Inverse Problems

PAPER • OPEN ACCESS

Nyström type subsampling analyzed as a regularized projection

To cite this article: Galyna Kriukova *et al* 2017 *Inverse Problems* **33** 074001

OPEN ACCESS
IOP Publishing

Inverse Problems **33** (2017) 074001 (19pp)

doi:10.1088/1361-6420/33/7/074001

Nyström type subsampling analyzed as a regularized projection

Galyna Kriukova[1,3], **Sergiy Pereverzyev Jr**[2] and **Pavlo Tkachenko**[1]

[1] Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Altenbergerstrae 69, A-4040 Linz, Austria
[2] Institute of Mathematics, University of Innsbruck, Technikestraße 13, A-6020 Innsbruck, Austria

E-mail: galyna.kriukova@ricam.oeaw.ac.at, sergiy.pereverzyev@uibk.ac.at and pavlo.tkachenko@ricam.oeaw.ac.at

Received 12 January 2016, revised 6 April 2016
Accepted for publication 4 May 2016
Published 21 June 2017

Abstract

In the statistical learning theory the Nyström type subsampling methods are considered as tools for dealing with big data. In this paper we consider Nyström subsampling as a special form of the projected Lavrentiev regularization, and study it using the approaches developed in the regularization theory. As a result, we prove that the same capacity independent learning rates that are guaranteed for standard algorithms running with quadratic computational complexity can be obtained with subquadratic complexity by the Nyström subsampling approach, provided that the subsampling size is chosen properly. We propose *a priori* rule for choosing the subsampling size and *a posteriori* strategy for dealing with uncertainty in the choice of it. The theoretical results are illustrated by numerical experiments.

Keywords: Nyström subsampling, big data, linear functional strategy, regularization, source condition, computational complexity

1. Introduction

Regularization based kernel methods, such as kernel ridge regression (KRR), provide an effective framework for the supervised learning [14, 15]. However, a standard implementation of these methods is infeasible when dealing with the so-called 'Big Data'.

[3] Author to whom any correspondence should be addressed.

The Big Data concept can be considered from different points of view. Here, by 'Big Data', we mean data sets exceeding the computational capacity of conventional learning systems. For example, in KRR, one receives a training data set $\mathbf{z}$ of N samples of the form $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^{N}$, where each input $x_i \in X \subset \mathbb{R}^d$ is related to the output $y_i \in \mathbb{R}$ by an unknown target function $f^*: X \to \mathbb{R}$, and the goal is to approximate this function by the minimizer $f_{\mathbf{z}}^{\alpha}$ of the regularized empirical risk functional:

$$T_{\mathbf{z}}^{\alpha}(f) := \frac{1}{|\mathbf{z}|} \sum_{(x_i, y_i) \in \mathbf{z}} (f(x_i) - y_i)^2 + \alpha \|f\|_{\mathcal{H}_{\mathsf{K}}}^2. \tag{1}$$

Here, $\mathcal{H}_{\mathsf{K}}$ denotes the reproducing kernel Hilbert space (RKHS) generated by a kernel K: $X \times X \to \mathbb{R}$, X is equipped with the same metric as $\mathbb{R}^d$, $|\mathbf{z}| = N$, and α is a regularization parameter.

By the representer theorem for RKHS [7], the minimizer of (1) is equal to

$$f_{\mathbf{z}}^{\alpha} = \sum_{x_i : (x_i, y_i) \in \mathbf{z}} c_i \mathsf{K}(\cdot, x_i),$$

where $\mathbf{c} = (c_i)_{i=1}^{|\mathbf{z}|} = (\mathbf{K} + \alpha|\mathbf{z}|\mathbf{I})^{-1}\mathbf{Y}$, $\mathbf{Y} = (y_i)_{i=1}^{|\mathbf{z}|}$, $\mathbf{I}$ is the $|\mathbf{z}| \times |\mathbf{z}|$ diagonal identity matrix, and $\mathbf{K}$ denotes the $|\mathbf{z}| \times |\mathbf{z}|$ kernel matrix with entries $\mathbf{K}_{ij} = \mathsf{K}(x_i, x_j)$.

Now, it is clear that KRR will suffer from at least quadratic computational complexity $\mathrm{O}(N^2)$ in the number of observations N, as this is the complexity of computing the kernel matrix $\mathbf{K}$. In the Big Data setting, where N is large, this is not acceptable. Therefore, learning schemes have been designed to avoid the computation of the exact minimizers $f_{\mathbf{z}}^{\alpha}$.

One family of such schemes, which we broadly refer to as the Nyström type subsampling, consists of methods replacing the kernel matrix $\mathbf{K}$ with a smaller matrix obtained by column subsampling [18, 19]. This can also be interpreted as a restriction of the minimization of $T_{\mathbf{z}}^{\alpha}(f)$ to the space

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

where $\mathbf{z}^{\nu} \subset \mathbf{z}$, and $|\mathbf{z}^{\nu}| = N_{\nu} \ll N$.

It can be shown [13] that the minimizer $f_{\mathbf{z},\mathbf{z}^{\nu}}^{\alpha}$ of $T_{\mathbf{z}}^{\alpha}(f)$ over the space $\mathcal{H}_{\mathsf{K}}^{\mathbf{z}^{\nu}}$ has the form

$$f_{\mathbf{z},\mathbf{z}^{\nu}}^{\alpha} = (\mathrm{P}_{\mathbf{z}^{\nu}} \mathrm{S}_{\mathbf{z}}^{*} \mathrm{S}_{\mathbf{z}} \mathrm{P}_{\mathbf{z}^{\nu}} + \alpha \mathrm{I})^{-1} \mathrm{P}_{\mathbf{z}^{\nu}} \mathrm{S}_{\mathbf{z}}^{*} \mathbf{Y}, \tag{2}$$

where $\mathrm{P}_{\mathbf{z}^{\nu}}$ is the orthogonal projection operator with range $\mathcal{H}_{\mathsf{K}}^{\mathbf{z}^{\nu}}$, $\mathrm{S}_{\mathbf{z}}$: $\mathcal{H}_{\mathsf{K}} \to \mathbb{R}^{|\mathbf{z}|}$ is the sampling operator, $\mathrm{S}_{\mathbf{z}} f = (f(x_1), f(x_2), \ldots, f(x_N))$, x_i: $(x_i, y_i) \in \mathbf{z}$, and $\mathrm{S}_{\mathbf{z}}^{*}$: $\mathbb{R}^{|\mathbf{z}|} \to \mathcal{H}_{\mathsf{K}}$ is the adjoint of $\mathrm{S}_{\mathbf{z}}$. If the norm $\|\cdot\|_{\mathbb{R}^{|\mathbf{z}|}}$ is defined as $|\mathbf{z}|^{-1}$ times the Euclidean norm, then

$$\mathrm{S}_{\mathbf{z}}^{*}\mathbf{u}\,(\cdot) = |\mathbf{z}|^{-1} \sum_{i=1}^{|\mathbf{z}|} u_i \mathsf{K}(\cdot, x_i), \qquad \mathbf{u} = (u_1, u_2, \ldots, u_N).$$

Observe that $f_{\mathbf{z},\mathbf{z}^{\nu}}^{\alpha}$ admits the representation

$$f_{\mathbf{z},\mathbf{z}^{\nu}}^{\alpha} = \sum_{x_i : (x_i, y_i) \in \mathbf{z}^{\nu}} c_i K(x_i, \cdot), \tag{3}$$

where the vector $\mathbf{c} = (c_i)_{i=1}^{|\mathbf{z}^{\nu}|}$ solves a linear system

$$(\mathbf{K}_{\mathbf{z},\mathbf{z}^{\nu}}^{\top} \mathbf{K}_{\mathbf{z},\mathbf{z}^{\nu}} + \alpha |\mathbf{z}| \mathbf{K}_{\mathbf{z}^{\nu},\mathbf{z}^{\nu}}) \mathbf{c} = \mathbf{K}_{\mathbf{z},\mathbf{z}^{\nu}}^{\top} \mathbf{Y} \tag{4}$$

with $|\mathbf{z}^{\nu}| \times |\mathbf{z}^{\nu}|$-matrix $\mathbf{K}_{\mathbf{z}^{\nu},\mathbf{z}^{\nu}} = (\mathbf{K}_{l,p} = \mathsf{K}(\tilde{x}_l, \tilde{x}_p))_{(\tilde{x}_l, \tilde{y}_l), (\tilde{x}_p, \tilde{y}_p) \in \mathbf{z}^{\nu}}$ and $|\mathbf{z}| \times |\mathbf{z}^{\nu}|$-matrix $\mathbf{K}_{\mathbf{z},\mathbf{z}^{\nu}} = (\mathbf{K}_{l,p} = \mathsf{K}(x_l, \tilde{x}_p))_{(x_l, y_l) \in \mathbf{z}, (\tilde{x}_p, \tilde{y}_p) \in \mathbf{z}^{\nu}}$. Now it is clear that the complexity of computing the minimizer $f_{\mathbf{z},\mathbf{z}^{\nu}}^{\alpha}$ of $T_{\mathbf{z}}^{\alpha}(f)$ over the space $\mathcal{H}_{\mathsf{K}}^{\mathbf{z}^{\nu}}$ is of order $\mathrm{O}(|\mathbf{z}| \cdot |\mathbf{z}^{\nu}|^2) = \mathrm{O}(N \cdot N_{\nu}^2)$.

Therefore, the main question about the Nyström type subsampling is the following: how big should N_ν be to incur no loss of the performance compared to the full kernel matrix $\mathbf{K}$; or, in other words, is it possible to implement the Nyström approach with a complexity that is subquadratic in the number of observations N without losing the performance?

Note that the answer on this question can be given in terms of the decay rate of the singular values of the kernel matrix $\mathbf{K}$. For example, if $\mathbf{K}$ is a low-rank matrix, then the number of Nystöm samples N_ν should be chosen related or even equal to the number of non-vanishing singular values of $\mathbf{K}$, which is assumed to be much smaller that N. Therefore, the idea of low-rank approximation has been widely used to obtain an approximation for kernel matrices $\mathbf{K}$.

Moreover, low-rank approximation is a building block in other kernel approximation strategies employed in the context of Nyström type subsampling, such as 'ensemble Nyström method' [9], 'pseudo landmark points' [6] and 'memory efficient kernel approximation' [16]. These strategies are proven to be efficient for the kernels maintaining special structures of matrices $\mathbf{K}$, such as shift-invariant kernels, for example. At the same time, the above-mentioned strategies are entirely based on the smoothness of the eigenspectrum of the approximated kernel matrices, but do not take into account the smoothness of the target function f^*, in spite of the fact that the latter one has an essential impact on the performance of KRR.

The studies of the Nyström approach accounting for the smoothness of the target functions have been recently given in [1, 13]. However, in [1], the error analysis is derived in a fixed design regression setting, such that x_i, $i = 1, 2, \ldots, |\mathbf{z}|$, are assumed to be uniformly sampled, for example. The study [13] extends the results of [1] to a general statistical learning setting. At the same time, the analysis of [13] is fairly technical and lengthy. In particular, it is based on the assumptions describing the capacity of the hypothesis space $\mathcal{H}_\mathrm{K}$ with respect to the unknown distribution ρ_X from which $\{x_i\}_{i=1}^{|\mathbf{z}|}$ is assumed to be sampled. Moreover, only the Hölder type of smoothness of the target functions is covered by the analysis of [13].

In the present study, we are going to analyze the so-called plain Nyström approach as a particular case of the regularized projection scheme. Therefore, we will use some arguments developed in the regularization theory for analyzing regularized projection approximations [11, 12]. Instead of the assumption on the capacity of the solution space, these arguments rely on the assumption on the approximation power of the projection method induced by the projector such as $\mathrm{P}_{\mathbf{z}^\nu}$ in (2). For the purpose of our study, the arguments developed in [11, 12] should be accompanied by the ones that take into account that in the context of learning, the regularized projection schemes, such as (2), operate only with noisy versions of the operators describing the learning tasks.

An analysis incorporating the above mentioned arguments is presented in the next section. Unlike [13], it gives capacity independent learning rates for the Nyström type subsampling. Moreover, it indicates a rather general *a priori* choice of the subsampling size $|\mathbf{z}^\nu|$ that allows a subquadratic complexity without loss of the performance. Such *a priori* choice of $|\mathbf{z}^\nu|$ requires a knowledge of the regularity of the unknown target function with respect to K and ρ_X and covers much more than just the Hölder type of smoothness. In section 3, we consider a situation when such *a priori* knowledge is not accurate, and may lead to uncertain parameter $|\mathbf{z}^\nu|$. In section 4, we discuss some simulations illustrating our theoretical results.

2. Approximation power, regularity and learning rate

A training data set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^{|\mathbf{z}|}$ is assumed to be sampled from the so-called sample space $Z = X \times Y$ endowed with a fixed but unknown probability distribution ρ, which can be factorized as $\rho(x, y) = \rho(y|x)\rho_X(x)$, where $\rho(\cdot|x)$ is the conditional distribution on $Y \subset \mathbb{R}$ given $x \in X$, and ρ_X is the so-called marginal distribution, from which the set of inputs $\{x_i\}_{i=1}^{|\mathbf{z}|}$ is supposed to be sampled.

A common assumption to simplify analysis is that $Y = [-D, D]$ for some $D > 0$. A weaker condition can be found in [3].

Given a training set $\mathbf{z} \subset Z$, the goal is to find an estimate $f = f_{\mathbf{z}}$ with a small expected risk

$$\mathcal{E}(f) = \int_{X\times Y} (f(x) - y)^2 \mathrm{d}\rho(x, y).$$

Once we choose $\mathcal{H}_{\mathsf{K}}$ as the so-called hypothesis space, the best possible risk value is clearly

$$\inf_{f\in\mathcal{H}_{\mathsf{K}}} \mathcal{E}(f).$$

As in [13], we assume that there exists $f^\dagger \in \mathcal{H}_{\mathsf{K}}$ such that

$$\mathcal{E}(f^\dagger) = \min_{f\in\mathcal{H}_{\mathsf{K}}} \mathcal{E}(f).$$

To formulate our further assumptions we need some operators, which are traditionally used in the context of regression learning. At first we consider the space $L_2(X, \rho_X)$ of square integrable functions with respect to ρ_X equipped with the usual norm $\|\cdot\|_\rho = \|\cdot\|_{L_2(X,\rho_X)}$. It is well-known [5] that for $f, f^\dagger \in \mathcal{H}_{\mathsf{K}}$ we have

$$\mathcal{E}(f) - \mathcal{E}(f^\dagger) = \|f - f^\dagger\|_\rho^2. \tag{5}$$

It is also known that if the kernel K is bounded then $\mathcal{H}_{\mathsf{K}}$ is continuously embedded in $L_2(X, \rho_X)$, such that the canonical embedding operator $\mathsf{J}_{\mathsf{K}} : \mathcal{H}_{\mathsf{K}} \to L_2(x, \rho_X)$ is continuous. Then we consider the adjoint operator $\mathsf{J}_{\mathsf{K}}^* : L_2(X, \rho_X) \to \mathcal{H}_{\mathsf{K}}$

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

and the covariance operator $\mathbf{C} = \mathsf{J}_{\mathsf{K}}^*\mathsf{J}_{\mathsf{K}} : \mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}$

$$\mathbf{C}f(\cdot) = \int_X \mathsf{K}(\cdot,x)\langle f, \mathsf{K}_x\rangle_{\mathsf{K}} \mathrm{d}\rho_X(x),$$

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

The operator $\mathbf{C}$ can be proved to be a positive trace class operator. Therefore, the operator $\mathbf{C}^{1/2} = \sqrt{\mathbf{C}}$ is well-defined and relates the norms of $f \in \mathcal{H}_{\mathsf{K}}$ in $\mathcal{H}_{\mathsf{K}}$ and $L_2(X, \rho_X)$ as follows

$$\|f\|_\rho = \|\mathbf{C}^{1/2} f\|_{\mathsf{K}}, \tag{6}$$

where $\|\cdot\|_{\mathsf{K}} = \|\cdot\|_{\mathcal{H}_{\mathsf{K}}}$.

We will measure the approximation power of the projection method induced by the projector $\mathrm{P}_{\mathbf{z}^\nu}$ in terms of the quantity $\|\mathbf{C}^{1/2}(\mathrm{I} - \mathrm{P}_{\mathbf{z}^\nu})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}$ that has been also studied in [13] (see lemma 6 [13]). At the same time, such kind of measure is usual in studying regularized projection methods [11, 12], and in spirit of that studies we assume that there is $\beta > 0$ such that the following holds with probability $1 - \delta$

$$\Delta_m \coloneqq \|\mathbf{C}^{1/2}(\mathrm{I} - \mathrm{P}_{\mathbf{z}^\nu})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} \leqslant d_{\delta,\beta} m^{-\beta},\ m = |\mathbf{z}^\nu|, \tag{7}$$

where $d_{\delta,\beta} = \mathrm{O}\left(\log^{\beta_1}\frac{1}{\delta}\right)$ and β_1 is a positive number depending only on β.

Note, that a probabilistic character of the assumption (7) is natural, because in the plain Nyström approach the points forming $\mathbf{z}^\nu$ are sampled uniformly at random without replacement from the training set $\mathbf{z}$.

As we have already mentioned, in [13], the Nyström subsampling approach was studied under assumptions on the capacity of $\mathcal{H}_\mathsf{K}$. These assumptions are formulated in [13] with the use of the quantity $\mathcal{N}_\infty(\lambda) = \sup\{\mathcal{N}_x(\lambda),\, x \in X\}$, where $\mathcal{N}_x(\lambda) = \langle \mathsf{K}_x, (\mathbf{C} + \lambda \mathrm{I})^{-1}\mathsf{K}_x\rangle_\mathsf{K}$. If in spirit of assumption 3 [13] we assume that $\mathcal{N}_\infty(\lambda) = \mathrm{O}(\lambda^{-\gamma})$, $0 < \gamma \leqslant 1$, then from lemma 6 [13] it follows that our assumption (7) is satisfied with any $\beta \in \left(0, \frac{1}{2\gamma}\right)$.

Our last assumption describes the regularity of $f^\dagger$ in terms of source condition concept that is fairly standard in the regularization theory [10]. In the context of the learning theory this concept has been introduced in [2]. Within this concept, we assume that $f^\dagger$ admits the representation

$$f^\dagger = \varphi(\mathbf{C})v^\dagger,\; v^\dagger \in \mathcal{H}_\mathsf{K},\; \|v^\dagger\|_\mathsf{K} \leqslant R, \tag{8}$$

where the function φ is operator monotone on $[0, d]$, $d > \|\mathbf{C}\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}$, and such that $\varphi(0) = 0$ and φ^2 is a concave function.

As it has been shown in [11] an important implication of operator monotonicity is that there is a number d_φ depending only on φ such that for any self-adjoint operators C, C_1 with spectra in $[0, d]$ it holds

$$\|\varphi(C) - \varphi(C_1)\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \leqslant d_\varphi \varphi(\|C - C_1\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}). \tag{9}$$

Moreover, as a consequence of the concavity of φ^2 we have (see proposition 2 [11])

$$\|(\mathrm{I} - \mathrm{P}_{\mathbf{z}^\nu})\varphi(\mathbf{C})\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \leqslant \varphi(\|\mathbf{C}^{1/2}(\mathrm{I} - \mathrm{P}_{\mathbf{z}^\nu})\|^2_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}}). \tag{10}$$

Note that our assumption (8) generalizes assumption 4 of [13], where only the case of operator monotone functions $\varphi(t) = t^s$, $0 < s \leqslant \frac{1}{2}$, has been studied.

In the sequel we extensively use the following bounds (see, e.g., [2]) that hold under the above assumptions with probability at least $1 - \delta$ and quantify the perturbation effect of random sampling:

$$\|\mathbf{C} - \mathrm{S}^*_\mathbf{z}\mathrm{S}_\mathbf{z}\|_{\mathcal{H}_\mathsf{K}\to\mathcal{H}_\mathsf{K}} \leqslant d_{1,\delta}|\mathbf{z}|^{-\frac{1}{2}}, \tag{11}$$

$$\|\mathrm{S}^*_\mathbf{z}\mathrm{S}_\mathbf{z} f^\dagger - \mathrm{S}^*_\mathbf{z}\mathbf{Y}\|_\mathsf{K} \leqslant d_{2,\delta}|\mathbf{z}|^{-\frac{1}{2}}, \tag{12}$$

where $d_{1,\delta}$ and $d_{2,\delta}$ are of order $\mathrm{O}\left(\log\frac{1}{\delta}\right)$ and depend only on K and ρ.

The following capacity independent learning rates have been proven in [2] for KRR (1)

Theorem 1 ([2]). *Consider a sampling space $Z = X \times [-D, D]$, where the input space $X \subset \mathbb{R}^d$ is closed. Consider also a bounded and continuous kernel* K *defined on X. If minimizer $f^\dagger$ of the expected risk $\mathcal{E}(f)$ over $\mathcal{H}_\mathsf{K}$ meets the assumption (8), then for $\alpha = \alpha_\mathbf{z} = \Theta^{-1}(|\mathbf{z}|^{-1/2})$, $\Theta(t) = \varphi(t)t$, we have with probability at least* $1 - \delta$ *that*

$$\|f^\dagger - f_\mathbf{z}^{\alpha_\mathbf{z}}\|_\rho = \mathrm{O}\left(\varphi(\Theta^{-1}(|\mathbf{z}|^{-1/2}))\sqrt{\Theta^{-1}(|\mathbf{z}|^{-1/2})}\log\frac{1}{\delta}\right). \tag{13}$$

Note that for $\varphi(t) = t^s$ the above theorem gives us the learning rate $\mathrm{O}\left(|\mathbf{z}|^{-\frac{s+\frac{1}{2}}{2(s+1)}}\right)$ that matches the result obtained in seminal paper by Smale and Zhou [17]. Moreover, for $\varphi(t) = t^s$ the rate (13) can be thought of as the limit case of the capacity dependent learning rate $O\left(|\mathbf{z}|^{-\frac{(s+\frac{1}{2})\mu}{2s\mu+\mu+1}}\right)$ obtained in [3] under the assumptions that the eigenvalues λ_i of the covariance operator $\mathbf{C}$ have polynomial decay $\lambda_i \asymp i^{-\mu}$ with $\mu > 1$.

Now we are going to prove that the same learning rate (13) can be achieved in Nyström type subsampling (2) if the approximation power of $\mathrm{P}_{\mathbf{z}^\nu}$ is high enough.

Theorem 2. *Assume the conditions of theorem 1, and let (7) be satisfied. If the size $m = |\mathbf{z}^\nu|$ of a subsampling $\mathbf{z}^\nu$ is chosen such that*

$$\Delta_m \leqslant \sqrt{\Theta_{1/2}^{-1}(|\mathbf{z}|^{-1/2})}, \; \Theta_{1/2}(t) = \varphi(t)\sqrt{t},$$

then with probability at least $1 - \delta$ we have

$$\|f^\dagger - f_{\mathbf{z},\mathbf{z}^\nu}^{\alpha_\mathbf{z}}\|_\rho = O\left(\varphi(\Theta^{-1}(|\mathbf{z}|^{-1/2}))\sqrt{\Theta^{-1}(|\mathbf{z}|^{-1/2})}\log^{\beta_2}\frac{1}{\delta}\right), \tag{14}$$

where $\beta_2 = \max\{1, \beta_1\}$, and β_1 is the same as in (7).

Before proving this statement, we first comment on the computational complexity of Nyström approximation (2) with a subsampling size $|\mathbf{z}^\nu|$ chosen according to theorem 2.

In view of the assumption (7) it is clear that the condition of the theorem can be satisfied with

$$|\mathbf{z}^\nu| \asymp [\Theta_{1/2}^{-1}(|\mathbf{z}|^{-1/2})]^{-\frac{1}{2\beta}}.$$

Let the assumption (8) be satisfied with

$$\varphi(t) = o(t^{\frac{1-\beta}{2\beta}})\text{ast} \to 0, \tag{15}$$

i.e. $\Theta_{1/2}(t) = o(t^{1/2\beta})$. Then

$$|\mathbf{z}|^{-\beta} = o(\Theta_{1/2}^{-1}(|\mathbf{z}|^{-1/2})) = o(|\mathbf{z}^\nu|^{-2\beta}),$$

which means that $|\mathbf{z}^\nu|^2 = o(|\mathbf{z}|)$ as $|\mathbf{z}| \to \infty$.

On the other hand, the computational complexity of (2) is of order $O(|\mathbf{z}||\mathbf{z}^\nu|^2)$ (see, e.g. [13]), and under the condition (15) it is subquadratic, because $|\mathbf{z}||\mathbf{z}^\nu|^2 = o(|\mathbf{z}|^2)$.

Thus, under the conditions of theorem 2 Nyström subsampling has the same learning rate as the one guaranteed by theorem 1 for KRR based on the whole sample $\mathbf{z}$. Moreover, theorem 2 allows an estimation of a regularity range, such as (15), for which the above mentioned learning rate can be achieved with subquadratic complexity. Note, that the condition (15) is automatically satisfied with $\beta \geqslant 1$, for example.

Proof of Theorem 2. It is known (see, e.g. [11]) that the following inequality holds true for functions φ mentioned in the assumption (8)

$$\sup_t |(1 - (\alpha + t)^{-1}t)\varphi(t)t^q| \leqslant h_{\varphi,q}\varphi(\alpha)\alpha^q, \; q \in [0, 1/2], \tag{16}$$

where $h_{\varphi,q}$ depends only on φ and q.

Note also that, by very definition, $\Theta_{1/2}(|\mathbf{z}|^{-1/2}) > \Theta(|\mathbf{z}|^{-1/2})$, and therefore

$$\Delta_m^2 = \Theta_{1/2}^{-1}(|\mathbf{z}|^{-1/2}) < \Theta^{-1}(|\mathbf{z}|^{-1/2}) = \alpha_\mathbf{z}. \tag{17}$$

Moreover, without loss of generality we can assume that $|\mathbf{z}|$ is so large that

$$\varphi(\max\{d_{1,\delta}, d_{2,\delta}\}|\mathbf{z}|^{-1/2}) < [\max\{d_{1,\delta}, d_{2,\delta}\}], \tag{18}$$

where $d_{1,\delta}$, $d_{2,\delta}$ are the numbers appearing in (11) and (12). This is not a real restriction, because the left-hand side of (18) tends to zero as $|\mathbf{z}| \to \infty$. A direct implication of (18) is that with probability at least $1 - \delta$

$$\alpha_{\mathbf{z}} = \Theta^{-1}(|\mathbf{z}|^{-1/2}) > \max\{\|\mathbf{C} - \mathbf{C}_{\mathbf{z}}\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}}, \|\mathbf{C}_{\mathbf{z}} f^{\dagger} - \mathrm{S}_{\mathbf{z}}^{*}\mathbf{Y}\|_{\mathsf{K}}\}. \tag{19}$$

Consider the decomposition

$$f^{\dagger} - f_{\mathbf{z},\mathbf{z}^{\nu}}^{\alpha_{\mathbf{z}}} = \sigma_1 + \sigma_2 + \sigma_3, \tag{20}$$

where

$$\begin{aligned}
\sigma_1 &= f^{\dagger} - \mathrm{P}_{\mathbf{z}^{\nu}} f^{\dagger}, \\
\sigma_2 &= \mathrm{P}_{\mathbf{z}^{\nu}} f^{\dagger} - (\alpha_{\mathbf{z}}\mathrm{I} + \mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}\mathrm{P}_{\mathbf{z}^{\nu}})^{-1}\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}\mathrm{P}_{\mathbf{z}^{\nu}} f^{\dagger}, \\
\sigma_3 &= (\alpha_{\mathbf{z}}\mathrm{I} + \mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}\mathrm{P}_{\mathbf{z}^{\nu}})^{-1}(\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}\mathrm{P}_{\mathbf{z}^{\nu}} f^{\dagger} - \mathrm{P}_{\mathbf{z}^{\nu}}\mathrm{S}_{\mathbf{z}}^{*}\mathbf{Y})
\end{aligned}$$

and we use notation $\mathbf{C}_{\mathbf{z}} = \mathrm{S}_{\mathbf{z}}^{*}\mathrm{S}_{\mathbf{z}}$.

Now we are going to bound each term of (20). From (6)–(8) and (10) we have

$$\begin{aligned}
\|\sigma_1\|_{\rho} &= \|\mathbf{C}^{1/2}(\mathrm{I} - \mathrm{P}_{\mathbf{z}^{\nu}})\varphi(\mathbf{C})v^{\dagger}\|_{\mathsf{K}} \\
&\leqslant R\|\mathbf{C}^{1/2}(\mathrm{I} - \mathrm{P}_{\mathbf{z}^{\nu}})\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \|(\mathrm{I} - \mathrm{P}_{\mathbf{z}^{\nu}})\varphi(\mathbf{C})\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \\
&\leqslant R\Delta_m \varphi(\Delta_m^2) = R\Theta_{1/2}(\Delta_m^2) \\
&\leqslant R\Theta_{1/2}(\Theta_{1/2}^{-1}(|\mathbf{z}|^{-1/2})) = R|\mathbf{z}|^{-1/2}.
\end{aligned} \tag{21}$$

To prove (14) we also need to bound σ_2, σ_3 in the norms $\|\cdot\|_{\mathsf{K}}$ and $\|\cdot\|_{\rho}$. We start with the decomposition

$$\sigma_2 = \sigma_{2,1} + \sigma_{2,2}, \tag{22}$$

where

$$\begin{aligned}
\sigma_{2,1} &= (\mathrm{I} - (\alpha_{\mathbf{z}}\mathrm{I} + \mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}\mathrm{P}_{\mathbf{z}^{\nu}})^{-1}\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}\mathrm{P}_{\mathbf{z}^{\nu}})\varphi(\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}\mathrm{P}_{\mathbf{z}^{\nu}})v^{\dagger}, \\
\sigma_{2,2} &= (\mathrm{I} - (\alpha_{\mathbf{z}}\mathrm{I} + \mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}\mathrm{P}_{\mathbf{z}^{\nu}})^{-1}\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}\mathrm{P}_{\mathbf{z}^{\nu}})\sigma_{2,2,1}, \\
\sigma_{2,2,1} &= (\mathrm{P}_{\mathbf{z}^{\nu}}\varphi(\mathbf{C}) - \mathrm{P}_{\mathbf{z}^{\nu}}\varphi(\mathbf{C})\mathrm{P}_{\mathbf{z}^{\nu}} + \mathrm{P}_{\mathbf{z}^{\nu}}\varphi(\mathbf{C})\mathrm{P}_{\mathbf{z}^{\nu}} \\
&\quad - \varphi(\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}\mathrm{P}_{\mathbf{z}^{\nu}}) + \varphi(\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}\mathrm{P}_{\mathbf{z}^{\nu}}) - \varphi(\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}\mathrm{P}_{\mathbf{z}^{\nu}}))v^{\dagger}.
\end{aligned}$$

From (16) it follows that

$$\|\sigma_{2,1}\|_{\mathsf{K}} \leqslant R\sup_t |(1 - (\alpha_{\mathbf{z}} + t)^{-1}t)\varphi(t)| \leqslant Rh_{\varphi,0}\varphi(\alpha_{\mathbf{z}}).$$

Moreover,

$$\begin{aligned}
\|\sigma_{2,1}\|_{\rho} &= \|\mathbf{C}^{1/2}\sigma_{2,1}\|_{\mathsf{K}} \\
&\leqslant \|\mathbf{C}_{\mathbf{z}}^{1/2}\mathrm{P}_{\mathbf{z}^{\nu}}\sigma_{2,1}\|_{\mathsf{K}} + \|(\mathbf{C}^{1/2} - \mathbf{C}_{\mathbf{z}}^{1/2})\mathrm{P}_{\mathbf{z}^{\nu}}\sigma_{2,1}\|_{\mathsf{K}}
\end{aligned}$$

and

$$\begin{aligned}
\|\mathbf{C}_{\mathbf{z}}^{1/2}\mathrm{P}_{\mathbf{z}^{\nu}}\sigma_{2,1}\|_{\mathsf{K}} &\leqslant \|(\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}\mathrm{P}_{\mathbf{z}^{\nu}})^{1/2}\sigma_{2,1}\|_{\mathsf{K}} \\
&\leqslant R\sup_t |(1 - (\alpha_{\mathbf{z}} + t)^{-1}t)t^{1/2}\varphi(t)| \leqslant Rh_{\varphi,\frac{1}{2}}\alpha_{\mathbf{z}}^{1/2}\varphi(\alpha_{\mathbf{z}}).
\end{aligned}$$

Keeping in mind that $\psi(t) = \sqrt{t}$ is an operator monotone function, from (9), (17) and (19), we have

$\|(\mathbf{C}^{1/2} - \mathbf{C}_{\mathbf{z}}^{1/2})\mathrm{P}_{\mathbf{z}^{\nu}}\sigma_{2,1}\|_{\mathsf{K}} \leqslant d_{1/2}\|\mathbf{C} - \mathbf{C}_{\mathbf{z}}\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}^{\frac{1}{2}}\|\sigma_{2,1}\|_{\mathsf{K}} \leqslant d_{1/2}Rh_{\varphi,0}\alpha_{\mathbf{z}}^{\frac{1}{2}}\varphi(\alpha_{\mathbf{z}}).$

All together this gives us the bound

$$\|\sigma_{2,1}\|_{\rho} = \mathrm{O}(\varphi(\alpha_{\mathbf{z}})\alpha_{\mathbf{z}}^{1/2}) = \mathrm{O}(\varphi(\Theta^{-1}(|\mathbf{z}|^{-1/2}))\sqrt{\Theta^{-1}(|\mathbf{z}|^{-1/2})}).$$

To estimate $\|\sigma_{2,2}\|_{\rho}$ we need to bound $\|\sigma_{2,2,1}\|_{\mathsf{K}}$. For this end, we use the following known estimate (see proposition 3 [11])

$$\|\mathrm{P}_{\mathbf{z}^{\nu}}\varphi(\mathbf{C})\mathrm{P}_{\mathbf{z}^{\nu}} - \varphi(\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}\mathrm{P}_{\mathbf{z}^{\nu}})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} \leqslant \bar{d}_{\varphi}\varphi(\|\mathbf{C}^{1/2}(\mathrm{I} - \mathrm{P}_{\mathbf{z}^{\nu}})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}^{2}).$$

Moreover, (9), (10), (17) and (19) give us

$$\|\varphi(\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}\mathrm{P}_{\mathbf{z}^{\nu}}) - \varphi(\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}\mathrm{P}_{\mathbf{z}^{\nu}})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} \leqslant d_{\varphi}\varphi(\|\mathbf{C} - \mathbf{C}_{\mathbf{z}}\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}) \leqslant d_{\varphi}\varphi(\alpha_{\mathbf{z}}),$$

and

$$\begin{aligned}&\|\mathrm{P}_{\mathbf{z}^{\nu}}\varphi(\mathbf{C}) - \mathrm{P}_{\mathbf{z}^{\nu}}\varphi(\mathbf{C})\mathrm{P}_{\mathbf{z}^{\nu}}\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} \leqslant \|\varphi(\mathbf{C})(\mathrm{I} - \mathrm{P}_{\mathbf{z}^{\nu}})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}\\ &\quad= \|(\mathrm{I} - \mathrm{P}_{\mathbf{z}^{\nu}})\varphi(\mathbf{C})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}} \leqslant \varphi(\|\mathbf{C}^{1/2}(\mathrm{I} - \mathrm{P}_{\mathbf{z}^{\nu}})\|_{\mathcal{H}_{\mathsf{K}}\to\mathcal{H}_{\mathsf{K}}}^{2}) \leqslant \varphi(\alpha_{\mathbf{z}}).\end{aligned}$$

Therefore, $\|\sigma_{2,2,1}\|_{\mathsf{K}} \leqslant R(\bar{d}_{\varphi} + d_{\varphi} + 1)\varphi(\alpha_{\mathbf{z}})$, and

$$\|\sigma_{2,2}\|_{\mathsf{K}} \leqslant \|\sigma_{2,2,1}\|_{\mathsf{K}} \sup_{t}|1 - \frac{t}{\alpha_{\mathbf{z}} + t}| \leqslant \|\sigma_{2,2,1}\|_{\mathsf{K}} = \mathrm{O}(\varphi(\alpha_{\mathbf{z}})).$$

Then, using the same argument as for $\|\sigma_{2,2,1}\|_{\rho}$ we obtain

$$\begin{aligned}\|\sigma_{2,2}\|_{\rho} &= \mathrm{O}(\varphi(\Theta^{-1}(|\mathbf{z}|^{-1/2}))\sqrt{\Theta^{-1}(|\mathbf{z}|^{-1/2})}),\\ \|\sigma_{2}\|_{\rho} &= \mathrm{O}(\varphi(\Theta^{-1}(|\mathbf{z}|^{-1/2}))\sqrt{\Theta^{-1}(|\mathbf{z}|^{-1/2})}).\end{aligned}$$

Finally, we need to estimate $\|\sigma_3\|_{\rho}$. Observe that

$$\begin{aligned}\|\sigma_3\|_{\mathsf{K}} &\leqslant \sup_{t}|(\alpha_{\mathbf{z}} + t)^{-1}|\|\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}\mathrm{P}_{\mathbf{z}^{\nu}}f^{\dagger} - \mathrm{P}_{\mathbf{z}^{\nu}}\mathrm{S}_{\mathbf{z}}^{*}\mathbf{Y}\|_{\mathsf{K}}\\ &\leqslant \frac{1}{\alpha_{\mathbf{z}}}(\|\mathrm{P}_{\mathbf{z}^{\nu}}(\mathbf{C}_{\mathbf{z}}f^{\dagger} - \mathrm{S}_{\mathbf{z}}^{*}\mathbf{Y})\|_{\mathsf{K}} + \|\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}f^{\dagger} - \mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}\mathrm{P}_{\mathbf{z}^{\nu}}f^{\dagger}\|_{\mathsf{K}})\end{aligned}$$

Then using (10)–(12) we obtain

$$\begin{aligned}&\|\mathrm{P}_{\mathbf{z}^{\nu}}(\mathbf{C}_{\mathbf{z}}f^{\dagger} - \mathrm{S}_{\mathbf{z}}^{*}\mathbf{Y})\|_{\mathsf{K}} \leqslant d_{2,\delta}|\mathbf{z}|^{-1/2},\\ &\|\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}f^{\dagger} - \mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}_{\mathbf{z}}\mathrm{P}_{\mathbf{z}^{\nu}}f^{\dagger}\|_{\mathsf{K}} \leqslant \|\mathrm{P}_{\mathbf{z}^{\nu}}(\mathbf{C}_{\mathbf{z}} - \mathbf{C})f^{\dagger}\|_{\mathsf{K}}\\ &\quad + \|\mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}f^{\dagger} - \mathrm{P}_{\mathbf{z}^{\nu}}\mathbf{C}\mathrm{P}_{\mathbf{z}^{\nu}}f^{\dagger}\|_{\mathsf{K}} + \|\mathrm{P}_{\mathbf{z}^{\nu}}(\mathbf{C} - \mathbf{C}_{\mathbf{z}})\mathrm{P}_{\mathbf{z}^{\nu}}f^{\dagger}\|_{\mathsf{K}}\\ &\leqslant 2d_{1,\delta}\|f^{\dagger}\|_{\mathsf{K}}|\mathbf{z}|^{-1/2} + \|\mathbf{C}(\mathrm{I} - \mathrm{P}_{\mathbf{z}^{\nu}})(\mathrm{I} - \mathrm{P}_{\mathbf{z}^{\nu}})\varphi(\mathbf{C})v^{\dagger}\|_{\mathsf{K}}\\ &\leqslant d_{3,\delta}(|\mathbf{z}|^{-1/2} + \Delta_{m}\varphi(\Delta_{m}^{2}))\\ &\leqslant d_{3,\delta}(|\mathbf{z}|^{-1/2} + \Theta_{1/2}(\Theta_{1/2}^{-1}(|\mathbf{z}|^{-1/2}))) = 2d_{3,\delta}|\mathbf{z}|^{-1/2},\end{aligned}$$

that allows us to write

$$\begin{aligned}\|\sigma_3\|_{\mathsf{K}} &= \mathrm{O}(\alpha_{\mathbf{z}}^{-1}|\mathbf{z}|^{-1/2}) = \mathrm{O}(\alpha_{\mathbf{z}}^{-1}\Theta(\Theta^{-1}(|\mathbf{z}|^{-1/2})))\\ &= \mathrm{O}([\Theta^{-1}(|\mathbf{z}|^{-1/2})]^{-1}\varphi(\Theta^{-1}(|\mathbf{z}|^{-1/2}))\Theta^{-1}(|\mathbf{z}|^{-1/2}))\\ &= \mathrm{O}(\varphi(\Theta^{-1}(|\mathbf{z}|^{-1/2}))).\end{aligned}$$

Using again the same argument as for $\|\sigma_{2,1}\|_{\rho}$ we obtain

$$\|\sigma_3\|_{\rho} = \mathrm{O}(\varphi(\Theta^{-1}(|\mathbf{z}|^{-1/2}))\sqrt{\Theta^{-1}(|\mathbf{z}|^{-1/2})}).$$

Summing up the above bounds for $\|\sigma_i\|$, $i = 1, 2, 3$, we prove the statement of the theorem. □

3. Dealing with uncertainty in the sampling size $|\mathbf{z}^{\nu}|$

Theorem 2 contains a recipe for choosing the subsampling size $|\mathbf{z}^{\nu}|$ depending on the regularity of the target function and on the approximation power of the corresponding projection method. Both of them, especially the first, may not be exactly given in the form described above. Then several subsampling sizes $|\mathbf{z}^{\nu_1}|$, $|\mathbf{z}^{\nu_2}|$,...,$|\mathbf{z}^{\nu_l}|$ may be tried in Nyström method, provided that $|\mathbf{z}^{\nu_i}| = o(|\mathbf{z}|^{1/2})$, $i = 1, 2, ..., l$. Of course, the number l of possible size candidates should not be too large to allow a calculation of all corresponding approximants $f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_1}}, f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_2}}, ..., f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_l}}$ with a subquadratic complexity. Nevertheless, the question appears of how to select a good approximant among the calculated ones, or how to use all of them. This question is similar to the one in the regularization theory, where some strategy for aggregating all calculated regularized approximants has been discussed recently [4]. In [8] the strategy [4] has been adjusted in the context of learning and presented in several versions.

According to the simplest version, the intention is to approximate the vector $c^* = (c_1^*, c_2^*, ..., c_l^*) \in \mathbb{R}^l$ solving the following minimization problem

$$\|f^{\dagger} - \sum_{i=1}^{l} c_i f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}\|_{\rho} \to \min. \tag{23}$$

Recall that $\|\cdot\|_{\rho}$ is the norm of the Hilbert space $L_2(X, \rho_X)$. Therefore, (23) is equivalent to the matrix problem

$$Gc = g^{\dagger}, \tag{24}$$

where G and $g^{\dagger}$ are respectively a Gram matrix and a vector of inner products $\langle \cdot, \cdot \rangle_{\rho}$ in $L_2(X, \rho_X)$, i.e.

$$G = (\langle f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}, f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_j}} \rangle_{\rho})_{i,j=1}^{l}, \quad g^{\dagger} = (\langle f^{\dagger}, f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_j}} \rangle_{\rho})_{i=1}^{l}. \tag{25}$$

Note that neither Gram matrix G nor the vector $g^{\dagger}$ is accessible, since the target function $f^{\dagger}$ is unknown and the marginal probability distribution ρ_X, which is involved in the definition of $\langle \cdot, \cdot \rangle_{\rho}$, is not assumed to be given.

On the other hand, $f^{\dagger}, f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}$, $i = 1, 2, ..., l$, belong to the space $\mathcal{H}_{\mathsf{K}}$. That is assumed to be continuously embedded into $L_2(X, \rho_X)$. Then, for example

$$\begin{aligned}\langle f^{\dagger}, f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}} \rangle_{\rho} &= \langle \mathsf{J}_{\mathsf{K}} f^{\dagger}, \mathsf{J}_{\mathsf{K}} f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}} \rangle_{\rho} = \langle \mathbf{C} f^{\dagger}, f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}} \rangle_{\mathsf{K}} \\ &= \langle (\mathbf{C} - \mathbf{C}_{\mathbf{z}}) f^{\dagger}, f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}} \rangle_{\mathsf{K}} + \langle \mathbf{C}_{\mathbf{z}} f^{\dagger} - \mathrm{S}^*_{\mathbf{z}} \mathbf{Y}, f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}} \rangle_{\mathsf{K}} + \langle \mathrm{S}^*_{\mathbf{z}} \mathbf{Y}, f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}} \rangle_{\mathsf{K}}.\end{aligned} \tag{26}$$

In view of (11) the first term of the last equality (26) can be estimated as follows:

$$\begin{aligned}|\langle (\mathbf{C} - \mathbf{C}_{\mathbf{z}}) f^{\dagger}, f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}} \rangle_{\mathsf{K}}| &\leqslant \|\mathbf{C} - \mathbf{C}_{\mathbf{z}}\|_{\mathcal{H}_{\mathsf{K}} \to \mathcal{H}_{\mathsf{K}}} \cdot \|f^{\dagger}\|_{\mathsf{K}} \cdot \|f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}\|_{\mathsf{K}} \\ &\leqslant d_{1,\delta} |\mathbf{z}|^{-1/2} \|f^{\dagger}\|_{\mathsf{K}} \cdot \|f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}\|_{\mathsf{K}}.\end{aligned} \tag{27}$$

Moreover, the norm $\|f^{\dagger}\|_{\mathsf{K}}$ does not depend on $|\mathbf{z}|$, $|\mathbf{z}^{\nu_i}|$, and the norm $\|f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}\|_{\mathsf{K}}$ can be controlled. So, with a high probability it holds

$$|\langle (\mathbf{C} - \mathbf{C}_{\mathbf{z}}) f^{\dagger}, f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}} \rangle_{\mathsf{K}}| = \mathrm{O}(|\mathbf{z}|^{-1/2}). \tag{28}$$

In the same way, with the use of (12) we have

$$|\langle \mathbf{C} f^{\dagger} - \mathrm{S}^*_{\mathbf{z}} \mathbf{Y}, f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}} \rangle_{\mathsf{K}}| = \mathrm{O}(|\mathbf{z}|^{-1/2}). \tag{29}$$

As to the third term of the last equality (26), it can be directly calculated from the training data since

$$\langle \mathrm{S}_{\mathbf{z}}^{*}\mathbf{Y}, f_{\mathbf{z},\mathbf{z}^{\nu_i}}^{\alpha} \rangle_{\mathrm{K}} = \langle \mathbf{Y}, \mathrm{S}_{\mathbf{z}} f_{\mathbf{z},\mathbf{z}^{\nu_i}}^{\alpha} \rangle_{\mathbb{R}^{|\mathbf{z}|}} = |z|^{-1}\sum_{k=1}^{|\mathbf{z}|} y_k f_{\mathbf{z},\mathbf{z}^{\nu_i}}^{\alpha}(x_k). \tag{30}$$

Therefore, from (26)–(30) we have with high probability

$$\langle f^{\dagger}, f_{\mathbf{z},\mathbf{z}^{\nu_i}}^{\alpha} \rangle_{\rho} = |z|^{-1}\sum_{k=1}^{|\mathbf{z}|} y_k f_{\mathbf{z},\mathbf{z}^{\nu_i}}^{\alpha}(x_k) + \mathrm{O}(|\mathbf{z}|^{-1/2}),\ i = 1, 2, \ldots, l. \tag{31}$$

Similar reasoning gives us the relations

$$\langle f_{\mathbf{z},\mathbf{z}^{\nu_i}}^{\alpha}, f_{\mathbf{z},\mathbf{z}^{\nu_j}}^{\alpha} \rangle_{\rho} = |z|^{-1}\sum_{k=1}^{|\mathbf{z}|} f_{\mathbf{z},\mathbf{z}^{\nu_i}}^{\alpha}(x_k) f_{\mathbf{z},\mathbf{z}^{\nu_j}}^{\alpha}(x_k) + \mathrm{O}(|\mathbf{z}|^{-1/2}),$$
$$i, j = 1, 2, \ldots, l. \tag{32}$$

In view of (31) and (32) the matrix

$$\tilde{G} = \left(|\mathbf{z}|^{-1}\sum_{k=1}^{|\mathbf{z}|} f_{\mathbf{z},\mathbf{z}^{\nu_i}}^{\alpha}(x_k) f_{\mathbf{z},\mathbf{z}^{\nu_j}}^{\alpha}(x_k) \right)_{i,j=1}^{l}$$

and the vector

$$\tilde{g} = \left(|\mathbf{z}|^{-1}\sum_{k=1}^{|\mathbf{z}|} y_k f_{\mathbf{z},\mathbf{z}^{\nu_i}}^{\alpha}(x_k) \right)_{i=1}^{l}$$

can be considered as approximations of G and $g^{\dagger}$ respectively. Moreover, with probability at least $1 - \delta$

$$\|G - \tilde{G}\|_{\mathbb{R}^l} = \mathrm{O}\left(|\mathbf{z}|^{-1/2} \log\frac{1}{\delta} \right), \quad \|g^{\dagger} - \tilde{g}\|_{\mathbb{R}^l} = \mathrm{O}\left(|\mathbf{z}|^{-1/2} \log\frac{1}{\delta} \right).$$

With the matrix $\tilde{G}$ in hand one can easily test whether or not $\tilde{G}^{-1}$ exists. For sufficiently large $|\mathbf{z}|$ in case of positive test result a standard perturbation argument (see, e.g. [8] for details) implies the invertibility of G^{-1}, the existence of the vectors $c^{*} = G^{-1}g^{\dagger}$, $\tilde{c} = \tilde{G}^{-1}\tilde{g}$ and the bound

$$\|c^{*} - \tilde{c}\|_{\mathbb{R}^l} = \mathrm{O}\left(|\mathbf{z}|^{-1/2} \log\frac{1}{\delta} \right)$$

that holds with probability at least $1 - \delta$.

Consider now the function

$$f_{\mathbf{z}}^{*} = \sum_{i=1}^{l} c_i^{*} f_{\mathbf{z},\mathbf{z}^{\nu_i}}^{\alpha},$$

that solves (23), and its approximation

$$\tilde{f}_{\mathbf{z}} = \sum_{i=1}^{l} \tilde{c}_i f_{\mathbf{z},\mathbf{z}^{\nu_i}}^{\alpha},$$

where $\tilde{c}_i$, $i = 1, 2, \ldots, l$, are the components of the vector $\tilde{c} = \tilde{G}^{-1}\tilde{g}$. Since $f_{\mathbf{z},\mathbf{z}^{\nu_i}}^{\alpha}$, $i = 1, 2, \ldots, l$, are up to our choice, their norms can be controlled such that

$$\|f_{\mathbf{z}}^{*} - \tilde{f}_{\mathbf{z}}\|_{\rho} \leqslant l \max_i \|f_{\mathbf{z},\mathbf{z}^{\nu_i}}^{\alpha}\|_{\rho} \|c^{*} - \tilde{c}\|_{\mathbb{R}^l} = \mathrm{O}\left(|\mathbf{z}|^{-1/2} \log\frac{1}{\delta} \right).$$

This gives us the following statement

Theorem 3. *Assume that $\tilde{G}$ is invertible and consider $\tilde{f}_{\mathbf{z}} = \sum_{i=1}^{l} \tilde{c}_i f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}$, $\tilde{c} = (\tilde{c}_i)_{i=1}^{l} = \tilde{G}^{-1}\tilde{g}$. Then under the conditions of theorem 2 for sufficiently large $|\mathbf{z}|$ we have with probability at least $1 - \delta$*

$$\|f^{\dagger} - \tilde{f}_{\mathbf{z}}\|_{\rho} = \min_{c_i} \|f^{\dagger} - \sum_{i=1}^{l} c_i f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}\|_{\rho} + O\left(|\mathbf{z}|^{-1/2} \log \frac{1}{\delta}\right),$$

where a coefficient implicit in O-symbol may depend on the cardinality l of the family $\{f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}\}$ and on the distribution ρ, but does not depend on $|\mathbf{z}|$ and δ.

Note that in theorem 3 the term $O\left(|\mathbf{z}|^{-1/2} \log \frac{1}{\delta}\right)$ is negligible because, as we know from [3], $|\mathbf{z}|^{-1/2}$ is of higher order than the best guaranteed accuracy of a reconstruction of the target function $f^{\dagger} \in \mathcal{H}_{\mathsf{K}}$ in $L_2(X, \rho_X)$ from a training set $\mathbf{z}$.

Thus, theorem 3 tells us that the effectively constructed linear combination of the candidates $f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}$, $i = 1, 2, ..., l$, is almost as accurate as the best linear aggregator of them.

A simple algorithmic sketch of constructing an almost best aggregator is provided below:

Algorithm. Algorithm of an aggregation of Nyström approximants

```
Input:    Dataset $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^{N}$,
          Subsampling datasets $\mathbf{z}^{\nu_1}, \mathbf{z}^{\nu_2}, ..., \mathbf{z}^{\nu_l} \subset \mathbf{z}$,
          Regularization parameter(s) $\alpha$, kernel $\mathsf{K}$
Output:   Nyström estimators $f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}$, $i = 1, 2, ..., l$,
          and their aggregator $\tilde{f}_{\mathbf{z}}$

for $j = 1$ to $l$ do
    calculate $f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu}}$ according to (3), (4), where $\mathbf{z}^{\nu} = \mathbf{z}^{\nu_j}$.
end for
for $j = 1$ to $l$ do (30)
  $\tilde{g}_j \leftarrow N^{-1}\sum_{i=1}^{N} y_i f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_j}}(x_i)$
end for
for $j = 1$ to $l$ do
  for $k = 1$ to $l$ do
    $\tilde{G}_{k,j} \leftarrow N^{-1}\sum_{i=1}^{N} f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_k}}(x_i) f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_j}}(x_i)$
  end for
end for
$\tilde{c} \leftarrow \tilde{G}^{-1}\tilde{g}$, where $\tilde{c} := (c_j)_{j=1}^{N}$, $\tilde{g} := (\tilde{g}_j)_{j=1}^{N}$, $\tilde{G} := (\tilde{G}_{k,j})_{k,j=1}^{N}$.
return $\tilde{f}_{\mathbf{z}} \leftarrow \sum_{j=1}^{N} \tilde{c}_j f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_j}}$
```

In the next section we present some numerical experiments illustrating the performance of the aggregator $\tilde{f}_{\mathbf{z}}$.

4. Numerical experiments

For our first experiment we simulate data in the same way as in [21], where another strategy for learning with big data called divide and conquer algorithm or distributed learning has been studied. Following that paper, we simulate training data sets $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^{|\mathbf{z}|}$, $|\mathbf{z}| \in \{2^8, 2^9, ..., 2^{13}\}$ from the regression model $y_i = f^{\dagger}(x_i) + \xi_i$, $i = 1, 2, ..., |\mathbf{z}|$, where $f^{\dagger}(x) = \min\{x, 1 - x\}$, the random samples x_i are uniformly distributed over [0, 1], and the noise random variables ξ_i are normally distributed with zero mean and variance $\sigma^2 = 1/5$. This simulated problem can be seen as a supervised learning with $X = [0, 1]$ and $\rho_X = \mathsf{Uni}[0, 1]$.

As in [21], all KRR estimators appearing in this experiment are constructed in $\mathcal{H}_\mathsf{K}$ with $\mathsf{K}(x, x') = 1 + \min\{x, x'\}$ and $\alpha = |\mathbf{z}|^{-2/3}$.

We perform plain Nyström subsampling and construct estimators $f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_1}}, f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_2}}$ with $|\mathbf{z}^{\nu_1}| = \lfloor |\mathbf{z}|^{4/10} \rfloor$ and $|\mathbf{z}^{\nu_2}| = \lfloor |\mathbf{z}|^{3/10} \rfloor$, such that the computational complexity of their construction is of order $o(|\mathbf{z}|^2)$, i.e. subquadratic. Then, as has been discussed in theorem 3, we construct the aggregator $\tilde{f}_\mathbf{z} = \tilde{c}_1 f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_1}} + \tilde{c}_2 f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_2}}$.

The accuracy of $\tilde{f}_\mathbf{z}$ is compared with the one of divide and conquer algoithm [21]. That algorithm is based on splitting a large training set $\mathbf{z}$ into p much smaller equal-sized subsets $\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_p$, $|\mathbf{z}_i| = \left\lfloor |\mathbf{z}|/p \right\rfloor$, $i = 1, 2, ..., p$; then, each data set $\mathbf{z}_i$ is used as a training set for constructing the minimizer $f^\alpha_{\mathbf{z}_i}$ of (1), where $\mathbf{z}$ is substituted for $\mathbf{z}_i$; finally, the approximations $f^\alpha_{\mathbf{z}_i}$, $i = 1, 2, ..., p$, are aggregated linearly with equal coefficients (averaged) into

$$f^\alpha_{\mathbf{z},p} = p^{-1} \sum_{i=1}^{p} f^\alpha_{\mathbf{z}_i}.$$

In our experiment we compare the errors $\|f^\dagger - \tilde{f}_\mathbf{z}\|$, $\|f^\dagger - f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_i}}\|$, $i = 1, 2$, and $\|f^\dagger - f^\alpha_{\mathbf{z},p}\|$. As in [21] we consider $p = 1, 4, 16, 64$, and execute each simulation 20 times to obtain average values of the errors. In figure 1 we plot these values versus the total number of samples $|\mathbf{z}|$, where the values corresponding to $\|f^\dagger - \tilde{f}_\mathbf{z}\|$, $\|f^\dagger - f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_i}}\|$, and $\|f^\dagger - f^\alpha_{\mathbf{z},p}\|$ are respectively depicted by dotted, dashed and solid lines. In addition, figure 2 shows the variation of errors over 20 simulations in terms of boxplots.

Figure 1 shows that in the considered case the aggregated approximation $\tilde{f}_\mathbf{z}$ outperforms all others, including the baseline KRR-solution $f^\alpha_{\mathbf{z},1}$ constructed for the full sample $\mathbf{z}$. It is also interesting to note, that the Nyström approximation $f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_2}}$, $|\mathbf{z}^{\nu_2}| = \lfloor |\mathbf{z}|^{3/10} \rfloor$, performs poorly, but the aggregated approximation $\tilde{f}_\mathbf{z}$ automatically uses the best of available options. This can be seen from figure 2 as well.

In our second experiment we follow the paper [13], where the dataset `pumadyn32nh` and `cpuSmall` have been used for an empirical study of the Nyström subsampling method. Dataset `pumadyn32nh` of 8192 instances is taken from the Pumadyn family[4] of realistic simulation of the dynamics of a Puma 560 robot arm. Number `32` in the name corresponds to 32 input attributes, `n`—to 'nonlinear' and `h`—to 'high unpredictability/noise'. Dataset `cpuSmall` (8192 instances) is from the collection of a computer systems activity measures comp-activ[5]. Prototask of `cpuSmall` is to predict portion of time, that cpus run in user mode from a restricted number (specifically, 12) of attributes. These datasets have been splitted in training and test sets and Gaussian kernels $\mathsf{K}(x, x') = \exp(-\|x - x'\|^2/2\sigma^2)$ have been used in construction of $f^\alpha_{\mathbf{z},\mathbf{z}^\nu}$. Moreover, 20% of the training points have been hold out for tuning such parameters as σ and α, and the performance of the selected models has been reported on the test sets.

In [13] the performance has been measured in particular by comparing the root-mean-square-errors (RMSE) of the approximations $f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_1}}, f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_2}}$ with large $|\mathbf{z}^{\nu_1}|$ and small $|\mathbf{z}^{\nu_2}|$.

It turns out (see [13] for details) that in the case of `cpuSmall` the effectiveness of the Nyström subsampling is not so high, since comparable values of RMSE of $f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_1}}, f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_2}}$ have been observed when both $|\mathbf{z}^{\nu_1}|$, $|\mathbf{z}^{\nu_2}|$, as well as $|\mathbf{z}|$, are of order of 10^3. Namely, the average

[4] http://www.cs.toronto.edu/delve/data/pumadyn/desc.html.

[5] http://www.cs.toronto.edu/delve/data/comp-activ/desc.html.

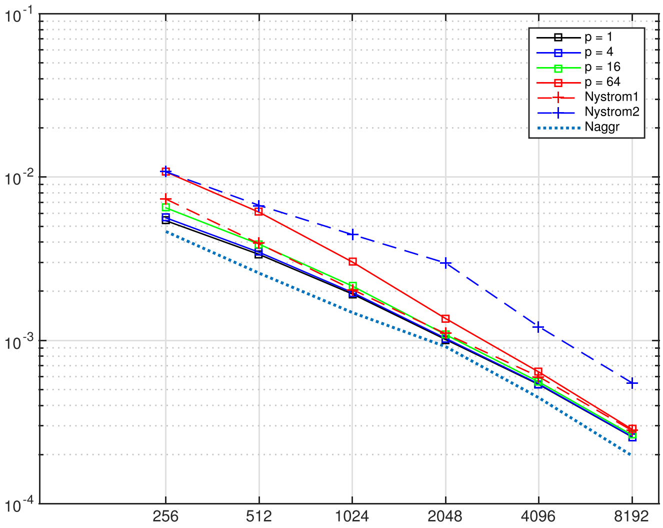

Figure 1. The mean square error between $f^\dagger$ and the averaged estimate $f^\alpha_{\mathbf{z},p}$ for $p = 1, 4, 16, 64$ (solid), Nyström solutions $f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_1}}$ ($|\mathbf{z}^{\nu_1}| = \lfloor|\mathbf{z}|^{4/10}\rfloor$) and $f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_2}}$, $|\mathbf{z}^{\nu_2}| = \lfloor|\mathbf{z}|^{3/10}\rfloor$ (dashed) and aggregated solution $\tilde{f}_\mathbf{z}$ (dotted).

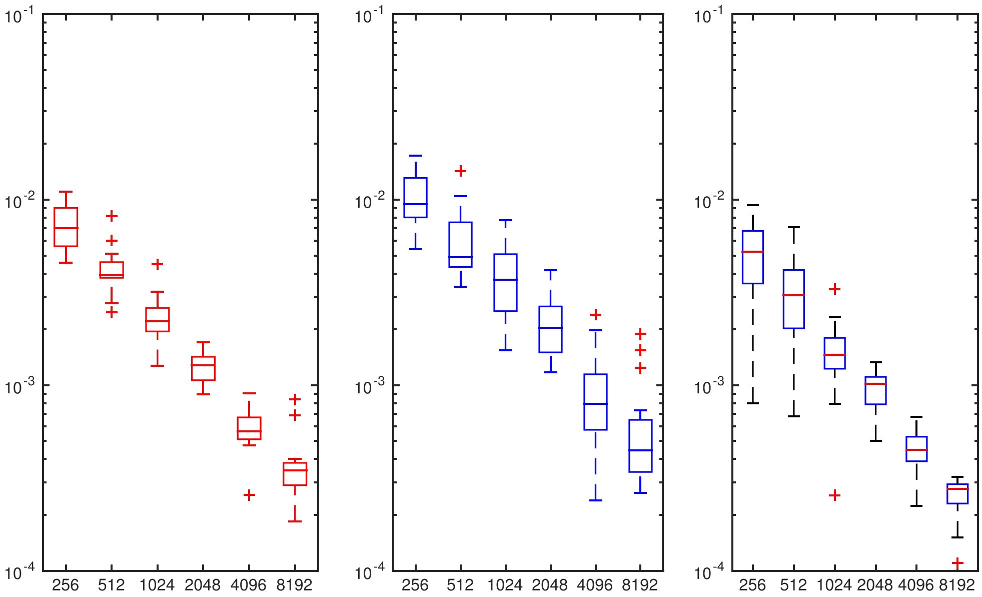

Figure 2. Variation of errors over 20 simulations for $f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_1}}$ (left red) and $f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_2}}$ (center blue) and aggregated solution $\tilde{f}_\mathbf{z}$ (right).

RMSE values over five simulations were respectively 12.2 and 13.3 for the subsampling sizes $|\mathbf{z}^{\nu_1}| = 5000$ and $|\mathbf{z}^{\nu_2}| = 2679$. Moreover, the average RMSE of $f^\alpha_{\mathbf{z},\mathbf{z}^\nu}$ was above 15 for $|\mathbf{z}^\nu| \leqslant 1000$. At the same time, in the case of `pumadyn32nh` the identical RMSE value of 0.033 has been observed for $f^\alpha_{\mathbf{z},\mathbf{z}^{\nu_i}}$, $i = 1, 2$, with $|\mathbf{z}^{\nu_1}| = 1000$ and $|\mathbf{z}^{\nu_2}| = 62$.

Table 1. Performance of Nyström approximants and their aggregator on a testing set of 4096 data points from `pumadyn32nh`.

Approximant	RMSE
$f_{\mathbf{z},\mathbf{z}^{\nu_1}}$	0.03381
$f_{\mathbf{z},\mathbf{z}^{\nu_2}}$	0.03325
$f_{\mathbf{z},\mathbf{z}^{\nu_3}}$	0.03442
Aggregator $\tilde{f}_{\mathbf{z}}$	0.03325

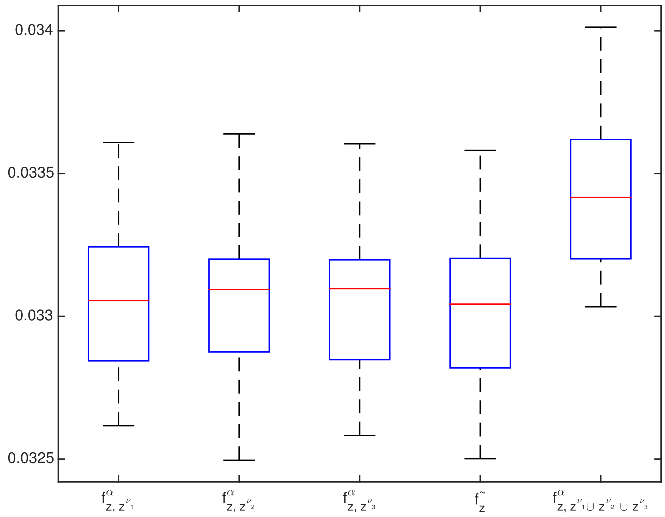

Figure 3. Variation of errors (RMSE) over 40 simulations for $f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_1}}$, $f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_2}}$, $f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_3}}$, aggregated solution $\tilde{f}_{\mathbf{z}}$, and $f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_1}\bigcup\mathbf{z}^{\nu_2}\bigcup\mathbf{z}^{\nu_3}}$ on a testing set from `pumadyn32nh`.

Such different performances may hardly be explained by different capacities of the used hypothesis spaces $\mathcal{H}_{\mathsf{K}}$, because in both considered cases they are generated by Gaussian kernels, and, moreover, the dimension of the input space X for `cpuSmall` is smaller that in case of `pumadyn32nh`.

In our theorem 2 one may find a plausible explanation of the above mentioned behaviour of Nyström approximations. Namely, that is because of the regularities of the target functions corresponding to `pumadyn32nh` and `cpuSmall` are described by source condition (8) with functions φ tending to zero with essentially different rates. This is an example of how theorem 2 can be used for interpreting empirical results and explaining limitations of the Nyström approach.

Now we use `pumadyn32nh` dataset for illustrating the performance of the arrgegators $\tilde{f}_{\mathbf{z}}$. As in [13] we construct the approximants $f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}$, $i = 1, 2, 3$, in $\mathcal{H}_{\mathsf{K}}$ generated by the Gaussian kernel of width $\sigma = 2.66$, and we use $\alpha = 10^{-7}$, $|\mathbf{z}| = 4096$, $|\mathbf{z}^{\nu_1}| = 200$, $|\mathbf{z}^{\nu_2}| = 60$, $|\mathbf{z}^{\nu_3}| = 20$. Table 1 reports the performance of $f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}$, $i = 1, 2, 3$, and $\tilde{f}_{\mathbf{z}}$.

Note that the performance of the Nyström method depends not only on the size $|\mathbf{z}^{\nu}|$ of a subsampling set $\mathbf{z}^{\nu} = \{(x_i, y_i)\}$, but also on the distribution of the inputs $\{x_i\}$, because the latter ones determine the space $\mathcal{H}_{\mathsf{K}}^{\mathbf{z}^{\nu}}$ containing approximants $f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu}}$. Therefore, even if the size

Table 2. Performance of Nyström approximants and their aggregator on a testing set from `breastcancer`, mean and standard deviation.

	RMSE	Accuracy	Precision	Sensitivity	Specificity	F1-score
$f_{\mathbf{z},\mathbf{z}^{\nu_1}}$	0.209 ± 0.016	0.964 ± 0.015	0.986 ± 0.017	0.915 ± 0.039	0.993 ± 0.009	0.948 ± 0.022
$f_{\mathbf{z},\mathbf{z}^{\nu_2}}$	0.228 ± 0.016	0.951 ± 0.015	0.988 ± 0.014	0.877 ± 0.042	0.994 ± 0.007	0.929 ± 0.023
$f_{\mathbf{z},\mathbf{z}^{\nu_3}}$	0.245 ± 0.017	0.940 ± 0.018	0.971 ± 0.027	0.862 ± 0.045	0.985 ± 0.015	0.912 ± 0.028
$\tilde{f}_{\mathbf{z}}$	0.208 ± 0.016	0.965 ± 0.015	0.986 ± 0.017	0.916 ± 0.038	0.993 ± 0.009	0.950 ± 0.022

$|\mathbf{z}^{\nu}|$ is *a priori* given, it is reasonable to aggregate the approximants $f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}$ corresponding to subsampling sets $\mathbf{z}^{\nu_i}$ of the same cardinality $|\mathbf{z}^{\nu_i}| = |\mathbf{z}^{\nu}|$, $i = 1, 2 ,\ldots.$

To illustrate this, we again consider `pumadyn32nh` dataset and choose $\mathbf{z}^{\nu_i}$, $i = 1, 2, 3$, randomly such that $|\mathbf{z}^{\nu_i}| = 50$. All other parameters are the same as in the experiment corresponding to table 1, but this time we execute simulations 40 times and report boxplot of the errors in figure 3.

Moreover, since the aggregator $\tilde{f}_{\mathbf{z}}$ of $f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}$, $i = 1, 2, 3$, belongs to the sum of $\mathcal{H}_{\mathsf{K}}^{\mathbf{z}^{\nu_i}}$, it is interesting to compare the performance of $\tilde{f}_{\mathbf{z}}$ with the one of the minimizer $f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_1}\cup\mathbf{z}^{\nu_2}\cup\mathbf{z}^{\nu_3}}$ of $T^{\alpha}_{\mathbf{z}}(f)$ over the space $\mathcal{H}_{\mathsf{K}}^{\mathbf{z}^{\nu_1}} + \mathcal{H}_{\mathsf{K}}^{\mathbf{z}^{\nu_2}} + \mathcal{H}_{\mathsf{K}}^{\mathbf{z}^{\nu_3}}$ that corresponds to the Nyström approximant constructed for $\mathcal{H}_{\mathsf{K}}^{\mathbf{z}^{\nu_1}\cup\mathbf{z}^{\nu_2}\cup\mathbf{z}^{\nu_3}}$. As it can be seen from figure 3, in the considered case the aggregation of the Nyström approximants $f^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}} \in \mathcal{H}_{\mathsf{K}}^{\mathbf{z}^{\nu_i}}$ outperforms a direct application of the Nyström approach to the sum of $\mathcal{H}_{\mathsf{K}}^{\mathbf{z}^{\nu_i}}$.

At the end we briefly discuss the performance of the aggregated Nyström approximants in the context of binary classification, that is, when y_i takes only two values, usually designated by zero and one. If a function, say f, is determined by means of a (regularized) least-squares regression from data $\{(x_i, y_i)\}$, then one can take a function $Cf(x) = 1_{\{f(x)>0.5\}}$ as a decision rule/classifier in discriminating the elements x of two classes.

Then in the context of binary classifications one usually consider the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) predictions, where TP and TN mean the numbers of cases when $Cf(x) = y = 1$ and $Cf(x) = y = 0$, while FP and FN mean, respectively, that $Cf(x) > y$ and $Cf(x) < y$. Then usual classification performance metrics are the accuracy $(\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})$, precision, also called positive predictive value, $\mathrm{PPV} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, recall, or sensitivity, $\mathrm{SE} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, and specificity $\mathrm{SE} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$. Moreover, to cumulate all the above mentioned terms the so-called balanced F1-score is also used, $F1 = 2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FN} + \mathrm{FP})$, that is the harmonic mean of precision and recall.

To illustrate the performance of the Nyström approach in classification we use the same dataset `breastcancer` as in [13]. This set consists of 569 instances with 30 features each, that describe characteristics of the digitized image of a breast mass. Outputs $y \in \{0, 1\}$ describe diagnosis (benign or malignant).

As in [13], the dataset `breastcancer` has been split in test sets consisting of 169 items, and training sets $\mathbf{z}$, $|\mathbf{z}| = 400$. Moreover, the values of the regularization parameter α and the width of Gaussian kernel σ have been chosen as $\alpha = 10^{-7}$ and $\sigma = 0.9$. Subsampling sets $\mathbf{z}^{\nu_i}$, $i = 1, 2, 3$, are randomly chosen from $\mathbf{z}$ such that $|\mathbf{z}^{\nu_1}| = 50$, $|\mathbf{z}^{\nu_2}| = 20$ $|\mathbf{z}^{\nu_2}| = 10$. The performance of Nyström classifiers $Cf^{\alpha}_{\mathbf{z},\mathbf{z}^{\nu_i}}$, $i = 1, 2, 3$, and their aggregations $C\tilde{f}^{\alpha}_{\mathbf{z}}$ over 40 random simulations is reported in table 2.

Moreover, we also consider[6] the dataset `default of credit card clients`, which consists of 30 000 instances with 23 features each, and default payment (Yes = 1, No = 0) is the binary response variable [20].

Dataset `default of credit card clients` has been split into training and testing sets consisting of 15 000 instances. The regularization parameter $\alpha = 10^{-7}$ and the width $\sigma = 0.9$ have been chosen. Subsampling sets $\mathbf{z}^{\nu_i}$, $i = 1, 2, 3$ are randomly chosen from training sets $\mathbf{z}$, $|\mathbf{z}| = 15\,000$, such that $|\mathbf{z}^{\nu_1}| = 100$, $|\mathbf{z}^{\nu_2}| = 70$, $|\mathbf{z}^{\nu_3}| = 50$. The performance of the corresponding Nyström classifiers and their aggregations over 20 simulations are reported in table 3.

[6] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

Table 3. Performance of Nyström approximants and their aggregator on a testing set, mean and standard deviation.

	RMSE	Accuracy	Precision	Sensitivity	Specificity	F1-score
$f_{\mathbf{z},\mathbf{z}^{\nu_1}}$	0.376 ± 0.002	0.810 ± 0.003	0.659 ± 0.011	0.291 ± 0.013	0.957 ± 0.002	0.403 ± 0.013
$f_{\mathbf{z},\mathbf{z}^{\nu_2}}$	0.378 ± 0.002	0.808 ± 0.003	0.658 ± 0.014	0.274 ± 0.015	0.960 ± 0.002	0.386 ± 0.016
$f_{\mathbf{z},\mathbf{z}^{\nu_3}}$	0.380 ± 0.002	0.805 ± 0.003	0.651 ± 0.013	0.253 ± 0.019	0.961 ± 0.003	0.364 ± 0.021
$\tilde{f}_{\mathbf{z}}$	0.375 ± 0.002	0.811 ± 0.003	0.662 ± 0.011	0.294 ± 0.011	0.961 ± 0.002	0.407 ± 0.011

To summarize our experiments, we have the following conclusion. Theorem 2 shows that a wide range of target functions can be learned with the best known capacity independent learning rates by means of Nyström subsampling algorithm of subquadratic complexity, provided that subsampling size is properly chosen. At the same time, a proper choice of the subsampling size requires knowledge that may not be available. Then several subsampling sizes can be tried, and in section 3 we have proposed a way of how to utilize/aggregate all of the corresponding Nyström approximants. The experiments reported above demonstrate that the aggregation approach described in section 3 automatically uses the best of the available options and can be recommended as a reliable strategy to be implemented together with the Nyström subsampling when dealing with uncertainty in the subsampling size.

Acknowledgments

The authors thank both referees for their constructive remarks. The authors affiliated with Johann Radon Institute for Computational and Applied Mathematics (RICAM) gratefully acknowledge the support of the Austrian Science Fund (FWF), projects P25424, I1669, and of the consortium AMMODIT funded within EU H2020-MSCA-RICE.

References

[1] Bach F 2013 Sharp analysis of low-rank kernel matrix approximations *JMLR: Workshop and Conf. Proc.* **30** 185–209

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

[3] Caponnetto A and De Vito E 2007 Optimal rates for the regularized least-squares algorithm *Found. Comput. Math.* **7** 331–68

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

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

[6] Hsieh C J, Si S and Dhillon I S 2014 Fast prediction for large-scale kernel machines *Advances in Neural Information Processing Systems (NIPS)* vol 27 ed Z Ghahramani *et al* (Curran Associates, Inc.) pp 3689–97

[7] Kimeldorf G S and Wahba G 1970 A correspondence between Bayesian estimation on stochastic processes and smoothing by splines *Ann. Math. Stat.* **41** 495–502

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

[9] Kumar S, Mohri M and Talwalkar A 2009 Ensemble nystrom method Advances in *Neural Information Processing Systems (NIPS)* vol 22 ed Y Bengio *et al* (Curran Associates, Inc.) pp 1060–8

[10] Mathé P and Hofmann B 2008 How general are general source conditions? *Inverse Problems* **24** 015009

[11] Mathé P and Pereverzev S V 2003 Discretization strategy for linear ill-posed problems in variable Hilbert scales *Inverse Problems* **19** 1263–77

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

[13] Rudi A, Camoriano R and Rosasco L 2015 Less is more: Nyström computational regularization *Advances in Neural Information Processing Systems* vol 28 ed C Cortes *et al* (Curran Associates, Inc.) pp 1648–56 arXiv:1507.04717

[14] Schölkopf B and Smola A J 2001 *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (Cambridge, MA: MIT Press)

[15] Shawe-Taylor J and Cristianini N 2004 *Kernel Methods for Pattern Analysis* (Cambridge: Cambridge University Press)

[16] Si S, Hsieh C J and Dhillon I 2014 Memory efficient kernel approximation *J. Mach. Learn. Res.* **32** 701–9

[17] Smale S and Zhou D X 2007 Learning theory estimates via integral operators and their approximations *Constructive Approx.* **26** 153–72

[18] Smola A J and Schölkopf B 2000 Sparse greedy matrix approximation for machine learning *Proc. 17th Int. Conf. on Machine Learning (ICML 2000) (Stanford University,Stanford, CA, USA, June 29–July 2, 2000)* pp 911–8

[19] Williams C and Seeger M 2001 Using the Nyström method to speed up kernel machines *Proc. 14th Annual Conf. on Neural Information Processing Systems* pp 682–8

[20] Yeh I C and Lien C H 2009 The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients *Expert Syst. Appl.* **36** 2473–80

[21] Zhang Y, Duchi J C and Wainwright M J 2013 Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates *JMLR: Workshop and Conf. Proc.* vol 30 pp 592–617 arXiv:1305.5029