Nr. 40
01. August 2017

Leopold-Franzens-Universität Innsbruck

Preprint-Series: Department of Mathematics - Applied Mathematics

Regularized Nyström subsampling in regression and ranking problems under general smoothness assumptions

G.L. Myleiko, S. Pereverzyev Jr., S.G. Solodky

Technikerstraße 13 - 6020 Innsbruck - Austria
Tel.: +43 512 507 53803 Fax: +43 512 507 53898
https://applied-math.uibk.ac.at

Regularized Nyström subsampling in regression and ranking problems under general smoothness assumptions

G.L. Myleiko[1], **S. Pereverzyev Jr.**[2], **S.G. Solodky**[1]

[1] Institute of Mathematics, National Academy of Sciences of Ukraine, Tereschenkivska street 3, 01004 Kyiv, Ukraine

[2] Department of Mathematics, University of Innsbruck, Technikestraße 13, A-6020 Innsbruck, Austria

E-mail: anna_mileyko@ukr.net, sergiy.pereverzyev@uibk.ac.at, solodky@imath.kiev.ua

July 2017

Abstract. In the supervised learning, the Nyström type subsampling is considered as a tool for reducing the computational complexity of regularized kernel methods in the big data setting. Up to now, the theoretical analysis of this approach has been done almost exclusively in the context of the regression learning and in the case where the smoothness of the target functions is restricted to the Hölder type source conditions. Such conditions do not cover the case of target functions with high and low smoothness, which are also of practical interest. Moreover, in the case of the Hölder source conditions, there is no need to consider a regularization with high enough qualification because order-optimal learning rates are achieved by the simple Tikhonov regularization known also as the kernel ridge regression. At the same time, this learning method does not improve its performance for any smoothness higher than Hölder ones. Therefore, in this paper, our goal is to extend previous analysis of the Nyström type subsampling to the case of the general source conditions, and to the regularization schemes with high enough qualification. We also show that under rather natural assumption, our results can be easily reformulated in the ranking learning setting.

1. Introduction

The application of the theory of kernel learning machines to the case of the so-called Big Data is limited by the need to deal with large kernel matrices $\mathbb{K}_n = \{\mathsf{K}(x_i, x_j)\}_{i,j=1}^n$ consisting of the values of a kernel $\mathsf{K}(x,t)$, employed in a learning machine, at given data points $\{x_i\}_{i=1}^n$.

To avoid dealing with the entire matrix $\mathbb{K}_n$, the Nyström family of algorithms selects randomly a subset of m columns and the corresponding rows from the kernel matrix $\mathbb{K}_n$ and uses them to construct a low rank approximation for the solution of the considered learning task. Then the question is whether there is a possibility to realize

the Nyström approach with subquadratic complexity in the number of observations n without loosing the learning rates of the methods that use the full kernel matrix $\mathbb{K}_n$.

A positive answer to this question has been recently given in [2, 28] (see also [17]) in the case of the so-called kernel ridge regression (KRR), where kernel learning machines are based on Tikhonov-type regularization with a priori chosen values of the regularization parameters and subsampling sizes m. Usually, such a priori choice requires a knowledge of the smoothness of a target function, and therefore, can seldom be implemented in practice.

At the same time, the authors of [28] argue that, in principle, optimal data-driven regularization parameter choices, e.g. based on hold-out estimates [6] can be used within the Nyström approach. But hold-out, as well as similar rules, would require the construction of a family of low rank approximations corresponding to different values of a regularization parameter, and select only one of them, while leaving others aside, in spite of the numerical expenses made for their construction. Therefore, the present study will employ the idea of optimal linear combination of the constructed approximations. This idea was proposed in the case of neural networks approximation [14]. The connection of this idea with the aggregation of regularized solutions by means of the linear functional strategy [8] has been recently established in [17]. In the present study the above connection is used to aggregate Nyström approximants corresponding to different values of subsampling sizes m and regularization parameters and then to estimate the aggregation performance by means of an oracle type inequality.

Due to [28] each of Nyström approximants can be seen as a result of a combination of regularization with a projection on the linear span of the kernel sections $\{\mathsf{K}(x_i, \cdot)\}$ at subsampled points x_i. Since the goal is for each value of the regularization parameter to keep the same learning rate as for the full kernel matrix $\mathbb{K}_n$, we need to properly relate the number of subsampled points m with the currently used value of the regularization parameter. In the case of the so-called Hölder type source conditions imposed on the target functions, such a relation has been established in [28]. In the present paper we extend the results of [28] for the general case. It allows us to guarantee that all aggregated Nyström approximants preserve the learning rates corresponding to the used values of the regularization parameters and the full kernel matrix $\mathbb{K}_n$. Then we will be able to claim that the approximant resulting from the above mentioned linear aggregation provides a learning rate which is at least as good as the one given by the best, but unknown, Nyström approximant involved in the aggregation procedure.

Moreover, since the use of the Nyström approach is reasonable only up to subsampling sizes m allowing a subquadratic complexity, the above mentioned relation between m and the regularization parameter gives a natural upper bound for m and indicates a range of regularization parameter values to perform the aggregation. This range contains the values corresponding to the first most stably calculated regularized approximants starting from large parameter values.

To the best of our knowledge, up to now, the Nyström approach has been studied only in the context of the so-called kernel ridge regression (KRR). One of the aims of the

present study is to analyze this approach in the context of another form of supervised learning, such as ranking. Toward this aim, we show in particular that if the target functions of regression and ranking tasks belong to a reproducing kernel Hilbert space (RKHS) then they solve linear operator equations with the same integral operator. This observation opens an opportunity to analyze regression and ranking problems uniformly in terms of that operator. For instance, the results on learning rates of the Nyström subsampling in KRR [28, 17] can be lifted to the corresponding statements about the rates of approximation of ranking target functions by Tikhonov regularization method combined with Nyström approach.

At the same time, it is well-known that the performance of the Tikhonov regularization scheme, which is a synonym of KRR, stops improving beyond some level that does not depend on the target functions. On the other hand, the Regularization theory tells us that this obstacle can be overcomed by employing methods of higher qualification. Such methods can be seen as particular forms of the so-called general regularization scheme that was first proposed in [3] and includes, for example, iterated Tikhonov regularization. Yet another aim of the present study is to estimate the learning rates resulting from a combination of the general regularization and the Nyström subsampling.

The paper is organized as follows. In the next section we recall the settings of regression and ranking problems and discuss the relation between them in case of least square losses. We also recall the concept of general regularization scheme and its qualification. Then we follow [28] and consider the Nyström approach as a regularized projection method. In Section 3 we extend the analysis of [28] to the case of general source conditions generated by concave functions. From the Regularization theory we know that the qualification of the Tikhonov regularization scheme is enough to cover all such conditions. Therefore, in Section 3 we restrict ourselves to the consideration of the Nyström approach combined with the Tikhonov scheme, i.e., KRR. We also consider aggregation of Tikhonov regularized approximants and illustrate it by a numerical test. The case of source conditions generated by functions increasing faster than concave ones is considered in Section 4. To cover such conditions a regularization scheme should have a higher qualification than that of the Tikhonov regularization. We present a simplified analysis of the Nyström approach for the schemes with such qualification. For the sake of diversity we do this in the ranking setting.

2. Problem Settings

Recall that the techniques becoming known as supervised learning refer to algorithms that are used to predict an output $y \in Y \subset \mathbb{R} = (-\infty, \infty)$ of a system under consideration from an input $x \in X \subset \mathbb{R}^d$, and a prediction is made on the basis of the so-called training set $\mathbf{z} = \{z_i = (x_i, y_i),\ i = 1, 2, \ldots, n\},\ |\mathbf{z}| = n$, of input-output pairs observed in the same system. Moreover, the input x and the output y are assumed to be related by a probabilistic relation, such that x determines a conditional probability

$\rho(y|x)$ of y given x, which is assumed to be unknown. The input x is also assumed to be random and governed by an unknown marginal probability ρ_X on X. Then there is an unknown probability measure $\rho(x,y) = \rho_X(x)\rho(y|x)$ on the set $Z = X \times Y$ from which a training set $\mathbf{z}$ is independently drawn. Thus, in supervised learning we are interested to learn from $\mathbf{z}$ a deterministic function $f = f_{\mathbf{z}} \colon X \to Y$ that will "mimic" the relation between the inputs x and outputs y.

The error introduced by a function $f_{\mathbf{z}}$ can be measured by the expected value of a chosen loss function. If such loss function is the squared loss $(y - f(x))^2$, then the supervised learning becomes the least squares regression problem, and is one of the most well-studied problems in machine learning.

In this setting the expected loss

$$\mathcal{E}^{\text{regr}}(f) = \int_Z (y - f(x))^2 \, d\rho(x,y)$$

is minimized by the so-called regression target function

$$f_\rho(x) = \int_Y y d\rho(y|x).$$

Moreover

$$\mathcal{E}^{\text{regr}}(f) - \mathcal{E}^{\text{regr}}(f_\rho) = \|f - f_\rho\|_\rho^2, \tag{1}$$

where $\|\cdot\|_\rho := \|\cdot\|_{L_2(X,\rho_X)}$ is the norm in the space $L_2(X,\rho_X)$ of square integrable functions with respect to the marginal probability measure.

In recent years another learning problem called ranking has been extensively studied in parallel to regression. Here we refer to [9, 7, 1, 26, 16, 34, 35], just to mention a few publications.

In the ranking setting for given true ranks (outputs) y, y' of the inputs $x,\ x' \in X$ the value $(y - y' - (f(x) - f(x')))^2$ is usually interpreted as the magnitude-preserving least squares loss [1, 7, 10, 34, 35] and the corresponding expected loss

$$\mathcal{E}^{\text{rank}}(f) = \int_{Z \times Z} (y - y' - (f(x) - f(x')))^2 \, d\rho(x,y) d\rho(x',y')$$

is minimized by a family of the ranking target functions

$$\mathcal{F}_\rho = \{f : f(x) = f_\rho(x) + c,\ c \in \mathbb{R}\}.$$

Note that in practice the minimization of the expected losses $\mathcal{E}^{\text{regr}}(f)$, $\mathcal{E}^{\text{rank}}(f)$ in the natural space $L_2(X,\rho_X)$ can not be directly performed, and the target functions $f_\rho,\ f_\rho + c \in \mathcal{F}_\rho$, can not be found, because the condition probability distribution $\rho(y|x)$ is unknown. Therefore, the goal might be to perform the minimization of the expected losses over some (accessible) hypothesis space $\mathcal{H} \subset L_2(X,\rho_X)$. A widely used choice of such a space is a Reproducing Kernel Hilbert space (RKHS) $\mathcal{H} = \mathcal{H}_{\mathsf{K}}$ generated by a kernel $\mathsf{K} \colon X \times X \to \mathbb{R}$.

In the sequel we will use the concept of universal kernels [25] and assume that the kernel K generating $\mathcal{H}_\mathsf{K}$ is universal on X. Then it is known [31, 6] that

$$f_\rho \in \mathcal{H}_\mathsf{K}.$$

Moreover, from [29] we know that f_ρ uniquely solves the integral equation

$$T_\mathsf{K} f = J_\mathsf{K}^* f_\rho, \tag{2}$$

where $J_\mathsf{K} : \mathcal{H}_\mathsf{K} \hookrightarrow L_2(X, \rho_X)$ is the inclusion operator, $J_\mathsf{K}^* : L_2(X, \rho_X) \to \mathcal{H}_\mathsf{K}$ is its adjoint, and $T_\mathsf{K} = J_\mathsf{K}^* J_\mathsf{K}$. The latter admit the representations

$$J_\mathsf{K}^* f(\cdot) = \int_X \mathsf{K}(\cdot, x) f(x) d\rho_X(x),$$

$$T_\mathsf{K} = \int_X \langle \cdot, \mathsf{K}_x \rangle_{\mathcal{H}_\mathsf{K}} \mathsf{K}_x d\rho_X(x),$$

where $\mathsf{K}_x = \mathsf{K}(\cdot, x) \in \mathcal{H}_\mathsf{K}$, and $\langle \cdot, \cdot \rangle_{\mathcal{H}_\mathsf{K}}$ is the inner product in $\mathcal{H}_\mathsf{K}$.

Note that the ranking target functions from $\mathcal{F}_\rho$ can also be described as the solutions of an integral equation of the first kind. Namely, from [7, 16] it follows that any $f \in \mathcal{F}_\rho$ solves the equation

$$L_\mathsf{K} f = L_\mathsf{K} f_\rho, \tag{3}$$

where $L_\mathsf{K} \colon \mathcal{H}_\mathsf{K} \to \mathcal{H}_\mathsf{K}$ is given as

$$L_\mathsf{K} f(\cdot) = \int_{X \times X} \left[\mathsf{K}(x, \cdot) - \mathsf{K}(x', \cdot) \right] f(x) d\rho_X(x) d\rho_X(x').$$

If from now on we assume that $\mathcal{H}_\mathsf{K}$ contains constant functions $f(x) \equiv c \in \mathbb{R}$, then for any $f \in \mathcal{H}_\mathsf{K}$

$$L_\mathsf{K} f = T_\mathsf{K} \left(f - \int_X f(x') d\rho_X(x') \right).$$

From this relation and (3) it is easy to see that the ranking target function

$$\overline{f}_\rho(x) = f_\rho(x) - \int_X f_\rho(x') d\rho_X(x')$$

uniquely solves the equation

$$T_\mathsf{K} f = L_\mathsf{K} f_\rho \tag{4}$$

with the operator T_K from the regression setting (2). Thus, if $\mathcal{H}_\mathsf{K}$ is generated by a universal kernel K and contains constant functions, then regression and ranking problems considered in such $\mathcal{H}_\mathsf{K}$ can be reduced to the equations (2), (4), which differ only by the right-hand sides, but involve the same operator T_K.

This observation allows a uniform approach to the regression and ranking problems with least squares losses.

Note that another possibility of a uniform approach to regression and ranking problems has been discussed in [15]. At the same time, as it can be seen from [29, 7],

in RKHS-setting these problems of supervised learning have been studied separately. Here we return to possibility of a uniform approach, because it will allow us to analyze the Nyström subsampling in both problem settings by employing some results of [28] obtained for the regression setting.

Note that the equations (2), (4) are ill-posed because the involved operator T_{K} is compact and its inverse cannot be a bounded operator in $\mathcal{H}_{\mathsf{K}}$. Therefore, these equations should be analyzed by methods of the Regularization theory. In the Hilbert space setting the analysis is usually started with relating the solutions, such as f_ρ and $\overline{f}_\rho$, to the equation operator, such as T_{K}, by the so-called source conditions. From [21] we know that there are always continuous, nondecreasing functions φ, $\overline{\varphi}$ such that $\varphi(0) = \overline{\varphi}(0) = 0$ and

$$f_\rho \in \mathrm{Range}(\varphi(T_{\mathsf{K}})), \quad \overline{f}_\rho \in \mathrm{Range}(\overline{\varphi}(T_{\mathsf{K}})). \tag{5}$$

Such functions are called index functions of the source conditions, such as (5), and they are used as terms for describing rates of convergence of regularization schemes.

Recall (see [3]) that regularization schemes can also be indexed by parameterized functions $g_\alpha\colon [0,T] \to \mathbb{R}$, $T > \|T_{\mathsf{K}}\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}$, $\alpha > 0$. The only requirements are that there are positive constants $\gamma_0, \overline{\gamma}, \tilde{\gamma}$ such that

$$\sup_{0<t\le T} |1 - t g_\alpha(t)| \le \gamma_0, \qquad \sup_{0<t\le T} \sqrt{t}|g_\alpha(t)| \le \frac{\overline{\gamma}}{\sqrt{\alpha}}, \qquad \sup_{0<t\le T} |g_\alpha(t)| \le \frac{\tilde{\gamma}}{\alpha}. \tag{6}$$

Further properties of a regularization scheme indexed by g_α can be characterized by the so-called qualification that is the maximal positive number p for which

$$\sup_{0<t\le T} t^p |1 - t g_\alpha(t)| \le \gamma_p \alpha^p, \tag{7}$$

where γ_p does not depend on α.

Recall that the popular Tikhonov regularization method, which is in the context of learning is also known as Kernel Regularized Least Squares (KRLS), corresponds to $g_\alpha(t) = (\alpha + t)^{-1}$ and has the qualification $p = 1$. Note that in the previous studies on Nyström subsampling [30, 33, 28, 17] it is mainly considered in a combination with KRLS.

At the same time, there are other methods admitting a straightforward combination with the Nyström approach. One example is the l-times iterated Tikhonov regularization method that corresponds to

$$g_\alpha(t) = g_{l,\alpha}(t) = \sum_{i=1}^{l} \alpha^{i-1}(\alpha + t)^{-i} = \frac{1}{t}\left(1 - \frac{\alpha^l}{(\alpha + t)^l}\right) \tag{8}$$

and has the qualification $p = l$.

The following definition [20, 22] describes an interplay between the qualification and source conditions.

Definition 1. *We say that the qualification p covers the index function φ, if the function $t \to t^p/\varphi(t)$ is nondecreasing for $t \in (0, T]$.*

The importance of this definition is that it allows us to extend the inequality (7) to all index functions covered by the qualification p. Namely, the following statement is known [20, 22].

Proposition 1. *Let a regularization method be indexed by $g_\alpha(t)$ and have the qualification p. If this qualification covers an index function φ then*

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

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

Observe now that the equations (2), (4) are not only ill-posed but also inaccessible, because the integral operators T_{K}, L_{K} depend on the unknown marginal distribution ρ_X. Therefore, any regularization scheme cannot be applied directly to (2), (4). On the other hand, it is known (see [29, 7]), that these equations can be discretized with the use of data from the training set $\mathbf{z}$.

Namely, if we consider the sampling operator

$$S_{\mathbf{z}} : \mathcal{H}_{\mathsf{K}} \to \mathbb{R}^{|\mathbf{z}|}, \qquad S_{\mathbf{z}} f = (f(x_i))_{i=1}^{|\mathbf{z}|} \in \mathbb{R}^{|\mathbf{z}|},$$

then following [29, 7] the discretized versions of (2), (4) can be written respectively as

$$S_{\mathbf{z}}^* S_{\mathbf{z}} f = S_{\mathbf{z}}^* \mathbf{Y} \qquad \text{and} \qquad S_{\mathbf{z}}^* S_{\mathbf{z}} f = S_{\mathbf{z}}^* \mathbf{D}\mathbf{Y}, \tag{9}$$

where $\mathbf{Y} = (y_i)_{i=1}^{|\mathbf{z}|} \in \mathbb{R}^{|\mathbf{z}|}$, $\quad \mathbf{D} = \mathbf{I} - |\mathbf{z}|^{-1}\mathbf{1} \times \mathbf{1}$, $\quad \mathbf{I}$, $\mathbf{1}$ are the $|\mathbf{z}|$-th order unit matrix and the vector of all ones, and $S_{\mathbf{z}}^* \colon \mathbb{R}^{|\mathbf{z}|} \to \mathcal{H}_{\mathsf{K}}$ is the adjoint of $S_{\mathbf{z}}$; if the norm in $\mathbb{R}^{|\mathbf{z}|}$ is defined as $|\mathbf{z}|^{-1}$-times the standard Euclidean norm, then

$$(S_{\mathbf{z}}^* \mathbf{u})(x) = \frac{1}{|\mathbf{z}|} \sum_{i=1}^{|\mathbf{z}|} u_i \mathsf{K}(x, x_i), \quad \mathbf{u} = (u_i)_{i=1}^{|\mathbf{z}|} \in \mathbb{R}^{|\mathbf{z}|}.$$

If we assume from now on that all possible outputs y are bounded by a constant, say B, then according to [29, 7] a perturbation of (2), (4) caused by the discretization (9) can be estimated with probability at least $1 - \delta$ as follows

$$\begin{gathered} \|T_{\mathsf{K}} - S_{\mathbf{z}}^* S_{\mathbf{z}}\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \leq \kappa_{1,\delta} |\mathbf{z}|^{-1/2}, \quad \|L_{\mathsf{K}} f_\rho - S_{\mathbf{z}}^* \mathbf{D}\mathbf{Y}\|_{\mathcal{H}_{\mathsf{K}}} \leq \kappa_{2,\delta} |\mathbf{z}|^{-1/2}, \\ \|J_{\mathsf{K}}^* f_\rho - S_{\mathbf{z}}^* \mathbf{Y}\|_{\mathcal{H}_{\mathsf{K}}} \leq \kappa_{3,\delta} |\mathbf{z}|^{-1/2}, \end{gathered} \tag{10}$$

where $\kappa_{i,\delta}$, $i = 1, 2, 3$, are of order $O(\log \frac{1}{\delta})$ and depend also only on K and B.

Applying a regularization method indexed by g_α to the discretized equations (9) we obtain the following approximations of regression and ranking functions f_ρ, $\overline{f}_\rho$ respectively

$$f_{\mathbf{z}}^\alpha = g_\alpha(S_{\mathbf{z}}^* S_{\mathbf{z}}) S_{\mathbf{z}}^* \mathbf{Y}, \qquad \overline{f}_{\mathbf{z}}^\alpha = g_\alpha(S_{\mathbf{z}}^* S_{\mathbf{z}}) S_{\mathbf{z}}^* \mathbf{D}\mathbf{Y}. \tag{11}$$

In view of the relation

$$g_\alpha\left(S_{\mathbf{z}}^* S_{\mathbf{z}}\right) S_{\mathbf{z}}^* = S_{\mathbf{z}}^* g_\alpha\left(S_{\mathbf{z}} S_{\mathbf{z}}^*\right)$$

(see [32, Lemma 3.1, p.34]) and the representation

$$S_{\mathbf{z}} S_{\mathbf{z}}^* = |\mathbf{z}|^{-1} \mathbb{K}_{|\mathbf{z}|}, \qquad \mathbb{K}_{|\mathbf{z}|} = \left\{\mathsf{K}(x_i, x_j)\right\}_{i,j=1}^{|\mathbf{z}|},$$

it is clear that a regularization method indexed by g_α is in fact operating with a Gram matrix of the kernel values. Then the computational complexity of regularized kernel learning methods (11) is at least quadratic in the number of observations $|\mathbf{z}|$, and in the Big Data setting, where $|\mathbf{z}|$ is large, this is not acceptable.

As we already mentioned, to avoid dealing with the entire matrix $\mathbb{K}_{|\mathbf{z}|}$, the Nyström algorithms replace $\mathbb{K}_{|\mathbf{z}|}$ by a smaller low-rank matrix obtained by random subsampling of columns of $\mathbb{K}_{|\mathbf{z}|}$ that correspond to $x_i : (x_i, y_i) \in \mathbf{z}^\nu \subset \mathbf{z}$, $|\mathbf{z}^\nu| \ll |\mathbf{z}|$. A deep observation made in [28] is that the Nyström approach can be seen as a combination of a regularization g_α with projection onto subspaces

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

More precisely, in the Nyström approach the target functions are approximated not by (11), but by

$$f_{\mathbf{z},\mathbf{z}^\nu}^\alpha = g_\alpha\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right) P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* \mathbf{Y}, \qquad \overline{f}_{\mathbf{z},\mathbf{z}^\nu}^\alpha = g_\alpha\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right) P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* \mathbf{D}\mathbf{Y}, \tag{12}$$

where $P_{\mathbf{z}^\nu} : \mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}^{\mathbf{z}^\nu}$, $\|P_{\mathbf{z}^\nu}\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} = 1$, is the orthogonal projection operator with the range $\mathcal{H}_{\mathsf{K}}^{\mathbf{z}^\nu}$.

Note that to compute (12) it is not necessary to construct the projector $P_{\mathbf{z}^\nu}$ explicitly. For example, from [28] it follows that KRLS-approximants (12) corresponding to $g_\alpha(t) = (\alpha + t)^{-1}$ have the form of a linear combination

$$f(\cdot) = \sum_{l=1}^{|\mathbf{z}^\nu|} c_l \mathsf{K}\left(\cdot, \tilde{x}_l\right), \qquad \tilde{x}_l = x_{i_l} : (x_{i_l}, y_{i_l}) \in \mathbf{z}^\nu \subset \mathbf{z}, \quad l = 1, 2, \ldots, |\mathbf{z}^\nu|, \tag{13}$$

where the coefficients c_l solve the linear least squares problem

$$|\mathbf{z}|^{-1} \sum_{i=1}^{|\mathbf{z}|} \left(u_i - \sum_{l=1}^{|\mathbf{z}^\nu|} c_l \mathsf{K}\left(x_i, \tilde{x}_l\right)\right)^2 + \alpha \sum_{l,j=1}^{|\mathbf{z}^\nu|} c_l c_j \mathsf{K}\left(\tilde{x}_l, \tilde{x}_j\right) \to \min,$$

and the vector $U = (u_i)_{i=1}^{|\mathbf{z}|}$ is either $\mathbf{Y}$ (in the case of regression problem) or $\mathbf{D}\mathbf{Y}$ (in the case of ranking problem).

Following [28], in the sequel we consider two types of Nyström methods. In the first one, called Plain Nyström, the points $\tilde{x}_l = x_{i_l}$, such as in (13), are sampled uniformly at random without replacement from the inputs x_i of the training data set $\mathbf{z} = \{(x_i, y_i), i = 1, 2, \ldots, |\mathbf{z}|\}$.

We also consider the so-called Approximate leverage scores (ALS) Nyström method, in which the points $\tilde{x}_l = x_{i_l}, l = 1, 2, \ldots, |\mathbf{z}^\nu|$, are sampled from $\{x_i\}$ independently with replacement, and with probability to be selected given by $P_\alpha(i) = \tilde{l}_\alpha(i)/\sum_{j=1}^{|\mathbf{z}|} \tilde{l}_\alpha(j)$, where $\tilde{l}_\alpha(j)$ are the so-called d-approximate leverage scores such that for all $\alpha \in (\alpha_0, T)$, $\alpha_0 > |\mathbf{z}|^{-1}$, we have $d^{-1} l_\alpha(j) \le \tilde{l}_\alpha(j) \le d l_\alpha(j)$, and $l_\alpha(j)$ is the j-th diagonal element of the matrix $\mathbb{K}_{|\mathbf{z}|}(\mathbb{K}_{|\mathbf{z}|} + \alpha|\mathbf{z}|\mathbf{I})^{-1}$. Note that the effective computation of d-approximate leverage scores has been discussed, for example, in [12]. A theoretical advantage of ALS-Nyström approach will be seen in the next section.

3. Nyström type subsampling in the Tikhonov regularization

In this section we extend the analysis of [28] to the case of general source conditions (5) and then use it for aggregating Tikhonov regularized approximants.

From [4, 18] we know that the smoothness of the target functions, such as f_ρ, $\overline{f}_\rho$ in (5), is fully utilized when the qualification of a regularization method g_α employed in solving learning tasks covers not only the index functions $\varphi(t)$, $\overline{\varphi}(t)$, but also their products with $\sqrt{t}$. Therefore, dealing with the Tikhonov regularization $g_\alpha(t) = (\alpha+t)^{-1}$ with the qualification $p = 1$ it is reasonable to assume that the source conditions (5) are indexed by functions φ, $\overline{\varphi}$ tending to zero not faster than $\sqrt{t}$. Moreover, without essential loss of generality we may even assume that φ^2, $\overline{\varphi}^2$ are concave functions. Furthermore, as in the previous studies on the Tikhonov regularization in supervised learning [29, 4, 18] we assume that the source conditions (5) are indexed by the so-called operator monotone functions φ, $\overline{\varphi}$ such that, for example, for any non-negative self-adjoint operators $A, B\colon \mathcal{H}_\mathsf{K} \to \mathcal{H}_\mathsf{K}$ with spectra in $[0, T]$ it holds

$$\|\varphi(A) - \varphi(B)\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \le C\varphi\left(\|A - B\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}\right). \tag{14}$$

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 K, ρ, φ, p or the constants γ_0, $\overline{\gamma}$, $\widetilde{\gamma}$, γ_p in (6), (7).

Further explanations about source conditions indexed by operator monotone functions can be found in [20, Section 2.7.2]. We denote by $\mathcal{F}_{OC}$ the class of all operator monotone index functions $\varphi\colon [0, T] \to \mathbb{R}_+$ such that $\varphi^2(t)$ are concave. It is known [23] that for $\varphi \in \mathcal{F}_{OC}$ it holds

$$\|(I - P_{\mathbf{z}^\nu})\varphi(T_\mathsf{K})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \le \varphi\left(\left\|T_\mathsf{K}^{1/2}(I - P_{\mathbf{z}^\nu})\right\|^2_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}\right). \tag{15}$$

In view of (15), it is natural to measure the approximation power of the regularized projection methods (12) induced by the projectors $P_{\mathbf{z}^\nu}$ in terms of the quantity $\left\|T_\mathsf{K}^{1/2}(I - P_{\mathbf{z}^\nu})\right\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}$. Note that such kind of measure is usual in studying regularized projection methods (see, e.g., [23, 27]). In the sequel, to simplify formulas, we use the notation $\Delta_{\mathbf{z}^\nu} := \left\|T_\mathsf{K}^{1/2}(I - P_{\mathbf{z}^\nu})\right\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}$.

It is clear that in the context of the Nyström subsampling approach, the value of $\Delta_{\mathbf{z}^\nu}$ has a probabilistic character and depends on the way we perform the subsampling $\mathbf{z}^\nu$. This dependence can be clarified by employing the results of [28] formulated with the use of the quantities

$$\mathcal{N}_x(\alpha) = \left\langle \mathsf{K}_x, (\alpha I + T_{\mathsf{K}})^{-1} \mathsf{K}_x \right\rangle_{\mathcal{H}_{\mathsf{K}}}, \quad \mathcal{N}_\infty(\alpha) = \sup \{ \mathcal{N}_x(\alpha),\ x \in X \},$$

and

$$\mathcal{N}(\alpha) = \int_X \mathcal{N}_x(\alpha) d\rho_X(x) = \operatorname{trace}\left((\alpha I + T_{\mathsf{K}})^{-1} T_{\mathsf{K}} \right).$$

The latter one is called the effective dimension of $\mathcal{H}_{\mathsf{K}}$ and itself plays an important role in analyzing regularized learning methods, as it has been observed in [5] (see also [13, 19]).

The following statement, proven in [28] as Lemmas 6, 7, is important for the analysis of this section.

Proposition 2. *If $\mathbf{z}^\nu$ is subsampled according to the plain Nyström method, and*

$$|\mathbf{z}^\nu| \geq C \log \frac{1}{\delta} \mathcal{N}_\infty(\alpha) \log \frac{1}{\alpha},$$

or $\mathbf{z}^\nu$ is subsampled according to the ALS-Nyström method and

$$|\mathbf{z}^\nu| \geq C \log \frac{1}{\delta} \mathcal{N}(\alpha) \log |\mathbf{z}|, \ \alpha \geq C \log \frac{1}{\delta} \left(|\mathbf{z}|^{-1} \log |\mathbf{z}| \right),$$

then with confidence at least $1 - \delta$, it holds

$$\left\| (I - P_{\mathbf{z}^\nu}) (\alpha I + T_{\mathsf{K}})^{1/2} \right\|^2_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \leq 3\alpha.$$

Corollary 1. *Under the conditions of Proposition 2, with confidence at least $1 - \delta$, it holds*

$$\Delta^2_{\mathbf{z}^\nu} \leq 3\alpha.$$

Proof. The proof of this corollary follows from the spectral calculus and Proposition 2:

$$\begin{aligned}
\Delta^2_{\mathbf{z}^\nu} &= \left\| T_{\mathsf{K}}^{1/2} (I - P_{\mathbf{z}^\nu}) \right\|^2_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} = \left\| (I - P_{\mathbf{z}^\nu}) (\alpha I + T_{\mathsf{K}})^{1/2} (\alpha I + T_{\mathsf{K}})^{-1/2} T_{\mathsf{K}} \right\|^2_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \\
&\leq \left\| (I - P_{\mathbf{z}^\nu}) (\alpha I + T_{\mathsf{K}})^{1/2} \right\|^2_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \cdot \left\| (\alpha I + T_{\mathsf{K}})^{-1/2} T_{\mathsf{K}} \right\|^2_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \\
&\leq 3\alpha \sup_t \left| \frac{t^{1/2}}{(\alpha + t)^{1/2}} \right|^2 \leq 3\alpha.
\end{aligned}$$

□

Proposition 3. *If $f_\rho \in \operatorname{Range}(\varphi(T_{\mathsf{K}}))$ and $\varphi \in \mathcal{F}_{OC}$, then under the condition of Proposition 2, with confidence at least $1 - \delta$, it holds*

$$\| (I - P_{\mathbf{z}^\nu}) f_\rho \|_{L_2(X, \rho_X)} \leq C \sqrt{\alpha} \varphi(\alpha).$$

Proof. Let $v \in \mathcal{H}_{\mathsf{K}}$ be such that $f_\rho = \varphi(T_{\mathsf{K}})\, v$. With the use of the polar decomposition of J_{K}, we have

$$\|(I - P_{\mathbf{z}^\nu})\, f_\rho\|_{L_2(X,\rho_X)} = \|J_{\mathsf{K}}\,(I - P_{\mathbf{z}^\nu})\, f_\rho\|_{\mathcal{H}_{\mathsf{K}}} = \left\|T_{\mathsf{K}}^{1/2}\,(I - P_{\mathbf{z}^\nu})\,\varphi(T_{\mathsf{K}})\, v\right\|_{\mathcal{H}_{\mathsf{K}}}.$$

Then Corollary 1 and (15) give us

$$\begin{aligned}\|(I - P_{\mathbf{z}^\nu})\, f_\rho\|_{L_2(X,\rho_X)} &\le \|v\|_{\mathcal{H}_{\mathsf{K}}} \cdot \left\|T_{\mathsf{K}}^{1/2}\,(I - P_{\mathbf{z}^\nu})\right\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} \cdot \|(I - P_{\mathbf{z}^\nu})\,\varphi(T_{\mathsf{K}})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} \\ &\le \|v\|_{\mathcal{H}_{\mathsf{K}}}\,\sqrt{3\alpha}\varphi(3\alpha) \le C\sqrt{\alpha}\varphi(\alpha),\end{aligned}$$

where we use the fact that φ is covered by the qualification $p = 1$ such that $t \mapsto t/\varphi(t)$ is a nondecreasing function and

$$\frac{\alpha}{\varphi(\alpha)} \le \frac{3\alpha}{\varphi(3\alpha)} \implies \varphi(3\alpha) \le 3\varphi(\alpha).$$

The proposition is proved. □

Remark 1. *Note that the restriction $\alpha > C\,|\mathbf{z}|^{-1}\log|\mathbf{z}|$ in Proposition 2 is natural because in view of* (10) *$\varepsilon = O\left(|\mathbf{z}|^{-1/2}\right)$ is the level of the noise in equations* (9) *as compared to* (2),(4). *Then the regularization theory (see, e.g., [20]) tells us that in general the accuracy guaranteed by any of the regularization methods cannot be better than $O\left(|\mathbf{z}|^{-1/2}\right)$. Moreover, the convergence of a regularization method indexed by g_α is guaranteed only if $\varepsilon^2 = o(\alpha)$, or that is the same $|\mathbf{z}|^{-1} = o(\alpha)$. Therefore, the assumption that $\alpha > C\,|\mathbf{z}|^{-1}\log|\mathbf{z}|$ can be made without essential loss of generality.*

We are now in a position to formulate the basic result of this section.

Theorem 1. *Let $f_\rho \in \mathrm{Range}\,(\varphi(T_{\mathsf{K}}))$, $\varphi \in \mathcal{F}_{OC}$. If $f^\alpha_{\mathbf{z},\mathbf{z}^\nu}$ is given by* (12) *with $g_\alpha(t) = (\alpha + t)^{-1}$, then under the conditions of Proposition 2, with confidence at least $1 - \delta$, it holds*

$$\left(\mathcal{E}^{\mathrm{regr}}(f_\rho) - \mathcal{E}^{\mathrm{regr}}\left(f^\alpha_{\mathbf{z},\mathbf{z}^\nu}\right)\right)^{1/2} = \left\|f_\rho - f^\alpha_{\mathbf{z},\mathbf{z}^\nu}\right\|_{L_2(X,\rho_X)} \le C\log\frac{1}{\delta}\left(\varphi(\alpha)\sqrt{\alpha} + \sqrt{\frac{\mathcal{N}(\alpha)}{|\mathbf{z}|}}\right).$$

Proof. Using again the polar decomposition of J_{K}, we have

$$\left\|f_\rho - f^\alpha_{\mathbf{z},\mathbf{z}^\nu}\right\|_{L_2(X,\rho_X)} = \left\|J_{\mathsf{K}}\left(f_\rho - f^\alpha_{\mathbf{z},\mathbf{z}^\nu}\right)\right\|_{\mathcal{H}_{\mathsf{K}}} = \left\|T_{\mathsf{K}}^{1/2}\left(f_\rho - f^\alpha_{\mathbf{z},\mathbf{z}^\nu}\right)\right\|_{\mathcal{H}_{\mathsf{K}}} \le I_1 + I_2, \qquad (16)$$

where

$$\begin{aligned}I_1 &= \left\|T_{\mathsf{K}}^{1/2}\left(I - g_\alpha\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}\right)f_\rho\right\|_{\mathcal{H}_{\mathsf{K}}}, \\ I_2 &= \left\|T_{\mathsf{K}}^{1/2}g_\alpha\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}\left(\mathbf{Y} - S_{\mathbf{z}}f_\rho\right)\right\|_{\mathcal{H}_{\mathsf{K}}}.\end{aligned}$$

The term I_2 has been estimated in [28] such that with confidence at least $1 - \delta$, it holds

$$I_2 \le C\log\frac{1}{\delta}\left(\frac{\sqrt{\mathcal{N}_\infty(\alpha)}}{|\mathbf{z}|} + \sqrt{\frac{\mathcal{N}(\alpha)}{|\mathbf{z}|}}\right).$$

Note that for any $x \in X$

$$\begin{aligned}\mathcal{N}_x(\alpha) &= \left\langle \mathsf{K}_x, (\alpha I + T_{\mathsf{K}})^{-1} \mathsf{K}_x \right\rangle_{\mathcal{H}_{\mathsf{K}}} \leq \|\mathsf{K}_x\|^2_{\mathcal{H}_{\mathsf{K}}} \cdot \left\|(\alpha I + T_{\mathsf{K}})^{-1}\right\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \\ &\leq \max_x \mathsf{K}(x,x) \cdot \sup_t \left|(\alpha + t)^{-1}\right| \leq \alpha^{-1} \max_x \mathsf{K}(x,x),\end{aligned}$$

and for $\alpha > C \log \frac{1}{\delta} |\mathbf{z}|^{-1} \log |\mathbf{z}| > C |\mathbf{z}|^{-1}$, due to the uniform boundedness of K, we have $\mathcal{N}_\infty(\alpha) < C |\mathbf{z}|$, and

$$\frac{\sqrt{\mathcal{N}_\infty(\alpha)}}{|\mathbf{z}|} < C |\mathbf{z}|^{-1/2} < C \sqrt{\frac{\mathcal{N}(\alpha)}{|\mathbf{z}|}}$$

such that

$$I_2 \leq C \log \frac{1}{\delta} \sqrt{\frac{\mathcal{N}(\alpha)}{|\mathbf{z}|}}. \tag{17}$$

The rest of the proof is about the estimation of I_1. It is easy to check that

$$I_1 \leq I_{11} + I_{12} + I_{13}, \tag{18}$$

where

$$\begin{aligned} I_{11} &= \left\| T_{\mathsf{K}}^{1/2} \left(I - P_{\mathbf{z}^\nu}\right) f_\rho \right\|_{\mathcal{H}_{\mathsf{K}}} = \left\| \left(I - P_{\mathbf{z}^\nu}\right) f_\rho \right\|_{L_2(X,\rho_X)}, \\ I_{12} &= \left\| T_{\mathsf{K}}^{1/2} \left(I - g_\alpha \left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right) P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right) P_{\mathbf{z}^\nu} f_\rho \right\|_{\mathcal{H}_{\mathsf{K}}}, \\ I_{13} &= \left\| T_{\mathsf{K}}^{1/2} g_\alpha \left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right) P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} \left(I - P_{\mathbf{z}^\nu}\right) f_\rho \right\|_{\mathcal{H}_{\mathsf{K}}}. \end{aligned}$$

The term I_{11} has been already estimated in Proposition 3 such that with confidence at least $1 - \delta$, it holds

$$I_{11} \leq C \sqrt{\alpha} \varphi(\alpha). \tag{19}$$

In the estimation I_{12}, I_{13}, we use the following inequalities proven in [28]:

$$\left\| \left(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}}\right)^{1/2} \left(\alpha I + P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right)^{-1} P_{\mathbf{z}^\nu} \left(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}}\right)^{1/2} \right\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \leq 1, \tag{20}$$

$$\left\| \left(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}}\right)^{-1/2} T_{\mathsf{K}}^{1/2} \right\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \leq \left\| \left(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}}\right)^{-1/2} \left(\alpha I + T_{\mathsf{K}}\right)^{1/2} \right\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \leq 2, \tag{21}$$

$$\left\| \left(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}}\right)^{1/2} \left(\alpha I + T_{\mathsf{K}}\right)^{-1/2} \right\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \leq 2. \tag{22}$$

These inequalities hold for $\alpha > C \log \frac{1}{\delta} \left(|\mathbf{z}|^{-1} \log |\mathbf{z}|\right)$ with confidence at least $1 - \delta$.

Observe that

$$\begin{aligned} I_{12} &= \left\| T_{\mathsf{K}}^{1/2} \left(I - \left(\alpha I + P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right)^{-1} P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right) P_{\mathbf{z}^\nu} f_\rho \right\|_{\mathcal{H}_{\mathsf{K}}} \\ &= \left\| T_{\mathsf{K}}^{1/2} \left(\alpha I + P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right)^{-1} \left(\alpha I + P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu} - P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right) P_{\mathbf{z}^\nu} f_\rho \right\|_{\mathcal{H}_{\mathsf{K}}} \\ &= \alpha \left\| T_{\mathsf{K}}^{1/2} \left(\alpha I + P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right)^{-1} P_{\mathbf{z}^\nu} f_\rho \right\|_{\mathcal{H}_{\mathsf{K}}} \\ &= \alpha \Big\| T_{\mathsf{K}}^{1/2} \left(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}}\right)^{-1/2} \left(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}}\right)^{1/2} \left(\alpha I + P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right)^{-1} P_{\mathbf{z}^\nu} \circ \\ &\qquad \left(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}}\right)^{1/2} \left(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}}\right)^{-1/2} \left(\alpha I + T_{\mathsf{K}}\right)^{1/2} \left(\alpha I + T_{\mathsf{K}}\right)^{-1/2} f_\rho \Big\|_{\mathcal{H}_{\mathsf{K}}} \end{aligned}$$

and using (20), (21), we can continue

$$I_{12} \leq \alpha \left\|(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}})^{-1/2} T_{\mathsf{K}}\right\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \left\|(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}})^{1/2} (\alpha I + P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu})^{-1} P_{\mathbf{z}^\nu} \circ \right.$$
$$\left. (\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}})^{1/2}\right\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \left\|(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}})^{-1/2} (\alpha I + T_{\mathsf{K}})^{1/2}\right\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \times$$
$$\left\|(\alpha I + T_{\mathsf{K}})^{-1/2} f_\rho\right\|_{\mathcal{H}_{\mathsf{K}}} \leq 4\alpha \left\|(\alpha I + T_{\mathsf{K}})^{-1/2} f_\rho\right\|_{\mathcal{H}_{\mathsf{K}}}. \tag{23}$$

Recall that $f_\rho = \varphi(T_{\mathsf{K}})\, v$, $v \in \mathcal{H}_{\mathsf{K}}$, and $\varphi(t)$ is such that the function $\varphi^2(t)$ is concave. Then $\varphi^2(t)$ is covered by the qualification $p = 1$ because due to the concavity of $\varphi^2(t)$, $\varphi^2(0) = 0$, for any $t < t_1$, the point $(t, \varphi^2(t)) \in \mathbb{R}^2$ is above the straight line $t \mapsto \frac{\varphi^2(t_1)}{t_1} t$ that interpolates the function $\varphi^2(t)$ at the points $t = 0$, $t = t_1$, i.e.

$$\varphi^2(t) \geq \frac{\varphi^2(t_1)}{t_1} t \implies \frac{t_1}{\varphi^2(t_1)} \geq \frac{t}{\varphi^2(t)}, \quad t_1 > t.$$

By Definition 1, this means that $\varphi^2(t)$ is covered by the qualification $p = 1$. This observation allows us to use Proposition 1 such that for $g_\alpha(t) = (\alpha + t)^{-1}$

$$\left\|(\alpha I + T_{\mathsf{K}})^{-1/2} f_\rho\right\|_{\mathcal{H}_{\mathsf{K}}} = \left\|(\alpha I + T_{\mathsf{K}})^{-1/2} \varphi(T_{\mathsf{K}})\, v\right\|_{\mathcal{H}_{\mathsf{K}}} \leq \|v\|_{\mathcal{H}_{\mathsf{K}}} \sup_t \left((\alpha + t)^{-1/2} \varphi(t)\right) \leq$$
$$C \sup_t \left((\alpha + t)^{-1} \varphi^2(t)\right)^{1/2} = C \sup_t \alpha^{-1/2} \left|(1 - g_\alpha(t) t)\, \varphi^2(t)\right|^{1/2} \leq$$
$$C\alpha^{-1/2} \left(\varphi^2(\alpha)\right)^{1/2} = C\alpha^{-1/2} \varphi(\alpha).$$

Then from (23), we have

$$I_{12} \leq C\sqrt{\alpha}\varphi(\alpha). \tag{24}$$

To complete the proof, we need to estimate I_{13}. Observe that $P_{\mathbf{z}^\nu}(I - P_{\mathbf{z}^\nu}) = 0$, and therefore, for any $\alpha > 0$,

$$P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} (I - P_{\mathbf{z}^\nu}) = P_{\mathbf{z}^\nu} (\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}}) (I - P_{\mathbf{z}^\nu}).$$

Then using the inequalities (20)-(22), we have

$$I_{13} = \left\|T_{\mathsf{K}}^{1/2} g_\alpha (P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}) P_{\mathbf{z}^\nu} (\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}}) (I - P_{\mathbf{z}^\nu}) f_\rho\right\|_{\mathcal{H}_{\mathsf{K}}} \leq$$
$$\left\|T_{\mathsf{K}}^{1/2} (\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}})^{-1/2}\right\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \left\|(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}})^{1/2} (\alpha I + P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu})^{-1} P_{\mathbf{z}^\nu} \circ \right.$$
$$\left. (\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}})^{1/2}\right\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \left\|(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}})^{1/2} (I - P_{\mathbf{z}^\nu}) f_\rho\right\|_{\mathcal{H}_{\mathsf{K}}} \leq$$
$$2 \left\|(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}})^{1/2} (I - P_{\mathbf{z}^\nu}) f_\rho\right\|_{\mathcal{H}_{\mathsf{K}}}. \tag{25}$$

Moreover, keeping in mind that $f_\rho = \varphi(T_{\mathsf{K}})\, v$, $\varphi \in \mathcal{F}_{OC}$, and using (15) together with Corollary 1, we continue as follows:

$$\left\|(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}})^{1/2} (I - P_{\mathbf{z}^\nu}) f_\rho\right\|_{\mathcal{H}_{\mathsf{K}}} \leq \left\|(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}})^{1/2} (I - P_{\mathbf{z}^\nu})\right\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \times$$
$$\|(I - P_{\mathbf{z}^\nu}) \varphi(T_{\mathsf{K}})\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \|v\|_{\mathcal{H}_{\mathsf{K}}} \leq C\varphi\left(\left\|T_{\mathsf{K}}^{1/2} (I - P_{\mathbf{z}^\nu})\right\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}}^2\right) \times$$
$$\left\|(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}})^{1/2} (I - P_{\mathbf{z}^\nu})\right\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \leq C\varphi(\alpha) \left\|(\alpha I + S_{\mathbf{z}}^* S_{\mathbf{z}})^{1/2} (I - P_{\mathbf{z}^\nu})\right\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}}.$$

Furthermore, Proposition 2 and (22) give us

$$\left\|\left(\alpha I+S_{\mathbf{z}}^{*}S_{\mathbf{z}}\right)^{1/2}\left(I-P_{\mathbf{z}^{\nu}}\right)\right\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}\leq\left\|\left(\alpha I+S_{\mathbf{z}}^{*}S_{\mathbf{z}}\right)^{1/2}\left(\alpha I+T_{\mathsf{K}}\right)^{-1/2}\right\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}\times$$
$$\left\|\left(\alpha I+T_{\mathsf{K}}\right)^{1/2}\left(I-P_{\mathbf{z}^{\nu}}\right)\right\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}\leq 2\left\|\left(I-P_{\mathbf{z}^{\nu}}\right)\left(\alpha I+T_{\mathsf{K}}\right)^{1/2}\right\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}\leq C\sqrt{\alpha}. \tag{26}$$

Then from (25)-(26), it follows that

$$I_{13}\leq C\sqrt{\alpha}\varphi(\alpha).$$

Summing up the estimates (17)-(19), (24), we obtain the statement of the theorem. □

Carefully repeating the steps of the proof of Theorem 1, one can show that under the conditions of Proposition 2, for $\overline{f}_{\rho}\in\operatorname{Range}\left(\overline{\varphi}\left(T_{\mathsf{K}}\right)\right)$, $\overline{\varphi}\in\mathcal{F}_{OC}$, and $\overline{f}_{\mathbf{z},\mathbf{z}^{\nu}}^{\alpha}$ given by (12) with $g_{\alpha}(t)=(\alpha+t)^{-1}$, with confidence at least $1-\delta$, it holds

$$\left\|\overline{f}_{\rho}-\overline{f}_{\mathbf{z},\mathbf{z}^{\nu}}^{\alpha}\right\|_{L_2(X,\rho_X)}\leq C\log\frac{1}{\delta}\left(\overline{\varphi}(\alpha)\sqrt{\alpha}+\sqrt{\frac{\mathcal{N}(\alpha)}{|\mathbf{z}|}}\right).$$

For further discussions, it is important to compare Theorem 1 with the known bounds of the learning rates of KRR corresponding to (11) with $g_{\alpha}(t)=(\alpha+t)^{-1}$. From [5, 19], it follows that under the conditions of Theorem 1, with confidence at least $1-\delta$, we have

$$\left(\mathcal{E}^{\mathrm{regr}}\left(f_{\rho}\right)-\mathcal{E}^{\mathrm{regr}}\left(f_{\mathbf{z}}^{\alpha}\right)\right)^{1/2}\leq C\log\frac{1}{\delta}\left(\varphi(\alpha)\sqrt{\alpha}+\sqrt{\frac{\mathcal{N}(\alpha)}{|\mathbf{z}|}}\right). \tag{27}$$

The comparison of this estimation with Theorem 1 shows that for any $\alpha>C\log\frac{1}{\delta}\left(|\mathbf{z}|^{-1}\log|\mathbf{z}|\right)$, the Nyström subsampling preserves the known order of the learning rate of KRR, provided that the subsampling size $|\mathbf{z}^{\nu}|=|\mathbf{z}^{\nu}(\alpha)|$ meets the conditions of Proposition 2. Now, we discuss the conditions under which the Nyström approach achieves the same learning rate as (27) with subquadratic computational complexity in the sample size $|\mathbf{z}|$.

From Proposition 2 and Theorem 1, it follows that in the case of the plain Nyström method, the bound (27) is preserved independently of the smoothness of the index function $\varphi\in\mathcal{F}_{OC}$ if for a given α

$$|\mathbf{z}^{\nu}|=|\mathbf{z}^{\nu}\left(\alpha\right)|=C\log\frac{1}{\delta}\,\mathcal{N}_{\infty}(\alpha)\log\frac{1}{\alpha}. \tag{28}$$

Recall that $\mathcal{N}_{\infty}(\alpha)$ is the exact upper bound for

$$\mathcal{N}_{x}(\alpha)=\left\langle\mathsf{K}_{x},\left(\alpha I+T_{\mathsf{K}}\right)^{-1}\mathsf{K}_{x}\right\rangle_{\mathcal{H}_{\mathsf{K}}}=\left\|\left(\alpha I+T_{\mathsf{K}}\right)^{-1/2}\mathsf{K}_{x}\right\|_{\mathcal{H}_{\mathsf{K}}}^{2}.$$

Referring again to [21], we note that for any $x \in X$, there is always an index function $\psi = \psi_x \colon [0, T] \to \mathbb{R}$, $\psi_x(0) = 0$, such that $\mathsf{K}_x \in \operatorname{Range}(\psi_x(T_{\mathsf{K}}))$. As in [11], one can even assume that for any $x \in X$, $\psi_x(t) = t^{s/2}$, $0 < s \leq 1$, and

$$\mathsf{K}_x = T_{\mathsf{K}}^{s/2} v_x, \quad \|v_x\|_{\mathcal{H}_{\mathsf{K}}} \leq C. \tag{29}$$

Then

$$\mathcal{N}_\infty(\alpha) = \sup_x \left\| (\alpha I + T_{\mathsf{K}})^{-1/2} \mathsf{K}_x \right\|_{\mathcal{H}_{\mathsf{K}}}^2 \leq C \sup_t \left| \frac{t^{s/2}}{(\alpha + t)^{1/2}} \right|^2 \leq C \alpha^{s-1},$$

and

$$|\mathbf{z}^\nu(\alpha)| \leq C \log \frac{1}{\delta}\, \alpha^{s-1} \log \frac{1}{\alpha}.$$

In view of Remark 1, it is reasonable to consider $\alpha > C\,|\mathbf{z}|^{-1} \log |\mathbf{z}|$. For such values of α, we have

$$|\mathbf{z}^\nu(\alpha)| \leq C \log \frac{1}{\delta}\, |\mathbf{z}|^{1-s} \log |\mathbf{z}|\,. \tag{30}$$

On the other hand, it is known (see, e.g., [28]) that direct methods for constructing $f^{\alpha}_{\mathbf{z},\mathbf{z}^\nu}$ in form (13) require $O\left(|\mathbf{z}| \cdot |\mathbf{z}^\nu|^2 + |\mathbf{z}^\nu|^3\right)$ arithmetic operations, where $O\left(|\mathbf{z}^\nu|^3\right)$ operations are required for solving a system of linear equations with respect to the coefficients $\{\, c_\ell \,\}_{\ell=1}^{|\mathbf{z}^\nu|}$ in (13), and $O\left(|\mathbf{z}| \cdot |\mathbf{z}^\nu|^2\right)$ operations are necessary to form the corresponding matrix. Then from Theorem 1 and (29),(30), we can conclude that for all smoothness index functions $\varphi \in \mathcal{F}_{OC}$ and all values of α from the range of interest, the plain Nyström subsampling preserves the order of the learning rate (27) of KRR and has a subquadratic computational complexity provided a rather mild condition

$$\mathsf{K}_x \in \operatorname{Range}\left(T_{\mathsf{K}}^{1/4+\varepsilon}\right)$$

holds, where ε is an arbitrary small positive number.

From Proposition 2, we see that in the case of the ALS-Nyström method, the corresponding subsampling size $|\mathbf{z}^\nu| = |\mathbf{z}^\nu(\alpha)|$ is governed mainly by the effective dimension $\mathcal{N}(\alpha)$. In the literature (see, e.g., [5, 6, 13]), it is usually assumed that

$$\mathcal{N}(\alpha) \leq C \alpha^{-\beta}, \tag{31}$$

for some $\beta \in (0, 1)$. Since $\mathcal{N}(\alpha) \leq \mathcal{N}_\infty(\alpha)$, from our discussion above, it is clear that under condition (29) introduced in [11], this assumption is satisfied with $\beta = 1 - s$. Therefore, for the ALS-Nyström subsampling, assumption (31) leads to the same conclusion as in the case of the plain Nyström method.

At the same time, as it has been noted in [19], for some universal kernels K, one may expect a logarithmic decay of $\mathcal{N}(\alpha)$ such that

$$\mathcal{N}(\alpha) \leq C \log^\beta \frac{1}{\alpha}. \tag{32}$$

This assumption, together with Proposition 2 and Theorem 1, tells us that for any $\varphi \in \mathcal{F}_{OC}$ and $\alpha > C\,|\mathbf{z}|^{-1}\log|\mathbf{z}|$, the ALS-Nyström method preserves the order of the learning rate of KRR given by (27) provided that the subsampling size $|\mathbf{z}^\nu|$ is chosen as

$$|\mathbf{z}^\nu| = C\log\frac{1}{\delta}\,\log^{1+\beta}|\mathbf{z}|\,. \tag{33}$$

For such subsampling size, the number of the arithmetic operations required for constructing the ALS-Nyström approximant $f^\alpha_{\mathbf{z},\mathbf{z}^\nu}$ is of order $O\left(|\mathbf{z}|\log^{2(1+\beta)}|\mathbf{z}|\right)$, i.e. it is not only subquadratic, but superlinear.

Observe that for the plain Nyström method a superlinear complexity can also be stated under the assumption that instead of (29), the following source condition is satisfied:

$$\mathsf{K}_x = \psi\left(T_{\mathsf{K}}\right)v_x,\ \|v_x\| \le C,\ \psi(t) = t^{1/2}\log^{\beta/2}\frac{1}{t},\ \beta > 0. \tag{34}$$

Then

$$\mathcal{N}_\infty(\alpha) \le C\sup_t \frac{t\log^\beta t}{\alpha + t} \le C\log^\beta\frac{1}{\alpha},$$

and (28) can be rewritten as

$$|\mathbf{z}^\nu| = |\mathbf{z}^\nu(\alpha)| = C\log\frac{1}{\delta}\,\log^{1+\beta}\frac{1}{\alpha}, \tag{35}$$

that leads to the same conclusion as (33). Note that the source condition (34) can be assumed for rather smooth kernels, such as Gaussian ones, or/and rather fast decay of the singular values of the operator T_{K}, which of course depends on the unknown marginal probability ρ_X.

The choice of the subsample size $|\mathbf{z}^\nu| = |\mathbf{z}^\nu(\alpha)|$ suggested by Theorem 1 keeps the order of the learning rate of KRR guaranteed by (27) for the whole sample size $|\mathbf{z}|$ and the considered value of the regularization parameter α. The choice $|\mathbf{z}^\nu| = |\mathbf{z}^\nu(\alpha)|$ does not require a knowledge of the smoothness index function φ and is governed only by $\mathcal{N}_\infty(\alpha)$ or by the effective dimension $\mathcal{N}(\alpha)$, which can be estimated from the data, as it has been discussed, for example, in [28, 19].

The construction of $f^\alpha_{\mathbf{z},\mathbf{z}^\nu(\alpha)}$ is computationally much less expensive that it is for the KRR-approximant $f^\alpha_{\mathbf{z}}$, but the performance of the both approximations is very dependent on the value of the regularization parameter α that needs to be carefully selected.

Note that due to Theorem 1, to determine α in $f^\alpha_{\mathbf{z},\mathbf{z}^\nu(\alpha)}$, one can, in principle, use the same rules as for $f^\alpha_{\mathbf{z}}$. Among these rules, we can mention cross-validation based adaptation [6], and the balancing principle [19]. These, as well as similar rules, require the construction of the approximants $f^\alpha_{\mathbf{z},\mathbf{z}^\nu(\alpha)}$ or $f^\alpha_{\mathbf{z}}$ for several values of $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_\ell > C\,|\mathbf{z}|^{-1}\log|\mathbf{z}|$, and then select one of them according to some criterion. After such a selection, the other approximants are left aside, in spite of the numerical expenses made for their construction.

Instead, similar to [8, 17], we propose to aggregate all constructed $f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)}$ in the form

$$\tilde{f}_{\mathbf{z}} = \sum_{i=1}^{\ell} d_i f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)}, \tag{36}$$

where the vector $\mathbf{d} = (d_1, d_2, \ldots, d_\ell) \in \mathbb{R}^\ell$ solves the system $\mathbf{G}\mathbf{d} = \mathbf{g}$ of linear equations with the matrix $\mathbf{G} = (G_{ij})_{i,j=1}^{\ell}$,

$$G_{ij} = \left\langle S_{\mathbf{z}} f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)}, S_{\mathbf{z}} f^{\alpha_j}_{\mathbf{z},\mathbf{z}^\nu(\alpha_j)} \right\rangle_{\mathbb{R}^{|\mathbf{z}|}} = |\mathbf{z}|^{-1} \sum_{k=1}^{|\mathbf{z}|} f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)}(x_k)\ f^{\alpha_j}_{\mathbf{z},\mathbf{z}^\nu(\alpha_j)}(x_k),$$

and the right-hand-side vector $\mathbf{g} = (g_1, g_2, \ldots, g_\ell) \in \mathbb{R}^\ell$,

$$g_i = \left\langle S_{\mathbf{z}} f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)}, \mathbf{Y} \right\rangle_{\mathbb{R}^{|\mathbf{z}|}} = |\mathbf{z}|^{-1} \sum_{k=1}^{|\mathbf{z}|} y_k f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)}(x_k).$$

From Theorem 1, it follows that $f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)}$, $i = 1, 2, \ldots, \ell$, guarantee the same order of the learning rates as the corresponding KRR-approximants $f^{\alpha_i}_{\mathbf{z}}$. Moreover, the following statement holds true.

Theorem 2. *If the matrix* $\mathbf{G}$ *is well-conditioned, then with confidence at least* $1 - \delta$*, it holds*

$$\left\| f_\rho - \tilde{f}_{\mathbf{z}} \right\|_{L_2(X,\rho_X)} = \min_{c_i} \left\| f_\rho - \sum_{i=1}^{\ell} c_i f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)} \right\|_{L_2(X,\rho_X)} + O\left(|\mathbf{z}|^{-1/2} \log \frac{1}{\delta} \right), \tag{37}$$

where the coefficient implicit in the O*-symbol may depend on* ℓ*,* K*,* ρ*, but does not depend on* $|\mathbf{z}|$*.*

Note that, as it has been explained in Remark 1, the term $O\left(|\mathbf{z}|^{-1/2} \log \frac{1}{\delta} \right)$ is negligible. Moreover, it can be checked directly whether or not the matrix $\mathbf{G}$ is well-conditioned. This matrix is not well-conditioned when $f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)}$, $i = 1, 2, \ldots, \ell$, are close to be linearly dependent. The linear independence of $\left\{ f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)} \right\}$ can be restored by excluding some of $f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)}$ that can be detected, and that do not effect the value of the first term in the right-hand-side of (37).

Also note that the number of additional arithmetic operations required for constructing aggregator (36) is of order $O\left(\ell^3 + \ell^2 |\mathbf{z}|\right)$, and it cannot exceed a subquadratic complexity threshold unless the number ℓ of the considered values of the regularization parameter $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_\ell$, is of order $\ell = O\left(N^{1/2}\right)$ that never happens in practice, especially in the bid data setting.

We omit the proof of Theorem 2 because it is similar to that of Theorem 3 in [17]. The difference is that in [17], the aggregation has been considered in the case where the Nyström approximants $f^{\alpha}_{\mathbf{z},\mathbf{z}^\nu}$ are constructed with different subsample sizes $|\mathbf{z}^\nu|$, but with the same value of the regularization parameter α. In this way, one cannot expect

to achieve an optimal bound of the learning rate unless α is properly chosen a priori, and this can seldom be the case.

At the same time, from Theorems 1 and 2, it follows that with confidence at least $1-\delta$, we have

$$\begin{aligned}\left\|f_\rho-\tilde{f}_{\mathbf{z}}\right\|_{L_2(X,\rho_X)} &\le \min_{\alpha_i}\left\|f_\rho-f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)}\right\|_{L_2(X,\rho_X)}+O\left(|\mathbf{z}|^{-1/2}\log\frac{1}{\delta}\right)\\ &\le C\log\frac{1}{\delta}\min_{\alpha_i}\left(\varphi\left(\alpha_i\right)\sqrt{\alpha_i}+\sqrt{\frac{\mathcal{N}\left(\alpha_i\right)}{|\mathbf{z}|}}\right).\end{aligned}$$

In view of (27), this means that the aggregator (36) guarantees at least the order of the best learning rate bound that in principle can be achieved by KRR-approximants $f^{\alpha}_{\mathbf{z}}$ with α from the considered set of values of the regularization parameter. A simple illustrative example below shows that $\left\|f_\rho-\tilde{f}_{\mathbf{z}}\right\|_{L_2(X,\rho_X)}$ can be really smaller that $\min\limits_{\alpha_i}\left\|f_\rho-f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)}\right\|_{L_2(X,\rho_X)}$.

For an illustration, we use the same example as in [24]. Following that paper, we simulate the training data set $\mathbf{z}=\{\,(x_i,y_i)\,\}_{i=1}^{|\mathbf{z}|}$, $|\mathbf{z}|=10^4$, from the regression model $y_i=f_\rho(x_i)+\xi_i$, where

$$\begin{aligned}&f_\rho(x)=0.1\big(\,x+2\big(\,\exp\left(-8\left(4\pi/3-x\right)^2\right)-\exp\left(-8\left(\pi/2-x\right)^2\right)-\\&\exp\left(-8\left(3\pi/2-x\right)^2\right)\,\big)\,\big),\ x\in[0,2\pi],\end{aligned}$$

the random samples x_i are uniformly distributed over $[0,2\pi]$, and the noise random variables ξ_i are uniformly distributed over $[-0.05,0.05]$. The RKHS $\mathcal{H}_{\mathsf{K}}$ is generated by the kernel $\mathsf{K}(x,t)=xt+\exp\left(-8(x-t)^2\right)$, and $f_\rho\in\mathcal{H}_{\mathsf{K}}$.

As it has been shown in [19], for the considered kernel K, the simulated data suggest a logarithmic decay of $\mathcal{N}_\infty(\alpha)$ and $\mathcal{N}(\alpha)$. In particular, (32) is valid with $c=2.7$, $\beta=1$. Therefore, in view of Proposition 2, we relate the subsample sizes $|\mathbf{z}^\nu|=|\mathbf{z}^\nu\left(\alpha_i\right)|$ with the values of the regularization parameters $\alpha=\alpha_i$ as follows:

$$|\mathbf{z}^\nu|=|\mathbf{z}^\nu\left(\alpha_i\right)|=\lfloor 10\log\left(1/\alpha_i\right)\rfloor. \tag{38}$$

Then we construct $f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)}$ with $\alpha_i=10^{-3}\cdot 3^{i-7}$, $i=1,2,\ldots,11$, and $|\mathbf{z}^\nu\left(\alpha_i\right)|$ chosen according (38), see Table 1.

In the considered simulation, we observe that

$$\min_i\left\|f_\rho-f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)}\right\|_{L_2(0,2\pi)}=\left\|f_\rho-f^{\alpha_5}_{\mathbf{z},\mathbf{z}^\nu(\alpha_5)}\right\|_{L_2(0,2\pi)}=3.55\cdot 10^{-3}. \tag{39}$$

At the same time, for $|\mathbf{z}|=10^4$, the subsample size $|\mathbf{z}^\nu\left(\alpha_5\right)|=91$ is too large to allow the construction of $f^{\alpha_5}_{\mathbf{z},\mathbf{z}^\nu(\alpha_5)}$ with a subquadratic computational complexity in $|\mathbf{z}|$. Such a complexity can be expected for $|\mathbf{z}^\nu\left(\alpha_i\right)|$, $i=7,8,\ldots,11$. Aggregating the corresponding approximants $f^{\alpha_i}_{\mathbf{z},\mathbf{z}^\nu(\alpha_i)}$, $i=7,8,\ldots,11$, according to (36), we observe the error

$$\left\|f_\rho-\tilde{f}_{\mathbf{z}}\right\|_{L_2(0,2\pi)}=3.47\cdot 10^{-3}$$

i	1	2	3	4	5	6	7	8	9	10	11
$\lvert\mathbf{z}^{\nu}(\alpha_i)\rvert$	134	124	113	102	91	80	69	58	47	36	25

Table 1. Chosen subsample sizes $|\mathbf{z}^{\nu}(\alpha_i)|$ according to (38).

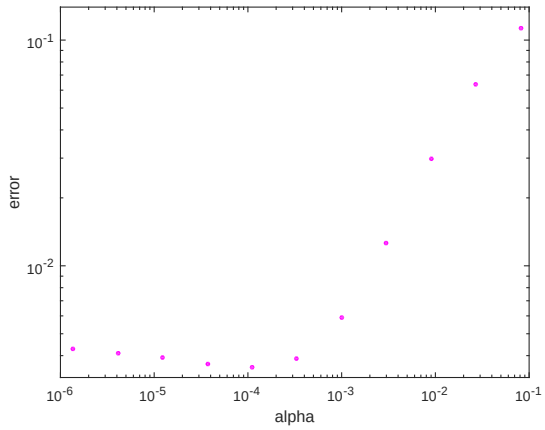

Figure 1. The errors $\left\|f_{\rho}-f^{\alpha_i}_{\mathbf{z},\mathbf{z}^{\nu}(\alpha_i)}\right\|_{L_2(0,2\pi)}$ in the considered example. The utmost left point corresponds to α_1, and the utmost right point — to α_{11}.

that is smaller than the error of the aggregated approximants $f^{\alpha_i}_{\mathbf{z},\mathbf{z}^{\nu}(\alpha_i)}$, $i=7,8,\dots,11$ (see Figure 1), and even a bit smaller than the minimal observed error (39).

Thus, the presented illustration supports our theoretical results and demonstrates a potential advantage of the aggregation (36).

4. Nyström type subsampling in a regularization with high enough qualification

The analysis in the previous section essentially uses the assumption that the index functions, describing smoothness in terms of the source conditions (5), do not tend to zero faster than $\sqrt{t}$. The qualification $p=1$ of the Tikhonov regularization $g_{\alpha}(t)=(\alpha+t)^{-1}$ is enough to cover all such index functions, as well as their products with $\sqrt{t}$.

In this section we discuss the case of higher smoothness, when index functions $\varphi(t)$, $\overline{\varphi}(t)=O(\sqrt{t})$, $0\le t\le 1$, may not be covered by the qualification of simple Tikhonov regularization.

To guarantee the best learning rates for such index functions one needs to combine the Nyström approach with regularization schemes $g_{\alpha}(t)$ having qualifications $p>1$.

It is known [4, 18] that under the source conditions (5) the best learning rates guaranteed with confidence $1-\delta$ independently of unknown distribution ρ are of order $O\left(\varphi^2(\Theta^{-1}(|\mathbf{z}|^{-\frac{1}{2}}))\Theta^{-1}(|\mathbf{z}|^{-\frac{1}{2}})\right)$, where $\Theta(t)=\varphi(t)t$, the coefficient implicit in the symbol O grows as $O(\log\frac{1}{\delta})$ with $\delta\to 0$, and for simplicity we use the same symbol φ for both index functions in (5).

From the results of this section it follows that, under mild assumptions, the above mentioned rates can be achieved with subquadratic computational complexity by a combination of a regularization scheme g_α with the qualification $p \geq 2$ and a Nyström type subsampling with a rather economical subsampling size $|\mathbf{z}|^\nu$ that can be chosen a priori without knowledge of smoothness index function φ.

To simplify the presentation we restrict our attention to the plain Nyström subsampling. Moreover, since in the previous section the Nyström approach has been discussed in the regression setting, here, for the sake of diversity, we consider the ranking setting.

We start with specifying a set of smoothness index functions φ that need a regularization with high enough qualification to be covered. Following [4], we consider the class $\mathcal{F}_{Lip}$ of index functions $\varphi\colon [0,T] \to \mathbb{R}_+$ allowing splitting $\varphi(t) = \psi_{Lip}(t)\cdot\psi_{mon}(t)$ into monotone Lipschitz parts ψ_{Lip}, $\psi_{Lip}(0) = 0$, and operator monotone parts $\psi_{mon}(t)$ such that $\psi_{mon}(0) = 0$, and $\psi^2_{mon}(t)$ is a concave function, and $\psi_{mon}(T) \leq 1$; the source conditions (5) indexed by $\varphi(t) = \psi_{mon}(t)$ have been considered in the previous section. Observe that for $\varphi \in \mathcal{F}_{Lip}$ the representation $\varphi(t) = \psi_{Lip}(t) \cdot \psi_{mon}(t)$ is not unique such that we implicitly assume that the Lipschitz constant for ψ_{Lip} is equal to 1, which means that for any non-negative self-adjoint operators $A, B : \mathcal{H}_\mathsf{K} \to \mathcal{H}_\mathsf{K}$ with spectra in $[0,T]$ it holds

$$\|\psi_{Lip}(A) - \psi_{Lip}(B)\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \leq \|A - B\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}. \tag{40}$$

Moreover, from [23] it follows that

$$\|P_{\mathbf{z}^\nu}\psi_{mon}(T_\mathsf{K})P_{\mathbf{z}^\nu} - \psi_{mon}(P_{\mathbf{z}^\nu}T_\mathsf{K}P_{\mathbf{z}^\nu})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \leq 2\psi_{mon}(\|T_\mathsf{K}^{1/2}(I - P_{\mathbf{z}^\nu})\|^2_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}). \tag{41}$$

Furthermore, for $\varphi \in \mathcal{F}_{Lip}$ we have

$$\|(I - P_{\mathbf{z}^\nu})\varphi(T_\mathsf{K})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \leq C\|T_\mathsf{K}^{1/2}(I - P_{\mathbf{z}^\nu})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}. \tag{42}$$

Here and in the sequel we continue to follow our convention about the use of the symbol C.

Now we are in a position to formulate the main result of this section.

Theorem 3. *Assume that $\overline{f}_\rho \in \mathrm{Range}(\varphi(T_\mathsf{K}))$ with $\varphi \in \mathcal{F}_{Lip}$. If the qualification p of a regularization method $g_\alpha(t)$ covers the function $\varphi(t)\sqrt{t}$, then for*

$$\alpha = \Theta^{-1}(|\mathbf{z}|^{-\frac{1}{2}}) \quad \textit{and} \quad \Delta_{\mathbf{z}^\nu} \leq \varphi\left(\Theta^{-1}(|\mathbf{z}|^{-1/2})\right) \tag{43}$$

with confidence $1 - \delta$ it holds

$$\mathcal{E}^{\mathrm{rank}}(\overline{f}^\alpha_{\mathbf{z},\mathbf{z}^\nu}) - \mathcal{E}^{\mathrm{rank}}(\overline{f}_\rho) = O\left(\varphi^2(\Theta^{-1}(|\mathbf{z}|^{-\frac{1}{2}}))\Theta^{-1}(|\mathbf{z}|^{-\frac{1}{2}})\log\frac{1}{\delta}\right),$$

where $\overline{f}^\alpha_{\mathbf{z},\mathbf{z}^\nu}$ is determined by (12).

To prove this theorem we need the following auxiliary assertion.

Lemma 1. *Assume that* $\overline{f}_\rho \in \mathrm{Range}(\varphi(T_{\mathsf{K}}))$, *where* $\varphi(t) = \psi_{Lip}(t) \cdot \psi_{mon}(t)$, $\varphi \in \mathcal{F}_{Lip}$. *If the qualification of a regularization method* $g_\alpha(t)$ *covers the index function* $\varphi(t)$, *then with confidence* $1-\delta$ *it holds*

$$\|\overline{f}_\rho - \overline{f}^{\alpha}_{\mathbf{z},\mathbf{z}^\nu}\|_{\mathcal{H}_{\mathsf{K}}} \le C \log\frac{1}{\delta}\Bigg\{\varphi(\alpha) + \psi_{Lip}(\alpha)\left[\psi_{mon}(\Delta^2_{\mathbf{z}^\nu}) + \psi_{mon}\Big(|\mathbf{z}|^{-1/2}\Big)\right] + \Delta_{\mathbf{z}^\nu} + \alpha^{-1/2}\Big(|\mathbf{z}|^{-1/4}\Delta_{\mathbf{z}^\nu} + \Delta^2_{\mathbf{z}^\nu}\Big) + \alpha^{-1}|\mathbf{z}|^{-1/2}\Bigg\}.$$

Proof. At first we observe that

$$\overline{f}_\rho - \overline{f}^{\alpha}_{\mathbf{z},\mathbf{z}^\nu} = \overline{f}_\rho - g_\alpha\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}\mathbf{DY} = \sigma_1 + \sigma_2 + \sigma_3, \tag{44}$$

where

$$\begin{aligned}
\sigma_1 &:= \overline{f}_\rho - g_\alpha\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\overline{f}_\rho,\\
\sigma_2 &:= g_\alpha\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\overline{f}_\rho - g_\alpha\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}\overline{f}_\rho,\\
\sigma_3 &:= g_\alpha\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}\overline{f}_\rho - g_\alpha\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}\mathbf{DY}.
\end{aligned}$$

Further, we use the notation $r_\alpha(t) = 1 - tg_\alpha(t)$ and estimate the norms of each terms σ_1, σ_2 and σ_3. In the condition of the lemma $\overline{f}_\rho = \varphi(T_{\mathsf{K}})v_\rho$, $v_\rho \in \mathcal{H}_{\mathsf{K}}$. Then

$$\begin{aligned}
\sigma_1 &= \left(1 - g_\alpha\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)\varphi(T_{\mathsf{K}})v_\rho\\
&= r_\alpha\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)\varphi\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)v_\rho\\
&+ r_\alpha\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)\left[\varphi(T_{\mathsf{K}}) - \varphi\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)\right]v_\rho = \sigma_{11} + \sigma_{12},
\end{aligned}$$

where

$$\begin{aligned}
\sigma_{11} &:= r_\alpha\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)\varphi\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)v_\rho,\\
\sigma_{12} &:= r_\alpha\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)\left[\varphi(T_{\mathsf{K}}) - \varphi\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)\right]v_\rho.
\end{aligned}$$

In view of Proposition 1 the norm of σ_{11} is immediately estimated as

$$\|\sigma_{11}\|_{\mathcal{H}_{\mathsf{K}}} \le C\varphi(\alpha). \tag{45}$$

Moreover, the bracketed part of σ_{12} can be rewritten as

$$\begin{aligned}
\varphi(T_{\mathsf{K}}) - \varphi\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right) &= \psi_{Lip}\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)\left[\psi_{mon}(T_{\mathsf{K}}) - \psi_{mon}\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)\right]\\
&+ \left[\psi_{Lip}(T_{\mathsf{K}}) - \psi_{Lip}\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)\right]\psi_{mon}(T_{\mathsf{K}}),
\end{aligned}$$

and due to Proposition 1, (6), (40) we have

$$\begin{aligned}
\|\sigma_{12}\|_{\mathcal{H}_{\mathsf{K}}} &\le C\psi_{Lip}(\alpha)\|\psi_{mon}(T_{\mathsf{K}}) - \psi_{mon}\left(P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\right)\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}\\
&+ C\|T_{\mathsf{K}} - P_{\mathbf{z}^\nu}S^*_{\mathbf{z}}S_{\mathbf{z}}P_{\mathbf{z}^\nu}\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}.
\end{aligned} \tag{46}$$

Further, we estimate the norm $\|\psi_{mon}(T_{\mathsf{K}}) - \psi_{mon}\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right)\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}}$:

$$\psi_{mon}(T_{\mathsf{K}}) - \psi_{mon}\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right) = P_{\mathbf{z}^\nu}\psi_{mon}(T_{\mathsf{K}}) - P_{\mathbf{z}^\nu}\psi_{mon}(T_{\mathsf{K}})P_{\mathbf{z}^\nu} + \psi_{mon}(T_{\mathsf{K}})$$
$$-P_{\mathbf{z}^\nu}\psi_{mon}(T_{\mathsf{K}}) + P_{\mathbf{z}^\nu}\psi_{mon}(T_{\mathsf{K}})P_{\mathbf{z}^\nu} - \psi_{mon}(P_{\mathbf{z}^\nu}T_{\mathsf{K}}P_{\mathbf{z}^\nu}) + \psi_{mon}(P_{\mathbf{z}^\nu}T_{\mathsf{K}}P_{\mathbf{z}^\nu})$$
$$-\psi_{mon}\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right) = P_{\mathbf{z}^\nu}\psi_{mon}(T_{\mathsf{K}})(I - P_{\mathbf{z}^\nu}) + (I - P_{\mathbf{z}^\nu})\psi_{mon}(T_{\mathsf{K}}) + P_{\mathbf{z}^\nu}\psi_{mon}(T_{\mathsf{K}})P_{\mathbf{z}^\nu}$$
$$-\psi_{mon}(P_{\mathbf{z}^\nu}T_{\mathsf{K}}P_{\mathbf{z}^\nu}) + \psi_{mon}(P_{\mathbf{z}^\nu}T_{\mathsf{K}}P_{\mathbf{z}^\nu}) - \psi_{mon}\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right).$$

Then,

$$\|\psi_{mon}(T_{\mathsf{K}}) - \psi_{mon}\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right)\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \le 2\|(I - P_{\mathbf{z}^\nu})\psi_{mon}(T_{\mathsf{K}})\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} + \|P_{\mathbf{z}^\nu}\psi_{mon}(T_{\mathsf{K}})P_{\mathbf{z}^\nu}$$
$$-\psi_{mon}(P_{\mathbf{z}^\nu}T_{\mathsf{K}}P_{\mathbf{z}^\nu})\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} + \|\psi_{mon}(P_{\mathbf{z}^\nu}T_{\mathsf{K}}P_{\mathbf{z}^\nu}) - \psi_{mon}\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right)\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}}. \tag{47}$$

Furthermore, it is clear that

$$\begin{aligned}\|P_{\mathbf{z}^\nu}T_{\mathsf{K}}P_{\mathbf{z}^\nu} - P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} &= \|P_{\mathbf{z}^\nu}\left(T_{\mathsf{K}}P_{\mathbf{z}^\nu} - S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right)\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \\ &\le \|\left(T_{\mathsf{K}} - S_{\mathbf{z}}^* S_{\mathbf{z}}\right)P_{\mathbf{z}^\nu}\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \le \|T_{\mathsf{K}} - S_{\mathbf{z}}^* S_{\mathbf{z}}\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}}.\end{aligned} \tag{48}$$

Now, from (47), (48), (41), (14), (15), we have

$$\begin{aligned}&\|\psi_{mon}(T_{\mathsf{K}}) - \psi_{mon}\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right)\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \\ &\qquad \le C\left(\psi_{mon}(\|T_{\mathsf{K}}^{\frac{1}{2}}(I - P_{\mathbf{z}^\nu})\|^2_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}}) + \psi_{mon}(\|T_{\mathsf{K}} - S_{\mathbf{z}}^* S_{\mathbf{z}}\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}})\right).\end{aligned}$$

Keeping in mind that $\psi_{mon}(t)$ is covered by the qualification $p = 1$, for any $C > 1$ we have

$$\frac{t}{\psi_{mon}(t)} \le \frac{Ct}{\psi_{mon}(Ct)} \qquad \Rightarrow \qquad \psi_{mon}(Ct) \le C\psi_{mon}(t).$$

In view of this and (10) we can conclude that with confidence $1 - \delta$ it holds

$$\|\psi_{mon}(T_{\mathsf{K}}) - \psi_{mon}\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right)\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \le C \log\frac{1}{\delta}\left(\psi_{mon}(\Delta^2_{\mathbf{z}^\nu}) + \psi_{mon}\Big(|\mathbf{z}|^{-1/2}\Big)\right). \tag{49}$$

Moreover, directly from (10) with confidence $1 - \delta$ one has

$$\|T_{\mathsf{K}} - P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \le \|P_{\mathbf{z}^\nu}T_{\mathsf{K}}(I - P_{\mathbf{z}^\nu})\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} + \|(I - P_{\mathbf{z}^\nu})T_{\mathsf{K}}\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}}$$
$$+\|T_{\mathsf{K}} - S_{\mathbf{z}}^* S_{\mathbf{z}}\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \le C \log\frac{1}{\delta}\left(\Delta_{\mathbf{z}^\nu} + |\mathbf{z}|^{-1/2}\right). \tag{50}$$

A substitution of the above bounds into (46) gives us

$$\|\sigma_{12}\|_{\mathcal{H}_{\mathsf{K}}} \le C \log\frac{1}{\delta}\Big(\psi_{Lip}(\alpha)\left[\psi_{mon}(\Delta^2_{\mathbf{z}^\nu}) + \psi_{mon}\Big(|\mathbf{z}|^{-1/2}\Big)\right] + \Delta_{\mathbf{z}^\nu} + |\mathbf{z}|^{-1/2}\Big). \tag{51}$$

Summing up (45) and (51), we get with confidence $1 - \delta$

$$\|\sigma_1\|_{\mathcal{H}_{\mathsf{K}}} \le C \log\frac{1}{\delta}\left(\varphi(\alpha) + \psi_{Lip}(\alpha)\left[\psi_{mon}(\Delta^2_{\mathbf{z}^\nu}) + \psi_{mon}\Big(|\mathbf{z}|^{-1/2}\Big)\right] + \Delta_{\mathbf{z}^\nu} + |\mathbf{z}|^{-1/2}\right).$$

To proceed further we need to estimate the norm of

$$\sigma_2 = g_\alpha(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}) S_{\mathbf{z}}^* S_{\mathbf{z}}(P_{\mathbf{z}^\nu} - I)\overline{f}_\rho.$$

With the use of polar decompositions we obtain the following inequality

$$\|\sigma_2\|_{\mathcal{H}_\mathsf{K}} \le \|g_\alpha(P_{\mathbf{z}^\nu} S^*_{\mathbf{z}} S_{\mathbf{z}} P_{\mathbf{z}^\nu})(P_{\mathbf{z}^\nu} S^*_{\mathbf{z}} S_{\mathbf{z}} P_{\mathbf{z}^\nu})^{1/2}\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \|(S^*_{\mathbf{z}} S_{\mathbf{z}})^{1/2}(I - P_{\mathbf{z}^\nu})\overline{f}_\rho\|_{\mathcal{H}_\mathsf{K}}.$$

Then with the use of (6) we can continue

$$\|\sigma_2\|_{\mathcal{H}_\mathsf{K}} \le \frac{C}{\sqrt{\alpha}} \left(\|T_\mathsf{K}^{1/2} - (S^*_{\mathbf{z}} S_{\mathbf{z}})^{1/2}\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} + \|T_\mathsf{K}^{1/2}(I - P_{\mathbf{z}^\nu})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \right) \|(I - P_{\mathbf{z}^\nu})\overline{f}_\rho\|_{\mathcal{H}_\mathsf{K}}.$$

Now (10), (42) and the fact that the function $\sqrt{t}$ is operator monotone, give us with confidence $1-\delta$

$$\begin{aligned}\|\sigma_2\|_{\mathcal{H}_\mathsf{K}} &\le \frac{C}{\sqrt{\alpha}} \left(\|T_\mathsf{K} - S^*_{\mathbf{z}} S_{\mathbf{z}}\|^{1/2}_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} + \|T_\mathsf{K}^{1/2}(I - P_{\mathbf{z}^\nu})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \right) \|(I - P_{\mathbf{z}^\nu})\varphi(T_\mathsf{K})v_\rho\|_{\mathcal{H}_\mathsf{K}} \\ &\le \frac{C}{\sqrt{\alpha}} \log\frac{1}{\delta} \left(|\mathbf{z}|^{-1/4}\Delta_{\mathbf{z}^\nu} + \Delta^2_{\mathbf{z}^\nu} \right).\end{aligned}$$

It remains to estimate the norm of σ_3. First, we rewrite σ_3 as follows

$$\begin{aligned}\sigma_3 &= g_\alpha(P_{\mathbf{z}^\nu} S^*_{\mathbf{z}} S_{\mathbf{z}} P_{\mathbf{z}^\nu}) P_{\mathbf{z}^\nu} \left[S^*_{\mathbf{z}} S_{\mathbf{z}} \overline{f}_\rho - S^*_{\mathbf{z}} \mathbf{DY} \right] \\ &= g_\alpha(P_{\mathbf{z}^\nu} S^*_{\mathbf{z}} S_{\mathbf{z}} P_{\mathbf{z}^\nu}) P_{\mathbf{z}^\nu} \left[(S^*_{\mathbf{z}} S_{\mathbf{z}} - T_\mathsf{K})\overline{f}_\rho + (T_\mathsf{K}\overline{f}_\rho - S^*_{\mathbf{z}} \mathbf{DY}) \right].\end{aligned}$$

In view of the relations $T_\mathsf{K}\overline{f}_\rho = L_\mathsf{K} f_\rho$, (6) and (10), with confidence $1-\delta$ we have

$$\begin{aligned}\|\sigma_3\|_{\mathcal{H}_\mathsf{K}} &\le \frac{C}{\alpha} \left(\|(S^*_{\mathbf{z}} S_{\mathbf{z}} - T_\mathsf{K})\overline{f}_\rho\|_{\mathcal{H}_\mathsf{K}} + \|S^*_{\mathbf{z}} \mathbf{DY} - L_\mathsf{K} f_\rho\|_{\mathcal{H}_\mathsf{K}} \right) \\ &\le \frac{C}{\alpha} \log\frac{1}{\delta} |\mathbf{z}|^{-1/2}.\end{aligned}$$

Summing up all estimates derived above, we finally obtain

$$\begin{aligned}\|\overline{f}_\rho - \overline{f}^\alpha_{\mathbf{z},\mathbf{z}^\nu}\|_{\mathcal{H}_\mathsf{K}} \le C \log\frac{1}{\delta} \Bigg\{ &\varphi(\alpha) + \psi_{Lip}(\alpha) \left[\psi_{mon}(\Delta^2_{\mathbf{z}^\nu}) + \psi_{mon}\left(|\mathbf{z}|^{-1/2} \right) \right] \\ &+ \Delta_{\mathbf{z}^\nu} + \alpha^{-1/2} \left(|\mathbf{z}|^{-1/4}\Delta_{\mathbf{z}^\nu} + \Delta^2_{\mathbf{z}^\nu} \right) + \alpha^{-1} |\mathbf{z}|^{-1/2} \Bigg\}.\end{aligned}$$

The lemma is proved. □

Now, we are ready to prove the main result of the section.

Proof. It is easy to see (e.g. [16]) that for any f

$$\mathcal{E}^{\text{rank}}(f) - \mathcal{E}^{\text{rank}}(f_\rho) = \int_{X\times X} \left(f(x) - f(x') - (f_\rho(x) - f_\rho(x')) \right)^2 \, d\rho_X(x) \, d\rho_X(x').$$

Recall that under the assumptions of the theorem $\overline{f}_\rho \in \mathcal{H}_\mathsf{K}$ and, moreover, $f = \overline{f}^{\alpha}_{\mathbf{z},\mathbf{z}^\nu} \in \mathcal{H}_\mathsf{K}$. Then due to the relations $\int_X \overline{f}_\rho(x)\ d\rho_X(x) = 0$ and $\|f\|_{L_2(X,\rho_X)} = \|T_\mathsf{K}^{\frac{1}{2}} f\|_{\mathcal{H}_\mathsf{K}}$ we obtain

$$\begin{aligned}0 \le \mathcal{E}^{\mathrm{rank}}(f) - \mathcal{E}^{\mathrm{rank}}(\overline{f}_\rho) &= \int_{X\times X} \Big(f^2(x) - 2f(x)\overline{f}_\rho(x) + \overline{f}^2_\rho(x) + f^2(x') \\ &-2f(x')\overline{f}_\rho(x') + \overline{f}^2_\rho(x') - 2f(x)f(x') \Big)\ d\rho_X(x)\ d\rho_X(x') \\ &= 2\|f - \overline{f}_\rho\|^2_{L_2(X,\rho_X)} - 2\left(\int_X f(x)\ d\rho_X(x)\right)^2 \le 2\|T_\mathsf{K}^{\frac{1}{2}}(f - \overline{f}_\rho)\|^2_{\mathcal{H}_\mathsf{K}}.\end{aligned}$$

Hence

$$\mathcal{E}^{\mathrm{rank}}(\overline{f}^{\alpha}_{\mathbf{z},\mathbf{z}^\nu}) - \mathcal{E}^{\mathrm{rank}}(\overline{f}_\rho) \le 2\|T_\mathsf{K}^{\frac{1}{2}}(\overline{f}^{\alpha}_{\mathbf{z},\mathbf{z}^\nu} - \overline{f}_\rho)\|^2_{\mathcal{H}_\mathsf{K}}. \tag{52}$$

Further, by (44) we have

$$\|T_\mathsf{K}^{\frac{1}{2}}(\overline{f}^{\alpha}_{\mathbf{z},\mathbf{z}^\nu} - \overline{f}_\rho)\|_{\mathcal{H}_\mathsf{K}} = \|T_\mathsf{K}^{\frac{1}{2}}(\sigma_1+\sigma_2+\sigma_3)\|_{\mathcal{H}_\mathsf{K}} \le \|T_\mathsf{K}^{\frac{1}{2}}\sigma_1\|_{\mathcal{H}_\mathsf{K}} + \|T_\mathsf{K}^{\frac{1}{2}}\sigma_2\|_{\mathcal{H}_\mathsf{K}} + \|T_\mathsf{K}^{\frac{1}{2}}\sigma_3\|_{\mathcal{H}_\mathsf{K}}, \tag{53}$$

where, σ_1, σ_2, σ_3 are the same as in Lemma 1, and we need to estimate each term in (53). First, we start with the decomposition

$$\begin{aligned}T_\mathsf{K}^{\frac{1}{2}} &= T_\mathsf{K}^{\frac{1}{2}}(I - P_{\mathbf{z}^\nu}) + (I - P_{\mathbf{z}^\nu})T_\mathsf{K}^{\frac{1}{2}}P_{\mathbf{z}^\nu} + \Big((P_{\mathbf{z}^\nu}T_\mathsf{K}P_{\mathbf{z}^\nu})^{\frac{1}{2}} - (P_{\mathbf{z}^\nu}S^*_\mathbf{z}S_\mathbf{z}P_{\mathbf{z}^\nu})^{\frac{1}{2}}\Big) \\ &+ \Big(P_{\mathbf{z}^\nu}T_\mathsf{K}^{\frac{1}{2}}P_{\mathbf{z}^\nu} - (P_{\mathbf{z}^\nu}T_\mathsf{K}P_{\mathbf{z}^\nu})^{\frac{1}{2}}\Big) + (P_{\mathbf{z}^\nu}S^*_\mathbf{z}S_\mathbf{z}P_{\mathbf{z}^\nu})^{\frac{1}{2}}.\end{aligned}$$

Then

$$\begin{aligned}\|T_\mathsf{K}^{\frac{1}{2}}\sigma_{11}\|_{\mathcal{H}_\mathsf{K}} \le \Big[&2\|T_\mathsf{K}^{\frac{1}{2}}(I - P_{\mathbf{z}^\nu})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} + \|(P_{\mathbf{z}^\nu}T_\mathsf{K}P_{\mathbf{z}^\nu})^{\frac{1}{2}} - (P_{\mathbf{z}^\nu}S^*_\mathbf{z}S_\mathbf{z}P_{\mathbf{z}^\nu})^{\frac{1}{2}}\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \\ &+ \|P_{\mathbf{z}^\nu}T_\mathsf{K}^{\frac{1}{2}}P_{\mathbf{z}^\nu} - (P_{\mathbf{z}^\nu}T_\mathsf{K}P_{\mathbf{z}^\nu})^{\frac{1}{2}}\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}\Big]\|\sigma_{11}\|_{\mathcal{H}_\mathsf{K}} + \|(P_{\mathbf{z}^\nu}S^*_\mathbf{z}S_\mathbf{z}P_{\mathbf{z}^\nu})^{\frac{1}{2}}\sigma_{11}\|_{\mathcal{H}_\mathsf{K}}.\end{aligned}$$

By means of (41) we have

$$\|P_{\mathbf{z}^\nu}T_\mathsf{K}^{\frac{1}{2}}P_{\mathbf{z}^\nu} - (P_{\mathbf{z}^\nu}T_\mathsf{K}P_{\mathbf{z}^\nu})^{\frac{1}{2}}\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \le 2\|T_\mathsf{K}^{\frac{1}{2}}(I - P_{\mathbf{z}^\nu})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}.$$

Since the function $\sqrt{t}$ is operator monotone it holds

$$\|(P_{\mathbf{z}^\nu}T_\mathsf{K}P_{\mathbf{z}^\nu})^{\frac{1}{2}} - (P_{\mathbf{z}^\nu}S^*_\mathbf{z}S_\mathbf{z}P_{\mathbf{z}^\nu})^{\frac{1}{2}}\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \le \|P_{\mathbf{z}^\nu}(T_\mathsf{K} - S^*_\mathbf{z}S_\mathbf{z})P_{\mathbf{z}^\nu}\|^{\frac{1}{2}}_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \le \|T_\mathsf{K} - S^*_\mathbf{z}S_\mathbf{z}\|^{\frac{1}{2}}_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}.$$

Thus,

$$\begin{aligned}\|T_\mathsf{K}^{\frac{1}{2}}\sigma_{11}\|_{\mathcal{H}_\mathsf{K}} \le &\left(4\|T_\mathsf{K}^{\frac{1}{2}}(I - P_{\mathbf{z}^\nu})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} + \|T_\mathsf{K} - S^*_\mathbf{z}S_\mathbf{z}\|^{\frac{1}{2}}_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}\right)\|\sigma_{11}\|_{\mathcal{H}_\mathsf{K}} \\ &+\|(P_{\mathbf{z}^\nu}S^*_\mathbf{z}S_\mathbf{z}P_{\mathbf{z}^\nu})^{\frac{1}{2}}\sigma_{11}\|_{\mathcal{H}_\mathsf{K}}.\end{aligned} \tag{54}$$

We have to estimate each of terms in the right-hand side of (54). Due to (45) we obtain

$$\|T_\mathsf{K}^{\frac{1}{2}}(I - P_{\mathbf{z}^\nu})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}\|\sigma_{11}\|_{\mathcal{H}_\mathsf{K}} \le C\Delta_{\mathbf{z}^\nu}\varphi(\alpha).$$

Further, we estimate the second term in the right-hand side of (54). Applying (10) and (45), we have with confidence $1-\delta$

$$\|T_{\mathsf{K}} - S_{\mathbf{z}}^* S_{\mathbf{z}}\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}^{\frac{1}{2}} \|\sigma_{11}\|_{\mathcal{H}_{\mathsf{K}}} \le C \log\frac{1}{\delta} |\mathbf{z}|^{-1/4} \varphi(\alpha).$$

By definition of σ_{11} we obtain

$$\|(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu})^{\frac{1}{2}} \sigma_{11}\|_{\mathcal{H}_{\mathsf{K}}} = \|(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu})^{\frac{1}{2}} r_\alpha\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right) \varphi\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right) v_\rho\|_{\mathcal{H}_{\mathsf{K}}},$$

and then Proposition 1 gives us

$$\|(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu})^{\frac{1}{2}} \sigma_{11}\|_{\mathcal{H}_{\mathsf{K}}} \le C\sqrt{\alpha}\varphi(\alpha).$$

Summing up the estimates obtained above, we have with confidence $1-\delta$ that

$$\|T_{\mathsf{K}}^{\frac{1}{2}} \sigma_{11}\|_{\mathcal{H}_{\mathsf{K}}} \le C \log\frac{1}{\delta} \left(\Delta_{\mathbf{z}^\nu} + |\mathbf{z}|^{-1/4} + \sqrt{\alpha}\right) \varphi(\alpha). \tag{55}$$

Now, we are going to estimate $\|T_{\mathsf{K}}^{\frac{1}{2}} \sigma_{12}\|_{\mathcal{H}_{\mathsf{K}}}$. Similarly to the previous case we begin with the decomposition

$$\begin{aligned}\|T_{\mathsf{K}}^{\frac{1}{2}} \sigma_{12}\|_{\mathcal{H}_{\mathsf{K}}} &\le 4\|T_{\mathsf{K}}^{\frac{1}{2}}(I - P_{\mathbf{z}^\nu})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} \|\sigma_{12}\|_{\mathcal{H}_{\mathsf{K}}} + \|T_{\mathsf{K}} - S_{\mathbf{z}}^* S_{\mathbf{z}}\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}^{\frac{1}{2}} \|\sigma_{12}\|_{\mathcal{H}_{\mathsf{K}}} \\ &+ \|(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu})^{\frac{1}{2}} \sigma_{12}\|_{\mathcal{H}_{\mathsf{K}}}. \end{aligned} \tag{56}$$

Directly from (51) it follows that

$$\begin{aligned}&\|T_{\mathsf{K}}^{\frac{1}{2}}(I - P_{\mathbf{z}^\nu})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} \|\sigma_{12}\|_{\mathcal{H}_{\mathsf{K}}} \\ &\le C \log\frac{1}{\delta} \Delta_{\mathbf{z}^\nu} \left(\psi_{Lip}(\alpha)\left[\psi_{mon}(\Delta_{\mathbf{z}^\nu}^2) + \psi_{mon}\left(|\mathbf{z}|^{-1/2}\right)\right] + \Delta_{\mathbf{z}^\nu} + |\mathbf{z}|^{-1/2}\right).\end{aligned}$$

Then, (10) gives us

$$\begin{aligned}&\|T_{\mathsf{K}} - S_{\mathbf{z}}^* S_{\mathbf{z}}\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}^{\frac{1}{2}} \|\sigma_{12}\|_{\mathcal{H}_{\mathsf{K}}} \\ &\quad\le C \log\frac{1}{\delta} |\mathbf{z}|^{-1/4} \left(\psi_{Lip}(\alpha)\left[\psi_{mon}(\Delta_{\mathbf{z}^\nu}^2) + \psi_{mon}\left(|\mathbf{z}|^{-1/2}\right)\right] + \Delta_{\mathbf{z}^\nu} + |\mathbf{z}|^{-1/2}\right).\end{aligned}$$

It remains to estimate the last term in (56). First, we write

$$\begin{aligned}&\|(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu})^{\frac{1}{2}} \sigma_{12}\|_{\mathcal{H}_{\mathsf{K}}} = \|(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu})^{\frac{1}{2}} r_\alpha\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right) \left[\varphi(T_{\mathsf{K}}) - \varphi\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right)\right] v_\rho\|_{\mathcal{H}_{\mathsf{K}}} \\ &\le \|(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu})^{\frac{1}{2}} r_\alpha\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right) \psi_{Lip}\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right) \|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} \times \\ &\| \left[\psi_{mon}(T_{\mathsf{K}}) - \psi_{mon}\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right)\right] v_\rho\|_{\mathcal{H}_{\mathsf{K}}} + \|(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu})^{\frac{1}{2}} r_\alpha\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right) \|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} \times \\ &\| \left[\psi_{Lip}(T_{\mathsf{K}}) - \psi_{Lip}\left(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu}\right)\right] \psi_{mon}(T_{\mathsf{K}}) v_\rho\|_{\mathcal{H}_{\mathsf{K}}}.\end{aligned}$$

Keeping in mind that the function $\psi_{Lip}(t)\sqrt{t}$ is covered by the qualification of g_α, in view of Proposition 1, (49) and (50) we have with confidence $1-\delta$

$$\begin{aligned}\|(P_{\mathbf{z}^\nu} S_{\mathbf{z}}^* S_{\mathbf{z}} P_{\mathbf{z}^\nu})^{\frac{1}{2}} \sigma_{12}\|_{\mathcal{H}_{\mathsf{K}}} \le C \log\frac{1}{\delta} \Bigg(&\sqrt{\alpha}\psi_{Lip}(\alpha)\left(\psi_{mon}(\Delta_{\mathbf{z}^\nu}^2) + \psi_{mon}\left(|\mathbf{z}|^{-1/2}\right)\right) \\ &+ \sqrt{\alpha}\left(\Delta_{\mathbf{z}^\nu} + |\mathbf{z}|^{-1/2}\right)\Bigg).\end{aligned}$$

Thus, summing up the above estimates, we get

$$\|T_{\mathsf{K}}^{\frac{1}{2}}\sigma_{12}\|_{\mathcal{H}_{\mathsf{K}}} \le C\log\frac{1}{\delta}\Bigg\{ \left(\Delta_{\mathbf{z}^\nu} + |\mathbf{z}|^{-1/4} + \sqrt{\alpha}\right)\psi_{Lip}(\alpha)\left[\psi_{mon}(\Delta^2_{\mathbf{z}^\nu}) + \psi_{mon}\Big(|\mathbf{z}|^{-1/2}\Big)\right] + \left(\Delta_{\mathbf{z}^\nu} + |\mathbf{z}|^{-1/4} + \sqrt{\alpha}\right)\left(\Delta_{\mathbf{z}^\nu} + |\mathbf{z}|^{-1/2}\right)\Bigg\}. \tag{57}$$

Combining (55) and (57), we have

$$\begin{aligned}\|T_{\mathsf{K}}^{\frac{1}{2}}\sigma_{1}\|_{\mathcal{H}_{\mathsf{K}}} \le C\log\frac{1}{\delta}\Bigg\{ & \left(\Delta_{\mathbf{z}^\nu} + |\mathbf{z}|^{-1/4} + \sqrt{\alpha}\right)\varphi(\alpha) \\ & + \left(\Delta_{\mathbf{z}^\nu} + |\mathbf{z}|^{-1/4} + \sqrt{\alpha}\right)\psi_{Lip}(\alpha)\left[\psi_{mon}(\Delta^2_{\mathbf{z}^\nu}) + \psi_{mon}\Big(|\mathbf{z}|^{-1/2}\Big)\right] \\ & + \left(\Delta_{\mathbf{z}^\nu} + |\mathbf{z}|^{-1/4} + \sqrt{\alpha}\right)\left(\Delta_{\mathbf{z}^\nu} + |\mathbf{z}|^{-1/2}\right)\Bigg\}.\end{aligned} \tag{58}$$

Recall, that according (43) it holds

$$\Delta_{\mathbf{z}^\nu} \le \varphi(\Theta^{-1}(|\mathbf{z}|^{-1/2})), \qquad \alpha = \Theta^{-1}(|\mathbf{z}|^{-\frac{1}{2}}),$$

with $\Theta(t) = \varphi(t)t$. Then for the chosen value of α

$$\frac{1}{\sqrt{|\mathbf{z}|}} = \alpha\varphi(\alpha),$$

and consequently, $|\mathbf{z}|^{-1/2} \le \alpha$ and $|\mathbf{z}|^{-1/2} \le \varphi(\alpha) \le \alpha$. Hence, in view of (58) it holds with confidence $1-\delta$

$$\|T_{\mathsf{K}}^{\frac{1}{2}}\sigma_{1}\|_{\mathcal{H}_{\mathsf{K}}} \le C\log\frac{1}{\delta}\sqrt{\alpha}\varphi(\alpha) = O(\sqrt{\Theta^{-1}(|\mathbf{z}|^{-1/2})}\varphi(\Theta^{-1}(|\mathbf{z}|^{-1/2}))\log\frac{1}{\delta}).$$

Using the same argument as for $\|T_{\mathsf{K}}^{\frac{1}{2}}\sigma_{1}\|_{\mathcal{H}_{\mathsf{K}}}$ it can be easily shown that for the chosen value of α all terms in (44) have the same order $O(\sqrt{\Theta^{-1}(|\mathbf{z}|^{-1/2})}\varphi(\Theta^{-1}(|\mathbf{z}|^{-1/2}))\log\frac{1}{\delta})$. Thus, finally we have

$$\|T_{\mathsf{K}}^{\frac{1}{2}}(\overline{f}^{\alpha}_{\mathbf{z},\mathbf{z}^\nu} - \overline{f}_{\rho})\|_{\mathcal{H}_{\mathsf{K}}} = O(\sqrt{\Theta^{-1}(|\mathbf{z}|^{-1/2})}\varphi(\Theta^{-1}(|\mathbf{z}|^{-1/2}))\log\frac{1}{\delta}),$$

which, by virtue of (52), proves the theorem. □

Remark 2. *Assume now that $f_\rho \in \mathrm{Range}(\varphi(T_{\mathsf{K}}))$ with $\varphi \in \mathcal{F}_{Lip}$. Then the proof of Theorem 3 can obviously be modified in such a way that for a regularization method $g_\alpha(t)$ with the qualification p covering $\varphi(t)\sqrt{t}$ and α, $\Delta_{\mathbf{z}^\nu}$ meeting (43) with confidence $1-\delta$ it holds*

$$\mathcal{E}^{\mathrm{regr}}(f^{\alpha}_{\mathbf{z},\mathbf{z}^\nu}) - \mathcal{E}^{\mathrm{regr}}(f_\rho) = O\left(\varphi^2(\Theta^{-1}(|\mathbf{z}|^{-\frac{1}{2}}))\Theta^{-1}(|\mathbf{z}|^{-\frac{1}{2}})\log\frac{1}{\delta}\right).$$

Remark 3. *Observe that if $\varphi \in \mathcal{F}_{Lip}$ and a qualification p covers the function $\varphi(t)\sqrt{t}$, then $\varphi(t)$ is covered by the qualification $p-1/2$, and*

$$|\mathbf{z}|^{-\frac{2p-1}{2(2p+1)}} \le C\varphi(\Theta^{-1}(|\mathbf{z}|^{-\frac{1}{2}})).$$

Therefore, if

$$\Delta_{\mathbf{z}^\nu} \le C|\mathbf{z}|^{-\frac{2p-1}{2(2p+1)}}, \tag{59}$$

then the second condition (43) is satisfied for any φ covered by the qualification $p-1/2$.

On the other hand, as it has been shown in the previous section, under the assumption [11] that

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

with confidence $1-\delta$ it holds

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

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

In view of (59) this means that for $|\mathbf{z}^\nu| = O\left(|\mathbf{z}|^{\frac{2p-1}{2\beta(2p+1)}}\right)$ the conditions (43) of Theorem 3 are satisfied, and the corresponding best distribution independent learning rate is achieved within the Nyström approach. Observe that for $\beta > \frac{2p-1}{2p+1}$ the computational complexity of the considered Nyström approximant $\overline{f}^{\alpha}_{\mathbf{z},\mathbf{z}^\nu}$ is subquadratic. For example, the combination of Plain Nyström subsampling with two-times iterated Tikhonov regularization (8), $l = p = 2$, allows us to achieve the corresponding best distribution independent learning rate with subquadratic computational complexity under a rather mild assumption that $\mathsf{K}_x \in \mathrm{Range}(T_{\mathsf{K}}^{\frac{1}{12}+\xi})$, where ξ is an arbitrary small positive number.

Acknowledgments

Sergiy Pereverzyev Jr. gratefully acknowledges the support of the Austrian Science Fund (FWF): project P 29514-N32. This work was done while the first and third authors were visiting Johann Radon Institute within the EU-Horizon 2020 MSC-RISE project AMMODIT.

References

[1] Agarwal S. and Niyogi P. 2009 Generalization bounds for ranking algorithms via algorithmic stability. *J. Mach. Learn. Res.* **10** 441–474

[2] Bach F. 2013 Sharp analysis of low-rank kernel matrix approximations. *Proceedings of the 26th Annual Conference on Learning Theory* **30** 185–209

[3] Bakushinski A.B. 1967 A general method of constructing regularizing algorithms for a linear ill-posed equation in Hilbert space. *USSR Comput. Math. Math. Phys.* **7** 279–287

[4] Bauer F., Pereverzev S. and Rosasco L. 2007 On regularization algorithms in learning theory. *J. of Complexity* **23** 52–72

[5] Caponnetto A., De Vito E. 2007 Optimal rates for the regularized least-squares algorithm. *Found. Comp. Math.* **7** 331–368

[6] Caponnetto A. and Yao Y. 2010 Cross-validation based adaptation for regularization operators in learning theory. *Anal. Appl.* **8** 161–183

[7] Chen H. 2012 The convergence rate of a regularized ranking algorithm. *J. Approx. Theory* **164** 1513–1519

[8] Chen J., Pereverzyev-Jr. S. and Xu Y. 2015 Aggregation of regularized solutions from multiple observation models. *Inverse Probl.* **31** 075005

[9] Cohen W., Schapire R. and Singer Y. 1999 Learning to order things. *J. Artif. Intell. Res.* **10** 243–270

[10] Cortes C., Mohri M. and Rastogi A. 2007 Magnitude preserving ranking algorithms. *Proc. of the 24th International Conference of Machine Learning* pp. 169–176

[11] De Vito E., Rosasco L. and Toigo A. 2014 Learning sets with separating kernels. *Appl. Comput. Harmon. Anal.* **37** 185–217

[12] Drineas P., Magdon-Ismail M., Mahoney M.W. and Woodruf D.P. 2012 Fast approximation of matrix coherence and statistical leverage. *J. Mach. Learn. Res.* **13** 3475–3506

[13] Guo Z.-Ch., Lin Sh.-B. and Zhou D.-X. 2017 Learning theory of distributed spectral algorithms. *Inverse Probl.* **33** 074009

[14] Hashem S. and Schmeiser B. 1993 Approximating a Function and its Derivatives Using MSE-Optimal Linear Combinations of Trained Feedforward Neural Networks. *Proceedings of the 1993 World Congress on Neural Networks, vol. 1, Lawrence Erlbaum Associates, Hillsdale, New Jersey* pp. 617–620

[15] Hu T., Fan J., Wu Q. and Zhou D.-X. 2013 Learning Theory Approach to Minimum Error Entropy Criterion. *J. Mach. Learn. Res.* **14** 377–397

[16] Kriukova G., Panasiuk O., Pereverzyev S.V. and Tkachenko P. 2016 A linear functional strategy for regularized ranking. *Neural Networks* **73** 26–35

[17] Kriukova G., Pereverzyev-Jr. S. and Tkachenko P. 2016 Nyström type subsampling analyzed as a regularized projection. *Inverse Probl.* **33** 074001

[18] Kriukova G., Tkachenko P. and Pereverzev S. 2015 On the convergence rate and some application of a regularized ranking algorithm. *J. Complexity* **33** 14–29

[19] Lu S., Mathe P. and Pereverzyev S. 2016 Balancing principle in supervised learning for a general regularization scheme. *RICAM-Report 2016-38. Preprint*

[20] Lu S. and Pereverzev S.V. 2013 *Regularization theory for ill-posed problems: selected topics.* Walter de Gruyter

[21] Mathe P. and Hofmann B. 2008 How general are general source conditions? *Inverse Probl.* **24** 015009

[22] Mathe P. and Pereverzev S.V. 2003 Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse Probl.* **19** 789–803

[23] Mathe P. and Pereverzev S.V. 2003 Discretization strategy for ill-posed problems in variable Hilbert scales. *Inverse Probl.* **19** 1263–1277

[24] Micchelli C. A. and Pontil M. 2005 Learning the kernel function via regularization. *J. Mach. Learn. Res.* **6** 1099–1125

[25] Micchelli C.A. , Xu Y. and Zhang H. 2006 Universal kernels. *J. Mach. Learn. Res.* **7** 2651–2667

[26] Mukherjee S. and Zhou D.-X. 2006 Learning coordinate covariances via gradients. *J. Mach. Learn. Res.* **7** 519–549

[27] Plato R. and Vainikko G.M. 1990 On the regularization of projection methods for solving ill-posed problems. *Numer. Math.* **57** 63–79

[28] Rudi A., Comoriano R. and Rosasco L. 2015 Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems* **28** ed Cortes C, Lawrence N, Lee

D, Sugiyama M, Garnett R and Garnett R (Curran Associates, Inc.) pp 1648–1656 also arXiv:1507.04717 [stat.ML]

[29] Smale S. and Zhou D.-X. 2007 Learning theory estimates via integral operators and their approimations. *Constr. Approx.* **26** 153–172

[30] Smola A.J. and Schölkopf B. 2000 Sparse greedy matrix approximation for machine learning. *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2* pp. 911–918

[31] Steinwart I. 2002 On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.* **2** 67–93

[32] Vainikko G. M. and Veretennikov A. Yu. 1986 *Iteration procedures in ill-posed problems.* Nauka, Moscow

[33] Williams C. and Seeger M. 2001 Using the Nyström method to speed up kernel machines. *Proceedings of the 14th Annual Conference on Neural Information Processing Systems* pp. 682–688

[34] Ying Y. and Zhou D.-X. 2016 Online pairwise learning algorithms. *Neural Comput.* **28** 743–777

[35] Zhao Y., Jun Fan J. and Shi L. 2017 Learning rates for regularized least squares ranking algorithm. *Anal. Appl.* to appear