above, it leads to saturation in the case of highly smooth solutions. To avoid this in [6] the authors applied iterative Tikhonov regularization, with the index function (2.6). Herewith, the IWRLS-approximant was performed as follows

$$f_{\mathbf{z}}^\lambda = g_\lambda(S_{X_S}^* B S_{X_S}) S_{X_S}^* B\overline{y}, \tag{2.7}$$

where $S_{X_S}^* B S_{X_S}$ is a self-adjoint, non-negative, and compact operator in RKHS $\mathcal{H}_\mathsf{K}$.

Note that the approximant (2.7) results from applying the regularization scheme to a finite-dimensional equation

$$S_{X_S}^* B S_{X_S} f = S_{X_S}^* B\overline{y},$$

which is the discretized version of (2.2). According to [6,23] a perturbation of (2.2) caused by the discretization (2.8) can be estimated with probability at least $1 - \delta$ as follows

$$\|L_T - S_{X_S}^* B S_{X_S}\|_{\mathcal{H}_\mathsf{K} \to \mathcal{H}_\mathsf{K}} \leq \kappa_1 \log^{\frac{1}{2}} \frac{1}{\delta} (m^{-\frac{1}{2}} + n^{-\frac{1}{2}}),$$

$$\|S_{X_S}^* B S_{X_S} f_q - S_{X_S}^* B\overline{y}\|_{\mathcal{H}_\mathsf{K}} \leq \kappa_2 \log^{\frac{1}{2}} \frac{1}{\delta} (m^{-\frac{1}{2}} + n^{-\frac{1}{2}}),$$

where $\kappa_1, \kappa_2 > 0$ are some constants.

It is known (cf. [19]) that KRR has at least quadratic computational cost $O(|\mathbf{z}|^2)$ in the number of observations $|\mathbf{z}|$ and this is the cost of computing the kernel matrix $\mathbb{K}_{|\mathbf{z}|} = |\mathbf{z}| S_{X_S} B S_{X_S}^*, \quad |\mathbf{z}| = n$, of values of the kernel $\mathsf{K}(x_i, x_j)$. Therefore in the big data setting where $|\mathbf{z}|$ is large enough, it is reasonable to avoid the computation of the minimizer $f_{\mathbf{z}}^\lambda$ (2.7). The Nyström subsampling overcome extra large computational costs by replacing $\mathbb{K}_{|\mathbf{z}|}$ by a smaller low-rank matrix obtained by a random subsample of columns $\mathbb{K}_{|\mathbf{z}|}$. An important analysis made in [19] shows that the Nyström subsampling can be considered as a combination of the regularization g_λ and a projection scheme on the subset

$$\mathcal{H}_{\mathsf{K}}^{\mathbf{z}^\nu} := \{f : f(\cdot) = \sum_{x_i :\, (x_i,y_i)\in\mathbf{z}^\nu} d_i \mathsf{K}(\cdot, x_i), \quad d_i \in \mathbb{R}\}.$$

To be more precise, according to the Nyström subsampling, the approximation of functions not carried out through (2.7), and

$$f^{\lambda}_{\mathbf{z},\mathbf{z}^\nu} = g_\lambda\left(P_{\mathbf{z}^\nu} S^*_{X_S} B S_{X_S} P_{\mathbf{z}^\nu}\right) \mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B\overline{y}, \tag{2.8}$$

where $\mathrm{P}_{\mathbf{z}^\nu} : \mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}^{\mathbf{z}^\nu}$, $\|\mathrm{P}_{\mathbf{z}^\nu}\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} = 1$ is the orthogonal projection operator with the range $\mathcal{H}_{\mathsf{K}}^{\mathbf{z}^\nu}$. Note (see [19]), to compute (2.8) it is not necessary to construct $\mathrm{P}_{\mathbf{z}^\nu}$ explicitly.

According to [6, 23] a perturbation of (2.2) caused by the discretization (2.8) can be estimated with probability at least $1-\delta$ as follows

$$\|L_T - S^*_{X_S} B S_{X_S}\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} \le \kappa_1 \log^{\frac{1}{2}} \frac{1}{\delta}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}}),$$

$$\|S^*_{X_S} B S_{X_S} f_q - S^*_{X_S} B\overline{y}\|_{\mathcal{H}_{\mathsf{K}}} \le \kappa_2 \log^{\frac{1}{2}} \frac{1}{\delta}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}}),$$

where $\kappa_1, \kappa_2 > 0$ are some constants.

Later on, we will need the following auxiliary statements (see [14])

$$\begin{aligned}\|P_{\mathbf{z}^\nu}\varphi(L_T)P_{\mathbf{z}^\nu} - \varphi(P_{\mathbf{z}^\nu} L_T P_{\mathbf{z}^\nu})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} \\ \le 2\varphi\left(\|L_T^{1/2}(I - P_{\mathbf{z}^\nu})\|^2_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}\right),\end{aligned} \tag{2.9}$$

$$\|(I - P_{\mathbf{z}^\nu})\varphi(L_T)\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} \le C\|L_T^{1/2}(I - P_{\mathbf{z}^\nu})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}. \tag{2.10}$$

Here and in the sequel, we adopt the convention that C denotes a generic positive coefficient, which can vary from inequality to inequality and may only depend on basic parameters such as $\rho_S, \rho_T, \kappa_0, \beta_0, y_0$ and others which may appear below.

Note that within the framework of the Nyström subsampling the value $\Delta_{T,\mathbf{z}^\nu} = \|L_T^{1/2}(I - P_{\mathbf{z}^\nu})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}$ has a probabilistic nature and depends on the way $\mathbf{z}^\nu$ is subsampled. Such dependence is considered in [9, 15, 18, 19].

3 Error Estimate of the Regularized Nyström Subsampling

In this section, we estimate an approximation accuracy of the minimizer f_q of the target expected risk $\mathcal{R}_q(f)$ by using the data $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n$, sampled from the source measure $\rho(x, y)$.

Theorem 3.1. *Let f_q satisfies the source condition (2.4) with $\varphi \in \mathcal{F}_{1/2}$, and the approximate solution $f^{\lambda_{m,n}}_{\mathbf{z},\mathbf{z}^\nu}$ is of the form (2.8), then for*

$$\lambda = \lambda_{m,n} = \theta^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}}) \text{ and } \Delta_{T,\mathbf{z}^\nu} \le \varphi\left(\theta^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})\right)$$

with probability $1-\delta$ it holds

$$\|f_q - f^{\lambda_{m,n}}_{\mathbf{z},\mathbf{z}^\nu}\|_{L_2,\rho_T} = O\left(\sqrt{\theta^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})}\varphi\left(\theta^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})\right)\log^{\frac{3}{4}}\frac{1}{\delta}\right),$$

$$\|f_q - f^{\lambda_{m,n}}_{\mathbf{z},\mathbf{z}^\nu}\|_{\mathcal{H}_{\mathsf{K}}} = O\left(\varphi\left(\theta^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})\right)\log^{\frac{1}{2}}\frac{1}{\delta}\right).$$

To prove this Theorem we will need the following statement.

Lemma 3.2. *For $\varphi \in \mathcal{F}_{1/2}$ it holds*

$$\|\varphi(L_T) - \varphi\left(P_{\mathbf{z}^\nu} S^*_{X_S} B S_{X_S} P_{\mathbf{z}^\nu}\right)\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \leq C \log^{\frac{1}{2}} \frac{1}{\delta}\Big(\Delta_{T,\mathbf{z}^\nu} + \varphi\left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}}\right)\Big).$$

Proof. We start with a decomposition

$$\varphi(L_T) - \varphi\left(P_{\mathbf{z}^\nu} S^*_{X_S} B S_{X_S} P_{\mathbf{z}^\nu}\right) = \mathrm{P}_{\mathbf{z}^\nu}\varphi(L_T)(I - \mathrm{P}_{\mathbf{z}^\nu}) + (I - \mathrm{P}_{\mathbf{z}^\nu})\varphi(L_T)$$

$$+\mathrm{P}_{\mathbf{z}^\nu}\varphi(L_T)\mathrm{P}_{\mathbf{z}^\nu} - \varphi(\mathrm{P}_{\mathbf{z}^\nu} L_T \mathrm{P}_{\mathbf{z}^\nu}) + \varphi(\mathrm{P}_{\mathbf{z}^\nu} L_T \mathrm{P}_{\mathbf{z}^\nu}) - \varphi\left(P_{\mathbf{z}^\nu} S^*_{X_S} B S_{X_S} P_{\mathbf{z}^\nu}\right).$$

Then by (2.9) and (2.10) we get

$$\|\varphi(L_T) - \varphi\left(P_{\mathbf{z}^\nu} S^*_{X_S} B S_{X_S} P_{\mathbf{z}^\nu}\right)\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \leq C\|L_T^{\frac{1}{2}}(I - \mathrm{P}_{\mathbf{z}^\nu})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}$$

$$+C\varphi\left(\|L_T^{\frac{1}{2}}(I - \mathrm{P}_{\mathbf{z}^\nu})\|^2_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}\right) + C\varphi\left(\|L_T - S^*_{X_S} B S_{X_S}\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}\right)$$

$$\leq C\left(\|L_T^{\frac{1}{2}}(I - \mathrm{P}_{\mathbf{z}^\nu})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} + \varphi\left(\|L_T - S^*_{X_S} B S_{X_S}\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}\right)\right).$$

By Proposition 2.1 and (2.4) with probability $1 - \delta$ we have

$$\|\varphi(L_T) - \varphi\left(P_{\mathbf{z}^\nu} S^*_{X_S} B S_{X_S} P_{\mathbf{z}^\nu}\right)\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \leq C \log^{\frac{1}{2}} \frac{1}{\delta}\Big(\Delta_{T,\mathbf{z}^\nu} + \varphi\left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}}\right)\Big).$$

Lemma is proved. □

Sketch of proof. The proof of Theorem is similar to Theorem 3 [16] for ranking and regression problems. Applying the techniques from [16] along with Lemma 3.2 one can get the statement of the theorem. We omit the proof of Theorem since it is fairly technical and lengthy.

Remark 3.1. *Note that under the conditions of Theorem 3.1 Nyström subsampling has the same learning rate as the one guaranteed by Theorem 1 [6] for algorithm based on the whole sample* $\mathbf{z}$*. Moreover, the application of the Nyström subsampling as some projection scheme for machine learning problems under the big data settings, such as regression and ranking, was considered earlier in the works [8, 15, 19]. In particular, in [8, 15] was established that such an approach allows to achieve the above mentioned learning rate with subquadratic cost. In addition, the effectiveness of this approach was proven in the works [3, 4, 9, 10, 12].*

4 Approximation of the Radon-Nikodym Derivate in RKHS

In this section, we approximate the Radon-Nikodym derivative $\beta = \frac{d\rho_T}{d\rho_S}$ which solves the integral equation

$$J^*_S \beta = \int_X \mathsf{K}(x, x')\beta(x')d\rho_S(x') = \int_X \mathsf{K}(x, x')d\rho_T(x') = J^*_T \mathbf{1}, \tag{4.1}$$

where $\mathbf{1}$ is the constant function that takes value 1 everywhere. Following [6] and [7], we assume that $\beta(x) \in \mathcal{H}_\mathsf{K}$. Without loss of generality, we assume that $\mathbf{1} \in \mathcal{H}_\mathsf{K}$. Then the equation (4.1) can be reduce to

$$J^*_S J_S \beta = J^*_T J_T \mathbf{1}. \tag{4.2}$$

Since, in practice, the amount of the unlabeled inputs is usually much greater than that of labeled ones, we assume that the sizes M and N of i.i.d. samples $(x_1^{'}, x_2^{'}, \dots, x_M^{'})$ and $(x_1, x_2, \dots, x_N)$ drawn respectively from ρ_T and ρ_S are much larger than m and n considered earlier. Then we define two sample operators

$$S_{M,T} f = (f(x_1^{'}), f(x_2^{'}), \cdots, f(x_M^{'})) \in \mathbb{R}^M,$$
$$S_{N,S} f = (f(x_1, f(x_2), \cdots, f(x_N)) \in \mathbb{R}^N,$$

and a finite-dimensional problem

$$S_{N,S}^* S_{N,S} \beta = S_{M,T}^* S_{M,T} \mathbf{1}, \tag{4.3}$$

which is an empirical version of the equation (4.2). Here, similar to the above notations the operators $S_{N,S}^* \colon \mathbb{R}^N \to \mathcal{H}_{\mathsf{K}}$ and $S_{M,T}^* \colon \mathbb{R}^M \to \mathcal{H}_{\mathsf{K}}$ are given as

$$S_{N,S}^* v(\cdot) = \frac{1}{N} \sum_{i=1}^{N} \mathsf{K}(\cdot, x_i^{'}) v_i, \quad v = (v_1, v_2, \dots, v_N) \in \mathbb{R}^N,$$

$$S_{M,T}^* u(\cdot) = \frac{1}{M} \sum_{j=1}^{M} \mathsf{K}(\cdot, x_j) u_j, \quad u = (u_1, u_2, \dots, u_M) \in \mathbb{R}^M.$$

Further, we implement the approach from the previous section. To obtain the Radon-Nikodym derivative from (4.3) we again apply the combination of the Nyström subsampling and the two-steps iterated Tikhonov regularization. Thus, the approximation to $\beta = \frac{d\rho_T}{d\rho_S}$ we will seek as

$$\tilde{\beta}_{M,N,\mathbf{z}^\nu}^{\alpha_{M,N}} = g_\alpha(\mathrm{P}_{\mathbf{z}^\nu} S_{N,S}^*, S_{N,S} \mathrm{P}_{\mathbf{z}^\nu}) \mathrm{P}_{\mathbf{z}^\nu} S_{M,T}^* S_{M,T} \mathrm{P}_{\mathbf{z}^\nu} \mathbf{1}. \tag{4.4}$$

We assume that $\beta = \frac{d\rho_T}{d\rho_S}$ satisfies the source condition

$$\beta = \varphi(L_S) \mu_\beta, \tag{4.5}$$

where $L_S = J_S^* J_S$, $\mu_\beta \in \mathcal{H}_{\mathsf{K}}$, $\varphi \in \mathcal{F}_{1/2}$.

According to [6, 23] with probability $1 - \delta$ it holds

$$\|S_{N,S}^* S_{N,S} \beta - S_{M,T}^* S_{M,T} \mathbf{1}\|_{\mathcal{H}_{\mathsf{K}}} \le C(N^{-1/2} + M^{-1/2}) \log^{\frac{1}{2}} \frac{1}{\delta}, \tag{4.6}$$

$$\|L_S - S_{N,S}^* S_{N,S}\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \le \frac{C \log^{\frac{1}{2}} \frac{1}{\delta}}{\sqrt{N}}, \tag{4.7}$$

$$\|L_T - S_{M,T}^* S_{M,T}\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \le \frac{C \log^{\frac{1}{2}} \frac{1}{\delta}}{\sqrt{M}}. \tag{4.8}$$

To simplify the subsequent presentation, let's denote $\Delta_{S,\mathbf{z}^\nu} := \|L_S^{\frac{1}{2}} (I - P_{\mathbf{z}^\nu})\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}}$.

Theorem 4.1. *Assume that $\beta = \frac{d\rho_T}{d\rho_S}$ satisfies the source condition (4.5) with $\varphi \in \mathcal{F}_{1/2}$ and the approximant $\tilde{\beta}_{M,N,\mathbf{z}^\nu}^{\alpha_{M,N}}$ given by (4.4), then for*

$$\begin{gathered} \alpha = \alpha_{M,N} = \theta_\varphi^{-1}\left(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\right), \\ \Delta_{S,\mathbf{z}^\nu} \le \varphi\left(\theta_\varphi^{-1}\left(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\right)\right), \quad \Delta_{T,\mathbf{z}^\nu} \le \varphi\left(\theta_\varphi^{-1}\left(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\right)\right), \end{gathered} \tag{4.9}$$

with probability at least $1-\delta$ *it holds*

$$\|\beta-\tilde{\beta}^{\alpha_{M,N}}_{M,N,\mathbf{z}^\nu}\|_{\mathcal{H}_\mathsf{K}}=O\left(\varphi\left(\theta_\varphi^{-1}\left(\frac{1}{\sqrt{N}}+\frac{1}{\sqrt{M}}\right)\right)\log^{\frac{1}{2}}\frac{1}{\delta}\right).$$

Proof. First, consider the decomposition

$$\begin{aligned}\beta-\tilde{\beta}^{\alpha_{M,N}}_{M,N,\mathbf{z}^\nu}&=\beta-g_\alpha(\mathrm{P}_{\mathbf{z}^\nu}S^*_{N,S}S_{N,S}\mathrm{P}_{\mathbf{z}^\nu})\mathrm{P}_{\mathbf{z}^\nu}S^*_{M,T}S_{M,T}\mathrm{P}_{\mathbf{z}^\nu}\mathbf{1}\\&=\omega_1+\omega_2+\omega_3+\omega_4,\end{aligned}\tag{4.10}$$

where

$$\begin{aligned}\omega_1&=\beta-g_\alpha(\mathrm{P}_{\mathbf{z}^\nu}S^*_{N,S}S_{N,S}\mathrm{P}_{\mathbf{z}^\nu})\mathrm{P}_{\mathbf{z}^\nu}S^*_{N,S}S_{N,S}\mathrm{P}_{\mathbf{z}^\nu}\beta;\\
\omega_2&=g_\alpha(\mathrm{P}_{\mathbf{z}^\nu}S^*_{N,S}S_{N,S}\mathrm{P}_{\mathbf{z}^\nu})\mathrm{P}_{\mathbf{z}^\nu}S^*_{N,S}S_{N,S}\mathrm{P}_{\mathbf{z}^\nu}\beta\\&\quad-g_\alpha(\mathrm{P}_{\mathbf{z}^\nu}S^*_{N,S}S_{N,S}\mathrm{P}_{\mathbf{z}^\nu})\mathrm{P}_{\mathbf{z}^\nu}S^*_{N,S}S_{N,S}\beta;\\
\omega_3&=g_\alpha(\mathrm{P}_{\mathbf{z}^\nu}S^*_{N,S}S_{N,S}\mathrm{P}_{\mathbf{z}^\nu})\mathrm{P}_{\mathbf{z}^\nu}S^*_{N,S}S_{N,S}\beta\\&\quad-g_\alpha(\mathrm{P}_{\mathbf{z}^\nu}S^*_{N,S}S_{N,S}\mathrm{P}_{\mathbf{z}^\nu})\mathrm{P}_{\mathbf{z}^\nu}S^*_{M,T}S_{M,T}\mathbf{1};\\
\omega_4&=g_\alpha(\mathrm{P}_{\mathbf{z}^\nu}S^*_{N,S}S_{N,S}\mathrm{P}_{\mathbf{z}^\nu})\mathrm{P}_{\mathbf{z}^\nu}S^*_{M,T}S_{M,T}\mathbf{1}\\&\quad-g_\alpha(\mathrm{P}_{\mathbf{z}^\nu}S^*_{N,S}S_{N,S}\mathrm{P}_{\mathbf{z}^\nu})\mathrm{P}_{\mathbf{z}^\nu}S^*_{M,T}S_{M,T}\mathrm{P}_{\mathbf{z}^\nu}\mathbf{1}.\end{aligned}$$

We are going to estimate the norm of each ω_i, $i=\overline{1,4}$. By Lemma 3.2 and (2.5) it is easy to show that

$$\|\omega_1\|_{\mathcal{H}_\mathsf{K}}\ \le C\log^{\frac{1}{2}}\tfrac{1}{\delta}\left(\varphi(\alpha)+\Delta_{S,\mathbf{z}^\nu}+\varphi\left(N^{-\frac{1}{2}}\right)\right).\tag{4.11}$$

Further, we estimate the norm of ω_2. Applying the polar decomposition and (2.5) we get

$$\begin{aligned}&\|\omega_2\|_{\mathcal{H}_\mathsf{K}}\le\|g_\alpha(\mathrm{P}_{\mathbf{z}^\nu}S^*_{N,S}S_{N,S}\mathrm{P}_{\mathbf{z}^\nu})\left(\mathrm{P}_{\mathbf{z}^\nu}S^*_{N,S}S_{N,S}\mathrm{P}_{\mathbf{z}^\nu}\right)^{\frac{1}{2}}\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}\\
&\times\|\left(S^*_{N,S}S_{N,S}\right)^{\frac{1}{2}}(I-\mathrm{P}_{\mathbf{z}^\nu})\beta\|_{\mathcal{H}_\mathsf{K}}\le\cdot\frac{C}{\sqrt{\alpha}}\Big(\|L_S^{1/2}-(S^*_{N,S}S_{N,S})^{1/2}\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}\\
&+\|L_S^{1/2}(I-P_{\mathbf{z}^\nu})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}\Big)\|(I-P_{\mathbf{z}^\nu})\varphi(L_S)\mu_\beta\|_{\mathcal{H}_\mathsf{K}}.\end{aligned}$$

Now, by means of (4.7), (2.10) and the fact that $\sqrt{t}$ is the monotone function with confidence $1-\delta$ we get

$$\begin{aligned}\|\omega_2\|_{\mathcal{H}_\mathsf{K}}&\le\frac{C}{\sqrt{\alpha}}\left(\|L_S-S^*_{N,S}S_{N,S}\|^{1/2}_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}+\|L_S^{1/2}(I-P_{\mathbf{z}^\nu})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}\right)\\
&\times\|L_S^{1/2}(I-P_{\mathbf{z}^\nu})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}\le\frac{C}{\sqrt{\alpha}}\log^{\frac{1}{4}}\frac{1}{\delta}\left(\left(\frac{1}{\sqrt{N}}\right)^{\frac{1}{2}}+\Delta_{S,\mathbf{z}^\nu}\right)\Delta_{S,\mathbf{z}^\nu}.\end{aligned}\tag{4.12}$$

Let's estimate the third term in the right-hand side of (4.10). By (2.5) and (4.6) with confidence $1-\delta$ we get

$$\|\omega_3\|_{\mathcal{H}_\mathsf{K}}\ \le\tfrac{C}{\alpha}\log^{\frac{1}{2}}\tfrac{1}{\delta}\Big(\tfrac{1}{\sqrt{N}}+\tfrac{1}{\sqrt{M}}\Big).\tag{4.13}$$

It remains to estimate the last term in the right-hand side of (4.10). First, we rewrite ω_4 as follows

$$\omega_4=g_\alpha(\mathrm{P}_{\mathbf{z}^\nu}S^*_{N,S}S_{N,S}\mathrm{P}_{\mathbf{z}^\nu})\mathrm{P}_{\mathbf{z}^\nu}\left[(S^*_{M,T}S_{M,T}-L_T)+L_T(I-\mathrm{P}_{\mathbf{z}^\nu})\right](I-\mathrm{P}_{\mathbf{z}^\nu})\mathbf{1}.$$

Since, $\mathcal{H}_{\mathsf{K}}$ is generated by the operator L_T and $\mathbf{1} \in \mathcal{H}_{\mathsf{K}}$, there is always $\mu_1 \in \mathcal{H}_{\mathsf{K}}$ such that

$$\mathbf{1} = L_T \mu_{\mathbf{1}}.$$

In view of this and the relations (2.5) and (4.8)

$$\begin{aligned}
&\|\omega_4\|_{\mathcal{H}_{\mathsf{K}}} \le \frac{C}{\alpha}\Big(\|L_T - S^*_{M,T}S_{M,T}\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} + \|L_T^{\frac{1}{2}}(I - \mathrm{P}_{\mathbf{z}^\nu})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}\Big)\\
&\times \|(I - \mathrm{P}_{\mathbf{z}^\nu})L_T\mu_{\mathbf{1}}\|_{\mathcal{H}_{\mathsf{K}}} \le \frac{C}{\alpha}\Big(\|L_T - S^*_{M,T}S_{M,T}\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}\\
&+ \|L_T^{\frac{1}{2}}(I - \mathrm{P}_{\mathbf{z}^\nu})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}\Big)\|L_T^{\frac{1}{2}}(I - \mathrm{P}_{\mathbf{z}^\nu})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}\\
&\le \frac{C}{\alpha}\log^{\frac{1}{2}}\frac{1}{\delta}\left(\frac{1}{\sqrt{M}} + \Delta_{T,\mathbf{z}^\nu}\right)\Delta_{T,\mathbf{z}^\nu}.
\end{aligned} \tag{4.14}$$

Summing up (4.11), (4.12), (4.13) and (4.14), we finally obtain

$$\begin{aligned}
\|\beta - \tilde{\beta}^{\alpha_{M,N}}_{M,N,\mathbf{z}^\nu}\|_{\mathcal{H}_{\mathsf{K}}} &\le C\log^{\frac{1}{2}}\frac{1}{\delta}\Big[\left(\varphi(\alpha) + \Delta_{S,\mathbf{z}^\nu} + \varphi\left(N^{-\frac{1}{2}}\right)\right)\\
&+ \frac{1}{\sqrt{\alpha}}\left(\left(\frac{1}{\sqrt{N}}\right)^{\frac{1}{2}} + \Delta_{T,\mathbf{z}^\nu}\right)\\
&+ \frac{1}{\alpha}\left(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\right) + \frac{1}{\alpha}\left(\frac{1}{\sqrt{M}} + \Delta_{T,\mathbf{z}^\nu}\right)\Delta_{T,\mathbf{z}^\nu}\Big].
\end{aligned}$$

The regularization parameter α is chosen according with (4.9), namely

$$\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}} = \alpha\varphi(\alpha).$$

and consequently $\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}} \le \alpha \Rightarrow \frac{1}{\sqrt{N}} \le \alpha$, $\frac{1}{\sqrt{M}} \le \alpha$ and $\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}} \le \varphi(\alpha) \Rightarrow \frac{1}{\sqrt{N}} \le \varphi(\alpha)$, $\frac{1}{\sqrt{M}} \le \varphi(\alpha)$. Hence, by $\Delta_{S,\mathbf{z}^\nu} \le \varphi(\alpha)$ and $\Delta_{T,\mathbf{z}^\nu} \le \varphi(\alpha)$ with confidence $1 - \delta$ we have

$$\begin{aligned}
\|\beta - \tilde{\beta}^{\alpha_{M,N}}_{M,N,\mathbf{z}^\nu}\|_{\mathcal{H}_{\mathsf{K}}} &\le C\log^{\frac{1}{2}}\frac{1}{\delta}\varphi(\alpha)\\
&= O\left(\varphi\left(\theta_\varphi^{-1}\left(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\right)\right)\log^{\frac{1}{2}}\frac{1}{\delta}\right).
\end{aligned} \tag{4.15}$$

Theorem is proved. □

Remark 4.1. *Previously, the approximation of the Radon-Nikodym derivative was considered in [6, 17] in the domain adaptation scenario, but without dealing with big data. The error estimate from [6, 17] coincides (by the order) with the known error estimate for problems considered under the standard machine learning settings. We can ensure the same order of accuracy (see (4.15)) by applying the Nyström subsampling for class of problems dealing with big data.*

5 Main Result

In this section to estimate learning rate of the approximant to the target function $f^* = f_q$, we apply again the approach based on the Reproducing Kernel Hilbert Space concept and also the algorithm which is a combination of the two-times iterated Tikhonov regularization and the Nyström subsampling. Herewith, we assume that the exact values of the Radon-Nikodym derivative $\beta = \frac{d\rho_T}{d\rho_S}$ are unknown. Due to this assumption, we need to deal with the matrix

$$B_{M,N} = diag(\beta^{\lambda_{M,N}}_{M,N}(x_1), \beta^{\lambda_{M,N}}_{M,N}(x_2), \ldots, \beta^{\lambda_{M,N}}_{M,N}(x_n))$$

instead of the matrix B. Then the approximate solution (2.8) has the form

$$f^{\lambda}_{\mathbf{z}^\nu,M,N} = g_\lambda(\mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B_{M,N} S_{X_S} \mathrm{P}_{\mathbf{z}^\nu}) \mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B_{M,N} \overline{y}.$$

Note (see [6]) that for any function $f \in \mathcal{H}_{\mathsf{K}}$ it holds

$$\|S^*_{X_S} B S_{X_S} f - S^*_{X_S} B_{M,N} S_{X_S} f\|_{\mathcal{H}_{\mathsf{K}}} \le \kappa_0^3 \|\beta - \tilde{\beta}^{\alpha_{M,N}}_{M,N,\mathbf{z}^\nu}\|_{\mathcal{H}_{\mathsf{K}}} \|f\|_{\mathcal{H}_{\mathsf{K}}},$$

where $\kappa_0 > 0$ is some constant.

Theorem 5.1. *Assume that $f_q \in \mathcal{F}_{1/2}$ satisfies the source condition (2.4) and the approximant $f^{\lambda}_{\mathbf{z}^\nu,M,N}$ is of the form (2.8), then for*

$$\begin{aligned} \lambda &= \theta^{-1}\left(\log^{\frac12}\frac{1}{\delta}\left(m^{-\frac12}+n^{-\frac12}+\varphi\left(\theta_\varphi^{-1}(M^{-\frac12}+N^{-\frac12})\right)\right)\right), \\ \Delta_{T,\mathbf{z}^\nu} &\le \varphi\left(\theta^{-1}\left(\log^{\frac12}\frac{1}{\delta}\left(m^{-\frac12}+n^{-\frac12}+\varphi\left(\theta_\varphi^{-1}(M^{-\frac12}+N^{-\frac12})\right)\right)\right)\right), \end{aligned} \tag{5.1}$$

where $\theta(t) = t\varphi(t)$, with probability $1-\delta$ it holds

$$\begin{aligned} &\|f_q - f^{\lambda}_{\mathbf{z},\mathbf{z}^\nu}\|_{L_2,\rho_T} \\ &= O\Big(\sqrt{\theta^{-1}\left(\log^{\frac12}\frac{1}{\delta}\left(m^{-\frac12}+n^{-\frac12}+\varphi\left(\theta_\varphi^{-1}(M^{-\frac12}+N^{-\frac12})\right)\right)\right)} \\ &\times \varphi\left(\theta^{-1}\left(\log^{\frac12}\frac{1}{\delta}\left(m^{-\frac12}+n^{-\frac12}+\varphi\left(\theta_\varphi^{-1}(M^{-\frac12}+N^{-\frac12})\right)\right)\right)\right)\Big), \end{aligned} \tag{5.2}$$

$$\begin{aligned} &\|f_q - f^{\lambda}_{\mathbf{z},\mathbf{z}^\nu}\|_{\mathcal{H}_{\mathsf{K}}} \\ &= O\left(\varphi\left(\theta^{-1}\left(\log^{\frac12}\frac{1}{\delta}\left(m^{-\frac12}+n^{-\frac12}+\varphi\left(\theta_\varphi^{-1}(M^{-\frac12}+N^{-\frac12})\right)\right)\right)\right)\right). \end{aligned}$$

Sketch of Proof. To estimate $\|f_q - f^{\lambda}_{\mathbf{z}^\nu,M,N}\|_{\mathcal{H}_{\mathsf{K}}}$ we use the decomposition

$$\begin{aligned} f_q - f^{\lambda}_{\mathbf{z}^\nu,M,N} &= f_q - g_\lambda(\mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B_{M,N} S_{X_S} \mathrm{P}_{\mathbf{z}^\nu}) \mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B_{M,N} \overline{y} \\ &= \overline{\sigma_1} + \overline{\sigma_2} + \overline{\sigma_3} + \overline{\sigma_4}, \end{aligned}$$

where

$$\begin{aligned} \overline{\sigma_1} &:= f_q - g_\lambda(\mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B S_{X_S} \mathrm{P}_{\mathbf{z}^\nu}) \mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B S_{X_S} \mathrm{P}_{\mathbf{z}^\nu} f_q, \\ \overline{\sigma_2} &:= g_\lambda(\mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B S_{X_S} \mathrm{P}_{\mathbf{z}^\nu}) \mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B S_{X_S} \mathrm{P}_{\mathbf{z}^\nu} f_q \\ &\quad - g_\lambda(\mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B_{M,N} S_{X_S} \mathrm{P}_{\mathbf{z}^\nu}) \mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B_{M,N} S_{X_S} \mathrm{P}_{\mathbf{z}^\nu} f_q, \\ \overline{\sigma_3} &:= g_\lambda(\mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B_{M,N} S_{X_S} \mathrm{P}_{\mathbf{z}^\nu}) \mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B_{M,N} S_{X_S} \mathrm{P}_{\mathbf{z}^\nu} f_q \\ &\quad - g_\lambda(\mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B_{M,N} S_{X_S} \mathrm{P}_{\mathbf{z}^\nu}) \mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B_{M,N} S_{X_S} f_q, \\ \overline{\sigma_4} &:= g_\lambda(\mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B_{M,N} S_{X_S} \mathrm{P}_{\mathbf{z}^\nu}) \mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B_{M,N} S_{X_S} f_q \\ &\quad - g_\lambda(\mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B_{M,N} S_{X_S} \mathrm{P}_{\mathbf{z}^\nu}) \mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B_{M,N} \overline{y}. \end{aligned}$$

First, we estimate the norm of $\overline{\sigma}_2$. It is easy to see that

$$\|\overline{\sigma}_2\|_{\mathcal{H}_\mathsf{K}} \le \lambda^r \sum_{j=0}^{r-1} \lambda^{-j-1} \|\mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B S_{X_S} \mathrm{P}_{\mathbf{z}^\nu} - \mathrm{P}_{\mathbf{z}^\nu} S^*_{X_S} B_{M,N} S_{X_S} \mathrm{P}_{\mathbf{z}^\nu}\|_{\mathcal{H}_\mathsf{K}} \lambda^{-r+j} \|f_q\|_{\mathcal{H}_\mathsf{K}}$$
$$\le \frac{1}{\lambda} \|S^*_{X_S} B S_{X_S} - S^*_{X_S} B_{M,N} S_{X_S}\|_{\mathcal{H}_\mathsf{K}} \|f_q\|_{\mathcal{H}_\mathsf{K}} \le \frac{1}{\lambda} \kappa_0^3 \|\beta - \tilde{\beta}^{\alpha_{M,N}}_{M,N,\mathbf{z}^\nu}\|_{\mathcal{H}_\mathsf{K}} \|f_q\|_{\mathcal{H}_\mathsf{K}}$$
$$\le \frac{C}{\lambda} \|\beta - \tilde{\beta}^{\alpha_{M,N}}_{M,N,\mathbf{z}^\nu}\|_{\mathcal{H}_\mathsf{K}}.$$

Further, applying technique performed in Theorem 3 [16] it is easy to obtain desired estimates. We omit full proof of Theorem since it repeats the reasoning from Theorem 3. In such a way, we give a general idea of the proof of Theorem.

Remark 5.1. *Let us analyze the error estimate (5.2). Put*

$$m = O(n), \quad m^{-\frac{1}{2}} = O\left(\varphi\left(\theta_\varphi^{-1}\left(M^{-\frac{1}{2}} + N^{-\frac{1}{2}}\right)\right)\right).$$

Then the condition (5.1) *can be rewritten as*

$$\lambda = \theta^{-1}(\log^{\frac{1}{2}} \frac{1}{\delta} m^{-\frac{1}{2}}), \quad \Delta_{T,\mathbf{z}^\nu} \le \varphi\left(\theta^{-1}\left(\log^{\frac{1}{2}} \frac{1}{\delta} m^{-\frac{1}{2}}\right)\right). \tag{5.3}$$

Note that for $\varphi \in \mathcal{F}_{1/2}$ the qualification 2 covers the function $\varphi(t)\sqrt{t}$, then $\varphi(t)$ is covered by the qualification $\frac{3}{2}$ and

$$m^{-\frac{3}{10}} \le C\varphi(\Theta^{-1}(m^{-\frac{1}{2}})).$$

Therefore, if

$$\Delta_{T,\mathbf{z}^\nu} \le C m^{-\frac{3}{10}}, \tag{5.4}$$

then the second assumption (5.3) holds for any $\varphi \in \mathcal{F}_{1/2}$.

On the other hand, as it has been shown in [15] under the assumption

$$\sup_x \|L_T^{-\frac{s}{2}} \mathsf{K}_x\|_{\mathcal{H}_\mathsf{K}} \le C, \quad 0 < s \le 1,$$

with probability $1-\delta$ it holds

$$\Delta_{T,\mathbf{z}^\nu} \le C \log \frac{1}{\delta} |\mathbf{z}^\nu|^{-\beta},$$

where $\beta \in \left[\frac{1}{2(1-s)} - \varepsilon, \frac{1}{2(1-s)}\right)$, and ε is an arbitrary small positive number.

In view of (5.4), this means that for $|\mathbf{z}^\nu| = O\left(|\mathbf{z}|^{\frac{3}{10\beta}}\right)$ the assumptions (5.1) of the Theorem 5.1 are satisfied, and the best distribution independent convergence rate is achieved within the Nyström subsampling. It is clear that for $\beta > \frac{3}{10}$ the computational complexity of the designed Nyström approximant $f^\lambda_{\mathbf{z},\mathbf{z}^\nu}$ is subquadratic in the number of observations m. In this case the combination of the Nyström subsampling and the two-times iterative Tikhonov regularization (2.6) allows us to achieve the best distribution independent learning rate with the subquadratic computational complexity under a rather mild condition then $\mathsf{K}_x \in \mathrm{Range}(L_T^{\frac{1}{12}+\xi})$, where ξ is an arbitrary small positive number.

6 Conclusions

The present study is focused on the implementation of the regularized Nyström subsampling to unsupervised domain adaptation problems in the big data settings. To overcome the difference between the source and the target probabilities distributions, which is typical for domain adaptation, the covariate shift assumption is imposed. Within the framework of the Reproducing Kernel Hilbert Space concept, an algorithm is constructed. This algorithm is a combination of the Nyström subsampling and the two-steps iterated Tikhonov regularization. Herewith, we assume that values of the Radon-Nikodym derivative are unknown. We prove that the proposed approach not only guarantees the same learning rate as algorithms based on the whole sample size, but also allows to achieve subquadratic computational complexity in the number of observations.

Acknowledgments

The authors acknowledge partial financial support due to the project "Mathematical modeling of complex dynamical systems and processes caused by the state security" (Reg. No. 0123U100853).

References

1. Bakushinski, A.B.: A general method of constructing regularizing algorithms for a linear ill-posed equation in Hilbert space. USSR Comput.Math. and Math.Phys. **7**, 279-287 (1967)
2. Bauer, F., Pereverzev, S., Rosasco, L.: On regularization algorithms in learning theory. J. of Complexity. **23**, 52-72 (2007)
3. Chen, H.: The convergence rate of a regularized ranking algorithm. J. of Approximation Theory. **164**, 1513-1519 (2012)
4. Cheng, W., Ting, H., Siyang, J.: Pairwise learning problems with regularization networks and Nyström subsampling approach. Neural Networks. **157**, 176-192 (2023)
5. Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., Smola, A.: Correcting Sample Selection Bias by Unlabeled Data. Advances in Neural Information Processing Systems. **19**, 601-608 (2006)
6. Gizewski, E.R., Mayer, L., Moser, B.A., Nguyen, D.H., Pereverzyev-Jr, S., Pereverzyev, S.V., Shepeleva, N., Zellinger, W.: On a regularization of unsupervised domain adaptation in RKHS. Applied and Computational Harmonic Analysis. **57**, 201-227 (2022)
7. Kanamori, T., Hido, S., Sugiyama, M.: A least-squares approach to direct importance estimation. Journal of Machine Learning Research. **10**, 1391-1445 (2009)
8. Kriukova, G., Tkachenko, P., Pereverzyev, S.: On the convergence rate and some application of a regularized ranking algorithm. J. of Complexity. **33**, 14-29 (2016)
9. Kriukova, G., Pereverzyev-Jr., S., Tkachenko, P.: Nyström type subsampling analyzed as a regularized projection. Inverse Problems. **33** (7):074001. (2017)
10. Lin, S.-B., Guo, X., Zhou, D.-X.: Distributed Learning with Regularized Least Squares. Journal of Machine Learning Research. **18**, 1-31 (2017)
11. Lu, S., Pereverzev, S.V.: Regularization Theory. Selected Topics Inverse and Ill-posed problems. Series 58. Walter de Gruyter GmbH., Berlin/Boston (2013)
12. Lu, S., Mathe, P., Pereverzyev-Jr., S.: Analysis of regularized Nyström subsampling for regression function of low smoothness. Analysis and Application. **17** (6), 931-946 (2019)
13. Mathe, P. Pereverzev, S.V.: Geometry of linear ill-posed problems in variable Hilbert scales. Inverse Problems. **19**, 789-803 (2003)
14. Mathe, P., Pereverzev, S.V.: Discretization strategy for ill-posed problems in variable Hilbert scales. Inverse Problems. **19** (6), 1263-1277 (2003)

15. Myleiko, G.L., Pereverzyev-Jr., S., Solodky, S.G.: Regularized Nyström subsampling in regression and ranking problems under general smoothness assumptions. Analysis and Applications. **17**, 453-475 (2019)
16. Myleiko, G.L., Pereverzyev-Jr., S., Solodky, S.G.: Regularized Nyström subsampling in regression and ranking problems under general smoothness assumptions. Preprint Series of Applied Mathematics Group, Preprint No. 40. University of Inssbruck, Austria. (2017)
17. Pereverzyev, S.: An Introduction to Artificial Intelligence Based on Reproducing Kernel Hilbert Spaces. Birkhäuser Cham (2022)
18. Plato, R., Vainikko, G.M.: On the regularization of projection methods for solving ill-posed problems. Numer. Math. **57**, 63-79 (1990)
19. Rudi, A., Comoriano, R., Rosasco, L.: Less is more: Nyström computational regularization. C. Cortes, N., Lawrence, D., Lee, M., Sugiyama, R. and Garnett, R. (eds.) Curran Associates, Inc. 1648-1656 (2015)
20. Shimodaria, H.: Improving Predictive Inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference. **90**, 227-244 (2000)
21. Smale, S., Zhou, D.-X.: Learning theory estimates via integral operators and their approximations. Constructive Approximation. **26**, 153-172 (2007)
22. Sugiyama, M., Müller, K.-R.: Input-dependent estimation of generalization error under covariate shift K.-R. Müller. Statistics & Decisions. **23**, 249-279 (2005)
23. Vito, E.D., Rosasco, L., Caponnetto, A., Giovannini, U.D., Odone, F.: Learning from examples as an inverse problem. Journal of Machine Learning Research. **6**, 883-904 (2005)

Received 17.10.2023
Revised 07.11.2023