Less is More: Nyström Computational Regularization

Alessandro Rudi[1], Raffaello Camoriano[2], Lorenzo Rosasco [1,3]

[1] Università degli Studi di Genova
DIBRIS, Via Dodecaneso 35, Genova, Italy

[2] Istituto Italiano di Tecnologia
iCub facility, Via Morego 30, Genova, Italy
raffaello.camoriano@iit.it

[3] Massachusetts Institute of Technology and Istituto Italiano di Tecnologia
Laboratory for Computational and Statistical Learning, Cambridge, MA 02139, USA
{*ale_rudi, lrosasco*}*@mit.edu*

November 6, 2015

Abstract

We study Nyström type subsampling approaches to large scale kernel methods, and prove learning bounds in the statistical learning setting, where random sampling and high probability estimates are considered. In particular, we prove that these approaches can achieve optimal learning bounds, provided the subsampling level is suitably chosen. These results suggest a simple incremental variant of Nyström Kernel Regularized Least Squares, where the subsampling level implements a form of computational regularization, in the sense that it controls at the same time regularization and computations. Extensive experimental analysis shows that the considered approach achieves state of the art performances on benchmark large scale datasets.

1 Introduction

Kernel methods provide an elegant and effective framework to develop nonparametric statistical approaches to learning [1]. However, memory requirements make these methods unfeasible when dealing with large datasets. Indeed, this observation has motivated a variety of computational strategies to develop large scale kernel methods [2–8].
In this paper we study subsampling methods, that we broadly refer to as Nyström approaches. These methods replace the empirical kernel matrix, needed by standard kernel methods, with a smaller matrix obtained by (column) subsampling [2,3]. Such procedures are shown to often dramatically reduce memory/time requirements while preserving good practical performances [9–12]. The goal of our study is two-fold. First, and foremost, we aim at providing a theoretical characterization of the generalization properties of such learning schemes in a statistical learning setting. Second, we wish to understand the role played by the subsampling level both from a statistical and a computational point of view.

As discussed in the following, this latter question leads to a natural variant of Kernel Regularized Least Squares (KRLS), where the subsampling level controls both regularization and computations.

From a theoretical perspective, the effect of Nyström approaches has been primarily characterized considering the discrepancy between a given empirical kernel matrix and its subsampled version [13–19]. While interesting in their own right, these latter results do not directly yield information on the generalization properties of the obtained algorithm. Results in this direction, albeit suboptimal, were first derived in [20] (see also [21, 22]), and more recently in [23, 24]. In these latter papers, sharp error analyses in expectation are derived in a fixed design regression setting for a form of Kernel Regularized Least Squares. In particular, in [23] a basic uniform sampling approach is studied, while in [24] a subsampling scheme based on the notion of leverage score is considered. The main technical contribution of our study is an extension of these latter results to the statistical learning setting, where the design is random and high probability estimates are considered. The more general setting makes the analysis considerably more complex. Our main result gives optimal finite sample bounds for both uniform and leverage score based subsampling strategies. These methods are shown to achieve the same (optimal) learning error as kernel regularized least squares, recovered as a special case, while allowing substantial computational gains. Our analysis highlights the interplay between the regularization and subsampling parameters, suggesting that the latter can be used to control simultaneously regularization and computations. This strategy implements a form of *computational regularization* in the sense that the computational resources are tailored to the generalization properties in the data. This idea is developed considering an incremental strategy to efficiently compute learning solutions for different subsampling levels. The procedure thus obtained, which is a simple variant of classical Nyström Kernel Regularized Least Squares with uniform sampling, allows for efficient model selection and achieves state of the art results on a variety of benchmark large scale datasets.

The rest of the paper is organized as follows. In Section 2, we introduce the setting and algorithms we consider. In Section 3, we present our main theoretical contributions. In Section 4, we discuss computational aspects and experimental results.

2 Supervised learning with KRLS and Nyström approaches

Let $X \times \mathbb{R}$ be a probability space with distribution ρ, where we view X and $\mathbb{R}$ as the input and output spaces, respectively. Let ρ_X denote the marginal distribution of ρ on X and $\rho(\cdot|x)$ the conditional distribution on $\mathbb{R}$ given $x \in X$. Given a hypothesis space $\mathcal{H}$ of measurable functions from X to $\mathbb{R}$, the goal is to minimize the *expected risk*,

$$\min_{f \in \mathcal{H}} \mathcal{E}(f), \qquad \mathcal{E}(f) = \int_{X \times \mathbb{R}} (f(x) - y)^2 d\rho(x, y), \tag{1}$$

provided ρ is known only through a training set of $(x_i, y_i)_{i=1}^n$ sampled identically and independently according to ρ. A basic example of the above setting is random design regression with the squared loss, in which case

$$y_i = f_*(x_i) + \epsilon_i, \quad i = 1, \dots, n, \tag{2}$$

with f_* a fixed *regression* function, $\epsilon_1, \dots, \epsilon_n$ a sequence of random variables seen as noise, and $x_1, \dots, x_n$ random inputs. In the following, we consider kernel methods, based on choosing a hypothesis space which is a separable reproducing kernel Hilbert space. The latter is a Hilbert space $\mathcal{H}$ of functions, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, such that there exists a function $K : X \times X \to \mathbb{R}$ with the following two properties: 1) for all $x \in X$, $K_x(\cdot) = K(x, \cdot)$ belongs to $\mathcal{H}$, and 2) the so called reproducing property holds: $f(x) = \langle f, K_x \rangle_{\mathcal{H}}$, for all $f \in \mathcal{H}$, $x \in X$ [25]. The function K, called reproducing kernel, is easily shown to be symmetric and positive definite, that is the kernel matrix $(K_N)_{i,j} = K(x_i, x_j)$ is positive semidefinite for all $x_1, \dots, x_N \in X$, $N \in \mathbb{N}$. A classical way to derive an empirical solution to problem (1) is to consider a Tikhonov regularization approach, based on the minimization of the penalized empirical functional,

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2, \lambda > 0. \tag{3}$$

The above approach is referred to as Kernel Regularized Least Squares (KRLS) or Kernel Ridge Regression (KRR). It is easy to see that a solution $\hat{f}_\lambda$ to problem (3) exists, it is unique and the representer theorem [1] shows that it can be written as

$$\hat{f}_\lambda(x) = \sum_{i=1}^{n} \hat{\alpha}_i K(x_i, x) \quad \text{with} \quad \hat{\alpha} = (K_n + \lambda n I)^{-1} y, \tag{4}$$

where $x_1, \dots, x_n$ are the training set points, $y = (y_1, \dots, y_n)$ and K_n is the empirical kernel matrix. Note that this result implies that we can restrict the minimization in (3) to the space,

$$\mathcal{H}_n = \{f \in \mathcal{H} \mid f = \sum_{i=1}^{n} \alpha_i K(x_i, \cdot),\ \alpha_1, \dots, \alpha_n \in \mathbb{R}\}.$$

Storing the kernel matrix K_n, and solving the linear system in (4), can become computationally unfeasible as n increases. In the following, we consider strategies to find more efficient solutions, based on the idea of replacing $\mathcal{H}_n$ with

$$\mathcal{H}_m = \{f \mid f = \sum_{i=1}^{m} \alpha_i K(\tilde{x}_i, \cdot),\ \alpha \in \mathbb{R}^m\},$$

where $m \leq n$ and $\{\tilde{x}_1, \dots, \tilde{x}_m\}$ is a subset of the input points in the training set. The solution $\hat{f}_{\lambda,m}$ of the corresponding minimization problem can now be written as,

$$\hat{f}_{\lambda,m}(x) = \sum_{i=1}^{m} \tilde{\alpha}_i K(\tilde{x}_i, x) \quad \text{with} \quad \tilde{\alpha} = (K_{nm}^\top K_{nm} + \lambda n K_{mm})^\dagger K_{nm}^\top y, \tag{5}$$

where $A^\dagger$ denotes the Moore-Penrose pseudoinverse of a matrix A, and $(K_{nm})_{ij} = K(x_i, \tilde{x}_j)$, $(K_{mm})_{kj} = K(\tilde{x}_k, \tilde{x}_j)$ with $i \in \{1, \dots, n\}$ and $j, k \in \{1, \dots, m\}$ [2]. The above approach is related to Nyström methods and different approximation strategies correspond to different ways to select the inputs subset. While our framework applies to a broader class of strategies, see Section C.1, in the following we primarily consider two techniques.

Plain Nyström. The points $\{\tilde{x}_1, \ldots, \tilde{x}_m\}$ are sampled uniformly at random without replacement from the training set.
Approximate leverage scores (ALS) Nyström. Recall that the *leverage scores* associated to the training set points $x_1, \ldots, x_n$ are

$$(l_i(t))_{i=1}^n, \quad l_i(t) = (K_n(K_n + tnI)^{-1})_{ii}, \quad i \in \{1, \ldots, n\} \tag{6}$$

for any $t > 0$, where $(K_n)_{ij} = K(x_i, x_j)$. In practice, leverage scores are onerous to compute and approximations $(\widehat{l}_i(t))_{i=1}^n$ can be considered [16,17,24] . In particular, in the following we are interested in suitable approximations defined as follows:

Definition 1 (T-approximate leverage scores)**.** *Let $(l_i(t))_{i=1}^n$ be the leverage scores associated to the training set for a given t. Let $\delta > 0$, $t_0 > 0$ and $T \geq 1$. We say that $(\widehat{l}_i(t))_{i=1}^n$ are T-approximate leverage scores with confidence δ, when with probability at least $1 - \delta$,*

$$\frac{1}{T} l_i(t) \leq \widehat{l}_i(t) \leq T l_i(t) \quad \forall i \in \{1, \ldots, n\}, t \geq t_0.$$

Given T-approximate leverage scores for $t > \lambda_0$, $\{\tilde{x}_1, \ldots, \tilde{x}_m\}$ are sampled from the training set independently with replacement, and with probability to be selected given by $P_t(i) = \widehat{l}_i(t) / \sum_j \widehat{l}_j(t)$.
In the next section, we state and discuss our main result showing that the KRLS formulation based on plain or approximate leverage scores Nyström provides optimal empirical solutions to problem (1).

3 Theoretical analysis

In this section, we state and discuss our main results. We need several assumptions. The first basic assumption is that problem (1) admits at least a solution.

Assumption 1. *There exists an $f_{\mathcal{H}} \in \mathcal{H}$ such that*

$$\mathcal{E}(f_{\mathcal{H}}) = \min_{f \in \mathcal{H}} \mathcal{E}(f).$$

Note that, while the minimizer might not be unique, our results apply to the case in which $f_{\mathcal{H}}$ is the unique minimizer with minimal norm. Also, note that the above condition is weaker than assuming the regression function in (2) to belong to $\mathcal{H}$. Finally, we note that the study of the paper can be adapted to the case in which minimizers do not exist, but the analysis is considerably more involved and left to a longer version of the paper.
The second assumption is a basic condition on the probability distribution.

Assumption 2. *Let z_x be the random variable $z_x = y - f_{\mathcal{H}}(x)$, with $x \in X$, and y distributed according to $\rho(y|x)$. Then, there exists $M, \sigma > 0$ such that $\mathbb{E}|z_x|^p \leq \frac{1}{2} p! M^{p-2} \sigma^2$ for any $p \geq 2$, almost everywhere on X.*

The above assumption is needed to control random quantities and is related to a *noise* assumption in the regression model (2). It is clearly weaker than the often considered bounded output assumption [25], and trivially verified in classification.

The last two assumptions describe the capacity (roughly speaking the *"size"*) of the hypothesis space induced by K with respect to ρ and the regularity of $f_{\mathcal{H}}$ with respect to K and ρ. To discuss them, we first need the following definition.

Definition 2 (Covariance operator and effective dimensions)**.** *We define the covariance operator as*

$$C : \mathcal{H} \to \mathcal{H}, \quad \langle f, Cg \rangle_{\mathcal{H}} = \int_X f(x)g(x)d\rho_X(x) \ , \quad \forall f, g \in \mathcal{H}.$$

Moreover, for $\lambda > 0$, we define the random variable $\mathcal{N}_x(\lambda) = \left\langle K_x, (C + \lambda I)^{-1} K_x \right\rangle_{\mathcal{H}}$ with $x \in X$ distributed according to ρ_X and let

$$\mathcal{N}(\lambda) = \mathbb{E}\mathcal{N}_x(\lambda), \qquad \mathcal{N}_\infty(\lambda) = \sup_{x \in X} \mathcal{N}_x(\lambda).$$

We add several comments. Note that C corresponds to the second moment operator, but we refer to it as the covariance operator with an abuse of terminology. Moreover, note that $\mathcal{N}(\lambda) = \mathrm{Tr}(C(C + \lambda I)^{-1})$ (see [26]). This latter quantity, called effective dimension or degrees of freedom, can be seen as a measure of the capacity of the hypothesis space. The quantity $\mathcal{N}_\infty(\lambda)$ can be seen to provide a uniform bound on the leverage scores in Eq. (6). Clearly, $\mathcal{N}(\lambda) \leq \mathcal{N}_\infty(\lambda)$ for all $\lambda > 0$.

Assumption 3. *The kernel K is measurable, C is bounded. Moreover, for all $\lambda > 0$ and a $Q > 0$,*

$$\mathcal{N}_\infty(\lambda) < \infty, \tag{7}$$
$$\mathcal{N}(\lambda) \leq Q\lambda^{-\gamma}, \quad 0 < \gamma \leq 1. \tag{8}$$

Measurability of K and boundedness of C are minimal conditions to ensure that the covariance operator is a well defined linear, continuous, self-adjoint, positive operator [25]. Condition (7) is satisfied if the kernel is bounded $\sup_{x \in X} K(x, x) = \kappa^2 < \infty$, indeed in this case $\mathcal{N}_\infty(\lambda) \leq \kappa^2/\lambda$ for all $\lambda > 0$. Conversely, it can be seen that condition (7) together with boundedness of C imply that the kernel is bounded, indeed [1]

$$\kappa^2 \leq 2\|C\|\mathcal{N}_\infty(\|C\|).$$

Boundedness of the kernel implies in particular that the operator C is trace class and allows to use tools from spectral theory. Condition (8) quantifies the capacity assumption and is related to covering/entropy number conditions (see [25] for further details). In particular, it is known that condition (8) is ensured if the eigenvalues $(\sigma_i)_i$ of C satisfy a polynomial decaying condition $\sigma_i \sim i^{-\frac{1}{\gamma}}$. Note that, since the operator C is trace class, Condition (8) always holds for $\gamma = 1$. Here, for space constraints and in the interest of clarity we restrict to such a polynomial condition, but the analysis directly applies to other conditions including exponential decay or a finite rank conditions [26]. Finally, we have the following regularity assumption.

Assumption 4. *There exists $s \geq 0$, $1 \leq R < \infty$, such that $\|C^{-s} f_{\mathcal{H}}\|_{\mathcal{H}} < R$.*

[1] If $\mathcal{N}_\infty(\lambda)$ is finite, then $\mathcal{N}_\infty(\|C\|) = \sup_{x \in X} \|(C + \|C\|I)^{-1} K_x\|^2 \geq 1/2\|C\|^{-1} \sup_{x \in X} \|K_x\|^2$, therefore $K(x, x) \leq 2\|C\|\mathcal{N}_\infty(\|C\|)$.

The above condition is fairly standard, and can be equivalently formulated in terms of classical concepts in approximation theory such as interpolation spaces [25]. Intuitively, it quantifies the degree to which $f_{\mathcal{H}}$ can be well approximated by functions in the RKHS $\mathcal{H}$ and allows to control the bias/approximation error of a learning solution. For $s = 0$, it is always satisfied. For larger s, we are assuming $f_{\mathcal{H}}$ to belong to subspaces of $\mathcal{H}$ that are the images of the fractional compact operators C^s. Such spaces contain functions which, expanded on a basis of eigenfunctions of C, have larger coefficients in correspondence to large eigenvalues. Such an assumption is natural in view of using techniques such as (4), which can be seen as a form of spectral filtering, that estimate stable solutions by discarding the contribution of small eigenvalues [27]. In the next section, we are going to quantify the quality of empirical solutions of Problem (1) obtained by schemes of the form (5), in terms of the quantities in Assumptions 2, 3, 4.

3.1 Main results

In this section, we state and discuss our main results, starting with optimal finite sample error bounds for regularized least squares based on plain and approximate leverage score based Nyström subsampling.

Theorem 1. *Under Assumptions 1, 2, 3, and 4, let* $\delta > 0$, $\nu = \min(s, 1/2)$, $p = 1+1/(2\nu+\gamma)$ *and assume*

$$n \geq 1655\kappa^2 + 223\kappa^2 \log \frac{6\kappa^2}{\delta} + \left(\frac{38p}{\|C\|} \log \frac{114\kappa^2 p}{\|C\|\delta} \right)^p .$$

Then, the following inequality holds with probability at least $1 - \delta$,

$$\mathcal{E}(\hat{f}_{\lambda,m}) - \mathcal{E}(f_{\mathcal{H}}) \leq q^2 n^{-\frac{2\nu+1}{2\nu+\gamma+1}}, \quad \text{with } q = 6R \left(2\|C\| + \frac{M\kappa}{\sqrt{\|C\|}} + \sqrt{\frac{Q\sigma^2}{\|C\|^\gamma}} \right) \log \frac{6}{\delta}, \tag{9}$$

with $\hat{f}_{\lambda,m}$ *as in* (5), $\lambda = \|C\| n^{-\frac{1}{2\nu+\gamma+1}}$ *and*

1. *for plain Nyström*
$$m \geq (67 \vee 5\mathcal{N}_\infty(\lambda)) \log \frac{12\kappa^2}{\lambda\delta};$$
2. *for ALS Nyström and* T*-approximate leverage scores with subsampling probabilities* P_λ, $t_0 \geq \frac{19\kappa^2}{n} \log \frac{12n}{\delta}$ *and*
$$m \geq (334 \vee 78T^2\mathcal{N}(\lambda)) \log \frac{48n}{\delta}.$$

We add several comments. First, the above results can be shown to be optimal in a minimax sense. Indeed, minimax lower bounds proved in [26, 28] show that the learning rate in (9) is optimal under the considered assumptions (see Thm. 2, 3 of [26], for a discussion on minimax lower bounds see Sec. 2 of [26]). Second, the obtained bounds can be compared to those obtained for other regularized learning techniques. Techniques known to achieve optimal error rates include Tikhonov regularization [26,28,29], iterative regularization by early stopping [30, 31], spectral cut-off regularization (a.k.a. principal

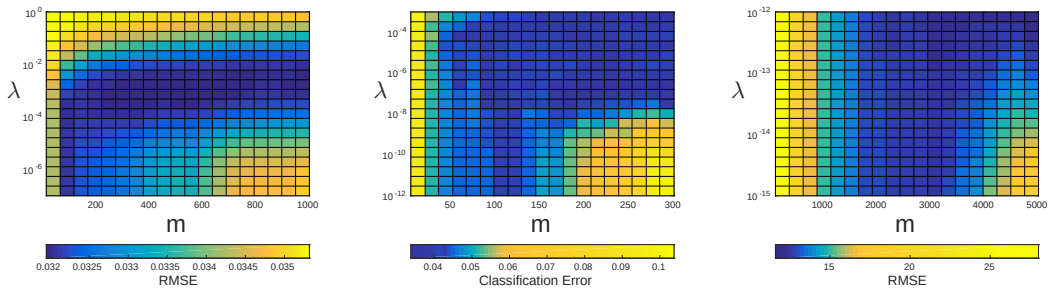

Figure 1: Validation errors associated to 20×20 grids of values for m (x axis) and λ (y axis) on `pumadyn32nh` (left), `breast cancer` (center) and `cpuSmall` (right).

component regression or truncated SVD) [30, 31], as well as regularized stochastic gradient methods [32]. All these techniques are essentially equivalent from a statistical point of view and differ only in the required computations. For example, iterative methods allow for a computation of solutions corresponding to different regularization levels which is more efficient than Tikhonov or SVD based approaches. The key observation is that all these methods have the same $O(n^2)$ memory requirement. In this view, our results show that randomized subsampling methods can break such a memory barrier, and consequently achieve much better time complexity, while preserving optimal learning guarantees. Finally, we can compare our results with previous analysis of randomized kernel methods. As already mentioned, results close to those in Theorem 1 are given in [23, 24] in a fixed design setting. Our results extend and generalize the conclusions of these papers to a general statistical learning setting. Relevant results are given in [8] for a different approach, based on averaging KRLS solutions obtained splitting the data in m groups (*divide and conquer RLS*). The analysis in [8] is only in expectation, but considers random design and shows that the proposed method is indeed optimal provided the number of splits is chosen depending on the effective dimension $\mathcal{N}(\lambda)$. This is the only other work we are aware of establishing optimal learning rates for randomized kernel approaches in a statistical learning setting. In comparison with Nyström computational regularization the main disadvantage of the divide and conquer approach is computational and in the model selection phase where solutions corresponding to different regularization parameters and number of splits usually need to be computed.
The proof of Theorem 1 is fairly technical and lengthy. It incorporates ideas from [26] and techniques developed to study spectral filtering regularization [30, 33]. In the next section, we briefly sketch some main ideas and discuss how they suggest an interesting perspective on regularization techniques including subsampling.

3.2 Proof sketch and a computational regularization perspective

A key step in the proof of Theorem 1 is an error decomposition, and corresponding bound, for any fixed λ and m. Indeed, it is proved in Theorem 2 and Proposition 2 that, for $\delta > 0$,

with probability at least $1-\delta$,

$$\left|\mathcal{E}(\hat{f}_{\lambda,m}) - \mathcal{E}(f_{\mathcal{H}})\right|^{1/2} \lesssim R\left(\frac{M\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{\sigma^2\mathcal{N}(\lambda)}{n}}\right)\log\frac{6}{\delta} + R\mathcal{C}(m)^{1/2+\nu} + R\lambda^{1/2+\nu}. \quad (10)$$

The first and last term in the right hand side of the above inequality can be seen as forms of *sample and approximation errors* [25] and are studied in Lemma 4 and Theorem 2. The mid term can be seen as a *computational error* and depends on the considered subsampling scheme. Indeed, it is shown in Proposition 2 that $\mathcal{C}(m)$ can be taken as,

$$\mathcal{C}_{\mathrm{pl}}(m) = \min\left\{t > 0 \;\middle|\; (67 \vee 5\mathcal{N}_\infty(t))\log\frac{12\kappa^2}{t\delta} \le m\right\},$$

for the plain Nyström approach, and

$$\mathcal{C}_{\mathrm{ALS}}(m) = \min\left\{\frac{19\kappa^2}{n}\log\frac{12n}{\delta} \le t \le \|C\| \;\middle|\; 78T^2\mathcal{N}(t)\log\frac{48n}{\delta} \le m\right\},$$

for the approximate leverage scores approach. The bounds in Theorem 1 follow by: 1) minimizing in λ the sum of the first and third term 2) choosing m so that the computational error is of the same order of the other terms. Computational resources and regularization are then tailored to the generalization properties of the data at hand. We add a few comments. First, note that the error bound in (10) holds for a large class of subsampling schemes, as discussed in Section C.1 in the appendix. Then specific error bounds can be derived developing computational error estimates. Second, the error bounds in Theorem 2 and Proposition 2, and hence in Theorem 1, easily generalize to a larger class of regularization schemes beyond Tikhonov approaches, namely spectral filtering [30]. For space constraints, these extensions are deferred to a longer version of the paper. Third, we note that, in practice, optimal data driven parameter choices, e.g. based on hold-out estimates [31], can be used to adaptively achieve optimal learning bounds.
Finally, we observe that a different perspective is derived starting from inequality (10), and noting that the role played by m and λ can also be exchanged. Letting m play the role of a regularization parameter, λ can be set as a function of m and m tuned adaptively. For example, in the case of a plain Nyström approach, if we set

$$\lambda = \frac{\log m}{m}, \quad \text{and} \quad m = 3n^{\frac{1}{2\nu+\gamma+1}}\log n,$$

then the obtained learning solution achieves the error bound in Eq. (9). As above, the subsampling level can also be chosen by cross-validation. Interestingly, in this case by tuning m we naturally control computational resources and regularization. An advantage of this latter parameterization is that, as described in the following, the solution corresponding to different subsampling levels is easy to update using Cholesky rank-one update formulas [34]. As discussed in the next section, in practice, a joint tuning over m and λ can be done starting from small m and appears to be advantageous both for error and computational performances.

Input: Dataset $(x_i, y_i)_{i=1}^n$, Subsampling $(\tilde{x}_j)_{j=1}^m$, Regularization Parameter λ.
Output: Nyström KRLS estimators $\{\tilde{\alpha}_1, \dots, \tilde{\alpha}_m\}$.
Compute γ_1; $R_1 \leftarrow \sqrt{\gamma_1}$;
for $t \in \{2, \dots, m\}$ **do**
  Compute A_t, u_t, v_t;
  $R_t \leftarrow \begin{pmatrix} R_{t-1} & 0 \\ 0 & 0 \end{pmatrix}$; $R_t \leftarrow \texttt{cholup}(R_t, u_t, '+')$;
  $R_t \leftarrow \texttt{cholup}(R_t, v_t, '-')$;
  $\tilde{\alpha}_t \leftarrow R_t^{-1}(R_t^{-\top}(A_t y))$;
end for

Algorithm 1: Incremental Nyström KRLS.

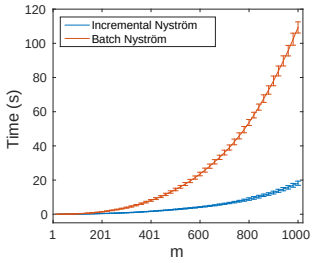

Figure 2: Model selection time on the `cpuSmall` dataset. $m \in [1, 1000]$ and $T = 50$, 10 repetitions.

4 Incremental updates and experimental analysis

In this section, we first describe an incremental strategy to efficiently explore different subsampling levels and then perform extensive empirical tests aimed in particular at: 1) investigating the statistical and computational benefits of considering varying subsampling levels, and 2) compare the performance of the algorithm with respect to state of the art solutions on several large scale benchmark datasets. Throughout this section, we only consider a plain Nyström approach, deferring to future work the analysis of leverage scores based sampling techniques. Interestingly, we will see that such a basic approach can often provide state of the art performances.

4.1 Efficient incremental updates

Algorithm 1 efficiently compute solutions corresponding to different subsampling levels, by exploiting rank-one Cholesky updates [34]. The proposed procedure allows to efficiently compute a whole regularization path of solutions, and hence perform fast model selection[2] (see Sect. A). In Algorithm 1, the function `cholup` is the Cholesky rank-one update formula available in many linear algebra libraries. The total cost of the algorithm is $O(nm^2 + m^3)$ time to compute $\tilde{\alpha}_2, \dots, \tilde{\alpha}_m$, while a naive non-incremental algorithm would require $O(nm^2T + m^3T)$ with T is the number of analyzed subsampling levels. The following are some quantities needed by the algorithm: $A_1 = a_1$ and $A_t = (A_{t-1}\ a_t) \in \mathbb{R}^{n \times t}$, for any $2 \le t \le m$. Moreover, for any $1 \le t \le m$, $g_t = \sqrt{1 + \gamma_t}$ and

$$u_t = (c_t/(1+g_t),\ g_t), \quad a_t = (K(\tilde{x}_t, x_1), \dots, K(\tilde{x}_t, x_n)), \quad c_t = A_{t-1}^\top a_t + \lambda n b_t,$$
$$v_t = (c_t/(1+g_t),\ -1), \quad b_t = (K(\tilde{x}_t, \tilde{x}_1), \dots, K(\tilde{x}_t, \tilde{x}_{t-1})), \quad \gamma_t = a_t^\top a_t + \lambda n K(\tilde{x}_t, \tilde{x}_t).$$

[2]The code for Algorithm 1 is available at `lcsl.github.io/NystromCoRe`.

4.2 Experimental analysis

We empirically study the properties of Algorithm 1, considering a Gaussian kernel of width σ. The selected datasets are already divided in a training and a test part[3]. We randomly split the training part in a training set and a validation set (80% and 20% of the n training points, respectively) for parameter tuning via cross-validation. The m subsampled points for Nyström approximation are selected uniformly at random from the training set. We report the performance of the selected model on the fixed test set, repeating the process for several trials.

Interplay between λ and m. We begin with a set of results showing that incrementally exploring different subsampling levels can yield very good performance while substantially reducing the computational requirements. We consider the `pumadyn32nh` ($n = 8192$, $d = 32$), the `breast cancer` ($n = 569$, $d = 30$), and the `cpuSmall` ($n = 8192$, $d = 12$) datasets[4]. In Figure 1, we report the validation errors associated to a 20×20 grid of values for λ and m. The λ values are logarithmically spaced, while the m values are linearly spaced. The ranges and kernel bandwidths, chosen according to preliminary tests on the data, are $\sigma = 2.66$, $\lambda \in [10^{-7}, 1]$, $m \in [10, 1000]$ for `pumadyn32nh`, $\sigma = 0.9$, $\lambda \in [10^{-12}, 10^{-3}]$, $m \in [5, 300]$ for `breast cancer`, and $\sigma = 0.1$, $\lambda \in [10^{-15}, 10^{-12}]$, $m \in [100, 5000]$ for `cpuSmall`. The main observation that can be derived from this first series of tests is that a small m is sufficient to obtain the same results achieved with the largest m. For example, for `pumadyn32nh` it is sufficient to choose $m = 62$ and $\lambda = 10^{-7}$ to obtain an average test RMSE of 0.33 over 10 trials, which is the same as the one obtained using $m = 1000$ and $\lambda = 10^{-3}$, with a 3-fold speedup of the joint training and validation phase. Also, it is interesting to observe that for given values of λ, large values of m can decrease the performance. This observation is consistent with the results in Section 3.1, showing that m can play the role of a regularization parameter. Similar results are obtained for `breast cancer`, where for $\lambda = 4.28 \times 10^{-6}$ and $m = 300$ we obtain a 1.24% average classification error on the test set over 20 trials, while for $\lambda = 10^{-12}$ and $m = 67$ we obtain 1.86%. For `cpuSmall`, with $m = 5000$ and $\lambda = 10^{-12}$ the average test RMSE over 5 trials is 12.2, while for $m = 2679$ and $\lambda = 10^{-15}$ it is only slightly higher, 13.3, but computing its associated solution requires less than half of the time and approximately half of the memory.

Regularization path computation. If the subsampling level m is used as a regularization parameter, the computation of a regularization path corresponding to different subsampling levels becomes crucial during the model selection phase. A naive approach, that consists in recomputing the solutions of Eq. 5 for each subsampling level, would require $O(m^2nT + m^3LT)$ computational time, where T is the number of solutions with different subsampling levels to be evaluated and L is the number of Tikhonov regularization parameters. On the other hand, by using the incremental Nyström algorithm the model selection time complexity is $O(m^2n + m^3L)$ for the whole regularization path. We experimentally verify this speedup on `cpuSmall` with 10 repetitions, setting $m \in [1, 5000]$ and $T = 50$. The model selection times, measured on a server with 12×2.10GHz Intel® Xeon® E5-2620 v2 CPUs and 132 GB of RAM, are reported in Figure 2. The result clearly confirms the

[3] In the following we denote by n the total number of points and by d the number of dimensions.

[4] `www.cs.toronto.edu/~delve` and `archive.ics.uci.edu/ml/datasets`

Table 1: Test RMSE comparison for exact and approximated kernel methods. The results for KRLS, Batch Nyström, RF and Fastfood are the ones reported in [6]. n_{tr} is the size of the training set.

Dataset	n_{tr}	d	*Incremental Nyström RBF*	*KRLS RBF*	*Batch Nyström RBF*	*RF RBF*	*Fastfood RBF*	*Fastfood FFT*	*KRLS Matern*	*Fastfood Matern*
Insurance Company	5822	85	$0.23180 \pm 4 \times 10^{-5}$	**0.231**	0.232	0.266	0.264	0.266	0.234	0.235
CPU	6554	21	$\mathbf{2.8466 \pm 0.0497}$	7.271	6.758	7.103	7.366	4.544	4.345	4.211
CT slices (axial)	42800	384	$\mathbf{7.1106 \pm 0.0772}$	NA	60.683	49.491	43.858	58.425	NA	14.868
Year Prediction MSD	463715	90	$\mathbf{0.10470 \pm 5 \times 10^{-5}}$	NA	0.113	0.123	0.115	0.106	NA	0.116
Forest	522910	54	0.9638 ± 0.0186	NA	**0.837**	0.840	0.840	0.838	NA	0.976

beneficial effects of incremental Nyström model selection on the computational time.

Predictive performance comparison. Finally, we consider the performance of the algorithm on several large scale benchmark datasets considered in [6], see Table 1. σ has been chosen on the basis of preliminary data analysis. m and λ have been chosen by cross-validation, starting from small subsampling values up to $m_{max} = 2048$, and considering $\lambda \in \left[10^{-12}, 1\right]$. After model selection, we retrain the best model on the entire training set and compute the RMSE on the test set. We consider 10 trials, reporting the performance mean and standard deviation. The results in Table 1 compare Nyström computational regularization with the following methods (as in [6]):

- **Kernel Regularized Least Squares (KRLS):** Not compatible with large datasets.
- **Random Fourier features (RF):** As in [4], with a number of random features $D = 2048$.
- **Fastfood RBF, FFT and Matern kernel:** As in [6], with $D = 2048$ random features.
- **Batch Nyström:** Nyström method [3] with uniform sampling and $m = 2048$.

The above results show that the proposed incremental Nyström approach behaves really well, matching state of the art predictive performances.

Acknowledgments

The work described in this paper is supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216; and by FIRB project RBFR12M3AC, funded by the Italian Ministry of Education, University and Research.

References

[1] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. MIT Press, 2002.

[2] Alex J. Smola and Bernhard Schölkopf. Sparse Greedy Matrix Approximation for Machine Learning. In *ICML*, pages 911–918. Morgan Kaufmann, 2000.

[3] Christopher Williams and Matthias Seeger. Using the Nyström Method to Speed Up Kernel Machines. In *NIPS*, pages 682–688. MIT Press, 2000.

[4] Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In *NIPS*, pages 1177–1184. Curran Associates, Inc., 2007.

[5] Jiyan Yang, Vikas Sindhwani, Haim Avron, and Michael W. Mahoney. Quasi-Monte Carlo Feature Maps for Shift-Invariant Kernels. In *ICML*, volume 32 of *JMLR Proceedings*, pages 485–493. JMLR.org, 2014.

[6] Quoc V. Le, Tamás Sarlós, and Alexander J. Smola. Fastfood - Computing Hilbert Space Expansions in loglinear time. In *ICML*, volume 28 of *JMLR Proceedings*, pages 244–252. JMLR.org, 2013.

[7] Si Si, Cho-Jui Hsieh, and Inderjit S. Dhillon. Memory Efficient Kernel Approximation. In *ICML*, volume 32 of *JMLR Proceedings*, pages 701–709. JMLR.org, 2014.

[8] Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Divide and Conquer Kernel Ridge Regression. In *COLT*, volume 30 of *JMLR Proceedings*, pages 592–617. JMLR.org, 2013.

[9] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Ensemble Nystrom Method. In *NIPS*, pages 1060–1068. Curran Associates, Inc., 2009.

[10] Mu Li, James T. Kwok, and Bao-Liang Lu. Making Large-Scale Nyström Approximation Possible. In *ICML*, pages 631–638. Omnipress, 2010.

[11] Kai Zhang, Ivor W. Tsang, and James T. Kwok. Improved Nyström Low-rank Approximation and Error Analysis. ICML, pages 1232–1239. ACM, 2008.

[12] Bo Dai, Bo Xie 0002, Niao He, Yingyu Liang, Anant Raj, Maria-Florina Balcan, and Le Song. Scalable Kernel Methods via Doubly Stochastic Gradients. In *NIPS*, pages 3041–3049, 2014.

[13] Petros Drineas and Michael W. Mahoney. On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-Based Learning. *JMLR*, 6:2153–2175, December 2005.

[14] Alex Gittens and Michael W. Mahoney. Revisiting the Nystrom method for improved large-scale machine learning. 28:567–575, 2013.

[15] Shusen Wang and Zhihua Zhang. Improving CUR Matrix Decomposition and the Nyström Approximation via Adaptive Sampling. *JMLR*, 14(1):2729–2769, 2013.

[16] Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *JMLR*, 13:3475–3506, 2012.

[17] Michael B. Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform Sampling for Matrix Approximation. In *ITCS*, pages 181–190. ACM, 2015.

[18] Shusen Wang and Zhihua Zhang. Efficient Algorithms and Error Analysis for the Modified Nystrom Method. In *AISTATS*, volume 33 of *JMLR Proceedings*, pages 996–1004. JMLR.org, 2014.

[19] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling Methods for the Nyström Method. *JMLR*, 13(1):981–1006, 2012.

[20] Corinna Cortes, Mehryar Mohri, and Ameet Talwalkar. On the Impact of Kernel Approximation on Learning Accuracy. In *AISTATS*, volume 9 of *JMLR Proceedings*, pages 113–120. JMLR.org, 2010.

[21] Rong Jin, Tianbao Yang, M. Mahdavi, Yu-Feng Li, and Zhi-Hua Zhou. Improved Bounds for the Nyström Method With Application to Kernel Classification. *Information Theory, IEEE Transactions on*, 59(10):6939–6949, Oct 2013.

[22] Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström Method vs Random Fourier Features: A Theoretical and Empirical Comparison. In *NIPS*, pages 485–493, 2012.

[23] Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *COLT*, volume 30 of *JMLR Proceedings*, pages 185–209. JMLR.org, 2013.

[24] Ahmed Alaoui and Michael W Mahoney. Fast Randomized Kernel Methods With Statistical Guarantees. *arXiv*, 2014.

[25] I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer New York, 2008.

[26] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[27] L. Lo Gerfo, Lorenzo Rosasco, Francesca Odone, Ernesto De Vito, and Alessandro Verri. Spectral Algorithms for Supervised Learning. *Neural Computation*, 20(7):1873–1897, 2008.

[28] I. Steinwart, D. R. Hush, and C. Scovel. Optimal Rates for Regularized Least Squares Regression. In *COLT*, 2009.

[29] S. Mendelson and J. Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, 2010.

[30] F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

[31] A. Caponnetto and Yuan Yao. Adaptive rates for regularization operators in learning theory. *Analysis and Applications*, 08, 2010.

[32] Y. Ying and M. Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

[33] Alessandro Rudi, Guillermo D Canas, and Lorenzo Rosasco. On the Sample Complexity of Subspace Learning. In *NIPS*, pages 2067–2075, 2013.

[34] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. 3rd edition, 1996.

A The incremental algorithm

Let $(x_i, y_i)_{i=1}^n$ be the dataset and $(\tilde{x}_i)_{i=1}^m$ be the selected Nyström points. We want to compute $\tilde{\alpha}$ of Eq. 5, incrementally in m. Towards this goal we compute an incremental Cholesky decomposition R_t for $t \in \{1, \dots, m\}$ of the matrix $G_t = K_{nt}^\top K_{nt} + \lambda n K_{tt}$, and the coefficients $\tilde{\alpha}_t$ by $\tilde{\alpha}_t = R_t^{-1} R_t^{-\top} K_{nt}^\top y$. Note that, for any $1 \leq t \leq m-1$, by assuming $G_t = R_t^\top R_t$ for an upper triangular matrix R_t, we have

$$G_{t+1} = \begin{pmatrix} G_t & c_{t+1} \\ c_{t+1}^\top & \gamma_{t+1} \end{pmatrix} = \begin{pmatrix} R_t & 0 \\ 0 & 0 \end{pmatrix}^\top \begin{pmatrix} R_t & 0 \\ 0 & 0 \end{pmatrix} + C_{t+1} \quad \text{with} \quad C_{t+1} = \begin{pmatrix} 0 & c_{t+1} \\ c_{t+1}^\top & \gamma_{t+1} \end{pmatrix},$$

and c_{t+1}, γ_{t+1} as in Section 4.1. Note moreover that $G_1 = \gamma_1$. Thus if we decompose the matrix C_{t+1} in the form $C_{t+1} = u_{t+1} u_{t+1}^\top - v_{t+1} v_{t+1}^\top$ we are able compute R_{t+1}, the Cholesky matrix of G_{t+1}, by updating a bordered version of R_t with two rank-one Cholesky updates. This is exactly Algorithm 1 with u_{t+1} and v_{t+1} as in Section 4.1. Note that the rank-one Cholesky update requires $O(t^2)$ at each call, while the computation of c_t requires $O(nt)$ and the ones of $\tilde{\alpha}_t$ requires to solve two triangular linear systems, that is $O(t^2 + nt)$. Therefore the total cost for computing $\tilde{\alpha}_2, \dots, \tilde{\alpha}_m$ is $O(nm^2 + m^3)$.

B Preliminary definitions

We begin introducing several operators that will be useful in the following. Let $z_1, \dots, z_m \in \mathcal{H}$ and for all $f \in \mathcal{H}$, $a \in \mathbb{R}^m$, let

$$\begin{aligned} Z_m &: \mathcal{H} \to \mathbb{R}^m, & Z_m f &= (\langle z_1, f \rangle_{\mathcal{H}}, \dots, \langle z_m, f \rangle_{\mathcal{H}}), \\ Z_m^* &: \mathbb{R}^m \to \mathcal{H}, & Z_m^* a &= \textstyle\sum_{i=1}^m a_i z_i. \end{aligned}$$

Let $S_n = \frac{1}{\sqrt{n}} Z_m$ and $S_n^* = \frac{1}{\sqrt{n}} Z_m^*$ the operators obtained taking $m = n$ and $z_i = K_{x_i}$, $\forall i = 1, \dots, n$ in the above definitions. Moreover, for all $f, g \in \mathcal{H}$ let

$$C_n : \mathcal{H} \to \mathcal{H}, \quad \langle f, C_n g \rangle_{\mathcal{H}} = \frac{1}{n} \sum_{i=1}^n f(x_i) g(x_i).$$

The above operators are linear and finite rank. Moreover $C_n = S_n^* S_n$ and $K_n = n S_n S_n^*$, and further $B_{nm} = \sqrt{n} S_n Z_m^* \in \mathbb{R}^{n \times m}$, $G_{mm} = Z_m Z_m^* \in \mathbb{R}^{m \times m}$ and $\tilde{K}_n = B_{nm} G_{mm}^\dagger B_{nm}^\top \in \mathbb{R}^{n \times n}$.

C Representer theorem for Nyström computational regularization and extensions

In this section we consider explicit representations of the estimator obtained via Nyström computational regularization and extensions. Indeed, we consider a general subspace $\mathcal{H}_m$ of $\mathcal{H}$, and the following problem

$$\hat{f}_{\lambda,m} = \operatorname*{argmin}_{f \in \mathcal{H}_m} \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2. \tag{11}$$

In the following lemmas, we show three different characterizations of $f_{\lambda,m}$.

Lemma 1. *Let $f_{\lambda,m}$ be the solution of the problem in Eq. (11). Then it is characterized by the following equation*

$$(P_m C_n P_m + \lambda I)\hat{f}_{\lambda,m} = P_m S_n^* \hat{y}_n, \tag{12}$$

with P_m the projection operator with range $\mathcal{H}_m$ and $\hat{y}_n = \frac{1}{\sqrt{n}} y$.

Proof. The proof proceeds in three steps. First, note that, by rewriting Problem (11) with the notation introduced in the previous section, we obtain,

$$\hat{f}_{\lambda,m} = \operatorname*{argmin}_{f \in \mathcal{H}_m} \|S_n f - \hat{y}_n\|^2 + \lambda \|f\|_{\mathcal{H}}^2. \tag{13}$$

This problem is strictly convex and coercive, therefore admits a unique solution. Second, we show that its solution coincide to the one of the following problem,

$$\hat{f}^* = \operatorname*{argmin}_{f \in \mathcal{H}} \|S_n P_m f - \hat{y}_n\|^2 + \lambda \|f\|_{\mathcal{H}}^2. \tag{14}$$

Note that the above problem is again strictly convex and coercive. To show that $\hat{f}_{\lambda,m} = \hat{f}^*$, let $\hat{f}^* = a + b$ with $a \in \mathcal{H}_m$ and $b \in \mathcal{H}_m^{\perp}$. A necessary condition for $\hat{f}^*$ to be optimal, is that $b = 0$, indeed, considering that $P_m b = 0$, we have

$$\|S_n P_m f^* - \hat{y}_n\|^2 + \lambda \|f^*\|_{\mathcal{H}}^2 = \|S_n P_m a - \hat{y}_n\|^2 + \lambda \|a\|_{\mathcal{H}}^2 + \lambda \|b\|_{\mathcal{H}}^2 \geq \|S_n P_m a - \hat{y}_n\|^2 + \lambda \|a\|_{\mathcal{H}}^2.$$

This means that $\hat{f}^* \in \mathcal{H}_m$, but on $\mathcal{H}_m$ the functionals defining Problem (13) and Problem (14) are identical because $P_m f = f$ for any $f \in \mathcal{H}_m$ and so $\hat{f}_{\lambda,m} = \hat{f}^*$. Therefore, by computing the derivative of the functional of Problem (14), we see that $\hat{f}_{\lambda,m}$ is given by Eq. (12). □

Using the above results, we can give an equivalent representations of the function $\hat{f}_{\lambda,m}$. Towards this end, let Z_m be a linear operator as in Sect. B such that the range of Z_m^* is exactly $\mathcal{H}_m$. Morever, let

$$Z_m = U \Sigma V^*$$

be the SVD of Z_m where $U : \mathbb{R}^t \to \mathbb{R}^m$, $\Sigma : \mathbb{R}^t \to \mathbb{R}^t$, $V : \mathbb{R}^t \to \mathcal{H}$, $t \leq m$ and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_t)$ with $\sigma_1 \geq \cdots \geq \sigma_t > 0$, $U^* U = I_t$ and $V^* V = I_t$. Then the orthogonal projection operator P_m is given by $P_m = VV^*$ and the range of P_m is exactly $\mathcal{H}_m$. In the following lemma we give a characterization of $\hat{f}_{\lambda,m}$ that will be useful in the proof of the main theorem.

Lemma 2. *Given the above definitions , $\hat{f}_{\lambda,m}$ can be written as*

$$\hat{f}_{\lambda,m} = V(V^* C_n V + \lambda I)^{-1} V^* S_n^* \hat{y}_n. \tag{15}$$

Proof. By Lemma 1, we know that $\hat{f}_{\lambda,m}$ is written as in Eq. (12). Now, note that $\hat{f}_{\lambda,m} = P_m \hat{f}_{\lambda,m}$ and Eq. (12) imply $(P_m C_m P_m + \lambda I) P_m \hat{f}_{\lambda,m} = P_m S_n^* \hat{y}_n$, that is equivalent to

$$V(V^* C_n V + \lambda I) V^* \hat{f}_{\lambda,m} = VV^* S_n^* \hat{y}_n,$$

by substituting P_m with VV^*. Thus by premultiplying the previous equation by V^* and dividing by $V^*C_mV + \lambda I$, we have

$$V^*\hat{f}_{\lambda,m} = (V^*C_mV + \lambda I)^{-1}V^*S_n^*\hat{y}_n.$$

Finally, by premultiplying by V,

$$\hat{f}_{\lambda,m} = P_m\hat{f}_{\lambda,m} = V(V^*C_mV + \lambda I)^{-1}V^*S_n^*\hat{y}_n.$$

□

Finally, the following result provide a characterization of the solution useful for computations.

Lemma 3 (Representer theorem for $\hat{f}_{\lambda,m}$)**.** *Given the above definitions, we have that* $\hat{f}_{\lambda,m}$ *can be written as*

$$\hat{f}_{\lambda,m}(x) = \sum_{i=1}^{m} \tilde{\alpha}_i z_i(x), \quad \textit{with } \tilde{\alpha} = (B_{nm}^\top B_{nm} + \lambda n G_{mm})^\dagger B_{nm}^\top y \qquad \forall\, x \in X. \tag{16}$$

Proof. According to the definitions of B_{nm} and G_{mm} we have that

$$\tilde{\alpha} = (B_{nm}^\top B_{nm} + \lambda n G_{mm})^\dagger B_{nm}^\top y = ((Z_mS_n^*)(S_nZ_m^*) + \lambda(Z_mZ_m^*))^\dagger(Z_mS_n^*)\hat{y}_n.$$

Moreover, according to the definition of Z_m we have

$$\hat{f}_{\lambda,m}(x) = \sum_{i=1}^{m} \tilde{\alpha}_i \langle z_i, K_x\rangle = \langle Z_mK_x, \tilde{\alpha}\rangle_{\mathbb{R}^m} = \langle K_x, Z_m^*\tilde{\alpha}\rangle_{\mathcal{H}} \quad \forall\, x \in X,$$

so that

$$\hat{f}_{\lambda,m} = Z_m^*((Z_mS_n^*)(S_nZ_m^*) + \lambda(Z_mZ_m^*))^\dagger(Z_mS_n^*)\hat{y}_n = Z_m^*(Z_mC_{n\lambda}Z_m^*)^\dagger(Z_mS_n^*)\hat{y}_n,$$

where $C_{n\lambda} = C_n + \lambda I$. Let $F = U\Sigma$, $G = V^*C_nV + \lambda I$, $H = \Sigma U^\top$, and note that F, GH, G and H are full-rank matrices, then we can perform the full-rank factorization of the pseudo-inverse (see Eq.24, Thm. 5, Chap. 1 of [1]) obtaining

$$(Z_mC_{n\lambda}Z_m^*)^\dagger = (FGH)^\dagger = H^\dagger(FG)^\dagger = H^\dagger G^{-1}F^\dagger = U\Sigma^{-1}(V^*C_nV + \lambda I)^{-1}\Sigma^{-1}U^*.$$

Finally, simplyfing U and Σ, we have

$$\begin{aligned}\hat{f}_{\lambda,m} &= Z_m^*(Z_mC_{n\lambda}Z_m^*)^\dagger(Z_mS_n^*)\hat{y}_n \\ &= V\Sigma U^*U\Sigma^{-1}(V^*C_nV + \lambda I)^{-1}\Sigma^{-1}U^*U\Sigma V^*S_n^*\hat{y}_n \\ &= V(V^*C_nV + \lambda I)^{-1}V^*S_n^*\hat{y}_n.\end{aligned}$$

□

C.1 Extensions

Inspection of the proof shows that our analysis extends beyond the class of subsampling schemes in Theorem 1. Indeed, the error decomposition Theorem 2 directly applies to a large family of approximation schemes. Several further examples are described next.

KRLS and Generalized Nyström In general we could choose an arbitrary $\mathcal{H}_m \subseteq \mathcal{H}$. Let $Z_m : \mathcal{H} \to \mathbb{R}^m$ be a linear operator such that

$$\mathcal{H}_m = \operatorname{ran} Z_m^* = \{f \mid f = Z_m^* \alpha,\ \alpha \in \mathbb{R}^m\}. \tag{17}$$

Without loss of generality, Z_m^* is expressible as $Z_m^* = (z_1, \dots, z_m)^\top$ with $z_1, \dots, z_m \in \mathcal{H}$, therefore, according to Section B and to Lemma 3, the solution of KRLS approximated with the generalized Nyström scheme is

$$\hat{f}_{\lambda,m}(x) = \sum_{i=1}^{m} \tilde{\alpha}_i z_i(x), \quad \text{with } \tilde{\alpha} = (B_{nm}^\top B_{nm} + \lambda n G_{mm})^\dagger B_{nm}^\top y \tag{18}$$

with $B_{nm} \in \mathbb{R}^{n\times m}$, $(B_{nm})_{ij} = z_j(x_i)$ and $G_{mm} \in \mathbb{R}^{m\times m}$, $(G_{mm})_{ij} = \langle z_i, z_j\rangle_{\mathcal{H}}$, or equivalently

$$\hat{f}_{\lambda,m}(x) = \sum_{i=1}^{m} \tilde{\alpha}_i z_i(x), \quad \tilde{\alpha} = G_{mm}^\dagger B_{nm}^\top (\tilde{K}_n + \lambda n I)^\dagger \hat{y}_n, \quad \tilde{K}_n = B_{nm} G_{mm}^\dagger B_{nm}^\top \tag{19}$$

The following are some examples of Generalized Nyström approximations.

Plain Nyström with various sampling schemes [2–4] For a realization $s : \mathbb{N} \to \{1, \dots, n\}$ of a given sampling scheme, we choose $Z_m = S_m$ with $S_m^* = (K_{x_{s(1)}}, \dots, K_{x_{s(m)}})^\top$ where $(x_i)_{i=1}^n$ is the training set. With such Z_m we obtain $\tilde{K}_n = K_{nm}(K_{mm})^\dagger K_{nm}^\top$ and so Eq. (18) becomes exactly Eq. (5).

Reduced rank Plain Nyström [5] Let $p \geq m$, S_p as in the previous example, the linear operator associated to p points of the dataset. Let $K_{pp} = S_p S_p^\top \in \mathbb{R}^{p\times p}$, that is $(K_{pp})_{ij} = K(x_i, x_j)$. Let $K_{pp} = \sum_{i=1}^{p} \sigma_i u_i u_i^\top$ its eigenvalue decomposition and $U_m = (u_1, \dots, u_m)$. Let $(K_{pp})_m = U_m^\top K_{pp} U_m$ be the m-rank approximation of K_{pp}. We approximate this family by choosing $Z_m = U_m^\top S_p$, indeed we obtain $\tilde{K_n} = K_{nm} U_m (U_m^\top K_{pp} U_m)^\dagger U_m^\top K_{nm}^\top = K_{nm}(K_{pp})_m^\dagger K_{nm}^\top$.

Nyström with sketching matrices [6] We cover this family by choosing $Z_m = R_m S_n$, where S_n is the same operator as in the plain Nyström case where we select all the points of the training set and R_m a $m \times n$ sketching matrix. In this way we have $\tilde{K_n} = K_n R_m^* (R_m K_n R_m^*)^\dagger R_m K_n$, that is exactly the SPSD sketching model.

D Probabilistic inequalities

In this section we collect five main probabilistic inequalities needed in the proof of the main result. We let ρ_X denote the marginal distribution of ρ on X and $\rho(\cdot|x)$ the conditional distribution on $\mathbb{R}$ given $x \in X$. Lemmas 6, 7 and especially Proposition 1 are new and of interest in their own right.

The first result is essentially taken from [7].

Lemma 4 (Sample Error)**.** *Under Assumptions 1, 2 and 3, for any $\delta > 0$, the following holds with probability $1-\delta$*

$$\|(C+\lambda I)^{-1/2}(S_n^*\widehat{y}_n - C_n f_{\mathcal{H}})\| \leq 2\left(\frac{M\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{\sigma^2\mathcal{N}(\lambda)}{n}}\right)\log\frac{2}{\delta}.$$

Proof. The proof is given in [7] for bounded kernels and the slightly stronger condition $\int(e^{\frac{|y-f_{\mathcal{H}}(x)|}{M}} - \frac{|y-f_{\mathcal{H}}(x)|}{M} - 1)d\rho(y|x) \leq \sigma^2/M^2$ in place of Assumption 2. More precisely, note that

$$(C+\lambda I)^{-1/2}(S_n^*\widehat{y}_n - C_n f_{\mathcal{H}}) = \frac{1}{n}\sum_{i=1}^n \zeta_i,$$

where $\zeta_1, \dots, \zeta_n$ are i.i.d. random variables, defined as $\zeta_i = (C+\lambda I)^{-1/2}K_{x_i}(y_i - f_{\mathcal{H}}(x_i))$. For any $1 \leq i \leq n$,

$$\begin{aligned}\mathbb{E}\zeta_i &= \int_{X\times\mathbb{R}} (C+\lambda I)^{-1/2}K_{x_i}(y_i - f_{\mathcal{H}}(x_i))d\rho(x_i,y_i)\\ &= \int_X (C+\lambda I)^{-1/2}K_{x_i}\int_{\mathbb{R}}(y_i - f_{\mathcal{H}}(x_i))d\rho(y_i|x_i)d\rho_X(x_i) = 0,\end{aligned}$$

almost everywhere by Assumption 1 (see Step 3.2 of Thm. 4 in [7]). In the same way we have

$$\begin{aligned}\mathbb{E}\|\zeta_i\|^p &= \int_{X\times\mathbb{R}} \|(C+\lambda I)^{-1/2}K_{x_i}(y_i - f_{\mathcal{H}}(x_i))\|^p d\rho(x_i,y_i)\\ &= \int_X \|(C+\lambda I)^{-1/2}K_{x_i}\|^p \int_{\mathbb{R}} |y_i - f_{\mathcal{H}}(x_i)|^p d\rho(y_i|x_i)d\rho_X(x_i)\\ &\leq \sup_{x\in X}\|(C+\lambda I)^{-1/2}K_x\|^{p-2}\int_X \|(C+\lambda I)^{-1/2}K_{x_i}\|^2\int_{\mathbb{R}}|y_i - f_{\mathcal{H}}(x_i)|^p d\rho(y_i|x_i)d\rho_X(x_i)\\ &\leq \frac{1}{2}p!\sqrt{\sigma^2\mathcal{N}(\lambda)}^2(M\sqrt{\mathcal{N}_\infty(\lambda)})^{p-2},\end{aligned}$$

where $\sup_{x\in X}\|(C+\lambda I)^{-1/2}K_x\| = \sqrt{\mathcal{N}_\infty(\lambda)}$ and $\int_X\|(C+\lambda I)^{-1/2}K_{x_i}\|^2 = \mathcal{N}(\lambda)$ by Assumption 3, while the bound on the moments of $y - f(x)$ is given in Assumption 2. Finally, to concentrate the sum of random vectors, we apply Prop. 11. □

The next result is taken from [8].

Lemma 5. *Under Assumption 3, for any $\delta \geq 0$ and $\frac{9\kappa^2}{n}\log\frac{n}{\delta} \leq \lambda \leq \|C\|$, the following inequality holds with probability at least $1-\delta$,*

$$\|(C_n+\lambda I)^{-1/2}C^{1/2}\| \leq \|(C_n+\lambda I)^{-1/2}(C+\lambda I)^{1/2}\| \leq 2.$$

Proof. Lemma 7 of [8] gives an the extended version of the above result. Our bound on λ is scaled by κ^2 because in [8] it is assumed $\kappa \leq 1$. □

Lemma 6 (plain Nyström approximation). *Under Assumption 3, let J be a partition of $\{1,\dots,n\}$ chosen uniformly at random from the partitions of cardinality m. Let $\lambda > 0$, for any $\delta > 0$, such that $m \geq 67 \log \frac{4\kappa^2}{\lambda\delta} \vee 5\mathcal{N}_\infty(\lambda) \log \frac{4\kappa^2}{\lambda\delta}$, the following holds with probability $1-\delta$*

$$\|(I - P_m)C^{1/2}\|^2 \leq 3\lambda,$$

where P_m is the projection operator on the subspace $\mathcal{H}_m = \operatorname{span}\{K_{x_j} \mid j \in J\}$.

Proof. Define the linear operator $C_m : \mathcal{H} \to \mathcal{H}$, as $C_m = \frac{1}{m}\sum_{j\in J} K_{x_j} \otimes K_{x_j}$. Now note that the range of C_m is exactly $\mathcal{H}_m$. Therefore, by applying Prop. 3 and 7, we have that

$$\|(I - P_m)C_\lambda^{1/2}\|^2 \leq \lambda\|(C_m + \lambda I)^{-1/2}C^{1/2}\|^2 \leq \frac{\lambda}{1-\beta(\lambda)},$$

with $\beta(\lambda) = \lambda_{\max}\left(C_\lambda^{-1/2}(C - C_m)C_\lambda^{-1/2}\right)$. To upperbound $\frac{\lambda}{1-\beta(\lambda)}$ we need an upperbound for $\beta(\lambda)$. Considering that, given the partition J, the random variables $\zeta_j = K_{x_j} \otimes K_{x_j}$ are i.i.d., then we can apply Prop. 8, to obtain

$$\beta(\lambda) \leq \frac{2w}{3m} + \sqrt{\frac{2w\mathcal{N}_\infty(\lambda)}{m}},$$

where $w = \log \frac{4\operatorname{Tr}(C)}{\lambda\delta}$ with probability $1-\delta$. Thus, by choosing $m \geq 67w \vee 5\mathcal{N}_\infty(\lambda)w$, we have that $\beta(\lambda) \leq 2/3$, that is

$$\|(I - P_m)C_\lambda^{1/2}\|^2 \leq 3\lambda.$$

Finally, note that by definition $\operatorname{Tr}(C) \leq \kappa^2$. □

Lemma 7 (Nyström approximation for ALS selection method). *Let $(\hat{l}_i(t))_{i=1}^n$ be the collection of approximate leverage scores. Let $\lambda > 0$ and P_λ be defined as $P_\lambda(i) = \hat{l}_i(\lambda)/\sum_{j\in N}\hat{l}_j(\lambda)$ for any $i \in N$ with $N = \{1,\dots,n\}$. Let $\mathcal{I} = (i_1,\dots,i_m)$ be a collection of indices independently sampled with replacement from N according to the probability distribution P_λ. Let P_m be the projection operator on the subspace $\mathcal{H}_m = \operatorname{span}\{K_{x_j} | j \in J\}$ and J be the subcollection of $\mathcal{I}$ with all the duplicates removed. Under Assumption 3, for any $\delta > 0$ the following holds with probability $1-2\delta$*

$$\|(I - P_m)(C + \lambda I)^{1/2}\| \leq 3\lambda,$$

when the following conditions are satisfied:

1. *there exists a $T \geq 1$ and a $\lambda_0 > 0$ such that $(\hat{l}_i(t))_{i=1}^n$ are T-approximate leverage scores for any $t \geq \lambda_0$ (see Def. 1),*
2. *$n \geq 1655\kappa^2 + 223\kappa^2 \log \frac{2\kappa^2}{\delta}$,*
3. *$\lambda_0 \vee \frac{19\kappa^2}{n}\log\frac{2n}{\delta} \leq \lambda \leq \|C\|$,*
4. *$m \geq 334 \log \frac{8n}{\delta} \vee 78T^2\mathcal{N}(\lambda)\log\frac{8n}{\delta}$.*

Proof. Define $\tau = \delta/4$. Next, define the diagonal matrix $H \in \mathbb{R}^{n\times n}$ with $(H)_{ii} = 0$ when $P_\lambda(i) = 0$ and $(H)_{ii} = \frac{nq(i)}{mP_\lambda(i)}$ when $P_\lambda(i) > 0$, where $q(i)$ is the number of times the index i is present in the collection $\mathcal{I}$. We have that

$$S_n^* H S_n = \frac{1}{m}\sum_{i=1}^{n} \frac{q(i)}{P_\lambda(i)} K_{x_i} \otimes K_{x_i} = \frac{1}{m}\sum_{j\in J} \frac{q(j)}{P_\lambda(j)} K_{x_j} \otimes K_{x_j}.$$

Now, considering that $\frac{q(j)}{P_\lambda(j)} > 0$ for any $j \in J$, thus $\operatorname{ran} S_n^* H S_n = \mathcal{H}_m$. Therefore, by using Prop. 3 and 7, we exploit the fact that the range of P_m is the same of $S_n^* H S_n$, to obtain

$$\|(I - P_m)(C + \lambda I)^{1/2}\|^2 \le \lambda \|(S_n^* H S_n + \lambda I)^{-1/2} C^{1/2}\|^2 \le \frac{\lambda}{1-\beta(\lambda)},$$

with $\beta(\lambda) = \lambda_{\max}\left(C_\lambda^{-1/2}(C - S_n^* H S_n)C_\lambda^{-1/2}\right)$. Considering that the function $(1-x)^{-1}$ is increasing on $-\infty < x < 1$, in order to bound $\lambda/(1-\beta(\lambda))$ we need an upperbound for $\beta(\lambda)$. Here we split $\beta(\lambda)$ in the following way,

$$\beta(\lambda) \le \underbrace{\lambda_{\max}\left(C_\lambda^{-1/2}(C - C_n)C_\lambda^{-1/2}\right)}_{\beta_1(\lambda)} + \underbrace{\lambda_{\max}\left(C_\lambda^{-1/2}(C_n - S_n^* H S_n)C_\lambda^{-1/2}\right)}_{\beta_2(\lambda)}.$$

Considering that C_n is the linear combination of independent random vectors, for the first term we can apply Prop. 8, obtaining a bound of the form

$$\beta_1(\lambda) \le \frac{2w}{3n} + \sqrt{\frac{2w\kappa^2}{\lambda n}},$$

with probability $1-\tau$, where $w = \log\frac{4\kappa^2}{\lambda\tau}$ (we used the fact that $\mathcal{N}_\infty(\lambda) \le \kappa^2/\lambda$). Then, after dividing and multiplying by $C_{n\lambda}^{1/2}$, we split the second term $\beta_2(\lambda)$ as follows:

$$\begin{aligned}
\beta_2(\lambda) &\le \|C_\lambda^{-1/2}(C_n - S_n^* H S_n)C_\lambda^{-1/2}\| \\
&\le \|C_\lambda^{-1/2} C_{n\lambda}^{1/2} C_{n\lambda}^{-1/2}(C_n - S_n^* H S_n)C_{n\lambda}^{-1/2} C_{n\lambda}^{1/2} C_\lambda^{-1/2}\| \\
&\le \|C_\lambda^{-1/2} C_{n\lambda}^{1/2}\|^2 \|C_{n\lambda}^{-1/2}(C_n - S_n^* H S_n)C_{n\lambda}^{-1/2}\|.
\end{aligned}$$

Let

$$\beta_3(\lambda) = \|C_{n\lambda}^{-1/2}(C_n - S_n^* H S_n)C_{n\lambda}^{-1/2}\| = \|C_{n\lambda}^{-1/2} S_n^*(I - H)S_n C_{n\lambda}^{-1/2}\|. \tag{20}$$

Note that $S_n C_{n\lambda}^{-1} S_n^* = K_n(K_n + \lambda n I)^{-1}$ indeed $C_{n\lambda}^{-1} = (S_n^* S_n + \lambda I)^{-1}$ and $K_n = n S_n S_n^*$. Therefore we have

$$S_n C_{n\lambda}^{-1} S_n^* = S_n(S_n^* S_n + \lambda I)^{-1} S_n^* = (S_n S_n^* + \lambda I)^{-1} S_n S_n^* = (K_n + \lambda n I)^{-1} K_n.$$

Thus, if we let $U\Sigma U^\top$ be the eigendecomposition of K_n, we have that $(K_n + \lambda n I)^{-1} K_n = U(\Sigma + \lambda n I)^{-1}\Sigma U^\top$ and thus $S_n C_{n\lambda}^{-1} S_n^* = U(\Sigma + \lambda n I)^{-1}\Sigma U^\top$. In particular this implies that $S_n C_{n\lambda}^{-1} S_n^* = U Q_n^{1/2} Q_n^{1/2} U^\top$ with $Q_n = (\Sigma + \lambda n I)^{-1}\Sigma$. Therefore we have

$$\beta_3(\lambda) = \|C_{n\lambda}^{-1/2} S_n^*(I - H)S_n C_{n\lambda}^{-1/2}\| = \|Q_n^{1/2} U^\top (I - H) U Q_n^{1/2}\|,$$

where we used twice the fact that $\|ABA^*\| = \|(A^*A)^{1/2}B(A^*A)^{1/2}\|$ for any bounded linear operators A, B.

Consider the matrix $A = Q_n^{1/2}U^\top$ and let a_i be the i-th column of A, and e_i be the i-th canonical basis vector for each $i \in \mathbb{N}$. We prove that $\|a_i\|^2 = l_i(\lambda)$, the true leverage score, since

$$\|a_i\|^2 = \|Q_n^{1/2}U^\top e_i\|^2 = e_i^\top U Q_n U^\top e_i = ((K_n + \lambda n I)^{-1}K_n)_{ii} = l_i(\lambda).$$

Noting that $\sum_{k=1}^n \frac{q(k)}{P_\lambda(k)} a_k a_k^\top = \sum_{i=\mathfrak{I}} \frac{1}{P_\lambda(i)} a_i a_i^\top$, we have

$$\beta_3(\lambda) = \|AA^\top - \frac{1}{m}\sum_{i\in\mathfrak{I}} \frac{1}{P_\lambda(i)} a_i a_i^\top\|.$$

Moreover, by the T-approximation property of the approximate leverage scores (see Def. 1), we have that for all $i \in \{1, \dots, n\}$, when $\lambda \geq \lambda_0$, the following holds with probability $1-\delta$

$$P_\lambda(i) = \frac{\hat{l}_i(\lambda)}{\sum_j \hat{l}_j(\lambda)} \geq T^{-2}\frac{l_i(\lambda)}{\sum_j l_j(\lambda)} = T^{-2}\frac{\|a_i\|^2}{\operatorname{Tr} AA^\top}.$$

Then, we can apply Prop. 9, so that, after a union bound, we obtain the following inequality with probability $1-\delta-\tau$:

$$\beta_3(\lambda) \leq \frac{2\|A\|^2 \log\frac{2n}{\tau}}{3m} + \sqrt{\frac{2\|A\|^2 T^2 \operatorname{Tr} AA^\top \log\frac{2n}{\tau}}{m}} \leq \frac{2\log\frac{2n}{\tau}}{3m} + \sqrt{\frac{2T^2\widehat{\mathcal{N}}(\lambda)\log\frac{2n}{\tau}}{m}},$$

where the last step follows from $\|A\|^2 = \|(K_n + \lambda n I)^{-1}K_n\| \leq 1$ and $\operatorname{Tr}(AA^\top) = \operatorname{Tr}(C_{n\lambda}^{-1}C_n) := \widehat{\mathcal{N}}(\lambda)$. Applying Proposition 1, we have that $\widehat{\mathcal{N}}(\lambda) \leq 1.3\mathcal{N}(\lambda)$ with probability $1-\tau$, when $\frac{19\kappa^2}{n}\log\frac{n}{4\tau} \leq \lambda \leq \|C\|$ and $n \geq 405\kappa^2 \vee 67\kappa^2\log\frac{\kappa^2}{2\tau}$. Thus, by taking a union bound again, we have

$$\beta_3(\lambda) \leq \frac{2\log\frac{2n}{\tau}}{3m} + \sqrt{\frac{5.3T^2\mathcal{N}(\lambda)\log\frac{2n}{\tau}}{m}},$$

with probability $1-2\tau-\delta$ when $\lambda_0 \vee \frac{19\kappa^2}{n}\log\frac{n}{\delta} \leq \lambda \leq \|C\|$ and $n \geq 405\kappa^2 \vee 67\kappa^2\log\frac{2\kappa^2}{\delta}$. The last step is to bound $\|C_\lambda^{-1/2}C_{n\lambda}^{1/2}\|^2$, as follows

$$\|C_\lambda^{-1/2}C_{n\lambda}^{1/2}\|^2 = \|C_\lambda^{-1/2}C_{n\lambda}C_\lambda^{-1/2}\| = \|I + C_\lambda^{-1/2}(C_n - C)C_\lambda^{-1/2}\| \leq 1 + \eta,$$

with $\eta = \|C_\lambda^{-1/2}(C_n - C)C_\lambda^{-1/2}\|$. Note that, by applying Prop. 8 we have that $\eta \leq \frac{2(\kappa^2+\lambda)\theta}{3\lambda n} + \sqrt{\frac{2\kappa^2\theta}{3\lambda n}}$ with probability $1-\tau$ and $\theta = \log\frac{8\kappa^2}{\lambda\tau}$. Finally, by collecting the above results and taking a union bound we have

$$\beta(\lambda) \leq \frac{2w}{3n} + \sqrt{\frac{2w\kappa^2}{\lambda n}} + (1+\eta)\left(\frac{2\log\frac{2n}{\tau}}{3m} + \sqrt{\frac{5.3T^2\mathcal{N}(\lambda)\log\frac{2n}{\tau}}{m}}\right),$$

with probability $1-4\tau-\delta = 1-2\delta$ when $\lambda_0 \vee \frac{19\kappa^2}{n}\log\frac{n}{\delta} \leq \lambda \leq \|C\|$ and $n \geq 405\kappa^2 \vee 67\kappa^2\log\frac{2\kappa^2}{\delta}$. Note that, if we select $n \geq 405\kappa^2 \vee 223\kappa^2\log\frac{2\kappa^2}{\delta}$, $m \geq 334\log\frac{8n}{\delta}$, $\lambda_0 \vee$

$\frac{19\kappa^2}{n}\log\frac{2n}{\delta} \le \lambda \le \|C\|$ and $\frac{78T^2\mathcal{N}(\lambda)\log\frac{8n}{\delta}}{m} \le 1$ the conditions are satisfied and we have $\beta(\lambda) \le 2/3$, so that

$$\|(I - P_m)C^{1/2}\|^2 \le 3\lambda,$$

with probability $1 - 2\delta$. □

Proposition 1 (Empirical Effective Dimension). *Let $\hat{\mathcal{N}}(\lambda) = \operatorname{Tr} C_n C_{n\lambda}^{-1}$. Under the Assumption 3, for any $\delta > 0$ and $n \ge 405\kappa^2 \vee 67\kappa^2 \log\frac{6\kappa^2}{\delta}$, if $\frac{19\kappa^2}{n}\log\frac{n}{4\delta} \le \lambda \le \|C\|$, then the following holds with probability $1 - \delta$,*

$$\frac{|\hat{\mathcal{N}}(\lambda) - \mathcal{N}(\lambda)|}{\mathcal{N}(\lambda)} \le 4.5q + (1 + 9q)\sqrt{\frac{3q}{\mathcal{N}(\lambda)}} + \frac{q + 13.5q^2}{\mathcal{N}(\lambda)} \le 1.65,$$

with $q = \frac{4\kappa^2 \log\frac{6}{\delta}}{3\lambda n}$.

Proof. Let $\tau = \delta/3$. Define $B_n = C_\lambda^{-1/2}(C - C_n)C_\lambda^{-1/2}$. Choosing λ in the range $\frac{19\kappa^2}{n}\log\frac{n}{4\tau} \le \lambda \le \|C\|$, Prop. 8 assures that $\lambda_{\max}(B_n) \le 1/3$ with probability $1 - \tau$. Then, using the fact that $C_{n\lambda}^{-1} = C_\lambda^{-1/2}(I - B_n)^{-1}C_\lambda^{-1/2}$ (see the proof of Prop. 7) we have

$$\begin{aligned}
|\hat{\mathcal{N}}(\lambda) - \mathcal{N}(\lambda)| &= |\operatorname{Tr} C_{n\lambda}^{-1}C_n - CC_\lambda^{-1} = \lambda \operatorname{Tr} C_{n\lambda}^{-1}(C_n - C)C_\lambda^{-1}| \\
&= |\lambda \operatorname{Tr} C_\lambda^{-1/2}\,(I - B_n)^{-1}\,C_\lambda^{-1/2}(C_n - C)C_\lambda^{-1/2}C_\lambda^{-1/2}| \\
&= |\lambda \operatorname{Tr} C_\lambda^{-1/2}\,(I - B_n)^{-1}\,B_n C_\lambda^{-1/2}|.
\end{aligned}$$

Considering that for any symmetric linear operator $X : \mathcal{H} \to \mathcal{H}$ the following identity holds

$$(I - X)^{-1}X = X + X(I - X)^{-1}X,$$

when $\lambda_{\max}(X) \le 1$, we have

$$\begin{aligned}
\lambda|\operatorname{Tr} C_\lambda^{-1/2}\,(I - B_n)^{-1}\,B_n C_\lambda^{-1/2}| \le{}& \underbrace{\lambda|\operatorname{Tr} C_\lambda^{-1/2}B_n C_\lambda^{-1/2}|}_{A} \\
&+ \underbrace{\lambda|\operatorname{Tr} C_\lambda^{-1/2}B_n\,(I - B_n)^{-1}\,B_n C_\lambda^{-1/2}|}_{B}.
\end{aligned}$$

To find an upperbound for A define the i.i.d. random variables $\eta_i = \left\langle K_{x_i}, \lambda C_\lambda^{-2}K_{x_i}\right\rangle \in \mathbb{R}$ with $i \in \{1, \dots, n\}$. By linearity of the trace and the expectation, we have $M = \mathbb{E}\eta_1 = \mathbb{E}\left\langle K_{x_i}, \lambda C_\lambda^{-2}K_{x_i}\right\rangle = \mathbb{E}\operatorname{Tr}(\lambda C_\lambda^{-2}K_{x_1} \otimes K_{x_1}) = \lambda \operatorname{Tr}(C_\lambda^{-2}C)$. Therefore,

$$\lambda|\operatorname{Tr} C_\lambda^{-1/2}B_n C_\lambda^{-1/2}| = \left|M - \frac{1}{n}\sum_{i=1}^{n}\eta_i\right|,$$

and we can apply the Bernstein inequality (Prop. 10) with

$$|M - \eta_1| \le \lambda\|C_\lambda^{-2}\|\|K_{x_1}\|^2 + M \le \frac{\kappa^2}{\lambda} + M \le \frac{2\kappa^2}{\lambda} = L,$$

$$\mathbb{E}(\eta_1 - M)^2 = \mathbb{E}\eta_1^2 - M^2 \leq \mathbb{E}\eta_1^2 \leq LM = \sigma^2.$$

An upperbound for M is $M = \operatorname{Tr}(\lambda C_\lambda^{-2} C) = \operatorname{Tr}((I - C_\lambda^{-1} C) C_\lambda^{-1} C) \leq \mathcal{N}(\lambda)$. Thus, we have

$$\lambda |\operatorname{Tr} C_\lambda^{-1/2} B_n C_\lambda^{-1/2}| \leq \frac{4\kappa^2 \log \frac{2}{\tau}}{3\lambda n} + \sqrt{\frac{4\kappa^2 \mathcal{N}(\lambda) \log \frac{2}{\tau}}{\lambda n}},$$

with probability $1 - \tau$.

To find an upperbound for B, let $\mathcal{L}$ be the space of Hilbert-Schmidt operators on $\mathcal{H}$. $\mathcal{L}$ is a Hilbert space with scalar product $\langle U, V \rangle_{HS} = \operatorname{Tr}(UV^*)$ for all $U, V \in \mathcal{L}$. Next, note that $B = \|Q\|_{HS}^2$ where $Q = \lambda^{1/2} C_\lambda^{-1/2} B_n (I - B_n)^{-1/2}$, moreover

$$\|Q\|_{HS}^2 \leq \|\lambda^{1/2} C_\lambda^{-1/2}\|^2 \|B_n\|_{HS}^2 \|(I - B_n)^{-1/2}\|^2 \leq 1.5 \|B_n\|_{HS}^2,$$

since $\|(I - B_n)^{-1/2}\|^2 = (1 - \lambda_{\max}(B_n))^{-1} \leq 3/2$ and $(1 - \sigma)^{-1}$ is increasing and positive on $[-\infty, 1)$.

To find a bound for $\|B_n\|_{HS}$ consider that $B_n = T - \frac{1}{n} \sum_{i=1}^n \zeta_i$ where ζ_i are i.i.d. random operators defined as $\zeta_i = C_\lambda^{-1/2}(K_{x_i} \otimes K_{x_i}) C_\lambda^{-1/2} \in \mathcal{L}$ for all $i \in \{1, \ldots, n\}$, and $T = \mathbb{E}\zeta_1 = C_\lambda^{-1} C \in \mathcal{L}$. Then we can apply the Bernstein's inequality for random vectors on a Hilbert space (Prop. 11), with the following L and σ^2:

$$\|T - \zeta_1\|_{HS} \leq \|C_\lambda^{-1/2}\|^2 \|K_{x_1}\|_{\mathcal{H}}^2 + \|T\|_{HS} \leq \frac{\kappa^2}{\lambda} + \|T\|_{HS} \leq \frac{2\kappa^2}{\lambda} = L,$$

$$\mathbb{E}\|\zeta_1 - T\|^2 = \mathbb{E} \operatorname{Tr}(\zeta_1^2 - T^2) \leq \mathbb{E} \operatorname{Tr}(\zeta_1^2) \leq L \mathbb{E} \operatorname{Tr}(\zeta_1) = \sigma^2,$$

where $\|T\|_{HS} \leq \mathbb{E} \operatorname{Tr}(\zeta_1) = \mathcal{N}(\lambda)$, obtaining

$$\|B_n\|_{HS} \leq \frac{4\kappa^2 \log \frac{2}{\tau}}{\lambda n} + \sqrt{\frac{4\kappa^2 \mathcal{N}(\lambda) \log \frac{2}{\tau}}{\lambda n}},$$

with probability $1 - \tau$. Then, by taking a union bound for the three events we have

$$|\widehat{\mathcal{N}}(\lambda) - \mathcal{N}(\lambda)| \leq q + \sqrt{3q\mathcal{N}(\lambda)} + 1.5 \left(3q + \sqrt{3q\mathcal{N}(\lambda)}\right)^2,$$

with $q = \frac{4\kappa^2 \log \frac{6}{\delta}}{3\lambda n}$, and with probability $1 - \delta$. Finally, if the second assumption on λ holds, then we have $q \leq 4/57$. Noting that $n \geq 405\kappa^2$, and that $\mathcal{N}(\lambda) \geq \|C C_\lambda^{-1}\| = \frac{\|C\|}{\|C\| + \lambda} \geq 1/2$, we have that

$$|\widehat{\mathcal{N}}(\lambda) - \mathcal{N}(\lambda)| \leq \left(\frac{q}{3\mathcal{N}(\lambda)} + \sqrt{\frac{q}{\mathcal{N}(\lambda)}} + 1.5 \left(\frac{q}{\sqrt{\mathcal{N}(\lambda)}} + \sqrt{q} \right)^2 \right) \mathcal{N}(\lambda) \leq 1.65 \mathcal{N}(\lambda).$$

□

E Proofs of main theorem

A key step to derive the proof of Theorem 1 is the error decomposition given by the following theorem, together with the probabilistic inequalities in the previous section.

Theorem 2 (Error decomposition for KRLS+Ny). *Under Assumptions 1, 3, 4, let* $\nu = \min(s, 1/2)$ *and* $\hat{f}_{\lambda,m}$ *a KRLS + generalized Nyström solution as in Eq.* (18). *Then for any* $\lambda, m > 0$ *the error is bounded by*

$$\left|\mathcal{E}(\hat{f}_{\lambda,m}) - \mathcal{E}(f_{\mathcal{H}})\right|^{1/2} \leq q\Big(\underbrace{\mathcal{S}(\lambda, n)}_{\text{Sample error}} + \underbrace{\mathcal{C}(m)^{1/2+\nu}}_{\text{Computational error}} + \underbrace{\lambda^{1/2+\nu}}_{\text{Approximation error}} \Big), \tag{21}$$

where $\mathcal{S}(\lambda, n) = \|(C + \lambda I)^{-1/2}(S_n^* \widehat{y}_n - C_n f_{\mathcal{H}})\|$ *and* $\mathcal{C}(m) = \|(I - P_m)(C + \lambda I)^{1/2}\|^2$ *with* $P_m = Z_m^*(Z_m Z_m^*)^{\dagger} Z_m$. *Moreover* $q = R(\beta^2 \vee (1 + \theta\beta))$, $\beta = \|(C_n + \lambda I)^{-1/2}(C + \lambda I)^{1/2}\|$, $\theta = \|(C_n + \lambda I)^{1/2}(C + \lambda I)^{-1/2}\|$.

Proof. Let $C_\lambda = C + \lambda I$ and $C_{n\lambda} = C_n + \lambda I$ for any $\lambda > 0$. Let $\hat{f}_{\lambda,m}$ as in Eq. (18). By Lemma 1, Lemma 2 and Lemma 3 we know that $\hat{f}_{\lambda,m}$ is characterized by $\hat{f}_{\lambda,m} = g_{\lambda m}(C_n) S_n^* \widehat{y}_n$ with $g_{\lambda,m}(C_n) = V(V^* C_n V + \lambda I)^{-1} V^*$. By using the fact that $\mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}) = \|C^{1/2}(f - f_{\mathcal{H}})\|_{\mathcal{H}}^2$ for any $f \in \mathcal{H}$ (see Prop. 1 Point 3 of [7]), we have

$$\begin{aligned}
|\mathcal{E}(\hat{f}_{\lambda,m}) - \mathcal{E}(f_{\mathcal{H}})|^{1/2} &= \|C^{1/2}(\hat{f}_{\lambda,m} - f_{\mathcal{H}})\|_{\mathcal{H}} = \|C^{1/2}(g_{\lambda,m}(C_n) S_n^* \widehat{y}_n - f_{\mathcal{H}})\|_{\mathcal{H}} \\
&= \|C^{1/2}(g_{\lambda,m}(C_n) S_n^*(\widehat{y}_n - S_n f_{\mathcal{H}} + S_n f_{\mathcal{H}}) - f_{\mathcal{H}})\|_{\mathcal{H}} \\
&\leq \underbrace{\|C^{1/2} g_{\lambda,m}(C_n) S_n^*(\widehat{y}_n - S_n f_{\mathcal{H}})\|_{\mathcal{H}}}_{A} + \underbrace{\|C^{1/2}(I - g_{\lambda,m}(C_n) C_n) f_{\mathcal{H}}\|_{\mathcal{H}}}_{B}.
\end{aligned}$$

Bound for the term A Multiplying and dividing by $C_{n\lambda}^{1/2}$ and $C_\lambda^{1/2}$ we have

$$A \leq \|C^{1/2} C_{n\lambda}^{-1/2}\| \|C_{n\lambda}^{1/2} g_{\lambda,m}(C_n) C_{n\lambda}^{1/2}\| \|C_{n\lambda}^{-1/2} C_\lambda^{1/2}\| \|C_\lambda^{-1/2} S_n^*(\widehat{y}_n - S_n f_{\mathcal{H}})\|_{\mathcal{H}} \leq \beta^2 \mathcal{S}(\lambda, n),$$

where the last step is due to Lemma 8 and the fact that

$$\|C^{1/2} C_{n\lambda}^{-1/2}\| \leq \|C^{1/2} C_\lambda^{-1/2}\| \|C_\lambda^{1/2} C_{n\lambda}^{-1/2}\| \leq \|C_\lambda^{1/2} C_{n\lambda}^{-1/2}\|.$$

Bound for the term B Noting that $g_{\lambda,m}(C_n) C_{n\lambda} V V^* = V V^*$, we have

$$\begin{aligned}
I - g_{\lambda,m}(C_n) C_n &= I - g_{\lambda,m}(C_n) C_{n\lambda} + \lambda g_{\lambda,m}(C_n) \\
&= I - g_{\lambda,m}(C_n) C_{n\lambda} V V^* - g_{\lambda,m}(C_n) C_{n\lambda}(I - V V^*) + \lambda g_{\lambda,m}(C_n) \\
&= (I - V V^*) + \lambda g_{\lambda,m}(C_n) - g_{\lambda,m}(C_n) C_{n\lambda}(I - V V^*).
\end{aligned}$$

Therefore, noting that by Ass. 4 we have $\|C_\lambda^{-\nu} f_{\mathcal{H}}\|_{\mathcal{H}} \leq \|C_\lambda^{-s} f_{\mathcal{H}}\|_{\mathcal{H}} \leq \|C^{-s} f_{\mathcal{H}}\|_{\mathcal{H}} \leq R$, then, by reasoning as in A, we have

$$\begin{aligned}
B &\leq \|C^{1/2}(I - g_{\lambda,m}(C_n) C_n) C_\lambda^{\nu}\| \|C_\lambda^{-\nu} f_{\mathcal{H}}\|_{\mathcal{H}} \\
&\leq R\|C^{1/2} C_\lambda^{-1/2}\| \|C_\lambda^{1/2}(I - V V^*) C_\lambda^{\nu}\| + R\lambda \|C^{1/2} C_{n\lambda}^{-1/2}\| \|C_{n\lambda}^{1/2} g_{\lambda,m}(C_n) C_\lambda^{\nu}\| \\
&\quad + R\|C^{1/2} C_{n\lambda}^{-1/2}\| \|C_{n\lambda}^{1/2} g_{\lambda,m}(C_n) C_{n\lambda}^{1/2}\| \|C_{n\lambda}^{1/2} C_\lambda^{-1/2}\| \|C_\lambda^{1/2}(I - V V^*) C_\lambda^{\nu}\| \\
&\leq R(1 + \beta\theta) \underbrace{\|C_\lambda^{1/2}(I - V V^*) C_\lambda^{\nu}\|}_{B.1} + R\beta \underbrace{\lambda \|C_{n\lambda}^{1/2} g_{\lambda,m}(C_n) C_\lambda^{\nu}\|}_{B.2},
\end{aligned}$$

where in the second step we applied the decomposition of $I - g_{\lambda m}(C_n)C_n$.

Bound for the term B.1 Since VV^* is a projection operator, we have that $(I - VV^*) = (I - VV^*)^s$, for any $s > 0$, therefore

$$B.1 = \|C_\lambda^{1/2}(I - VV^*)^2 C_\lambda^\nu\| \leq \|C_\lambda^{1/2}(I - VV^*)\|\|(I - VV^*)C_\lambda^\nu\|.$$

By applying Cordes inequality (Prop. 4) to $\|(I - VV^*)C_\lambda^\nu\|$ we have,

$$\|(I - VV^*)C_\lambda^\nu\| = \|(I - VV^*)^{2\nu} C_\lambda^{\frac{1}{2}2\nu}\| = \|(I - VV^*)C_\lambda^{1/2}\|^{2\nu}.$$

Bound for the term B.2 We have

$$\begin{aligned}
B.2 &\leq \lambda\|C_{n\lambda}^{1/2} g_{\lambda,m}(C_n) C_{n\lambda}^\nu\|\|C_{n\lambda}^{-\nu} C_\lambda^\nu\| \\
&\leq \lambda\|C_{n\lambda}^{1/2} g_{\lambda,m}(C_n) C_{n\lambda}^\nu\|\|C_{n\lambda}^{-1/2} C_\lambda^{1/2}\|^{2\nu} \\
&\leq \beta^{2\nu}\lambda\|(V^* C_{n\lambda} V)^{1/2}(V^* C_{n\lambda} V)^{-1}(V^* C_{n\lambda} V)^\nu\| \\
&= \beta^{2\nu}\lambda\|(V^* C_n V + \lambda I)^{-(1/2-\nu)}\| \leq \beta\lambda^{1/2+\nu},
\end{aligned}$$

where the first step is obtained multipling and dividing by $C_{n\lambda}^\nu$, the second step by applying Cordes inequality (see Prop. 4), the third step by Prop. 6. □

Proposition 2 (Bounds for plain and ALS Nyström). *For any $\delta > 0$, let $n \geq 1655\kappa^2 + 223\kappa^2 \log\frac{6\kappa^2}{\delta}$, let $\frac{19\kappa^2}{n}\log\frac{6n}{\delta} \leq \lambda \leq \|C\|$ and define*

$$\begin{aligned}
\mathcal{C}_{pl}(m) &= \min\left\{ t > 0 \;\middle|\; (67 \vee 5\mathcal{N}_\infty(t)) \log\frac{12\kappa^2}{t\delta} \leq m \right\}, \\
\mathcal{C}_{ALS}(m) &= \min\left\{ \frac{19\kappa^2}{n}\log\frac{12n}{\delta} \leq t \leq \|C\| \;\middle|\; 78T^2\mathcal{N}(t)\log\frac{48n}{\delta} \leq m \right\}.
\end{aligned}$$

Under the assumptions of Thm. 2 and Assumption 2, 3, if one of the following two conditions hold

1. *plain Nyström is used,*
2. *ALS Nyström is used with*
   (a) *T-approximate leverage scores, for any $t \geq \frac{19\kappa^2}{n}\log\frac{12n}{\delta}$ (see Def. 1),*
   (b) *resampling probabilities P_t where $t = \mathcal{C}_{ALS}(m)$ (see Sect. 2),*
   (c) *$m \geq 334\log\frac{48n}{\delta}$,*

then the following holds with probability $1 - \delta$

$$\left|\mathcal{E}(\hat{f}_{\lambda,m}) - \mathcal{E}(f_{\mathcal{H}})\right|^{1/2} \leq 6R\left(\frac{M\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{\sigma^2\mathcal{N}(\lambda)}{n}}\right)\log\frac{6}{\delta} + 3R\mathcal{C}(m)^{1/2+\nu} + 3R\lambda^{1/2+\nu} \tag{22}$$

where $\mathcal{C}(m) = \mathcal{C}_{pl}(m)$ in case of plain Nyström and $\mathcal{C}(m) = \mathcal{C}_{ALS}(m)$ in case of ALS Nyström.

Proof. In order to get explicit bounds from Thm. 2, we have to control four quantities that are $\beta, \theta, \mathcal{S}(\lambda, n)$ and $\mathcal{C}(m)$. In the following we bound such quantities in probability and then take a union bound. Let $\tau = \delta/3$. We can control both β and θ, by bounding $b(\lambda) = \|C_\lambda^{-1/2}(C_n - C)C_\lambda^{-1/2}\|$. Indeed, by Prop. 7, we have that $\beta \leq 1/(1 - b(\lambda))$, while

$$\theta^2 = \|C_\lambda^{-1/2} C_{n\lambda} C_\lambda^{-1/2}\| = \|I + C_\lambda^{-1/2}(C_n - C)C_\lambda^{-1/2}\| \leq 1 + b(\lambda).$$

Exploiting Prop. 8, with the fact that $\mathcal{N}(\lambda) \leq \mathcal{N}_\infty(\lambda) \leq \frac{\kappa^2}{\lambda}$ and $\operatorname{Tr} C \leq \kappa^2$, we have that $b(\lambda) \leq \frac{2(\kappa^2+\lambda)w}{3\lambda n} + \sqrt{\frac{2w\kappa^2}{\lambda n}}$ for $w = \log \frac{4\kappa^2}{\tau\lambda}$ with probability $1-\tau$. Simple computations show that with n and λ as in the statement of this corollary, we have $b(\lambda) \leq 1/3$. Therefore $\beta \leq 1.5$, while $\theta \leq 1.16$ and $q = R(\beta^2 \vee (1 + \theta\beta)) < 2.75R$ with probability $1 - \tau$. Next, we bound $\mathcal{S}(\lambda, n)$. Here we exploit Lemma 4 which gives, with probability $1 - \tau$,

$$\mathcal{S}(\lambda, n) \leq 2\left(\frac{M\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{\sigma^2\mathcal{N}(\lambda)}{n}}\right)\log\frac{2}{\tau}.$$

To bound $\mathcal{C}(m)$ for plain Nyström, Lemma 6 gives $\mathcal{C}(m) \leq 3t$ with probability $1 - \tau$, for a $t > 0$ such that $(67 \vee 5\mathcal{N}_\infty(t)) \log \frac{4\kappa^2}{t\tau} \leq m$. In particular, we choose $t = \mathcal{C}_{\mathrm{pl}}(m)$ to satisfy the condition. Next we bound $\mathcal{C}(m)$ for ALS Nyström. Using Lemma 7 with $\lambda_0 = \frac{19\kappa^2}{n}\log\frac{2n}{\tau}$, we have $\mathcal{C}(m) \leq 3t$ with probability $1 - \tau$ under some conditions on t, m, n, on the approximate leverage scores and on the resampling probability. Here again the requirement on n is satisfied by the hypotesis on n of this proposition, while the condition on the approximate leverage scores and on the resampling probabilities are satisfied by conditions (a), (b) of this proposition. The remaining two conditions are $\frac{19\kappa^2}{n}\log\frac{4n}{\tau} \leq t \leq \|C\|$ and $(334 \vee 78T^2\mathcal{N}(t))\log\frac{16n}{\tau} \leq m$. They are satisfied by choosing $t = \mathcal{C}_{\mathrm{ALS}}(m)$ and by assuming that $m \geq 334\log\frac{16n}{\tau}$. Finally, the proposition is obtained by substituting each of the four quantities $\beta, \theta, \mathcal{S}(\lambda, n), \mathcal{C}(m)$ with the corresponding upperbounds in Eq. (21), and by taking the union bounds on the associated events. □

Proof of Theorem 1. By exploiting the results of Prop. 2, obtained from the error decomposition of Thm. 2 we have that

$$\left|\mathcal{E}(\hat{f}_{\lambda,m}) - \mathcal{E}(f_{\mathcal{H}})\right|^{1/2} \leq 6R\left(\frac{M\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{\sigma^2\mathcal{N}(\lambda)}{n}}\right)\log\frac{6}{\delta} + 3R\mathcal{C}(m)^{1/2+\nu} + 3R\lambda^{1/2+\nu} \tag{23}$$

with probability $1 - \delta$, under conditions on λ, m, n, on the resampling probabilities and on the approximate leverage scores. The last is satisfied by condition (a) in this theorem. The conditions on λ, n are $n \geq 1655\kappa^2 + 223\kappa^2\log\frac{6\kappa^2}{\delta}$ and $\frac{19\kappa^2}{n}\log\frac{12n}{\delta} \leq \lambda \leq \|C\|$. If we assume that $n \geq 1655\kappa^2 + 223\kappa^2\log\frac{6\kappa^2}{\delta} + \left(\frac{38p}{\|C\|}\log\frac{114\kappa^2 p}{\|C\|\delta}\right)^p$ we satisfy the condition on n and at the same time we are sure that $\lambda = \|C\|n^{-1/(2\nu+\gamma+1)}$ satisfies the condition on λ. In the plain Nyström case, if we assume that $m \geq 67\log\frac{12\kappa^2}{\lambda\delta} + 5\mathcal{N}_\infty(\lambda)\log\frac{12\kappa^2}{\lambda\delta}$, then $\mathcal{C}(m) = \mathcal{C}_{\mathrm{pl}}(m) \leq \lambda$. In the ALS Nyström case, if we assume that $m \geq (334 \vee 78T^2\mathcal{N}(\lambda))\log\frac{48n}{\delta}$ the condition on m is satisfied, then $\mathcal{C}(m) = \mathcal{C}_{\mathrm{ALS}}(m) \leq \lambda$, moreover the conditions on the

resampling probabilities is satisfied by condition (b) of this theorem. Therefore, by setting $\lambda = \|C\| n^{-1/(2\nu+\gamma+1)}$ in Eq. (23) and considering that $\mathcal{N}_\infty(\lambda) \leq \kappa^2 \lambda^{-1}$ we easily obtain the result of this theorem. □

The following lemma is a technical result needed in the error decomposition (Thm. 2).

Lemma 8. *For any $\lambda > 0$, let V be such that $V^*V = I$ and C_n be a positive self-adjoint operator. Then, the following holds,*

$$\|(C_n + \lambda I)^{1/2} V (V^* C_n V + \lambda I)^{-1} V^* (C_n + \lambda I)^{1/2}\| \leq 1.$$

Proof. Let $C_{n\lambda} = C_n + \lambda I$ and $g_{\lambda m}(C_n) = V(V^* C_n V + \lambda I)^{-1} V^*$, then

$$\begin{aligned}
\|C_{n\lambda}^{1/2} g_{\lambda m}(C_n) C_{n\lambda}^{1/2}\|^2 &= \|C_{n\lambda}^{1/2} g_{\lambda m}(C_n) C_{n\lambda} g_{\lambda m}(C_n) C_{n\lambda}^{1/2}\| \\
&= \|C_{n\lambda}^{1/2} V (V^* C_{n\lambda} V)^{-1} (V^* C_{n\lambda} V)(V^* C_{n\lambda} V)^{-1} V^* C_{n\lambda}^{1/2}\| \\
&= \|C_{n\lambda}^{1/2} g_{\lambda m}(C_n) C_{n\lambda}^{1/2}\|,
\end{aligned}$$

and therefore the only possible values for $\|C_{n\lambda}^{1/2} g_{\lambda m}(C_n) C_{n\lambda}^{1/2}\|$ are 0 or 1. □

F Auxiliary results

Proposition 3. *Let $\mathcal{H}, \mathcal{K}, \mathcal{F}$ three separable Hilbert spaces, let $Z : \mathcal{H} \to \mathcal{K}$ be a bounded linear operator and let W be a projection operator on $\mathcal{H}$ such that $\operatorname{ran} P = \overline{\operatorname{ran} Z^*}$. Then for any bounded linear operator $F : \mathcal{F} \to \mathcal{H}$ and any $\lambda > 0$ we have*

$$\|(I - P)X\| \leq \lambda^{1/2} \|(Z^* Z + \lambda I)^{-1/2} X\|.$$

Proof. First of all note that $\lambda(Z^*Z + \lambda I)^{-1} = I - Z^*(ZZ^* + \lambda I)^{-1} Z$, that $Z = ZP$ and that $\|Z^*(ZZ^* + \lambda I)^{-1} Z\| \leq 1$ for any $\lambda > 0$. Then for any $v \in \mathcal{H}$ we have

$$\begin{aligned}
\left\langle v, Z^*(ZZ^* + \lambda I)^{-1} Z v \right\rangle = \left\langle v, PZ^*(ZZ^* + \lambda I)^{-1} ZP v \right\rangle &= \|(ZZ^* + \lambda I)^{-1/2} ZP v\|^2 \\
&\leq \|(ZZ^* + \lambda I)^{-1/2} Z\|^2 \|Pv\|^2 \leq \|Pv\|^2 = \langle v, Pv \rangle
\end{aligned}$$

therefore $P - Z^*(ZZ^* + \lambda I)^{-1} Z$ is a positive operator, and $(I - Z^*(ZZ^* + \lambda I)^{-1} Z) - (I - P)$ too. Now we can apply Prop. 5. □

Proposition 4 (Cordes Inequality [9])**.** *Let A, B two positive semidefinite bounded linear operators on a separable Hilbert space. Then*

$$\|A^s B^s\| \leq \|AB\|^s \quad \textit{when } 0 \leq s \leq 1.$$

Proposition 5. *Let $\mathcal{H}, \mathcal{K}, \mathcal{F}, \mathcal{G}$ be three separable Hilbert spaces and let $X : \mathcal{H} \to \mathcal{K}$ and $Y : \mathcal{H} \to \mathcal{F}$ be two bounded linear operators. For any bounded linear operator $Z : \mathcal{G} \to \mathcal{H}$, if $Y^*Y - X^*X$ is a positive self-adjoint operator then $\|XZ\| \leq \|YZ\|$.*

Proof. If $Y^*Y - X^*X$ is a positive operator then $Z^*(Y^*Y - X^*X)Z$ is positive too. Thus for all $f \in \mathcal{H}$ we have that $\langle f, (Q-P)f\rangle \geq 0$, where $Q = Z^*Y^*YZ$ and $P = Z^*X^*XZ$. Thus, by linearity of the inner product, we have

$$\|Q\| = \sup_{f\in\mathcal{G}} \langle f, Qf\rangle = \sup_{f\in\mathcal{G}} \{\langle f, Pf\rangle + \langle f, (Q-P)f\rangle\} \geq \sup_{f\in\mathcal{G}} \langle f, Pf\rangle = \|P\|.$$

□

Proposition 6. *Let $\mathcal{H}, \mathcal{K}$ be two separable Hilbert spaces, let $A : \mathcal{H} \to \mathcal{H}$ be a positive linear operator, $V : \mathcal{H} \to \mathcal{K}$ a partial isometry and $B : \mathcal{K} \to \mathcal{K}$ a bounded operator. Then $\|A^rVBV^*A^s\| \leq \|(V^*AV)^rB(V^*AV)^s\|$, for all $0 \leq r, s \leq 1/2$.*

Proof. By Hansen's inequality (see [10]) we know that $(V^*AV)^{2t} - V^*A^{2t}V$ is positive selfadjoint operator for any $0 \leq t \leq 1/2$, therefore we can apply Prop. 5 two times, obtaining

$$\|A^rV(BV^*A^s)\| \leq \|(V^*AV)^r(BV^*A^s)\| = \|((V^*AV)^rB)V^*A^s\| \leq \|((V^*AV)^rB)(V^*AV)^s\|.$$

□

Proposition 7. *Let $\mathcal{H}$ be a separable Hilbert space, let A, B two bounded self-adjoint positive linear operators and $\lambda > 0$. Then*

$$\|(A+\lambda I)^{-1/2}B^{1/2}\| \leq (1-\beta)^{-1/2}$$

when

$$\beta = \lambda_{\max}\left[(B+\lambda I)^{-1/2}(B-A)(B+\lambda I)^{-1/2}\right] < 1.$$

Proof. Let $B_\lambda = B + \lambda I$. Note that

$$\begin{aligned}(A+\lambda I)^{-1} &= [(B+\lambda I) - (B-A)]^{-1} \\ &= \left[B_\lambda^{1/2}\left(I - B_\lambda^{-1/2}(B-A)B_\lambda^{-1/2}\right)B_\lambda^{1/2}\right]^{-1} \\ &= B_\lambda^{-1/2}\left[I - B_\lambda^{-1/2}(B-A)B_\lambda^{-1/2}\right]^{-1}B_\lambda^{-1/2}.\end{aligned}$$

Now let $X = (I - B_\lambda^{-1/2}(B-A)B_\lambda^{-1/2})^{-1}$. We have that,

$$\begin{aligned}\|(A+\lambda I)^{-1/2}B^{1/2}\| &= \|B^{1/2}(A+\lambda I)^{-1}B^{1/2}\|^{1/2} \\ &= \|B^{1/2}B_\lambda^{-1/2}XB_\lambda^{-1/2}B^{1/2}\|^{1/2} \\ &= \|X^{1/2}B_\lambda^{-1/2}B^{1/2}\|,\end{aligned}$$

because $\|Z\| = \|Z^*Z\|^{1/2}$ for any bounded operator Z. Note that

$$\|X^{1/2}B_\lambda^{-1/2}B^{1/2}\| \leq \|X\|^{1/2}\|B_\lambda^{-1/2}B^{1/2}\| \leq \|X\|^{1/2}.$$

Finally let $Y = B_\lambda^{-1/2}(B-A)B_\lambda^{-1/2}$ and assume that $\lambda_{\max}(Y) < 1$, then

$$\|X\| = \|(I-Y)^{-1}\| = (1-\lambda_{\max}(Y))^{-1},$$

since $X = w(Y)$ with $w(\sigma) = (1-\sigma)^{-1}$ for $-\infty \leq \sigma < 1$, and w is positive and monotonically increasing on the domain. □

G Tail bounds

Let $\|\cdot\|_{HS}$ denote the Hilbert-Schmidt norm.

Proposition 8. *Let $v_1, \dots, v_n$ with $n \geq 1$, be independent and identically distributed random vectors on a separable Hilbert spaces $\mathcal{H}$ such that $Q = \mathbb{E}\, v \otimes v$ exists, is trace class, and for any $\lambda > 0$ there exists a constant $\mathcal{N}_\infty(\lambda) < \infty$ such that $\langle v, (Q+\lambda I)^{-1} v\rangle \leq \mathcal{N}_\infty(\lambda)$ almost everywhere. Let $Q_n = \frac{1}{n}\sum_{i=1}^n v_i \otimes v_i$ and take $0 < \lambda \leq \|Q\|$. Then for any $\delta \geq 0$, the following holds*

$$\|(Q+\lambda I)^{-1/2}(Q-Q_n)(Q+\lambda I)^{-1/2}\| \leq \frac{2\beta(1+\mathcal{N}_\infty(\lambda))}{3n} + \sqrt{\frac{2\beta\mathcal{N}_\infty(\lambda)}{n}}$$

with probability $1-2\delta$. Here $\beta = \log \frac{4\operatorname{Tr} Q}{\lambda\delta}$. Moreover it holds that

$$\lambda_{\max}\left((Q+\lambda I)^{-1/2}(Q-Q_n)(Q+\lambda I)^{-1/2}\right) \leq \frac{2\beta}{3n} + \sqrt{\frac{2\beta\mathcal{N}_\infty(\lambda)}{n}}$$

with probability $1-\delta$.

Proof. Let $Q_\lambda = Q + \lambda I$. Here we apply Prop. 12 on the random variables $Z_i = M - Q_\lambda^{-1/2} v_i \otimes Q_\lambda^{-1/2} v_i$ with $M = Q_\lambda^{-1/2} Q Q_\lambda^{-1/2}$ for $1 \leq i \leq n$. Note that the expectation of Z_i is 0. The random vectors are bounded by

$$\|Q_\lambda^{-1/2} Q Q_\lambda^{-1/2} - Q_\lambda^{-1/2} v_i \otimes Q_\lambda^{-1/2} v_i\| \leq \left\langle v, Q_\lambda^{-1} v\right\rangle + \|Q_\lambda^{-1/2} Q Q_\lambda^{-1/2}\| \leq \mathcal{N}_\infty(\lambda) + 1$$

and the second orded moment is

$$\begin{aligned}\mathbb{E}(Z_1)^2 &= \mathbb{E}\ \left\langle v_1, Q_\lambda^{-1} v_1\right\rangle\ Q_\lambda^{-1/2} v_1 \otimes Q_\lambda^{-1/2} v_1 \quad - \quad Q_\lambda^{-2} Q^2 \\ &\leq \mathcal{N}_\infty(\lambda)\mathbb{E}\, Q_\lambda^{-1/2} v_1 \otimes Q_\lambda^{-1/2} v_1 = \mathcal{N}_\infty(\lambda) Q = S.\end{aligned}$$

Now we can apply Prop. 12. Now some considerations on β. It is $\beta = \log \frac{4\operatorname{Tr} S}{\|S\|\delta} = \frac{4\operatorname{Tr} Q_\lambda^{-1} Q}{\|Q_\lambda^{-1} Q\|\delta}$, now $\operatorname{Tr} Q_\lambda^{-1} Q \leq \frac{1}{\lambda}\operatorname{Tr} Q$. We need a lowerbound for $\|Q_\lambda^{-1} Q\| = \frac{\sigma_1}{\sigma_1+\lambda}$ where $\sigma_1 = \|Q\|$ is the biggest eigenvalue of Q, now $\lambda \leq \sigma_1$ thus $\frac{\operatorname{Tr} Q}{\lambda\delta}$.

For the second bound of this proposition, the analysis remains the same except for L, indeed

$$\sup_{f\in\mathcal{H}} \langle f, Z_1 f\rangle = \sup_{f\in\mathcal{H}} \left\langle f, Q_\lambda^{-1} Q f\right\rangle - \left\langle f, Q_\lambda^{-1/2} v_i\right\rangle^2 \leq \sup_{f\in\mathcal{H}} \left\langle f, Q_\lambda^{-1} Q f\right\rangle \leq 1.$$

□

Remark 1. *In Prop. 8, let define $\kappa^2 = \inf_{\lambda>0} \mathcal{N}_\infty(\lambda)(\|Q\|+\lambda)$. When $n \geq 405\kappa^2 \vee 67\kappa^2 \log \frac{\kappa^2}{2\delta}$ and $\frac{9\kappa^2}{n}\log\frac{n}{2\delta} \leq \lambda \leq \|Q\|$ we have that*

$$\lambda_{\max}\left((Q+\lambda I)^{-1/2}(Q-Q_n)(Q+\lambda I)^{-1/2}\right) \leq \frac{1}{2},$$

with probability $1-\delta$, while it is less than $1/3$ with the same probability, if $\frac{19\kappa^2}{n}\log\frac{n}{4\delta} \leq \lambda \leq \|Q\|$.

Proposition 9 (Theorem 2 [11]. Approximation of matrix products.). *Let n, m be positive integers. Consider a matrix $A \in \mathbb{R}^{n\times m}$ and denote by a_i the i-th column of A. Let $m \leq n$ and $I = \{i_1, \dots, i_m\}$ be a subset of $N = \{1, \dots, n\}$ formed by m elements chosen randomly with replacement, according to a distribution that associates the probability $P(i)$ to the element $i \in N$. Assume that there exists a $\beta \in (0, 1]$ such that the probabilities $P(1), \dots, P(n)$ satisfy $P(i) \geq \beta \frac{\|a_i\|^2}{\operatorname{Tr} AA^\top}$ for all $i \in N$. For any $\delta > 0$ the following holds*

$$\|AA^\top - \frac{1}{m}\sum_{i\in I}\frac{1}{P(i)}a_i a_i^\top\| \leq \frac{2L\log\frac{2n}{\delta}}{3m} + \sqrt{\frac{2LS\log\frac{2n}{\delta}}{m}}$$

with probability $1-\delta$. Here $L = \|A\|^2$ and $S = \frac{1}{\beta}\operatorname{Tr} AA^\top$.

Proposition 10 (Bernstein's inequality for sum of random variables). *Let $x_1, \dots, x_n$ be a sequence of independent and identically distributed random variables on $\mathbb{R}$ with zero mean. If there exists an $L, S \in \mathbb{R}$ such that $x_1 \leq L$ almost everywhere and $\mathbb{E}x_1^2 \leq S$, then for any $\delta > 0$ the following holds with probability $1-\delta$:*

$$\frac{1}{n}\sum_{i=1}^{n} x_i \leq \frac{2L\log\frac{1}{\delta}}{3n} + \sqrt{\frac{2S\log\frac{1}{\delta}}{n}}.$$

If there exists an $L' \geq |x_1|$ almost everywhere, then the same bound, computed with L' instead of L, holds for the for the absolute value of the left hand side, with probability $1-2\delta$.

Proof. It is a restatement of Theorem 3 of [12]. □

Proposition 11 (Bernstein's inequality for sum of random vectors). *Let $z_1, \dots, z_n$ be a sequence of independent identically distributed random vectors on a separable Hilbert space $\mathcal{H}$. Assume $\mu = \mathbb{E}z_1$ exists and let $\sigma, M \geq 0$ such that*

$$\mathbb{E}\|z_1 - \mu\|_{\mathcal{H}}^p \leq \frac{1}{2}p!\sigma^2 L^{p-2}$$

for all $p \geq 2$. Then for any $\tau \geq 0$:

$$\|\frac{1}{n}\sum_{i=1}^{n} z_i - \mu\|_{\mathcal{H}} \leq \frac{2L\log\frac{2}{\delta}}{n} + \sqrt{\frac{2\sigma^2\log\frac{2}{\delta}}{n}}$$

with probability greater or equal $1-\delta$.

Proof. restatement of Theorem 3.3.4 of [13]. □

Proposition 12 (Bernstein's inequality for sum of random operators). *Let $\mathcal{H}$ be a separable Hilbert space and let $X_1, \dots, X_n$ be a sequence of independent and identically distributed self-adjoint positive random operators on $\mathcal{H}$. Assume that there exists $\mathbb{E}X_1 = 0$ and $\lambda_{\max}(X_1) \leq L$ almost everywhere for some $L > 0$. Let S be a positive operator such that $\mathbb{E}(X_1)^2 \leq S$. Then for any $\delta \geq 0$ the following holds*

$$\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \leq \frac{2L\beta}{3n} + \sqrt{\frac{2\|S\|\beta}{n}}$$

with probability $1 - \delta$. *Here* $\beta = \log \frac{2 \operatorname{Tr} S}{\|S\| \delta}$.

If there exists an L' *such that* $L' \geq \|X_1\|$ *almost everywhere, then the same bound, computed with* L' *instead of* L*, holds for the operatorial norm with probability* $1 - 2\delta$.

Proof. The theorem is a restatement of Theorem 7.3.1 of [14] generalized to the separable Hilbert space case by means of the technique in Section 4 of [15]. □

References

[1] Adi Ben-Israel and Thomas NE Greville. *Generalized inverses: theory and applications*. Springer, 2003.

[2] Alex J. Smola and Bernhard Schölkopf. Sparse Greedy Matrix Approximation for Machine Learning. In *ICML*, pages 911–918. Morgan Kaufmann, 2000.

[3] Christopher Williams and Matthias Seeger. Using the Nyström Method to Speed Up Kernel Machines. In *NIPS*, pages 682–688. MIT Press, 2000.

[4] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling Methods for the Nyström Method. *JMLR*, 13(1):981–1006, 2012.

[5] Petros Drineas and Michael W. Mahoney. On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-Based Learning. *JMLR*, 6:2153–2175, December 2005.

[6] Alex Gittens and Michael W. Mahoney. Revisiting the Nystrom method for improved large-scale machine learning. 28:567–575, 2013.

[7] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[8] Alessandro Rudi, Guillermo D Canas, and Lorenzo Rosasco. On the Sample Complexity of Subspace Learning. In *NIPS*, pages 2067–2075, 2013.

[9] Junichi Fujii, Masatoshi Fujii, Takayuki Furuta, and Ritsuo Nakamoto. Norm inequalities equivalent to Heinz inequality. *Proceedings of the American Mathematical Society*, 118(3), 1993.

[10] F. Hansen. An operator inequality. *Mathematische Annalen*, 246(3), 1980.

[11] Ahmed Alaoui and Michael W Mahoney. Fast Randomized Kernel Methods With Statistical Guarantees. *arXiv*, 2014.

[12] Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Advanced Lectures on Machine Learning*. 2004.

[13] Vadim Vladimirovich Yurinsky. *Sums and Gaussian vectors*. 1995.

[14] Joel A Tropp. User-Friendly Tools for Random Matrices: An Introduction. 2012.

[15] Stanislav Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *arXiv*, 2011.