FALKON: An Optimal Large Scale Kernel Method

Alessandro Rudi *
INRIA – Sierra Project-team,
École Normale Supérieure, Paris

Luigi Carratino
University of Genoa
Genova, Italy

Lorenzo Rosasco
University of Genoa,
LCSL, IIT & MIT

Abstract

Kernel methods provide a principled way to perform non linear, nonparametric learning. They rely on solid functional analytic foundations and enjoy optimal statistical properties. However, at least in their basic form, they have limited applicability in large scale scenarios because of stringent computational requirements in terms of time and especially memory. In this paper, we take a substantial step in scaling up kernel methods, proposing FALKON, a novel algorithm that allows to efficiently process millions of points. FALKON is derived combining several algorithmic principles, namely stochastic subsampling, iterative solvers and preconditioning. Our theoretical analysis shows that optimal statistical accuracy is achieved requiring essentially $O(n)$ memory and $O(n\sqrt{n})$ time. An extensive experimental analysis on large scale datasets shows that, even with a single machine, FALKON outperforms previous state of the art solutions, which exploit parallel/distributed architectures.

1   Introduction

The goal in supervised learning is to learn from examples a function that predicts well new data. Nonparametric methods are often crucial since the functions to be learned can be non-linear and complex Kernel methods are probably the most popular among nonparametric learning methods, but despite excellent theoretical properties, they have limited applications in large scale learning because of time and memory requirements, typically at least quadratic in the number of data points.
Overcoming these limitations has motivated a variety of practical approaches including gradient methods, as well accelerated, stochastic and preconditioned extensions, to improve time complexity [1, 2, 3, 4, 5, 6]. Random projections provide an approach to reduce memory requirements, popular methods including Nyström [7, 8], random features [9], and their numerous extensions. From a theoretical perspective a key question has become to characterize statistical and computational trade-offs, that is if, or under which conditions, computational gains come at the expense of statistical accuracy. In particular, recent results considering least squares, show that there are large class of problems for which, by combining Nyström or random features approaches [10, 11, 12, 13, 14, 15] with ridge regression, it is possible to substantially reduce computations, while preserving the same optimal statistical accuracy of exact kernel ridge regression (KRR). While statistical lower bounds exist for this setting, there are no corresponding computational lower bounds. The state of the art approximation of KRR, for which optimal statistical bounds are known, typically requires complexities that are roughly $O(n^2)$ in time and memory (or possibly $O(n)$ in memory, if kernel computations are made on the fly).

In this paper, we propose and study FALKON, a new algorithm that, to the best of our knowledge, has the best known theoretical guarantees. At the same time FALKON provides an efficient approach to apply kernel methods on millions of points, and tested on a variety of large scale problems

---

*E-mail: `alessandro.rudi@inria.fr`. This work was done when A.R. was working at Laboratory of Computational and Statistical Learning (Istituto Italiano di Tecnologia).

outperform previously proposed methods while utilizing only a fraction of computational resources. More precisely, we take a substantial step in provably reducing the computational requirements, showing that, up to logarithmic factors, a time/memory complexity of $O(n\sqrt{n})$ and $O(n)$ is sufficient for optimal statistical accuracy. Our new algorithm, exploits the idea of using Nyström methods to approximate the KRR problem, but also to efficiently compute a preconditioning to be used in conjugate gradient. To the best of our knowledge this is the first time all these ideas are combined and put to fruition. Our theoretical analysis derives optimal statistical rates both in a basic setting and under benign conditions for which fast rates are possible. The potential benefits of different sampling strategies are also analyzed. Most importantly, the empirical performances are thoroughly tested on available large scale data-sets. Our results show that, even on a single machine, FALKON can outperforms state of the art methods on most problems both in terms of time efficiency and prediction accuracy. In particular, our results suggest that FALKON could be a viable kernel alternative to deep fully connected neural networks for large scale problems.

The rest of the paper is organized as follows. In Sect. 2 we give some background on kernel methods. In Sect. 3 we introduce FALKON, while in Sect. 4 we present and discuss the main technical results. Finally in Sect. 5 we present experimental results.

2 Statistical and Computational Trade-offs in Kernel Methods

We consider the supervised learning problem of estimating a function from random noisy samples. In statistical learning theory, this can be formalized as the problem of solving

$$\inf_{f \in \mathcal{H}} \mathcal{E}(f), \qquad \mathcal{E}(f) = \int (f(x) - y)^2 d\rho(x, y), \tag{1}$$

given samples $(x_i, y_i)_{i=1}^n$ from ρ, which is fixed but unknown and where, $\mathcal{H}$ is a space of candidate solutions. Ideally, a good empirical solution $\widehat{f}$ should have small *excess risk*

$$\mathcal{R}(\widehat{f}) = \mathcal{E}(\widehat{f}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f), \tag{2}$$

since this implies it will generalize/predict well new data. In this paper, we are interested in both computational and statistical aspects of the above problem. In particular, we investigate the computational resources needed to achieve optimal statistical accuracy, i.e. minimal excess risk. Our focus is on the most popular class of nonparametric methods, namely kernel methods.

Kernel methods and ridge regression. Kernel methods consider a space $\mathcal{H}$ of functions

$$f(x) = \sum_{i=1}^n \alpha_j K(x, x_i), \tag{3}$$

where K is a positive definite kernel [2]. The coefficients $\alpha_1, \ldots, \alpha_n$ are typically derived from a convex optimization problem, that for the square loss is

$$\widehat{f}_{n,\lambda} = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2, \tag{4}$$

and defines the so called kernel ridge regression (KRR) estimator [16]. An advantage of least squares approaches is that they reduce computations to a linear system

$$(K_{nn} + \lambda n I) \, \alpha = \widehat{y}, \tag{5}$$

where K_{nn} is an $n \times n$ matrix defined by $(K_{nn})_{ij} = K(x_i, x_j)$ and $\widehat{y} = (y_1, \ldots y_n)$. We next comment on computational and statistical properties of KRR.

Computations. Solving Eq. (5) for large datasets is challenging. A direct approach requires $O(n^2)$ in space, to allocate K_{nn}, $O(n^2)$ kernel evaluations, and $O(n^2 c_K + n^3)$ in time, to compute and invert K_{nn} (c_K is the kernel evaluation cost assumed constant and omitted throughout).

Statistics. Under basic assumptions, KRR achieves an error $\mathcal{R}(\widehat{f}_{\lambda_n}) = O(n^{-1/2})$, for $\lambda_n = n^{-1/2}$, which is optimal in a minimax sense and can be improved *only* under more stringent assumptions [17, 18].

---

[2] K is positive definite, if the matrix with entries $K(x_i, x_j)$ is positive semidefinite $\forall x_1, \ldots, x_N, N \in \mathbb{N}$ [16]

The question is then if it is possible to achieve the statistical properties of KRR, with less computations.

Gradient methods and early stopping. A natural idea is to consider iterative solvers and in particular gradient methods, because of their simplicity and low iteration cost. A basic example is computing the coefficients in (3) by

$$\alpha_t = \alpha_{t-1} + \tau \left[(K_{nn}\alpha_{t-1} - \widehat{y}) + \lambda n \alpha_{t-1} \right], \tag{6}$$

for a suitable step-size choice τ.

Computations. In this case, if t is the number of iterations, gradient methods require $O(n^2 t)$ in time, $O(n^2)$ in memory and $O(n^2)$ in kernel evaluations, if the kernel matrix is stored. Note that, the kernel matrix can also be computed on the fly with only $O(n)$ memory, but $O(n^2 t)$ kernel evaluations are required. We note that, beyond the above simple iteration, several variants have been considered including accelerated [1, 19] and stochastic extensions [20].

Statistics. The statistical properties of iterative approaches are well studied and also in the case where λ is set to zero, and regularization is performed by choosing a suitable stopping time [21]. In this latter case, the number of iterations can roughly be thought of $1/\lambda$ and $O(\sqrt{n})$ iterations are needed for basic gradient descent, $O(n^{1/4})$ for accelerated methods and possible $O(1)$ iterations/epochs for stochastic methods. Importantly, we note that unlike most optimization studies, here we are considering the number of iterations needed to solve (1), rather than (4).

While the time complexity of these methods dramatically improves over KRR, and computations can be done in blocks, memory requirements (or number of kernel evaluations) still makes the application to large scale setting cumbersome. Randomization provides an approach to tackle this challenge.

Random projections. The rough idea is to use random projections to compute K_{nn} only approximately. The most popular examples in this class of approaches are Nyström [7, 8] and random features [9] methods. In the following we focus in particular on a basic Nyström approach based on considering functions of the form

$$\widetilde{f}_{\lambda,M}(x) = \sum_{i=1}^{M} \widetilde{\alpha}_i K(x, \widetilde{x}_i), \quad \text{with} \quad \{\widetilde{x}_1, \ldots, \widetilde{x}_M\} \subseteq \{x_1, \ldots, x_n\}, \tag{7}$$

defined considering only a subset of M training points sampled uniformly. In this case, there are only M coefficients that, following the approach in (4), can be derived considering the linear system

$$H\widetilde{\alpha} = z, \quad \text{where} \quad H = K_{nM}^\top K_{nM} + \lambda n K_{MM}, \quad z = K_{nM}^\top \widehat{y}. \tag{8}$$

Here K_{nM} is the $n \times M$ matrix with $(K_{nM})_{ij} = K(x_i, \widetilde{x}_j)$ and K_{MM} is the $M \times M$ matrix with $(K_{MM})_{ij} = K(\widetilde{x}_i, \widetilde{x}_j)$. This method consists in subsampling the columns of K_{nn} and can be seen as a particular form of random projections.

Computations. Direct methods for solving (8) require $O(nM^2)$ in time to form $K_{nM}^\top K_{nM}$ and $O(M^3)$ for solving the linear system, and only $O(nM)$ kernel evaluations. The naïve memory requirement is $O(nM)$ to store K_{nM}, however if $K_{nM}^\top K_{nM}$ is computed in blocks of dimension at most $M \times M$ only $O(M^2)$ memory is needed. Iterative approaches as in (6) can also be combined with random projections [22, 23, 24] to slightly reduce time requirements (see Table. 1, or Sect. F in the appendix, for more details).

Statistics. The key point though, is that random projections allow to dramatically reduce memory requirements as soon as $M \ll n$ and the question arises of whether this comes at expenses of statistical accuracy. Interestingly, recent results considering this question show that there are large classes of problems for which $M = \tilde{O}(\sqrt{n})$ suffices for the same optimal statistical accuracy of the exact KRR [11, 12, 13].

In summary, in this case the computations needed for optimal statistical accuracy are reduced from $O(n^2)$ to $O(n\sqrt{n})$ kernel evaluations, but the best time complexity is basically $O(n^2)$. In the rest of the paper we discuss how this requirement can indeed be dramatically reduced.

3 FALKON

Our approach is based on a novel combination of randomized projections with iterative solvers plus preconditioning. The main novelty is that we use random projections to approximate both the problem *and* the preconditioning.

Preliminaries: preconditioning and KRR. We begin recalling the basic idea behind preconditioning. The key quantity is the condition number, that for a linear system is the ratio between the largest and smallest singular values of the matrix defining the problem [25]. For example, for problem (5) the condition number is given by

$$\text{cond}(K_{nn} + \lambda nI) = (\sigma_{\max} + \lambda n)/(\sigma_{\min} + \lambda n),$$

with $\sigma_{\max}, \sigma_{\min}$ largest and smallest eigenvalues of K_{nn}, respectively. The importance of the condition number is that it captures the time complexity of iteratively solving the corresponding linear system. For example, if a simple gradient descent (6) is used, the number of iterations needed for an ϵ accurate solution of problem (5) is

$$t = O(\text{cond}(K_{nn} + \lambda nI) \log(1/\epsilon)).$$

It is shown in [23] that in this case $t = \sqrt{n} \log n$ are needed to achieve a solution with good statistical properties. Indeed, it can be shown that roughly $t \approx 1/\lambda \log(\frac{1}{\epsilon})$ are needed where $\lambda = 1/\sqrt{n}$ and $\epsilon = 1/n$. The idea behind preconditioning is to use a suitable matrix B to define an equivalent linear system with better condition number. For (5), an ideal choice is B such that

$$BB^\top = (K_{nn} + \lambda nI)^{-1} \tag{9}$$

and $B^\top(K_{nn} + \lambda nI)B\,\beta = B^\top \hat{y}$. Clearly, if β_* solves the latter problem, $\alpha_* = B\beta_*$ is a solution of problem (5). Using a preconditioner B as in (9) one iteration is sufficient, but computing the B is typically as hard as the original problem. The problem is to derive preconditioning such that (9) might hold only approximately, but that can be computed efficiently. Derivation of efficient preconditioners for the exact KRR problem (5) has been the subject of recent studies, [3, 4, 26, 5, 6]. In particular, [4, 26, 5, 6] consider random projections to approximately compute a preconditioner. Clearly, while preconditioning (5) leads to computational speed ups in terms of the number of iterations, requirements in terms of memory/kernel evaluation are the same as standard kernel ridge regression.

The key idea to tackle this problem is to consider an efficient preconditioning approach for problem (8) rather than (5).

Basic FALKON algorithm. We begin illustrating a basic version of our approach. The key ingredient is the following preconditioner for Eq. (8),

$$BB^\top = \left(\frac{n}{M} K_{MM}^2 + \lambda n K_{MM} \right)^{-1}, \tag{10}$$

which is itself based on a Nyström approximation[3]. The above preconditioning is a natural approximation of the ideal preconditioning of problem (8) that is $BB^\top = (K_{nM}^\top K_{nM} + \lambda n K_{MM})^{-1}$ and reduces to it if $M = n$. Our theoretical analysis, shows that $M \ll n$ suffices for deriving optimal statistical rates. In its basic form FALKON is derived combining the above preconditioning and gradient descent,

$$\widehat{f}_{\lambda,M,t}(x) = \sum_{i=1}^{M} \alpha_{t,i} K(x, \widetilde{x}_i), \quad \text{with} \quad \alpha_t = B\beta_t \quad \text{and} \tag{11}$$

$$\beta_k = \beta_{k-1} - \frac{\tau}{n} B^\top \left[K_{nM}^\top (K_{nM}(B\beta_{k-1}) - \widehat{y}) + \lambda n K_{MM}(B\beta_{k-1}) \right], \tag{12}$$

for $t \in \mathbb{N}$, $\beta_0 = 0$ and $1 \le k \le t$ and a suitable chosen τ. In practice, a refined version of FALKON is preferable where a faster gradient iteration is used and additional care is taken in organizing computations.

FALKON. The actual version of FALKON we propose is Alg. 1 (see Sect. A, Alg. 2 for the complete algorithm). It consists in solving the system $B^\top HB\beta = B^\top z$ via conjugate gradient [25], since it is a fast gradient method and does not require to specify the step-size. Moreover, to compute B quickly, with reduced numerical errors, we consider the following strategy

$$B = \frac{1}{\sqrt{n}} T^{-1} A^{-1}, \qquad T = \text{chol}(K_{MM}), \quad A = \text{chol}\left(\frac{1}{M} T T^\top + \lambda I \right), \tag{13}$$

where chol() is the Cholesky decomposition (in Sect. A the strategy for non invertible K_{MM}).

---

[3] For the sake of simplicity, here we assume K_{MM} to be invertible and the Nyström centers selected with uniform sampling from the training set, see Sect. A and Alg. 2 in the appendix for the general algorithm.

Algorithm 1 `MATLAB` code for FALKON. It requires $O(nMt + M^3)$ in time and $O(M^2)$ in memory. See Sect. A and Alg. 2 in the appendixes for the complete algorithm.

Input: Dataset $X = (x_i)_{i=1}^n \in \mathbb{R}^{n \times D}$, $\hat{y} = (y_i)_{i=1}^n \in \mathbb{R}^n$, centers $C = (\tilde{x}_j)_{j=1}^M \in \mathbb{R}^{M \times D}$, KernelMatrix computing the kernel matrix given two sets of points, regularization parameter λ, number of iterations t.
Output: Nyström coefficients α.

```
function alpha = FALKON(X, C, Y, KernelMatrix, lambda, t)
    n = size(X,1); M = size(C,1); KMM = KernelMatrix(C,C);
    T = chol(KMM + eps*M*eye(M));
    A = chol(T*T'/M + lambda*eye(M));

    function w = KnM_times_vector(u, v)
        w = zeros(M,1); ms = ceil(linspace(0, n, ceil(n/M)+1));
        for i=1:ceil(n/M)
            Kr = KernelMatrix( X(ms(i)+1:ms(i+1),:), C );
            w = w + Kr'*(Kr*u + v(ms(i)+1:ms(i+1),:));
        end
    end

    BHB = @(u) A'\(T'\(KnM_times_vector(T\(A\u), zeros(n,1))/n) + lambda*(A\u));
    r = A'\(T'\KnM_times_vector(zeros(M,1), Y/n));
    alpha = T\(A\conjgrad(BHB, r, t));
end
```

Computations. in Alg. 1, B is never built explicitly and A, T are two upper-triangular matrices, so $A^{-\top}u$, $A^{-1}u$ for a vector u costs M^2, and the same for T. The cost of computing the preconditioner is only $\frac{4}{3}M^3$ floating point operations (consisting in two Cholesky decompositions and one product of two triangular matrices). Then FALKON requires $O(nMt + M^3)$ in time and the same $O(M^2)$ memory requirement of the basic Nyström method, if matrix/vector multiplications at each iteration are performed in blocks. This implies $O(nMt)$ kernel evaluations are needed.

The question remains to characterize M and the number of iterations needed for good statistical accuracy. Indeed, in the next section we show that roughly $O(n\sqrt{n})$ computations and $O(n)$ memory are sufficient for optimal accuracy. This implies that FALKON is currently the most efficient kernel method with the same optimal statistical accuracy of KRR, see Table 1.

4 Theoretical Analysis

In this section, we characterize the generalization properties of FALKON showing it achieves the optimal generalization error of KRR, with dramatically reduced computations. This result is given in Thm. 3 and derived in two steps. First, we study the difference between the excess risk of FALKON and that of the basic Nyström (8), showing it depends on the condition number induced by the preconditioning, hence on M (see Thm.1). Deriving these results requires some care, since differently to standard optimization results, our goal is to solve (1) i.e. achieve small excess risk, not to minimize the empirical error. Second, we show that choosing $M = \widetilde{O}(1/\lambda)$ allows to make this difference as small as $e^{-t/2}$ (see Thm.2). Finally, recalling that the basic Nyström for $\lambda = 1/\sqrt{n}$ has essentially the same statistical properties of KRR [13], we answer the question posed at the end of the last section and show that roughly $\log n$ iterations are sufficient for optimal statistical accuracy. Following the discussion in the previous section this means that the computational requirements for optimal accuracy are $\widetilde{O}(n\sqrt{n})$ in time/kernel evaluations and $\widetilde{O}(n)$ in space. Later in this section faster rates under further regularity assumptions are also derived and the effect of different selection methods for the Nyström centers considered. The proofs for this section are provided in Sect. E of the appendixes.

4.1 Main Result

The first result is interesting in its own right since it corresponds to translating optimization guarantees into statistical results. In particular, we derive a relation the excess risk of the FALKON algorithm $\widehat{f}_{\lambda,M,t}$ from Alg. 1 and the Nyström estimator $\widetilde{f}_{\lambda,M}$ from Eq. (8) with uniform sampling.

Algorithm	train time	kernel evaluations	memory	test time
SVM / KRR + direct method	n^3	n^2	n^2	n
KRR + iterative [1, 2]	$n^2 \sqrt[4]{n}$	n^2	n^2	n
Doubly stochastic [22]	$n^2 \sqrt{n}$	$n^2 \sqrt{n}$	n	n
Pegasos / KRR + sgd [27]	n^2	n^2	n	n
KRR + iter + precond [3, 28, 4, 5, 6]	n^2	n^2	n	n
Divide & Conquer [29]	n^2	$n\sqrt{n}$	n	n
Nyström, random features [7, 8, 9]	n^2	$n\sqrt{n}$	n	$\sqrt{n}$
Nyström + iterative [23, 24]	n^2	$n\sqrt{n}$	n	$\sqrt{n}$
Nyström + sgd [20]	n^2	$n\sqrt{n}$	n	$\sqrt{n}$
FALKON (see Thm. 3)	$\boldsymbol{n\sqrt{n}}$	$\boldsymbol{n\sqrt{n}}$	$\boldsymbol{n}$	$\boldsymbol{\sqrt{n}}$

Table 1: Computational complexity required by different algorithms, for optimal generalization. Logarithmic terms are not showed.

Theorem 1. *Let $n, M \geq 3$, $t \in \mathbb{N}$, $0 < \lambda \leq \lambda_1$ and $\delta \in (0, 1]$. Assume there exists $\kappa \geq 1$ such that $K(x, x) \leq \kappa^2$ for any $x \in X$. Then, the following inequality holds with probability $1 - \delta$*

$$\mathcal{R}(\widehat{f}_{\lambda,M,t})^{1/2} \;\; \leq \;\; \mathcal{R}(\widetilde{f}_{\lambda,M})^{1/2} \;\; + \;\; 4\widehat{v}\, e^{-\nu t} \sqrt{1 + \frac{9\kappa^2}{\lambda n} \log \frac{n}{\delta}},$$

where $\widehat{v}^2 = \frac{1}{n} \sum_{i=1}^{n} y_i^2$ and $\nu = \log(1 + 2/(\mathrm{cond}\,(B^\top HB)^{1/2} - 1))$, with $\mathrm{cond}\,(B^\top HB)$ the condition number of $B^\top HB$. Note that $\lambda_1 > 0$ is a constant not depending on λ, n, M, δ, t.

The additive term in the bound above decreases exponentially in the number of iterations. If the condition number of $B^\top HB$ is smaller than a small universal constant (e.g. 17), then $\nu > 1/2$ and the additive term decreases as $e^{-\frac{t}{2}}$. Next, theorems derive a condition on M that allows to control $\mathrm{cond}\,(B^\top HB)$, and derive such an exponential decay.

Theorem 2. *Under the same conditions of Thm. 1, if*

$$M \geq 5 \left[1 + \frac{14\kappa^2}{\lambda}\right] \log \frac{8\kappa^2}{\lambda\delta}.$$

then the exponent ν in Thm. 1 satisfies $\nu \geq 1/2$.

The above result gives the desired exponential bound showing that after $\log n$ iterations the excess risk of FALKON is controlled by that of the basic Nyström, more precisely

$$\mathcal{R}(\widehat{f}_{\lambda,M,t}) \leq 2\mathcal{R}(\widetilde{f}_{\lambda,M}) \quad \text{when} \quad t \geq \log \mathcal{R}(\widetilde{f}_{\lambda,M}) \;\; + \;\; \log\left(1 + \frac{9\kappa^2}{\lambda n} \log \frac{n}{\delta}\right) \;\; + \;\; \log\left(16\widehat{v}^2\right).$$

Finally, we derive an excess risk bound for FALKON. By the no-free-lunch theorem, this requires some conditions on the learning problem. We first consider a standard basic setting where we only assume it exists $f_{\mathcal{H}} \in \mathcal{H}$ such that $\mathcal{E}(f_{\mathcal{H}}) = \inf_{f \in \mathcal{H}} \mathcal{E}(f)$.

Theorem 3. *Let $\delta \in (0, 1]$. Assume there exists $\kappa \geq 1$ such that $K(x, x) \leq \kappa^2$ for any $x \in X$, and $y \in [-\frac{a}{2}, \frac{a}{2}]$, almost surely, $a > 0$. There exist $n_0 \in \mathbb{N}$ such that for any $n \geq n_0$, if*

$$\lambda = \frac{1}{\sqrt{n}}, \qquad M \geq 75 \sqrt{n} \, \log \frac{48\kappa^2 n}{\delta}, \qquad t \geq \frac{1}{2} \log(n) \; + \; 5 + 2\log(a + 3\kappa),$$

then with probability $1 - \delta$,

$$\mathcal{R}(\widehat{f}_{\lambda,M,t}) \leq \frac{c_0 \log^2 \frac{24}{\delta}}{\sqrt{n}}.$$

In particular n_0, c_0 do not depend on λ, M, n, t and c_0 do not depend on δ.

The above result provides the desired bound, and all the constants are given in the appendix. The obtained learning rate is the same as the full KRR estimator and is known to be optimal in a minmax sense [17], hence not improvable. As mentioned before, the same bound is also achieved by the

basic Nyström method but with much worse time complexity. Indeed, as discussed before, using a simple iterative solver typically requires $O(\sqrt{n} \log n)$ iterations, while we need only $O(\log n)$. Considering the choice for M this leads to a computational time of $O(nMt) = O(n\sqrt{n})$ for optimal generalization (omitting logarithmic terms). To the best of our knowledge FALKON currently provides the best time/space complexity to achieve the statistical accuracy of KRR. Beyond the basic setting considered above, in the next section we show that FALKON can achieve much faster rates under refined regularity assumptions and also consider the potential benefits of leverage score sampling.

4.2 Fast learning rates and Nyström with approximate leverage scores

Considering fast rates and Nyström with more general sampling is considerably more technical and a heavier notation is needed. Our analysis apply to any approximation scheme (e.g. [30, 12, 31]) satisfying the definition of q-approximate leverage scores [13], satisfying $q^{-1} l_i(\lambda) \leq \widehat{l}_i(\lambda) \leq q l_i(\lambda), \quad \forall\, i \in \{1, \ldots, n\}$. Here $\lambda > 0$, $l_i(\lambda) = (K_{nn}(K_{nn} + \lambda nI)^{-1})_{ii}$ are the leverage scores and $q \geq 1$ controls the quality of the approximation. In particular, given λ, the Nyström points are sampled independently from the dataset with probability $p_i \propto \widehat{l}_i(\lambda)$. We need a few more definitions. Let $K_x = K(x, \cdot)$ for any $x \in X$ and $\mathcal{H}$ the reproducing kernel Hilbert space [32] of functions with inner product defined by $\mathcal{H} = \overline{\mathrm{span}\{K_x \mid x \in X\}}$ and closed with respect to the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ defined by $\langle K_x, K_{x'} \rangle_{\mathcal{H}} = K(x, x')$, for all $x, x' \in X$. Define $C : \mathcal{H} \to \mathcal{H}$ to be the linear operator $\langle f, Cg \rangle_{\mathcal{H}} = \int_X f(x)g(x)d\rho_X(x)$, for all $f, g \in \mathcal{H}$. Finally define the following quantities,

$$\mathcal{N}_\infty(\lambda) = \sup_{x \in X} \|(C + \lambda I)^{-1/2} K_x\|_{\mathcal{H}}, \quad \mathcal{N}(\lambda) = \mathrm{Tr}(C(C + \lambda I)^{-1}).$$

The latter quantity is known as *degrees of freedom* or *effective dimension*, can be seen as a measure of the *size* of $\mathcal{H}$. The quantity $\mathcal{N}_\infty(\lambda)$ can be seen to provide a uniform bound on the leverage scores. In particular note that $\mathcal{N}(\lambda) \leq \mathcal{N}_\infty(\lambda) \leq \frac{\kappa^2}{\lambda}$ [13]. We can now provide a refined version of Thm. 2.

Theorem 4. *Under the same conditions of Thm. 1, the exponent ν in Thm. 1 satisfies $\nu \geq 1/2$, when*

1. *either Nyström uniform sampling is used with $M \geq 70\left[1 + \mathcal{N}_\infty(\lambda)\right] \log \frac{8\kappa^2}{\lambda\delta}$.*

2. *or Nyström q-approx. lev. scores [13] is used, with $\lambda \geq \frac{19\kappa^2}{n} \log \frac{n}{2\delta}$, $n \geq 405\kappa^2 \log \frac{12\kappa^2}{\delta}$,*

$$M \geq 215\left[2 + q^2 \mathcal{N}(\lambda)\right] \log \frac{8\kappa^2}{\lambda\delta}.$$

We then recall the standard, albeit technical, assumptions leading to fast rates [17, 18]. The *capacity condition* requires the existence of $\gamma \in (0, 1]$ and $Q \geq 0$, such that $\mathcal{N}(\lambda) \leq Q^2 \lambda^{-\gamma}$. Note that this condition is always satisfied with $Q = \kappa$ and $\gamma = 1$. The *source condition* requires the existence of $r \in [1/2, 1]$ and $g \in \mathcal{H}$, such that $f_{\mathcal{H}} = C^{r-1/2}g$. Intuitively, the *capacity condition* measures the size of $\mathcal{H}$, if γ is small then $\mathcal{H}$ is small and rates are faster. The *source condition* measures the regularity of $f_{\mathcal{H}}$, if r is big $f_{\mathcal{H}}$ is regular and rates are faster. The case $r = 1/2$ and $\gamma = D/(2s)$ (for a kernel with smoothness s and input space $\mathbb{R}^D$) recovers the classic Sobolev condition. For further discussions on the interpretation of the conditions above see [17, 18, 11, 13]. We can then state our main result on fast rates

Theorem 5. *Let $\delta \in (0, 1]$. Assume there exists $\kappa \geq 1$ such that $K(x, x) \leq \kappa^2$ for any $x \in X$, and $y \in [-\frac{a}{2}, \frac{a}{2}]$, almost surely, with $a > 0$. There exist an $n_0 \in \mathbb{N}$ such that for any $n \geq n_0$ the following holds. When*

$$\lambda = n^{-\frac{1}{2r+\gamma}}, \qquad t \;\geq\; \log(n) \;+\; 5 + 2\log(a + 3\kappa^2),$$

1. *and either Nyström uniform sampling is used with $M \geq 70\left[1 + \mathcal{N}_\infty(\lambda)\right] \log \frac{8\kappa^2}{\lambda\delta}$,*

2. *or Nyström q-approx. lev. scores [13] is used with $M \geq 220\left[2 + q^2 \mathcal{N}(\lambda)\right] \log \frac{8\kappa^2}{\lambda\delta}$,*

then with probability $1 - \delta$,

$$\mathcal{R}(\widehat{f}_{\lambda, M, t}) \;\leq\; c_0 \log^2 \frac{24}{\delta} \; n^{-\frac{2r}{2r+\gamma}}.$$

where $\widehat{f}_{\lambda, M, t}$ is the FALKON estimator (Sect. 3, Alg. 1 and Sect. A, Alg. 2 in the appendix for the complete version). In particular n_0, c_0 do not depend on λ, M, n, t and c_0 do not depend on δ.
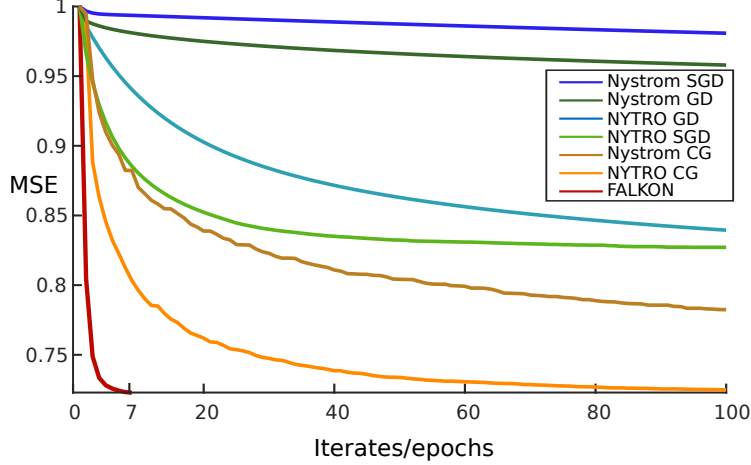
Figure 1: Falkon is compared to stochastic gradient, gradient descent and conjugate gradient applied to Problem (8), while NYTRO refer to the variants described in [23]. The graph shows the test error on the HIGGS dataset (1.1×10^7 examples) with respect to the number of iterations (epochs for stochastic algorithms).

The above result shows that FALKON achieves the same fast rates as KRR, under the same conditions [17]. For $r = 1/2, \gamma = 1$, the rate in Thm. 3 is recovered. If $\gamma < 1, r > 1/2$, FALKON achieves a rate close to $O(1/n)$. By selecting the Nyström points with uniform sampling, a bigger M could be needed for fast rates (albeit always less than n). However, when approximate leverage scores are used M, smaller than $n^{\gamma/2} \ll \sqrt{n}$ is always enough for optimal generalization. This shows that FALKON with approximate leverage scores is the first algorithm to achieve fast rates with a computational complexity that is $O(n\mathcal{N}(\lambda)) = O(n^{1+\frac{\gamma}{2r+\gamma}}) \leq O(n^{1+\frac{\gamma}{2}})$ in time.

5 Experiments

We present FALKON's performance on a range of large scale datasets. As shown in Table 2, 3, FALKON achieves state of the art accuracy and typically outperforms previous approaches in all the considered large scale datasets including IMAGENET. This is remarkable considering FALKON required only a fraction of the competitor's computational resources. Indeed we used a single machine equipped with two Intel Xeon E5-2630 v3, one NVIDIA Tesla K40c and 128 GB of RAM and a basic MATLAB FALKON implementation, while typically the results for competing algorithm have been performed on clusters of GPU workstations (accuracies, times and used architectures are cited from the corresponding papers).

A minimal MATLAB implementation of FALKON is presented in Appendix G. The code necessary to reproduce the following experiments, plus a FALKON version that is able to use the GPU, is available on GitHub at `https://github.com/LCSL/FALKON_paper` . The error is measured with MSE, RMSE or relative error for regression problems, and with classification error (c-err) or AUC for the classification problems, to be consistent with the literature. For datasets which do not have a fixed test set, we set apart 20% of the data for testing. For all datasets, but YELP and IMAGENET, we normalize the features by their z-score. From now on we denote with n the cardinality of the dataset, d the dimensionality. A comparison of FALKON with respect to other methods to compute the Nyström estimator, in terms of the MSE test error on the HIGGS dataset, is given in Figure 1.

MillionSongs [36] (Table 2, $n = 4.6 \times 10^5$, $d = 90$, regression). We used a Gaussian kernel with $\sigma = 6$, $\lambda = 10^{-6}$ and 10^4 Nyström centers. Moreover with 5×10^4 center, FALKON achieves a 79.20 MSE, and 4.49×10^{-3} rel. error in 630 *sec*.

TIMIT (Table 2, $n = 1.2 \times 10^6$, $d = 440$, multiclass classification). We used the same preprocessed dataset of [6] and Gaussian Kernel with $\sigma = 15$, $\lambda = 10^{-9}$ and 10^5 Nyström centers.

YELP (Table 2, $n = 1.5 \times 10^6$, $d = 6.52 \times 10^7$, regression). We used the same dataset of [24]. We extracted the 3-grams from the plain text with the same pipeline as [24], then we mapped

them in a sparse binary vector which records if the 3-gram is present or not in the example. We used a linear kernel with 5×10^4 Nyström centers. With 10^5 centers, we get a RMSE of 0.828 in 50 minutes.

Table 2: Architectures: ‡ cluster 128 EC2 r3.2xlarge machines, † cluster 8 EC2 r3.8xlarge machines, ℓ single machine with two Intel Xeon E5-2620, one Nvidia GTX Titan X GPU, 128GB RAM, ⋆ cluster with IBM POWER8 12-core processor, 512 GB RAM, ∗ unknown platform.

	MillionSongs			YELP		TIMIT	
	MSE	Relative error	Time(s)	RMSE	Time(m)	c-err	Time(h)
FALKON	**80.10**	$\mathbf{4.51 \times 10^{-3}}$	**55**	**0.833**	**20**	32.3%	**1.5**
Prec. KRR [4]	-	4.58×10^{-3}	$289^{\dagger}$	-	-	-	-
Hierarchical [33]	-	4.56×10^{-3}	$293^{\star}$	-	-	-	-
D&C [29]	80.35	-	$737^{\star}$	-	-	-	-
Rand. Feat. [29]	80.93	-	$772^{\star}$	-	-	-	-
Nyström [29]	80.38	-	$876^{\star}$	-	-	-	-
ADMM R. F.[4]	-	5.01×10^{-3}	$958^{\dagger}$	-	-	-	-
BCD R. F. [24]	-	-	-	0.949	$42^{\ddagger}$	34.0%	$1.7^{\ddagger}$
BCD Nyström [24]	-	-	-	0.861	$60^{\ddagger}$	33.7%	$1.7^{\ddagger}$
EigenPro [6]	-	-	-	-	-	32.6%	3.9^{ℓ}
KRR [33] [24]	-	4.55×10^{-3}	-	0.854	$500^{\ddagger}$	33.5%	$8.3^{\ddagger}$
Deep NN [34]	-	-	-	-	-	32.4%	-
Sparse Kernels [34]	-	-	-	-	-	**30.9%**	-
Ensemble [35]	-	-	-	-	-	33.5%	-

Table 3: Architectures: † cluster with IBM POWER8 12-core cpu, 512 GB RAM, ℓ single machine with two Intel Xeon E5-2620, one Nvidia GTX Titan X GPU, 128GB RAM, ‡ single machine [37]

	SUSY			HIGGS		IMAGENET	
	c-err	AUC	Time(m)	AUC	Time(h)	c-err	Time(h)
FALKON	**19.6%**	0.877	**4**	0.833	**3**	20.7%	**4**
EigenPro [6]	19.8%	-	6^{ℓ}	-	-	-	-
Hierarchical [33]	20.1%	-	$40^{\dagger}$	-	-	-	-
Boosted Decision Tree [38]	-	0.863	-	0.810	-	-	-
Neural Network [38]	-	0.875	-	0.816	-	-	-
Deep Neural Network [38]	-	**0.879**	$4680^{\ddagger}$	**0.885**	$78^{\ddagger}$	-	-
Inception-V4 [39]	-	-	-	-	-	**20.0%**	-

SUSY (Table 3, $n = 5 \times 10^6$, $d = 18$, binary classification). We used a Gaussian kernel with $\sigma = 4$, $\lambda = 10^{-6}$ and 10^4 Nyström centers.

HIGGS (Table 3, $n = 1.1 \times 10^7$, $d = 28$, binary classification). Each feature has been normalized subtracting its mean and dividing for its variance. We used a Gaussian kernel with diagonal matrix width learned with cross validation on a small validation set, $\lambda = 10^{-8}$ and 10^5 Nyström centers. If we use a single $\sigma = 5$ we reach an AUC of 0.825.

IMAGENET (Table 3, $n = 1.3 \times 10^6$, $d = 1536$, multiclass classification). We report the top 1 c-err over the validation set of ILSVRC 2012 with a single crop. The features are obtained from the convolutional layers of pre-trained Inception-V4 [39]. We used Gaussian kernel with $\sigma = 19$, $\lambda = 10^{-9}$ and 5×10^4 Nyström centers. Note that with linear kernel we achieve c-err $= 22.2\%$.

References

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

[2] 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.

[3] Gregory E Fasshauer and Michael J McCourt. Stable evaluation of gaussian radial basis function interpolants. *SIAM Journal on Scientific Computing*, 34(2):A737–A762, 2012.

[4] Haim Avron, Kenneth L Clarkson, and David P Woodruff. Faster kernel ridge regression using sketching and preconditioning. *arXiv preprint arXiv:1611.03220*, 2016.

[5] Alon Gonen, Francesco Orabona, and Shai Shalev-Shwartz. Solving ridge regression using sketched preconditioned svrg. *arXiv preprint arXiv:1602.02350*, 2016.

[6] Siyuan Ma and Mikhail Belkin. Diving into the shallows: a computational perspective on large-scale shallow learning. *arXiv preprint arXiv:1703.10622*, 2017.

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

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

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

[10] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pages 1313–1320, 2009.

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

[12] Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems 28*, pages 775–783. 2015.

[13] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1648–1656, 2015.

[14] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. *arXiv preprint arXiv:1602.04474*, 2016.

[15] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.

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

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

[18] Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In *COLT*, 2009.

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

[20] Aymeric Dieuleveut and Francis Bach. Non-parametric stochastic approximation with large step sizes. *arXiv preprint arXiv:1408.0361*, 2014.

[21] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[22] Bo Dai, Bo Xie, Niao He, Yingyu Liang, Anant Raj, Maria-Florina F Balcan, and Le Song. Scalable kernel methods via doubly stochastic gradients. In *Advances in Neural Information Processing Systems*, pages 3041–3049, 2014.

[23] Raffaello Camoriano, Tomás Angles, Alessandro Rudi, and Lorenzo Rosasco. Nytro: When subsampling meets early stopping. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1403–1411, 2016.

[24] Stephen Tu, Rebecca Roelofs, Shivaram Venkataraman, and Benjamin Recht. Large scale kernel learning using block coordinate descent. *arXiv preprint arXiv:1602.05310*, 2016.

[25] Yousef Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

[26] Kurt Cutajar, Michael Osborne, John Cunningham, and Maurizio Filippone. Preconditioning kernel matrices. In *International Conference on Machine Learning*, pages 2529–2538, 2016.

[27] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

[28] Yun Yang, Mert Pilanci, and Martin J Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *arXiv preprint arXiv:1501.06195*, 2015.

[29] 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.

[30] 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.

[31] 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.

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

[33] Jie Chen, Haim Avron, and Vikas Sindhwani. Hierarchically compositional kernels for scalable nonparametric learning. *CoRR*, abs/1608.00860, 2016.

[34] Avner May, Alireza Bagheri Garakani, Zhiyun Lu, Dong Guo, Kuan Liu, Aurelien Bellet, Linxi Fan, Michael Collins, Daniel J. Hsu, Brian Kingsbury, Michael Picheny, and Fei Sha. Kernel approximation methods for speech recognition. *CoRR*, abs/1701.03577, 2017.

[35] Po-Sen Huang, Haim Avron, Tara N. Sainath, Vikas Sindhwani, and Bhuvana Ramabhadran. Kernel methods match deep neural networks on timit. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 205–209, 2014.

[36] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *ISMIR*, 2011.

[37] Alexandre Alves. Stacking machine learning classifiers to identify higgs bosons at the lhc. *CoRR*, abs/1612.07725, 2016.

[38] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5, 2014.

[39] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. pages 4278–4284, 2017.

[40] Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics: Vol.: 1.: Functional Analysis*. Academic press, 1980.

[41] Ernesto D Vito, Lorenzo Rosasco, Andrea Caponnetto, Umberto D Giovannini, and Francesca Odone. Learning from examples as an inverse problem. In *Journal of Machine Learning Research*, pages 883–904, 2005.

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

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

FALKON: An Optimal Large Scale Kernel Method
Supplementary Materials

A   FALKON: General Algorithm

In this section we define a generalized version of FALKON. In particular we provide a preconditioner able to deal with non invertible K_{MM} and with Nyström centers selected by using approximate leverage scores. In Def. 3 we state the properties that such preconditioner must satisfy. In Example 1 we show that the preconditioner in Sect. 3 satisfies Def. 3 when K_{MM}.

First we recall some ways to sample Nyström centers, from the training set.

Nyström with uniform sampling. Let $n, M \in \mathbb{N}$ with $1 \le M \le n$. Let $x_1, \dots, x_n$ be the training set. The Nyström centers $\widetilde{x}_1, \dots, \widetilde{x}_M$ are a random subset of cardinality M uniformly sampled from the training set.

Nyström with approximate leverage scores. We recall the definition of approximate leverage scores, from [13] and then the sampling method based on them. Let $n \in \mathbb{N}$, $\lambda > 0$. Let $x_1, \dots, x_n$ be the training points and define $K_{nn} \in \mathbb{R}^{n \times n}$ as $(K_{nn})_{ij} = K(x_i, x_j)$ for $1 \le i, j \le n$. The exact leverage scores are defined by

$$l_\lambda(i) = \left(K_{nn}(K_{nn} + \lambda nI)^{-1} \right)_{ii},$$

for any $i \in 1, \dots, n$. Any bi-lipschitz approximation of the exact leverage scores, satisfying the following definition is denoted as approximate leverage scores.

Definition 1 (Nyström with (q, λ_0, δ)-approximate leverage scores [13])**.** *Let $\delta \in (0, 1]$ and $\lambda_0 > 0$ and $q \in [1, \infty)$. A (random) sequence $(\widehat{l}_\lambda(i))_{i=1}^n$ is denoted as (q, λ_0, δ)-approximate leverage scores when the following holds with probability at least $1 - \delta$*

$$\frac{1}{q} \, l_\lambda(i) \; \le \; \widehat{l}_\lambda(i) \; \le \; q \, l_\lambda(i), \qquad \forall \lambda \ge \lambda_0, t \in \{1, \dots, n\}.$$

In particular, given $n \in \mathbb{N}$ training points $x_1, \dots, x_n$, and a sequence of approximate leverage scores $(\widehat{l}_\lambda(i))_{i=1}^n$, the Nyström centers are selected in the following way. Let $p_i = \frac{\widehat{l}_\lambda(i)}{\sum_{j=1}^n \widehat{l}_\lambda(j)}$, with $1 \le i \le n$. Let $i_1, \dots, i_M$ be independently sampled from $\{1, \dots, n\}$ with probability $(p_i)_{i=1}^n$. Then $\widetilde{x}_1 := x_{i_1}, \dots, \widetilde{x}_M := x_{i_M}$.

Now we define a diagonal matrix depending on the used sampling scheme that will be needed for the general preconditioner.

Definition 2. *Let $D \in \mathbb{R}^{M \times M}$ be a diagonal matrix. If the Nyström centers are selected via uniform sampling, then $D_{jj} = 1$, for $1 \le j \le M$.*
Otherwhise, let $i_1, \dots, i_M \in \{1, \dots, n\}$ be the indexes of the training points sampled via approximate leverage scores. Then for $1 \le j \le M$,

$$D_{jj} = \sqrt{\frac{1}{np_{i_j}}}.$$

We note here that by definition D is a diagonal matrix with strictly positive and finite diagonal. Indeed it is true in the uniform case. In the leverage scores case, let $1 \leq j \leq M$. Note that since the index i_j has been sampled, it implies that the probability p_{i_j} is strictly larger than zero. Then, since $0 < p_{i_j} \leq 1$ then $0 < D_{jj} < \infty$ a.s. .

A.1 Generalized FALKON Algorithm

We now introduce some matrices needed for the definition of a generalized version of FALKON, able to deal with non invertible K_{MM} and with different sampling schemes, for the Nyström centers. Finally in Def. 4, we define a general form of the algorithm, that will be used in the rest of the appendix.

Definition 3 (The generalized preconditioner). *Let $M \in \mathbb{N}$. Let $\widetilde{x}_1, \ldots, \widetilde{x}_M \in X$ and $K_{MM} \in \mathbb{R}^{M \times M}$ with $(K_{MM})_{ij} = K(\widetilde{x}_i, \widetilde{x}_j)$, for $1 \leq i, j \leq M$. Let $D \in \mathbb{R}^{M \times M}$ be a diagonal matrix with strictly positive diagonal, defined according to Def 2.*

Let $\lambda > 0$, $q \leq M$ be the rank of K_{MM}, $Q \in \mathbb{R}^{M \times q}$ a partial isometry such that $Q^\top Q = I$ and $T \in \mathbb{R}^{q \times q}$ a triangular matrix. Moreover Q, T satisfy the following equation

$$DK_{MM}D = QT^\top TQ^\top.$$

Finally let $A \in \mathbb{R}^{q \times q}$ be a triangular matrix such that

$$A^\top A = \frac{1}{M}TT^\top + \lambda I.$$

Then the generalized preconditioner *is defined as*

$$B = \frac{1}{\sqrt{n}} DQT^{-1}A^{-1}.$$

Note that B is right invertible, indeed D is invertible, since is a diagonal matrix, with strictly positive diagonal, T, A are invertible since they are square and full rank and Q is a partial isometry, so $B^{-1} = \sqrt{n}ATQ^\top D^{-1}$ and $BB^{-1} = I$. Now we provide two ways to compute Q, T, A. We recall that the Cholesky algorithm, denoted by chol, given a square positive definite matrix, $B \in \mathbb{R}^{M \times M}$, produces an upper triangular matrix $R \in \mathbb{R}^{M \times M}$ such that $B = R^\top R$. While the pivoted (or rank revealing) QR decomposition, denoted by qr, given a square matrix B, with rank q, produces a partial isometry $Q \in \mathbb{R}^{M \times q}$ with the same range of M and an upper trapezoidal matrix $R \in \mathbb{R}^{q \times M}$ such that $B = QR$.

Example 1 (precoditioner satisfying Def. 3). *Let $\lambda > 0$, and K_{MM}, D as in Def. 3.*

1. *When K_{MM} is full rank ($q = M$), then the following Q, T, A satisfy Def. 3*

$$Q = I, \quad T = \text{chol}(DK_{MM}D), \quad A = \text{chol}\left(\frac{1}{M}TT^\top + \lambda I\right).$$

2. *When K_{MM} is of any rank ($q \leq M$), then the following Q, T, A satisfy Def. 3*

$$(Q, R) = \text{qr}(DK_{MM}D), \quad T = \text{chol}(Q^\top DK_{MM}DQ), \quad A = \text{chol}\left(\frac{1}{M}TT^\top + \lambda I\right).$$

Proof. In the first case, Q, T, A satisfy Def. 3 by construction. In the second case, since $QQ^\top$ is the projection matrix on the range of $DK_{MM}D$, then $QQ^\top DK_{MM}D = DK_{MM}D$ and, since $DK_{MM}D$ is symmetric, $DK_{MM}DQQ^\top = DK_{MM}D$, so

$$QT^\top TQ^\top = QQ^\top DK_{MM}DQQ^\top = DK_{MM}D.$$

Moreover note that, since the rank of K_{MM} is q, then the range of $DK_{MM}D$ is q, and so $Q^\top Q = I$, since it is a partial isometry with dimension $\mathbb{R}^{M \times q}$. Finally A satisfies Def. 3 by construction. $\square$

Instead of rank-revealing QR decomposition, eigen-decomposition can be used.

Example 2 (precoditioner for the deficient rank case, using eig instead of qr). *Let $\lambda > 0$, and K_{MM}, D as in Def. 3. Let $(\lambda_i, u_i)_{1 \leq i \leq M}$ be respectively the eigenvalues and the associated eigenvectors from the eigendecomposition of $DK_{MM}D$, with $\lambda_1 \geq \cdots \geq \lambda_M \geq 0$. So the following Q, T, A satisfy Def. 3, $Q = (u_1, \ldots, u_q)$ and $T = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_q})$, while $A = \text{diag}\left(\sqrt{\lambda + \frac{1}{M}\lambda_1}, \ldots, \sqrt{\lambda + \frac{1}{M}\lambda_q}\right)$.*

We recall that this approach to compute Q, T, A is conceptually simpler than the one with QR decomposition, but slower, since the hidden constants in the eigendecomposition are larger than the one of QR.

The following is the general form of the algorithm.

Definition 4 (Generalized FALKON algorithm). *Let $\lambda > 0$, $t \in \mathbb{N}$ and q, Q, T, A as in Def. 3. The generalized FALKON estimator is defined as follows*

$$\widehat{f}_{\lambda,M,t}(x) = \sum_{i=1}^{M} \alpha_i K(x, \tilde{x}_i), \quad with \quad \alpha = B\beta_t,$$

and $\beta_t \in \mathbb{R}^q$ denotes the vector resulting from t iterations of the conjugate gradient algorithm applied to the following linear system

$$W\beta = b, \quad where \quad W = B^\top(K_{nM}^\top K_{nM} + \lambda n K_{MM})B, \quad b = B^\top K_{nM}^\top \widehat{y}. \tag{14}$$

B Definitions and Notation

Here we recall some basic facts on linear operators and give some notation that will be used in the rest of the appendix, then we define the necessary operators to deal with the excess risk of FALKON via functional analytic tools.

Notation   Let $\mathcal{H}$ be an Hilbert space, we denote with $\|\cdot\|_{\mathcal{H}}$, the associated norm and with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ the associated inner product. We denote with $\|\cdot\|$ the operator norm for a bounded linear operator A, defined as $\|A\| = \sup_{\|f\|_{\mathcal{H}}=1} \|Af\|$. Moreover we will denote with $\otimes$ the tensor product, in particular

$$(u \otimes v)z = u \langle v, z \rangle_{\mathcal{H}}, \quad \forall u, v, z \in \mathcal{H}.$$

In the rest of the appendix $A + \lambda I$ is often denoted by A_λ where A is linear operator and $\lambda \in \mathbb{R}$, moreover we denote with A^* the adjoint of the linear operator A, we will use $A^\top$ if A is a matrix. When $\mathcal{H}$ is separable, we denote with Tr the trace, that is $\mathrm{Tr}(A) = \sum_{j=1}^{D} \langle u_i, Au_i \rangle_{\mathcal{H}}$ for any linear operator $A : \mathcal{H} \to \mathcal{H}$, where $(u_i)_{i=1}^{D}$ is an orthogonal basis for $\mathcal{H}$ and $D \in \mathbb{N} \cup \{\infty\}$ is the dimensionality of $\mathcal{H}$. Moreover we denote with $\|\cdot\|_{\mathrm{HS}}$ the Hilbert-Schmidt norm, that is $\|A\|_{\mathrm{HS}}^2 = \mathrm{Tr}(A^*A)$, for a linear operator A.

In the next proposition we recall the spectral theorem for compact self-adjoint operators on a Hilbert space.

Proposition 1 (Spectral Theorem for compact self-adjoint operators). *Let A be a compact self-adjoint operator on a separable Hilbert space $\mathcal{H}$. Then there exists a sequence $(\lambda_j)_{j=1}^{D}$ with $\lambda_j \in \mathbb{R}$, and an orthogonal basis of $\mathcal{H}$ $(u_j)_{j=1}^{D}$ where $D \in \mathbb{N} \cup \{\infty\}$ is the dimensionality of $\mathcal{H}$, such that*

$$A = \sum_{j=1}^{D} \lambda_j u_j \otimes u_j. \tag{15}$$

Proof. Thm. VI.16, pag. 203 of [40]. □

Let $\mathcal{H}$ be a separable Hilbert space (for the sake of simplicity assume $D = \infty$), and A be a bounded self-adjoint operator on $\mathcal{H}$ that admits a spectral decomposition as in Eq. 15. Then the largest and the smallest eigenvalues of A are denoted by

$$\lambda_{\max}(A) = \sup_{j \geq 1} \lambda_j, \quad \lambda_{\min}(A) = \inf_{j \geq 1} \lambda_j.$$

In the next proposition we recall a basic fact about bounded symmetric linear operators on a separable Hilbert space $\mathcal{H}$.

Proposition 2. *Let A be a bounded self-adjoint operator on $\mathcal{H}$, that admits a spectral decomposition as in Eq. 15. Then*

$$-\|A\| \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq \|A\|.$$

Proof. By definition of operator norm, we have that $\|Ax\|_{\mathcal{H}}^2 \leq \|A\|^2 \|x\|_{\mathcal{H}}^2 \quad \forall x \in \mathcal{H}$. Let $(\lambda_j, u_j)_{j=1}^{D}$ be an eigendecomposition of A, with D the dimensionality of $\mathcal{H}$, according to Prop. 1, then, for any $j \geq 1$, we have

$$\lambda_j^2 = \langle Au_j, Au_j \rangle = \|Au_j\|_{\mathcal{H}}^2 \leq \|A\|^2,$$

where we used the fact that $Au_j = \lambda_j u_j$ and that $\|u_j\|_{\mathcal{H}} = 1$. □

B.1 Definitions

Let X be a measurable and separable space and $\mathcal{Y} = \mathbb{R}$. Let ρ be a probability measure on $X \times \mathbb{R}$. We denote with ρ_X the marginal probability of ρ on X and with $\rho(y|x)$ the conditional probability measure on $\mathcal{Y}$ given X. Let $L^2(X, \rho_X)$ be the Lebesgue space of ρ_X square integrable functions, endowed with the inner product

$$\langle \phi, \psi \rangle_\rho = \int \phi(x)\psi(x)d\rho_X(x), \quad \forall \phi, \psi \in L^2(X, \rho_X),$$

and norm $\|\psi\|_\rho = \sqrt{\langle \psi, \psi \rangle_\rho}$ for any $\psi \in L^2(X, \rho_X)$. We now introduce the kernel and its associated space of functions. Let $K : X \times X \to \mathbb{R}$ be a positive definite kernel, measurable and uniformly bounded, i.e. there exists $\kappa \in (0, \infty)$, for which $K(x, x) \leq \kappa^2$ almost surely. We denote with K_x the function $K(x, \cdot)$ and with $(\mathcal{H}, \langle \cdot, \cdot \rangle_\mathcal{H})$, the Hilbert space of functions with the associated inner product induced by K, defined by

$$\mathcal{H} = \overline{\text{span}\{K_x \mid x \in X\}}, \quad \langle K_x, K_{x'} \rangle_\mathcal{H} = K(x, x'), \quad \forall x, x' \in X.$$

Now we define the linear operators used in the rest of the appendix

Definition 5. *Under the assumptions above, for any $f \in \mathcal{H}, \phi \in L^2(X, \rho_X)$*

- *$S : \mathcal{H} \to L^2(X, \rho_X)$, such that $Sf = \langle f, K_{(\cdot)} \rangle_\mathcal{H} \in L^2(X, \rho_X)$, with adjoint*

- *$S^* : L^2(X, \rho_X) \to \mathcal{H}$, such that $S^*\phi = \int \phi(x)K_x d\rho_X(x) \in \mathcal{H}$.*

- *$L : L^2(X, \rho_X) \to L^2(X, \rho_X)$, such that $L = SS^*$ and*

- *$C : \mathcal{H} \to \mathcal{H}$, such that $C = S^*S$.*

Let $x_i \in X$ with $1 \leq i \leq n$ and $n \in \mathbb{N}$, and $\widetilde{x}_j \in X$ for $1 \leq j \leq M$ and $M \in \mathbb{N}$. We define the following linear operators

Definition 6. *Under the assumptions above, for any $f \in \mathcal{H}, v \in \mathbb{R}^n, w \in \mathbb{R}^M$,*

- *$\widehat{S}_n : \mathcal{H} \to \mathbb{R}^n$, such that $\widehat{S}_n f = \frac{1}{\sqrt{n}}(\langle f, K_{x_i} \rangle)_{i=1}^n \in \mathbb{R}^n$, with adjoint*

- *$\widehat{S}_n^* : \mathbb{R}^n \to \mathcal{H}$, such that $\widehat{S}_n^* v = \frac{1}{\sqrt{n}} \sum_{i=1}^n v_i K_{x_i} \in \mathcal{H}$.*

- *$\widehat{C}_n : \mathcal{H} \to \mathcal{H}$, such that $\widehat{C}_n = \widehat{S}_n^* \widehat{S}_n$.*

- *$\widehat{S}_M : \mathcal{H} \to \mathbb{R}^M$, such that $\widehat{S}_M f = \frac{1}{\sqrt{M}}(\langle f, K_{\widetilde{x}_i} \rangle)_{i=1}^M \in \mathbb{R}^M$, with adjoint*

- *$\widehat{S}_M^* : \mathbb{R}^M \to \mathcal{H}$, such that $\widehat{S}_M^* w = \frac{1}{\sqrt{M}} \sum_{i=1}^M v_i K_{\widetilde{x}_i} \in \mathcal{H}$.*

- *$\widehat{C}_M : \mathcal{H} \to \mathcal{H}$, such that $\widehat{C}_M = \widehat{S}_M^* \widehat{S}_M$.*

- *$\widehat{G}_M : \mathcal{H} \to \mathcal{H}$, such that $\widehat{G}_M = \widehat{S}_M^* D^2 \widehat{S}_M$, with D defined in Def. 3 (see also Def. 2).*

We now recall some basic facts about $L, C, S, K_{nn}, \widehat{C}_n, \widehat{S}_n, K_{nM}$ and K_{MM}.

Proposition 3. *With the notation introduced above,*

1. $K_{nM} = \sqrt{nM}\widehat{S}_n\widehat{S}_M^*, \qquad\qquad K_{MM} = M\,\widehat{S}_M\widehat{S}_M^*, \qquad\qquad K_{nn} = n\,\widehat{S}_n\widehat{S}_n^*$

2. $C = \displaystyle\int_X K_x \otimes K_x d\rho_X(x), \qquad \text{Tr}(C) = \text{Tr}(L) = \|S\|_{HS}^2 = \int_X \|K_x\|_\mathcal{H}^2 d\rho_X(x) \leq \kappa^2,$

3. $\widehat{C}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^n K_{x_i} \otimes K_{x_i}, \qquad \text{Tr}(\widehat{C}_n) = \text{Tr}(K_{nn}/n) = \|\widehat{S}_n\|_{HS}^2 = \dfrac{1}{n}\sum_{i=1}^n \|K_{x_i}\|_\mathcal{H}^2 \leq \kappa^2,$

4. $\widehat{C}_M = \dfrac{1}{M}\displaystyle\sum_{i=1}^M K_{\widetilde{x}_i} \otimes K_{\widetilde{x}_i}, \qquad \text{Tr}(\widehat{C}_M) = \text{Tr}(K_{MM}/M) = \|\widehat{S}_M\|_{HS}^2 = \dfrac{1}{M}\sum_{i=1}^m \|K_{\widetilde{x}_i}\|_\mathcal{H}^2 \leq \kappa^2,$

5. $\widehat{G}_M = \dfrac{1}{M}\displaystyle\sum_{i=1}^M D_{ii}^2 K_{\widetilde{x}_i} \otimes K_{\widetilde{x}_i}.$

where $\otimes$ denotes the tensor product.

Proof. Note that $(K_{nM})_{ij} = K(x_i, \tilde{x}_j) = \langle K_{x_i}, K_{\tilde{x}_j} \rangle_{\mathcal{H}} = (\sqrt{nM}\widehat{S}_n \widehat{S}_M^*)_{ij}$, for any $1 \leq i \leq n, 1 \leq j \leq M$, thus $K_{nM} = \sqrt{nM}\widehat{S}_n \widehat{S}_M^*$. The same reasoning holds for K_{MM} and K_{nn}. For the second equation, by definition of $C = S^*S$ we have that, for each $h, h' \in \mathcal{H}$,

$$\langle h, Ch' \rangle_{\mathcal{H}} = \langle Sh, Sh' \rangle_{\rho} = \int_X \langle h, K_x \rangle_{\mathcal{H}} \langle K_x, h' \rangle_{\mathcal{H}} d\rho_X(x) = \int_X \left\langle h, \left(K_x \langle K_x, h' \rangle_{\mathcal{H}} \right) \right\rangle_{\mathcal{H}} d\rho_X(x)$$

$$= \int_X \left\langle h, \left(K_x \otimes K_x \right)h' \right\rangle_{\mathcal{H}} d\rho_X(x) = \left\langle h, \left(\int_X K_x \otimes K_x d\rho_X(x) \right)h' \right\rangle_{\mathcal{H}}.$$

Note that, since K is bounded almost surely, then $\|K_x\|_{\mathcal{H}} \leq \kappa$ for any $x \in X$, thus

$$\mathrm{Tr}(C) = \int_X \mathrm{Tr}(K_x \otimes K_x) d\rho_X(x) = \int_X \|K_x\|_{\mathcal{H}}^2 d\rho_X(x) \leq \kappa^2$$

by linearity of the trace. Thus $\mathrm{Tr}(C) < \infty$ and so

$$\mathrm{Tr}(C) = \mathrm{Tr}(S^*S) = \|S\|_{HS}^2 = \mathrm{Tr}(SS^*) = \mathrm{Tr}(L).$$

The proof for the rest of equations is analogous to the one for the second. $\qquad\square$

Now we recall a standard characterization of the excess risk

Proposition 4. *When $\int_{\mathcal{Y}} y^2 d\rho < \infty$, then there exist $f_\rho \in L^2(X, \rho_X)$ defined by*

$$f_\rho(x) = \int y d\rho(y|x),$$

almost everywhere. Moreover, for any $\hat{f} \in \mathcal{H}$ we have,

$$\mathcal{E}(\hat{f}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) = \|S\hat{f} - Pf_\rho\|_{\rho_X}^2,$$

where $P : L^2(X, \rho_X) \to L^2(X, \rho_X)$ is the projection operator whose range is the closure in $L^2(X, \rho_X)$ of the range of S.

Proof. Page 890 of [41]. $\qquad\square$

C Analytic results

The section of analytic results is divided in two subsections, where we bound the condition number of the FALKON preconditioned linear system (14) and we decompose the excess risk of FALKON, with respect to analytical quantities that will be controlled in probability in the following sections.

C.1 Analytic results (I): Controlling condition number of W

First we characterize the matrix W defining the FALKON preconditioned linear system (14), with respect to the operators defined in Def. 6 (see next lemma) and in particular we characterize its condition number with respect to the norm of an auxiliary operator defined in Lemma 2. Finally we bound the norm of such operator with respect to analytical quantities more amenable to be bounded in probability (Lemma 3).

Lemma 1 (Characterization of W). *Let $\lambda \in \mathbb{R}$. The matrix W in Def. 4 is characterized by*

$$W = A^{-\top}V^*(\widehat{C}_n + \lambda I)VA^{-1}, \quad with \quad V = \sqrt{nM}\widehat{S}_M^* BA.$$

Moreover V is a partial isometry such that $V^*V = I_{q \times q}$ and VV^* with the same range of $\widehat{S}_M^*$.

Proof. By the characterization of K_{nM}, K_{MM} and $\widehat{C}_n$ in Prop. 3, we have

$$K_{nM}^\top K_{nM} + \lambda K_{MM} = nM \, (\widehat{S}_M \widehat{S}_n^* \widehat{S}_n \widehat{S}_M^* + \lambda \widehat{S}_M \widehat{S}_M^*)$$
$$= nM \, \widehat{S}_M(\widehat{S}_n^* \widehat{S}_n + \lambda I)\widehat{S}_M^* = nM \, \widehat{S}_M(\widehat{C}_n + \lambda I)\widehat{S}_M^*.$$

Now note that, by definition of B in Def. 3 and of V, we have

$$\sqrt{nM}\widehat{S}_M^* B = \sqrt{nM}\widehat{S}_M^* BAA^{-1} = VA^{-1},$$

so

$$W = B^\top(K_{nM}^\top K_{nM} + \lambda K_{MM})B = nM \, B^\top \widehat{S}_M(\widehat{C}_n + \lambda I)\widehat{S}_M^* B = A^{-\top}V^*(\widehat{C}_n + \lambda I)VA^{-1}.$$

16

The last step is to prove that V is a partial isometry. First we need a characterization of V that is obtained by expanding the definition of B,

$$V = \sqrt{nM}\widehat{S}_M^* BA = \sqrt{nM}\widehat{S}_M^* \frac{1}{\sqrt{n}}DQT^{-1}A^{-1}A = \sqrt{M}\widehat{S}_M^* DQT^{-1}. \tag{16}$$

By the characterization of V, the characterization of K_{MM} in Prop. 3 and the definition of Q, T in terms of $DK_{MM}D$ in Def. 3 , we have

$$V^*V = MT^{-\top}Q^\top D\,\widehat{S}_M\widehat{S}_M^*\,DQT^{-1} = T^{-\top}Q^\top\,DK_{MM}D\,QT^{-1} = T^{-\top}Q^\top QT^\top TQ^\top QT^{-1} = I.$$

Moreover, by the characterization of V, of $DK_{MM}D$ with respect to $\widehat{S}_M$, and of Q, T (Prop. 3 and Def. 3),

$$VV^*\widehat{S}_M^* D = M\,\widehat{S}_M^* DQT^{-1}T^{-\top}Q^\top D\widehat{S}_M\widehat{S}_M^* = \widehat{S}_M^* DQT^{-1}T^{-\top}Q^\top\,DK_{MM}D$$

$$= \widehat{S}_M^* DQT^{-1}T^{-\top}Q^\top QT^\top TQ^\top = \widehat{S}_M^* DQQ^\top = \widehat{S}_M^* D,$$

where the last step is due to the fact that the range of $QQ^\top$ is the one of $DK_{MM}D$ by definition (see Def. 3), and since $DK_{MM}D = MD\widehat{S}_M\widehat{S}_M^* D$ by Prop. 3, it is the same of $D\widehat{S}_M$. Note finally that the range of $\widehat{S}_M^* D$ is the same of $\widehat{S}_M^*$ since D is a diagonal matrix with strictly positive elements on the diagonal (see Def. 3). □

Lemma 2. *Let $\lambda > 0$ and W be as in Eq. 14. Let $E = A^{-\top}V^*(\widehat{C}_n - \widehat{G}_M)VA^{-1}$, with V defined in Lemma 1. Then W is characterized by*

$$W = I + E.$$

In particular, when $\|E\| < 1$,

$$\mathrm{cond}\,(W) \leq \frac{1 + \|E\|}{1 - \|E\|}.$$

Proof. Let Q, T, A, D as in Def. 3, and V as in Lemma 1. According to Lemma 1 we have

$$W = A^{-\top}V^*(\widehat{C}_n + \lambda I)VA^{-1} = A^{-\top}(V^*\widehat{C}_n V + \lambda I)A^{-1}.$$

Now we bound the largest and the smallest eigenvalue of W. First of all note that

$$A^{-\top}(V^*\widehat{C}_n V + \lambda I)A^{-1} = A^{-\top}(V^*\widehat{G}_M V + \lambda I)A^{-1} + A^{-\top}V^*(\widehat{C}_n - \widehat{G}_M)VA^{-1}, \tag{17}$$

where $\widehat{G}_M$ is defined in Def. 6. To study the first term, we need a preliminary result, which simplifies $\widehat{S}_M V$. By using the definition of V, the characterization of K_{MM} in terms of $\widehat{S}_M$ (Prop. 3), the definition of B (Def. 3), and finally the characterization of $DK_{MM}D$ in terms of Q, T (Def. 3), we have

$$D\widehat{S}_M V = \sqrt{nM}D\widehat{S}_M\widehat{S}_M^* BA = \sqrt{\frac{n}{M}}DK_{MM}BA = \frac{1}{\sqrt{M}}DK_{MM}D\,QT^{-1}$$

$$= \frac{1}{\sqrt{M}}QT^\top TQ^\top QT^{-1} = \frac{1}{\sqrt{M}}QT^\top.$$

Now we can simplify the first term. We express $\widehat{G}_M$ with respect to $\widehat{S}_M$, then we apply the identity above on $D\widehat{S}_M V$ and on its transpose, finally we recall the identity $A^\top A = \frac{1}{M}TT^\top + \lambda I$ from Def. 3, obtaining

$$A^{-\top}(V^*\widehat{G}_M V + \lambda I)A^{-1} = A^{-\top}(V^*\widehat{S}_M^* D^2\widehat{S}_M V + \lambda I)A^{-1} = A^{-\top}(\frac{1}{M}TQ^\top QT^\top + \lambda I)A^{-1} \tag{18}$$

$$= A^{-\top}(\frac{1}{M}TT^\top + \lambda I)A^{-1} = A^{-\top}A^\top AA^{-1} = I. \tag{19}$$

So, by defining $E := A^{-\top}V^*(\widehat{C}_n - \widehat{G}_M)VA^{-1}$, we have

$$W = I + E.$$

Note that E is compact and self-adjoint, by definition. Then, by Prop. 1, 2 we have that W admits a spectral decomposition as in Eq. 15. Let $\lambda_{\max}(W)$ and $\lambda_{\min}(W)$ be respectively the largest and the smallest eigenvalues of W, by Prop. 2, and considering that $-\|E\| \leq \lambda_j(E) \leq \|E\|$ (see Prop. 1) we have

$$\lambda_{\max}(W) = \sup_{j \in \mathbb{N}} 1 + \lambda_j(E) = 1 + \sup_{j \in \mathbb{N}}\lambda_j(E) = 1 + \lambda_{\max}(E) \leq 1 + \|E\|,$$

$$\lambda_{\min}(W) = \inf_{j \in \mathbb{N}} 1 + \lambda_j(E) = 1 + \inf_{j \in \mathbb{N}}\lambda_j(E) = 1 + \lambda_{\min}(E) \geq 1 - \|E\|.$$

Since W is self-adjoint and positive, when $\|E\| < 1$, by definition of condition number, we have

$$\mathrm{cond}\,(W) = \frac{\lambda_{\max}(W)}{\lambda_{\min}(W)} \leq \frac{1 + \|E\|}{1 - \|E\|}.$$

□

Lemma 3. *Let E be defined as in Lemma 2 and let $\widehat{G}_M$ as in Def. 6, then*

$$\|E\| \leq \|\widehat{G}_{M\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{G}_{M\lambda}^{-1/2}\|. \tag{20}$$

Proof. By multiplying and dividing by $\widehat{G}_{M\lambda} = \widehat{G}_M + \lambda I$ we have

$$\|E\| = \|A^{-\top}V^*\widehat{G}_{M\lambda}^{1/2}\ \widehat{G}_{M\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{G}_{M\lambda}^{-1/2}\ \widehat{G}_{M\lambda}^{1/2}VA^{-1}\|$$

$$\leq \|A^{-\top}V^*\widehat{G}_{M\lambda}^{1/2}\|^2\|\widehat{G}_{M\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{G}_{M\lambda}^{-1/2}\|.$$

Now, considering that $V^*V = I$ and the identity in Eq. (18), we have

$$\|A^{-\top}V^*\widehat{G}_{M\lambda}^{1/2}\|^2 = \|A^{-\top}V^*(\widehat{G}_M + \lambda I)VA^{-1}\| = \|A^{-\top}(V^*\widehat{G}_MV + \lambda I)A^{-1}\| = 1. \tag{21}$$

$\square$

C.2 Analytic results (II): The computational oracle inequality

In this subsection (Lemma 8) we bound the excess risk of FALKON with respect to the one of the exact Nyström estimator. First we prove that FALKON is equal to the exact Nyström estimator as the iterations go to infinity (Lemma 4, 5). Then in Lemma 8 (via Lemma 6, 7) we use functional analytic tools, together with results from operator theory to relate the weak convergence result of the conjugate gradient method on the chosen preconditioned problem, with the excess risk.

Lemma 4 (Representation of the FALKON estimator as vector in $\mathcal{H}$)**.** *Let $\lambda > 0$, $M, t \in \mathbb{N}$ and B as in Def. 3. The FALKON estimator as in Def. 4 is characterized by the vector $\widehat{f} \in \mathcal{H}$ as follows,*

$$\widehat{f}_{\lambda,M,t} = \sqrt{M}\ \widehat{S}_M^*B\beta_t, \tag{22}$$

where $\beta_t \in \mathbb{R}^q$ denotes the vector resulting from t iterations of the conjugate gradient algorithm applied to the linear system in Def. 4.

Proof. According to the definition of $\widehat{f}_{\lambda,M,t}(\cdot)$ in Def.4 and the definition of the operator $\widehat{S}_M$ in Def. 6, denoting with $\alpha \in \mathbb{R}^M$ the vector $B\beta_t$, we have that

$$\widehat{f}_{\lambda,M,t}(x) = \sum_{i=1}^{M}\alpha_i K(x, \tilde{x}_i) = \left\langle K_x, \sum_{i=1}^{M}\alpha_i K_{\tilde{x}_i}\right\rangle_{\mathcal{H}} = \left\langle K_x, \sqrt{M}\ \widehat{S}_M^*\alpha\right\rangle_{\mathcal{H}},$$

for any $x \in X$. Then the vector in $\mathcal{H}$ representing the function $\widehat{f}_{\lambda,M,t}(\cdot)$ is

$$\widehat{f}_{\lambda,M,t} = \sqrt{M}\ \widehat{S}_M^*\alpha = \sqrt{M}\ \widehat{S}_M^*B\beta_t.$$

$\square$

Lemma 5 (Representation of the Nyström estimator as a vector in $\mathcal{H}$)**.** *Let $\lambda > 0, M \in \mathbb{N}$, and B as in Def. 3. The exact Nyström estimator, in Eq.(7) and Eq. (8) is characterized by the vector $\widetilde{f} \in \mathcal{H}$ as follows*

$$\widetilde{f}_{\lambda,M} = \sqrt{M}\ \widehat{S}_M^*B\beta_\infty, \tag{23}$$

where $\beta_\infty = W^{-1}B^\top K_{nM}^\top\widehat{y}$ is the vector resulting from infinite iterations of the conjugate gradient algorithm applied to the linear system in Eq. (14).

Proof. For the same reasoning in the proof of Lemma 4, we have that the FALKON estimator with infinite iterations is characterized by the following vector in $\mathcal{H}$

$$\widetilde{f}_{\lambda,M} = \sqrt{M}\ \widehat{S}_M^*B\beta_\infty.$$

To complete the proof, we need to prove 1) that $\beta_\infty = W^{-1}B^\top K_{nM}\widehat{y}$ and 2) that $\widetilde{f}_{\lambda,M}$ above, corresponds to the exact Nyström estimator, as in Eq. (8).

Now we characterize β_∞. First, by the characterization of W in Lemma 1 and the fact that $V^*V = I$, we have

$$W = A^{-\top}V^*(\widehat{C}_n + \lambda I)VA^{-1} = A^{-\top}(V^*\widehat{C}_nV + \lambda I)A^{-1}. \tag{24}$$

Since $\widehat{C}_n$ is a positive operator (see Def. 6) A is invertible and $\lambda > 0$, then W is a symmetric and positive definite matrix. The positive definiteness of W implies that it is invertible and that is has a finite condition number, making the conjugate gradient algorithm to converge to the solution of the system in Eq. (14) (Thm. 6.6 of [25] and Eq. 6.107). So we can explicitly characterize β_∞, by the solution of the system in Eq. (14), that is

$$\beta_\infty = W^{-1}B^\top K_{nM}^\top\widehat{y}. \tag{25}$$

So we proved that $\widetilde{f}_{\lambda,M} \in \mathcal{H}$, with the above characterization of β_∞, corresponds to FALKON with infinite iterations. Now we show that $\widetilde{f}_{\lambda,M}$ is equal to the Nyström estimator given in [13]. First we need to study $\widehat{S}_M^* BW^{-1}B^\top \widehat{S}_M$. By the characterization of W in Eq. (24), the identity $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$, valid for any A, B, C bounded invertible operators, and the definition of V (Lemma 1),

$$\widehat{S}_M^* BW^{-1}B^\top \widehat{S}_M = \widehat{S}_M^* B \left(A^{-\top}(V^*\widehat{C}_n V + \lambda I)A^{-1} \right)^{-1} B^\top \widehat{S}_M \tag{26}$$

$$= \widehat{S}_M^* BA(V^*\widehat{C}_n V + \lambda I)^{-1}A^\top B^\top \widehat{S}_M \tag{27}$$

$$= \frac{1}{Mn} V(V^*\widehat{C}_n V + \lambda I)^{-1}V^*. \tag{28}$$

By expanding β_∞, K_{nM} (see Lemma 3) in $\widetilde{f}_{\lambda,M}$,

$$\widetilde{f}_{\lambda,M} = \sqrt{M}\,\widehat{S}_M^* B\beta_\infty = \sqrt{M}\,\widehat{S}_M^* BW^{-1}B^\top K_{nM}^\top \widehat{y} = \sqrt{n}M\,\widehat{S}_M^* BW^{-1}B^\top \widehat{S}_M \widehat{S}_n^* \widehat{y} \tag{29}$$

$$= \frac{1}{\sqrt{n}}\,V(V^*\widehat{C}_n V + \lambda I)^{-1}V^* \widehat{S}_n^* \widehat{y}. \tag{30}$$

Now by Lemma 2 of [13] with $Z_m = \widehat{S}_M$, we know that the exact Nyström solution is characterized by the vector $\bar{f} \in \mathcal{H}$ defined as follows

$$\bar{f} = \frac{1}{\sqrt{n}}\bar{V}(\bar{V}^*\widehat{C}_n \bar{V} + \lambda I)^{-1}\bar{V}^* \widehat{S}_n^* \widehat{y},$$

with $\bar{V}$ a partial isometry, such that $\bar{V}^*\bar{V} = I$ and $\bar{V}\bar{V}^*$ with the same range of $\widehat{S}_M^*$. Note that, by definition of V in Lemma 1, we have that it is a partial isometry such that $V^*V = I$ and VV^* with the same range of $\widehat{S}_M^*$. This implies that $\bar{V} = VG$, for an orthogonal matrix $G \in \mathbb{R}^{q \times q}$. Finally, exploiting the fact that $G^{-1} = G^\top$, that $GG^\top = G^\top G = I$ and that for three invertible matrices A, B, C we have $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$,

$$\bar{f} = \frac{1}{\sqrt{n}}\bar{V}(\bar{V}^*\widehat{C}_n \bar{V} + \lambda I)^{-1}\bar{V}^* \widehat{S}_n^* \widehat{y} = \frac{1}{\sqrt{n}}VG\left(G^\top(V^*\widehat{C}_n V + \lambda I)G\right)^{-1}G^\top V^* \widehat{S}_n^* \widehat{y}$$

$$= \frac{1}{\sqrt{n}}VGG^\top\left(V^*\widehat{C}_n V + \lambda I\right)^{-1}GG^\top V^* \widehat{S}_n^* \widehat{y} = \frac{1}{\sqrt{n}}V\left(V^*\widehat{C}_n V + \lambda I\right)^{-1}V^* \widehat{S}_n^* \widehat{y} = \widetilde{f}_{\lambda,M}.$$

$\square$

The next lemma is necessary to prove Lemma 8.

Lemma 6. *When $\lambda > 0$ and B is as in Def. 3. then*

$$\sqrt{M}\|S\widehat{S}_M^* BW^{-1/2}\| \leq n^{-1/2}\|S\widehat{C}_{n\lambda}^{-1/2}\|.$$

Proof. By the fact that identity $\|Z\|^2 = \|ZZ^*\|$ valid for any bounded operator Z and the identity in Eq. 26, we have

$$M\|S\widehat{S}_M^* BW^{-1/2}\|^2 = M\|S\widehat{S}_M^* BW^{-1}B^\top \widehat{S}_M S^*\| = \frac{1}{n}\|SV(V^*\widehat{C}_n V + \lambda I)^{-1}V^* S^*\|$$

$$= \frac{1}{n}\|SV(V^*\widehat{C}_n V + \lambda I)^{-1/2}\|^2.$$

Denote with $\widehat{C}_{n\lambda}$ the operator $\widehat{C}_n + \lambda I$, by dividing and multiplying for $\widehat{C}_{n\lambda}^{-1/2}$, we have

$$SV(V^*\widehat{C}_n V + \lambda I)^{-1/2} = S\widehat{C}_{n\lambda}^{-1/2}\ \widehat{C}_{n\lambda}^{1/2}V(V^*\widehat{C}_n V + \lambda I)^{-1/2}.$$

The second term is equal to 1, indeed, since $V^*\widehat{C}_{n\lambda}V = V^*\widehat{C}_n V + \lambda I$, and $\|Z\|^2 = \|Z^*Z\|$, for any bounded operator Z, we have

$$\|\widehat{C}_{n\lambda}^{1/2}V(V^*\widehat{C}_n V + \lambda I)^{-1/2}\|^2 = \|(V^*\widehat{C}_n V + \lambda I)^{-1/2}V^*\widehat{C}_{n\lambda}V(V^*\widehat{C}_n V + \lambda I)^{-1/2}\| \tag{31}$$

$$= \|(V^*\widehat{C}_n V + \lambda I)^{-1/2}(V^*\widehat{C}_n V + \lambda I)(V^*\widehat{C}_n V + \lambda I)^{-1/2}\| \tag{32}$$

$$= 1. \tag{33}$$

Finally

$$\sqrt{M}\|S\widehat{S}_M^* BW^{-1/2}\| = \frac{1}{\sqrt{n}}\|SV(V^*\widehat{C}_n V + \lambda I)^{-1/2}\|$$

$$\leq \frac{1}{\sqrt{n}}\|S\widehat{C}_{n\lambda}^{-1/2}\|\|\widehat{C}_{n\lambda}^{1/2}V(V^*\widehat{C}_n V + \lambda I)^{-1/2}\|$$

$$\leq n^{-1/2}\|S\widehat{C}_{n\lambda}^{-1/2}\|.$$

$\square$

The next lemma is necessary to prove Lemma 8.

Lemma 7. *For any $\lambda > 0$, let β_∞ be the vector resulting from infinite iterations of the conjugate gradient algorithm applied to the linear system in Eq.* (14). *Then*

$$\|W^{1/2}\beta_\infty\|_{\mathbb{R}^q} \leq \|\widehat{y}\|_{\mathbb{R}^n}.$$

Proof. First we recall the characterization of β_∞ from Lemma 5,

$$\beta_\infty = W^{-1}B^\top K_{nM}^\top \widehat{y}.$$

So, by the characterization of K_{nM} in terms of $\widehat{S}_n, \widehat{S}_M$ (Prop. 3),

$$W^{1/2}\beta_\infty = W^{1/2}W^{-1}B^\top K_{nM}^\top \widehat{y} = \sqrt{nM}\, W^{-1/2}B^\top \widehat{S}_M \widehat{S}_n^* \widehat{y}.$$

Then, by applying the characterization of $\widehat{S}_M^* BW^{-1}B^\top \widehat{S}_M$ in terms of V, in Eq. 26

$$\|W^{1/2}\beta_\infty\|_{\mathbb{R}^q}^2 = nM\, \|W^{-1/2}B^\top \widehat{S}_M \widehat{S}_n^* \widehat{y}\|_{\mathbb{R}^q}^2 = nM\, \widehat{y}^\top \widehat{S}_n \widehat{S}_M^* BW^{-1}B^\top \widehat{S}_M \widehat{S}_n^* \widehat{y}$$

$$= \widehat{y}^\top \widehat{S}_n V (V^* \widehat{C}_n V + \lambda I)^{-1} V^* \widehat{S}_n^* \widehat{y} = \|(V^* \widehat{C}_n V + \lambda I)^{-1/2} V^* \widehat{S}_n^* \widehat{y}\|_{\mathbb{R}^q}^2.$$

Finally

$$\|(V^* \widehat{C}_n V + \lambda I)^{-1/2} \widehat{S}_n^* \widehat{y}\|_{\mathbb{R}^q} \leq \|(V^* \widehat{C}_n V + \lambda I)^{-1/2} \widehat{S}_n^*\| \|\widehat{y}\|_{\mathbb{R}^n}.$$

Note that

$$\|(V^* \widehat{C}_n V + \lambda I)^{-1/2} \widehat{S}_n^*\| \leq 1,$$

indeed

$$\|(V^* \widehat{C}_n V + \lambda I)^{-1/2} \widehat{S}_n^*\| \leq \|(V^* \widehat{C}_n V + \lambda I)^{-1/2} \widehat{C}_{n\lambda}^{1/2}\| \|\widehat{C}_{n\lambda}^{-1/2} \widehat{S}_n^*\|,$$

and the first term is equal to 1 by Eq. (31), moreover by definition of $\widehat{C}_n$ (Def. 6),

$$\|\widehat{C}_{n\lambda}^{-1/2} \widehat{S}_n^*\|^2 = \|\widehat{C}_{n\lambda}^{-1/2} \widehat{C}_n \widehat{C}_{n\lambda}^{-1/2}\| = \|\widehat{C}_{n\lambda}^{-1/2} \widehat{C}_n^{1/2}\|^2 = \sup_{\sigma \in \sigma(\widehat{C}_n)} \frac{\sigma}{\sigma + \lambda} \leq 1,$$

where $\sigma(\widehat{C}_n) \subset [0, \|\widehat{C}_n\|]$ is the set of eigenvalues of $\widehat{C}_n$. $\qquad\qquad\square$

Lemma 8. *Let $M \in \mathbb{N}$, $\lambda > 0$ and B satisfying Def. 4. Let $\widehat{f}_{\lambda,M,t}$ be the FALKON estimator after $t \in \mathbb{N}$ iterations and $\widetilde{f}_{\lambda,M}$ the exact Nyström estimator as in Eq. 7, 8. Let $c_0 \geq 0$ such that*

$$\|S\widehat{C}_{n\lambda}^{-1/2}\| \leq c_0,$$

then

$$\mathcal{R}(\widehat{f}_{\lambda,M,t})^{1/2} \leq \mathcal{R}(\widetilde{f}_{\lambda,M})^{1/2} + 2c_0\,\widehat{v}\left(1 - \frac{2}{\sqrt{cond(W)} + 1}\right)^t,$$

where $\widehat{v}^2 = \frac{1}{n}\sum_{i=1}^n y_i^2$.

Proof of Lemma 8. By Prop. 4 we have that for any $f \in \mathcal{H}$

$$(\mathcal{E}(f) - \inf_{f \in \mathcal{H}} \mathcal{E}(f))^{1/2} = \|Sf - Pf_\rho\|_{\rho_X},$$

with $P : L^2(X, \rho_X) \to L^2(X, \rho_X)$ the orthogonal projection operator whose range is the closure of the range of S in $L^2(X, \rho_X)$. Let $\widehat{f}_{\lambda,M,t} \in \mathcal{H}$ and $\widetilde{f}_{\lambda,M} \in \mathcal{H}$ be respectively the Hilbert vector representation of the FALKON estimator and of the exact Nyström estimator (Lemma 4 and Lemma 5). By adding and subtracting $\widetilde{f}_{\lambda,M}$ we have

$$|\mathcal{E}(\widehat{f}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)|^{1/2} = \|S\widehat{f}_{\lambda,M,t} - Pf_\rho\|_{\rho_X} = \|S(\widehat{f}_{\lambda,M,t} - \widetilde{f}_{\lambda,M}) + (S\widetilde{f}_{\lambda,M} - Pf_\rho)\|_{\rho_X}$$

$$\leq \|S(\widehat{f}_{\lambda,M,t} - \widetilde{f}_{\lambda,M})\|_{\rho_X} + \|S\widetilde{f}_{\lambda,M} - Pf_\rho\|_{\rho_X}$$

$$= \|S(\widehat{f}_{\lambda,M,t} - \widetilde{f}_{\lambda,M})\|_{\rho_X} + |\mathcal{E}(\widetilde{f}_{\lambda,M}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)|^{1/2}.$$

In particular, by expanding the definition of $\widehat{f}_{\lambda,M,t}, \widetilde{f}_{\lambda,M}$ from Lemma 4 and Lemma 5, we have

$$\|S(\widehat{f}_{\lambda,M,t} - \widetilde{f}_{\lambda,M})\|_{\rho_X} = \sqrt{M}\|S\widehat{S}_M^* B(\beta_t - \beta_\infty)\|_{\rho_X},$$

where $\beta_t \in \mathbb{R}^q$ and $\beta_\infty \in \mathbb{R}^q$ denote respectively the vector resulting from t iterations and infinite iterations of the conjugate gradient algorithm applied to the linear system in Eq. (14). Since W is symmetric positive definite

when $\lambda > 0$ (see proof of Lemma 5), we can apply the standard convergence results for the conjugate gradient algorithm (Thm. 6.6 of [25], in particular Eq. 6.107), that is the following

$$\|W^{1/2}(\beta_t - \beta_\infty)\|_{\mathbb{R}^q} \le q(W,t)\|W^{1/2}\beta_\infty\|_{\mathbb{R}^q}, \quad \text{with} \quad q(W,t) = 2\left(1 - \frac{2}{\sqrt{\text{cond}(W)} + 1}\right)^t.$$

So by dividing and multiplying by $W^{1/2}$ we have

$$\|S(\widehat{f}_{\lambda,M,t} - \widetilde{f}_{\lambda,M})\|_{\rho_X} = \sqrt{M}\|S\widehat{S}_M^* B(\beta_t - \beta_\infty)\|_{\rho_X} = \sqrt{M}\|S\widehat{S}_M^* BW^{-1/2}W^{1/2}(\beta_t - \beta_\infty)\|_{\rho_X}$$
$$\le \sqrt{M}\|S\widehat{S}_M^* BW^{-1/2}\|\|W^{1/2}(\beta_t - \beta_\infty)\|_{\mathbb{R}^q}$$
$$\le q(W,t)\sqrt{M}\|S\widehat{S}_M^* BW^{-1/2}\|\|W^{1/2}\beta_\infty\|_{\mathbb{R}^q}.$$

Finally, the term $\sqrt{M}\|S\widehat{S}_M^* BW^{-1/2}\|$ is bounded in Lemma 6 as

$$\sqrt{M}\|S\widehat{S}_M^* BW^{-1/2}\| \le \frac{1}{\sqrt{n}}\|S\widehat{C}_{n\lambda}^{-1/2}\| \le \frac{c_0}{\sqrt{n}},$$

while, for the term $\|W^{1/2}\beta_\infty\|_{\mathbb{R}^q}$, by Lemma 7, we have

$$\|W^{1/2}\beta_\infty\|_{\mathbb{R}^q} \le \|\widehat{y}\|_{\mathbb{R}^n} = (\sum_{i=1}^n y_i^2)^{1/2} = \sqrt{n}\sqrt{\frac{\sum_{i=1}^n y_i^2}{n}} = \sqrt{n}\widehat{v}.$$

$\square$

D   Probabilistic Estimates

In Lemma 9, 10 we provide probabilistic estimates of $\|E\|$, the quantity needed to bound the condition number of the preconditioned linear system of FALKON (see Lemma 1, 3). In particular Lemma 9, analyzes the case when the Nyström centers are selected with uniform sampling, while Lemma 10, considers the case when the Nyström centers are selected via approximate leverage scores sampling.

Now we are ready to provide probabilistic estimates for uniform sampling.

Lemma 9. *Let $\eta \in [0,1)$ and $\delta \in (0,1]$. When $\tilde{x}_1, \ldots, \tilde{x}_M$ are selected via Nyström uniform sampling (see Sect. A), $0 < \lambda \le \|C\|$, $M \le n$ and*

$$M \ge 4\left[\frac{1}{2} + \frac{1}{\eta} + \left(\frac{3 + 7\eta}{3 + 3\eta}\right)\left(1 + \frac{2}{\eta}\right)^2 \mathcal{N}_\infty(\lambda)\right]\log\frac{8\kappa^2}{\lambda\delta}, \tag{34}$$

then the following hold with probability at least $1 - \delta$,

$$\|C_\lambda^{-1/2}(C - \widehat{C}_n)C_\lambda^{-1/2}\| < \eta, \quad \|\widehat{G}_{M\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{G}_{M\lambda}^{-1/2}\| < \eta.$$

Proof. First of all, note that since the Nyström centers are selected by uniform sampling. Then $\tilde{x}_1, \ldots, \tilde{x}_M$ are independently and identically distributed according to ρ_X and moreover D is the identity matrix. So

$$\widehat{G}_M = \widehat{S}_M^* D^2 \widehat{S}_M = \widehat{S}_M^* \widehat{S}_M = \widehat{C}_M.$$

Note that, by multiplying and dividing by C_λ,

$$\|\widehat{G}_{M\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{G}_{M\lambda}^{-1/2}\| = \|\widehat{C}_{M\lambda}^{-1/2}(\widehat{C}_n - \widehat{C}_M)\widehat{C}_{M\lambda}^{-1/2}\|$$
$$= \|\widehat{C}_{M\lambda}^{-1/2}C_\lambda^{1/2}C_\lambda^{-1/2}(\widehat{C}_n - \widehat{C}_M)C_\lambda^{-1/2}C_\lambda^{1/2}\widehat{C}_{M\lambda}^{-1/2}\|$$
$$\le \|\widehat{C}_{M\lambda}^{-1/2}C_\lambda^{1/2}\|^2\|C_\lambda^{-1/2}(\widehat{C}_n - \widehat{C}_M)C_\lambda^{-1/2}\|$$
$$\le (1 - \lambda_{\max}(C_\lambda^{-1/2}(C - \widehat{C}_M)C_\lambda^{-1/2}))^{-1}\|C_\lambda^{-1/2}(\widehat{C}_n - \widehat{C}_M)C_\lambda^{-1/2}\|$$

where the last step is due to Prop. 9 of [14]. Moreover note that

$$\lambda_{\max}(C_\lambda^{-1/2}(C - \widehat{C}_M)C_\lambda^{-1/2}) \le \|C_\lambda^{-1/2}(C - \widehat{C}_M)C_\lambda^{-1/2}\|.$$

Let $\mu = \frac{\delta}{2}$. Note that $\widehat{C}_M = \frac{1}{M}\sum_{i=1}^M v_i \otimes v_i$ with v_i the random variable $v_i = K_{\tilde{x}_i}$ (see Prop. 3) and, since $\tilde{x}_1, \ldots, \tilde{x}_M$ are i.i.d. w.r.t. ρ_X, by the characterization of C in Prop. 3, we have for any $1 \le i \le M$,

$$\mathbb{E}v_i \otimes v_i = \int_X K_x \otimes K_x d\rho_X(x) = C.$$

Then, by considering that $\|v\| = \|K_x\| \le \kappa^2$ a. e., we can apply Prop. 7 of [14], obtaining

$$\|C_\lambda^{-1/2}(C - \widehat{C}_M)C_\lambda^{-1/2}\| \le \frac{2d(1 + \mathcal{N}_\infty(\lambda))}{3M} + \sqrt{\frac{2d\mathcal{N}_\infty(\lambda)}{3M}}, \quad d = \log\frac{4\kappa^2}{\lambda\mu},$$

with probability at least $1 - \mu$. Note that, when M satisfies Eq (34), we have $\|C_\lambda^{-1/2}(C - \widehat{C}_M)C_\lambda^{-1/2}\| < \eta/(2 + \eta)$. By repeating the same reasoning for C_n, we have

$$\|C_\lambda^{-1/2}(C - \widehat{C}_n)C_\lambda^{-1/2}\| \le \frac{2d(1 + \mathcal{N}_\infty(\lambda))}{3n} + \sqrt{\frac{2d\mathcal{N}_\infty(\lambda)}{3n}}, \quad d = \log\frac{4\kappa^2}{\lambda\mu},$$

with probability $1 - \mu$. Since $n \ge M$ and M satisfying Eq. (34), we have automatically that $\|C_\lambda^{-1/2}(C - \widehat{C}_n)C_\lambda^{-1/2}\| < \eta/(2 + \eta)$.

Finally note that, by adding and subtracting C,

$$\|C_\lambda^{-1/2}(\widehat{C}_n - \widehat{C}_M)C_\lambda^{-1/2}\| = \|C_\lambda^{-1/2}((\widehat{C}_n - C) + (C - \widehat{C}_M))C_\lambda^{-1/2}\|$$
$$\le \|C_\lambda^{-1/2}(C - \widehat{C}_n)C_\lambda^{-1/2}\| + \|C_\lambda^{-1/2}(C - \widehat{C}_M)C_\lambda^{-1/2}\|.$$

So by performing the intersection bound of the two previous events, we have

$$\|\widehat{C}_{M\lambda}^{-1/2}(\widehat{C}_n - \widehat{C}_M)\widehat{C}_{M\lambda}^{-1/2}\| \le (1 - \|C_\lambda^{-1/2}(C - \widehat{C}_n)C_\lambda^{-1/2}\|)^{-1} \times$$
$$\times \left(\|C_\lambda^{-1/2}(C - \widehat{C}_n)C_\lambda^{-1}\| + \|C_\lambda^{-1/2}(C - \widehat{C}_M)C_\lambda^{-1/2}\|\right) < \eta,$$

with probability at least $1 - 2\mu$. The last step consists in substituting μ with $\delta/2$. $\qquad\square$

The next lemma gives probabilistic estimates for $\|E\|$, that is the quantity needed to bound the condition number of the preconditioned linear system of FALKON (see Lemma 1, 3), when the Nyström centers are selected via approximate leverage scores sampling.

Lemma 10. *Let $\eta > 0$, $\delta \in (0, 1]$, $n, M \in \mathbb{N}$, $q \ge 1$ and $\lambda_0 > 0$. Let $x_1, \ldots, x_n$ be independently and identically distributed according to ρ_X. Let $\tilde{x}_1, \ldots, \tilde{x}_M$ be randomly selected from $x_1 \ldots, x_n$, by using the (q, λ_0, δ)-approximate leverage scores (see Def. 1 and discussion below), with $\lambda_0 \vee \frac{19\kappa^2}{n}\log\frac{n}{2\delta} \le \lambda \le \|C\|$. When $n \ge 405\kappa^2 \vee 67\kappa^2\log\frac{12\kappa^2}{\delta}$ and*

$$M \ge \left[2 + \frac{2}{\eta} + \frac{18(\eta^2 + 5\eta + 4)q^2}{\eta^2}\mathcal{N}(\lambda)\right]\log\frac{8\kappa^2}{\lambda\delta}, \tag{35}$$

then the following hold with probability at least $1 - \delta$,

$$\|\widehat{G}_{M\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{G}_{M\lambda}^{-1/2}\| < \eta, \quad \|C_\lambda^{-1/2}(C - \widehat{C}_n)C_\lambda^{-1/2}\| < \eta.$$

Proof. By multiplying and dividing by $\widehat{C}_{n\lambda} = \widehat{C}_n + \lambda I$, we have

$$\|\widehat{G}_{M\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{G}_{M\lambda}^{-1/2}\| = \|\widehat{G}_{M\lambda}^{-1/2}\widehat{C}_{n\lambda}^{1/2}\widehat{C}_{n\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{C}_{n\lambda}^{-1/2}\widehat{C}_{n\lambda}^{1/2}\widehat{G}_{M\lambda}^{-1/2}\|$$
$$\le \|\widehat{G}_{M\lambda}^{-1/2}\widehat{C}_{n\lambda}^{1/2}\|^2\|\widehat{C}_{n\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{C}_{n\lambda}^{-1/2}\|$$
$$\le (1 - \lambda_{\max}(\widehat{C}_{n\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{C}_{n\lambda}^{-1/2}))^{-1}\|\widehat{C}_{n\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{C}_{n\lambda}^{-1/2}\|$$

where the last step is due to Prop. 9 of [14]. Note that

$$\lambda_{\max}(\widehat{C}_{n\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{C}_{n\lambda}^{-1/2}) \le \|\widehat{C}_{n\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{C}_{n\lambda}^{-1/2}\|,$$

thus

$$\|\widehat{G}_{M\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{G}_{M\lambda}^{-1/2}\| \le \frac{t}{1 - t},$$

with $t = \|\widehat{C}_{n\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{C}_{n\lambda}^{-1/2}\|$. Now we bound t. We denote with $l_\lambda(j), \widehat{l}_\lambda(j)$, respectively the leverage scores and the (q, λ_0, δ)-approximate leverage score associated to the point x_j, as in Def. 1 and discussion above. First we need some considerations on the leverage scores. By the spectral theorem and the fact that $K_{nn} = n\,\widehat{S}_n\widehat{S}_n^*$ (see Prop. 3), we have

$$l_\lambda(j) = (K_{nn}(K_{nn} + \lambda nI)^{-1})_{jj} = e_j^\top \widehat{S}_n\widehat{S}_n^*(\widehat{S}_n\widehat{S}_n^* + \lambda I)^{-1}e_j = e_j^\top \widehat{S}_n(\widehat{S}_n^*\widehat{S}_n + \lambda I)^{-1}\widehat{S}_n^* e_j$$
$$= \frac{1}{n}\left\langle K_{x_j}, \widehat{C}_{n\lambda}^{-1}K_{x_j}\right\rangle = \frac{1}{n}\|\widehat{C}_{n\lambda}^{-1/2}K_{x_j}\|^2.$$

22

for any $1 \leq j \leq n$. Moreover, by the characterization of $\widehat{C}_n$ in Prop. 3, we have

$$\frac{1}{n} \sum_{j=1}^{n} l_\lambda(j) = \frac{1}{n} \sum_{j=1}^{n} \left\langle K_{x_j}, (\widehat{C}_n + \lambda)^{-1} K_{x_j} \right\rangle_{\mathcal{H}} = \frac{1}{n} \sum_{j=1}^{n} \mathrm{Tr}((\widehat{C}_n + \lambda)^{-1} (K_{x_j} \otimes K_{x_j}))$$

$$= \mathrm{Tr}((\widehat{C}_n + \lambda)^{-1} \frac{1}{n} \sum_{j=1}^{n} (K_{x_j} \otimes K_{x_j})) = \mathrm{Tr}(\widehat{C}_{n\lambda}^{-1} \widehat{C}_n).$$

Since the Nyström points are selected by using the (q, λ_0, δ)-approximate leverage scores, then $\widetilde{x}_t = x_{i_t}$ for $1 \leq t \leq M$, where $i_1, \ldots, i_M \in \{1, \ldots, n\}$ is the sequence of indexes obtained by approximate leverage scores sampling (see Sect. A). Note that $i_1, \ldots, i_M$ are independent random indexes, distributed as follows: for $1 \leq t \leq M$,

$$i_t = j, \quad \text{with probability} \quad p_j = \frac{\widehat{l}_\lambda(j)}{\sum_{h=1}^{n} \widehat{l}_\lambda(h)}, \quad \forall\, 1 \leq j \leq n.$$

Then, by recalling the definition of $\widehat{G}_M$ with respect to the matrix D defined as in Def. 2 and by Prop. 3 we have,

$$\widehat{G}_M = \widehat{S}_M^* D^2 \widehat{S}_M = \frac{1}{M} \sum_{t=1}^{M} \frac{1}{np_{i_t}} K_{x_{i_t}} \otimes K_{x_{i_t}}.$$

Consequently $\widehat{G}_M = \frac{1}{M} \sum_{i=1}^{M} v_i \otimes v_i$, where $(v_i)_{i=1}^{M}$ are independent random variables distributed in the following way

$$v_i = \frac{1}{\sqrt{p_j n}} K_{x_j}, \quad \text{with probability} \ p_j, \quad \forall\, 1 \leq j \leq n.$$

Now we study the moments of $\widehat{G}_M$ as a sum of independent random matrices, to apply non-commutative Bernstein inequality (e.g. Prop. 7 of [14]). We have that, for any $1 \leq i \leq M$

$$\mathbb{E} v_i \otimes v_i = \sum_{j=1}^{n} p_j \left(\frac{1}{p_j n} K_{x_j} \otimes K_{x_j} \right) = \widehat{C}_n,$$

$$\left\langle v_i, \widehat{C}_{n\lambda}^{-1} v_i \right\rangle_{\mathcal{H}} \leq \sup_{1 \leq j \leq n} \frac{\|\widehat{C}_{n\lambda}^{-1/2} K_{x_j}\|^2}{p_j n} = \sup_{1 \leq j \leq n} \frac{l_\lambda(j)}{p_j n} = \sup_{1 \leq j \leq n} \frac{l_\lambda(j)}{\widehat{l}_\lambda(j)} \frac{1}{n} \sum_{h=1}^{n} \widehat{l}_\lambda(h)$$

$$\leq q \frac{1}{n} \sum_{h=1}^{n} \widehat{l}_\lambda(h) \leq q^2 \frac{1}{n} \sum_{h=1}^{n} l_\lambda(h) = q^2 \,\mathrm{Tr}(\widehat{C}_{n\lambda}^{-1} \widehat{C}_n),$$

for all $1 \leq j \leq n$. Denote with $\widehat{\mathcal{N}}(\lambda)$, the quantity $\mathrm{Tr}(\widehat{C}_{n\lambda}^{-1} \widehat{C}_n)$, by applying Prop. 7 of [14], we have

$$\|\widehat{C}_{n\lambda}^{-1/2} (\widehat{C}_n - \widehat{G}_M) \widehat{C}_{n\lambda}^{-1/2}\| \leq \frac{2d(1 + q^2 \widehat{\mathcal{N}}(\lambda))}{3M} + \sqrt{\frac{2dq^2 \widehat{\mathcal{N}}(\lambda)}{M}}, \quad d = \log \frac{\kappa^2}{\lambda \mu}.$$

with probability at least $1 - \mu$. The final step consist in bounding the empirical intrinsic dimension $\widehat{\mathcal{N}}(\lambda)$ with respect to intrinsic dimension $\mathcal{N}(\lambda)$, for which we use Prop. 1 of [13], obtaining

$$\widehat{\mathcal{N}}(\lambda) \leq 2.65 \mathcal{N}(\lambda),$$

with probability at least $1 - \mu$, when $n \geq 405\kappa^2 \vee 67\kappa^2 \log \frac{6\kappa^2}{\mu}$ and $\frac{19\kappa^2}{n} \log \frac{n}{4\mu} \leq \lambda \leq \|C\|$. By intersecting the events, we have

$$\|\widehat{C}_{n\lambda}^{-1/2} (\widehat{C}_n - \widehat{G}_M) \widehat{C}_{n\lambda}^{-1/2}\| \leq \frac{5.3d(1 + q^2 \mathcal{N}(\lambda))}{3M} + \sqrt{\frac{5.3dq^2 \mathcal{N}(\lambda)}{M}}, \quad d = \log \frac{\kappa^2}{\lambda \mu}.$$

with probability at least $1 - 2\mu$. The last step consist in substituting μ with $\mu = \delta/2$. Thus, by selecting M as in Eq. 35, we have

$$t = \|\widehat{C}_{n\lambda}^{-1/2} (\widehat{C}_n - \widehat{G}_M) \widehat{C}_{n\lambda}^{-1/2}\| < \frac{\eta}{1 + \eta}.$$

That implies,

$$\|\widehat{G}_{M\lambda}^{-1/2} (\widehat{C}_n - \widehat{G}_M) \widehat{G}_{M\lambda}^{-1/2}\| < \frac{t}{1 - t} < \eta.$$

$\square$

E   Proof of Main Results

In this section we prove the main results of the paper. This section is divided in three subsections. In the first, we specify the computational oracle inequality for Nyström with uniform sampling, in the second we specify the computational oracle inequality for Nyström with approximate leverage scores sampling (see Sect. D for a definition), while the third subsection contains the proof of the main theorem presented in the paper.

Now we give a short sketch of the structure of the proofs. The definition of the general version of the FALKON algorithm (taking into account leverage scores and non invertible K_{MM}) is given in Sect. A. In Sect. B the notation and basic definition required for the rest of the analisys are provided.

Our starting point is the analysis of the basic Nyström estimator given in [13]. The key novelty is the quantification of the approximations induced by the preconditioned iterative solver by relating its excess risk to the one of the basic Nyström estimator.

A computational oracle inequality. First we prove that FALKON is equal to the exact Nyström estimator as the iterations go to infinity (Lemma 5, Sect. C). Then, in Lemma 8 (see also Lemma 6, 7, Sect. C) we show how optimization guarantees can be used to derive statistical results. More precisely, while optimization results in machine learning typically derives guarantees on empirical minimization problems, we show, using analytic and probabilistic tools, how these results can be turned into guarantees on the expected risks. Finally, in the proof of Thm. 1 we concentrate the terms of the inequality. The other key point is the study of the behavior of the condition number of $B^\top H B$ with B given in (10).

Controlling the condition number of $B^\top H B$. Let C_n, C_M be the empirical correlation operators in $\mathcal{H}$ associated respectively to the training set and the Nyström points $C_n = \frac{1}{n}\sum_{i=1}^n K_{x_i} \otimes K_{x_i}, C_M = \frac{1}{M}\sum_{j=1}^M K_{\tilde{x}_j} \otimes K_{\tilde{x}_j}$. In Lemma 1, Sect. C, we prove that $B^\top H B$ is equivalent to $A^{-\top} V^*(C_n + \lambda I)V A^{-1}$ for a suitable partial isometry V. Then in Lemma 2, Sect. C, we split it in two components

$$B^\top H B = A^{-\top} V^*(C_M + \lambda I)V A^{-1} \; + \; A^{-\top} V^*(C_n - C_M)V A^{-1}, \qquad (36)$$

and prove that the first component is just the identity matrix. By denoting the second component with E, Eq. (36), Sect. C, implies that the condition number of $B^\top H B$ is bounded by $(1 + \|E\|)/(1 - \|E\|)$, when $\|E\| < 1$. In Lemma 3 we prove that $\|E\|$ is analytically bounded by a suitable distance between $C_n - C_M$ and in Lemma 9, 10, Sect. D, we bound in probability such distance, when the Nyström centers are selected uniformly at random and with approximate leverage scores. Finally in Lemma 11, 12, Sect. D, we give a condition on M for the two kind of sampling, such that the condition number is controlled and the error term in the oracle inequality decays as $e^{-t/2}$, leading to Thm. 2, 4.

Now we provide the preliminary result necessary to prove a computational oracle inequality for FALKON.

Theorem 1 *Let $0 \leq \lambda \leq \|C\|$, B as in Def. 3 and $n, M, t \in \mathbb{N}$. Let $\widehat{f}_{\lambda,M,t}$ be the FALKON estimator, with preconditioner B, after t iterations Def. 4 and let $\widetilde{f}_{\lambda,M}$ be the exact Nyström estimator as in Eq. 8. Let $\delta \in (0,1]$ and $n \geq 3$, then following holds with probability $1 - \delta$*

$$\mathcal{R}(\widehat{f}_{\lambda,M,t})^{1/2} \;\; \leq \;\; \mathcal{R}(\widetilde{f}_{\lambda,M})^{1/2} \;\; + \;\; 4\widehat{v}\, e^{-\nu t}\sqrt{1 + \frac{9\kappa^2}{\lambda n}\log\frac{n}{\delta}},$$

where $\widehat{v}^2 = \frac{1}{n}\sum_{i=1}^n y_i^2$ and $\nu = \log\frac{\sqrt{\mathrm{cond}\,(W)}+1}{\sqrt{\mathrm{cond}\,(W)}-1}$. In particular $\nu \geq 1/2$, when $\mathrm{cond}\,(W) \leq (\frac{e^{1/2}+1}{e^{1/2}-1})^2$.

Proof. By applying Lemma 8, we have

$$\mathcal{R}(\widehat{f}_{\lambda,M,t})^{1/2} \;\; \leq \;\; \mathcal{R}(\widetilde{f}_{\lambda,M})^{1/2} \;\; + \;\; 2c_0\|S\widehat{C}_{n\lambda}^{-1/2}\|\,\widehat{v}\, e^{-\nu t}.$$

To complete the theorem we need to study the quantity $\|S\widehat{C}_{n\lambda}^{-1/2}\|$. In particular, define $\lambda_0 = \frac{9\kappa^2}{n}\log\frac{n}{\delta}$. By dividing and multiplying for $C_{n\lambda_0}^{1/2}$, we have

$$\|S\widehat{C}_{n\lambda}^{-1/2}\| = \|SC_{n\lambda_0}^{-1/2}C_{n\lambda_0}^{1/2}C_{n\lambda}^{-1/2}\| \leq \|SC_{n\lambda_0}^{-1/2}\|\|C_{n\lambda_0}^{1/2}C_{n\lambda}^{-1/2}\|.$$

Now, for the first term, since $\|Z\|^2 = \|Z^*Z\|$, and the fact that $C = S^*S$ (see Prop. 3), we have

$$\|SC_{n\lambda_0}^{-1/2}\|^2 = \|C_{n\lambda_0}^{-1/2}CC_{n\lambda_0}^{-1/2}\| = \|C^{1/2}C_{n\lambda_0}^{-1/2}\|,$$

moreover by Lemma 5 of [13] (or Lemma 7.6 of [42]), we have

$$\|C^{1/2}C_{n\lambda_0}^{-1/2}\| \leq 2,$$

with probability $1 - \delta$. Finally, by denoting with $\sigma(C)$ the set of eigenvalues of the positive operator C, recalling that $\sigma(C) \subset [0, \kappa^2]$ (see Prop. 3), we have

$$\|C_{n\lambda_0}^{1/2}C_{n\lambda}^{-1/2}\| = \sup_{\sigma \in \sigma(C)}\sqrt{\frac{\sigma + \lambda_0}{\sigma + \lambda}} \leq \sup_{\sigma \in [0,\kappa^2]}\sqrt{\frac{\sigma + \lambda_0}{\sigma + \lambda}} \leq \sqrt{1 + \frac{\lambda_0}{\lambda}}.$$

$\square$

24

E.1 Main Result (I): computational oracle inequality for FALKON with uniform sampling

Lemma 11. *Let $\delta \in (0,1]$, $0 < \lambda \leq \|C\|$, $n, M \in \mathbb{N}$, the matrix W as in Eq. 14 with B satisfying Def. 3 and the Nyström centers selected via uniform sampling. When*

$$M \geq 5\left[1 + 14\mathcal{N}_\infty(\lambda)\right] \log \frac{8\kappa^2}{\lambda\delta}, \tag{37}$$

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

$$\operatorname{cond}(W) \leq \left(\frac{e^{1/2}+1}{e^{1/2}-1}\right)^2.$$

Proof. By Lemma 1 we have that

$$\operatorname{cond}(W) \leq \frac{1 + \|E\|}{1 - \|E\|},$$

with the operator E defined in the same lemma. By Lemma 3, we have

$$\|E\| \leq \|\widehat{G}_{M\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{G}_{M\lambda}^{-1/2}\|.$$

Lemma 9 proves that when the Nyström centers are selected with uniform sampling and M satisfies Eq. (34) for a given parameter $\eta \in (0,1]$, then $\|\widehat{G}_{M\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{G}_{M\lambda}^{-1/2}\| \leq \eta$, with probability $1 - \delta$. In particular we select $\eta = \frac{2e^{1/2}}{e+1}$. The condition on M in Eq. (37) is derived by Eq. (34) by substituting η with $\frac{2e^{1/2}}{e+1}$. □

Theorem 6. *Let $\delta \in (0,1]$, $0 < \lambda \leq \|C\|$, $n, M \in \mathbb{N}$ and the Nyström centers be selected via uniform sampling. Let $\widehat{f}_{\lambda,M,t}$ be the FALKON estimator, after t iterations (Def. 4) and let $\widetilde{f}_{\lambda,M}$ be the exact Nyström estimator in Eq. (8). When*

$$M \geq 5\left[1 + 14\mathcal{N}_\infty(\lambda)\right] \log \frac{8\kappa^2}{\lambda\delta},$$

then, with probability $1 - 2\delta$,

$$\mathcal{R}(\widehat{f}_{\lambda,M,t})^{1/2} \;\leq\; \mathcal{R}(\widetilde{f}_{\lambda,M})^{1/2} \;+\; 4\widehat{v}\, e^{-\frac{t}{2}} \sqrt{1 + \frac{9\kappa^2}{\lambda n} \log \frac{n}{\delta}},$$

Proof. By applying Lemma 11 we have that

$$\operatorname{cond}(W) \leq (e^{1/2}+1)^2/(e^{1/2}-1)^2,$$

with probability $1 - \delta$ under the condition on M. Then apply the computational oracle inequality in Thm. 1 and take the union bound of the two events. □

Theorem 2. *Under the same conditions of Thm. 1, the exponent ν in Thm. 1 satisfies $\nu \geq 1/2$, with probability $1 - 2\delta$, when the Nyström centers are selected via uniform sampling (see Sect. A), and*

$$M \geq 5\left[1 + \frac{14\kappa^2}{\lambda}\right] \log \frac{8\kappa^2}{\lambda\delta}.$$

Proof. It is a direct application of Thm. 6. Indeed note that $\mathcal{N}_\infty(\lambda) \leq \frac{\kappa^2}{\lambda}$ by definition. □

E.2 Main Result (II): computational oracle inequality for FALKON with leverage scores

Lemma 12. *Let $\delta \in (0,1]$ and the matrix W be as in Eq. 14 with B satisfying Def. 3 and the Nyström centers selected via (q, λ_0, δ)-approximated leverage scores sampling (see Def. 1 and discussion below), with $\lambda_0 = \frac{19\kappa^2}{n} \log \frac{n}{2\delta}$. When $\lambda_0 \leq \lambda \leq \|C\|$, $n \geq 405\kappa^2 \vee 67\kappa^2 \log \frac{12\kappa^2}{\delta}$ and*

$$M \geq 5\left[1 + 43q^2\mathcal{N}(\lambda)\right] \log \frac{8\kappa^2}{\lambda\delta}, \tag{38}$$

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

$$\operatorname{cond}(W) \leq \left(\frac{e^{1/2}+1}{e^{1/2}-1}\right)^2.$$

Proof. By Lemma 1 we have that

$$\operatorname{cond}(W) \leq \frac{1 + \|E\|}{1 - \|E\|},$$

with the operator E defined in the same lemma. By Lemma 3 we have

$$\|E\| \leq \|\widehat{G}_{M\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{G}_{M\lambda}^{-1/2}\|.$$

Lemma 10 proves that when the Nyström centers are selected via q-approximate leverage scores and M satisfies Eq. (35) for a given parameter $\eta \in (0, 1]$, then $\|\widehat{G}_{M\lambda}^{-1/2}(\widehat{C}_n - \widehat{G}_M)\widehat{G}_{M\lambda}^{-1/2}\| \leq \eta$, with probability $1 - \delta$. In particular we select $\eta = \frac{2e^{1/2}}{e+1}$. The condition on M in Eq. (38) is derived by Eq. (35) by substituting η with $\frac{2e^{1/2}}{e+1}$. $\square$

Theorem 7. *Let $\delta \in (0, 1]$, $M, n \in \mathbb{N}$ and the Nyström centers be selected via (q, λ_0, δ)-approximated leverage scores sampling (see Def. 1 and discussion below), with $\lambda_0 = \frac{19\kappa^2}{n} \log \frac{n}{2\delta}$. Let $t \in \mathbb{N}$. Let $\widehat{f}_{\lambda,M,t}$ be the FALKON estimator, after t iterations (Def. 4) and let $\widetilde{f}_{\lambda,M}$ be the exact Nyström estimator in Eq. (8). When $\lambda_0 \leq \lambda \leq \|C\|$, $n \geq 405\kappa^2 \vee 67\kappa^2 \log \frac{12\kappa^2}{\delta}$ and*

$$M \geq 5\left[1 + 43q^2\mathcal{N}(\lambda)\right]\log\frac{8\kappa^2}{\lambda\delta},$$

then, with probability $1 - 2\delta$,

$$\mathcal{R}(\widehat{f}_{\lambda,M,t})^{1/2} \leq \mathcal{R}(\widetilde{f}_{\lambda,M})^{1/2} + 4\widehat{v}\, e^{-\frac{t}{2}} \sqrt{1 + \frac{9\kappa^2}{\lambda n}\log\frac{n}{\delta}},$$

Proof. By applying Lemma 12 we have that

$$\operatorname{cond}(W) \leq (e^{1/2} + 1)^2/(e^{1/2} - 1)^2,$$

with probability $1 - \delta$ under the conditions on λ, n, M. Then apply the computational oracle inequality in Thm. 1 and take the union bound of the two events. $\square$

Theorem 4. *Under the same conditions of Thm. 1, the exponent ν in Thm. 1 satisfies $\nu \geq 1/2$, with probability $1 - 2\delta$, when*

1. *either Nyström uniform sampling (see Sect. A) is used with $M \geq 70\left[1 + \mathcal{N}_\infty(\lambda)\right]\log\frac{8\kappa^2}{\lambda\delta}$.*

2. *or Nyström (q, λ_0, δ)-appr. lev. scores (see Sect. A) is used, with $\lambda \geq \frac{19\kappa^2}{n}\log\frac{n}{2\delta}$, $n \geq 405\kappa^2 \log\frac{12\kappa^2}{\delta}$, and*

$$M \geq 215\left[2 + q^2\mathcal{N}(\lambda)\right]\log\frac{8\kappa^2}{\lambda\delta}.$$

Proof. It is a merge of Thm. 6 and Thm. 7. $\square$

E.3 Main Results (III): Optimal Generalization Bounds

First we recall the standard assumptions to study generalization rates for the non-parametric supervised learning setting, with square loss function. Then we provide Thm. 8, from which we obtain Thm. 3 and Thm. 5.

There exists $\kappa \geq 1$ such that $K(x, x) \leq \kappa^2$ for any $x \in X$. There exists $f_\mathcal{H} \in \mathcal{H}$, such that $\mathcal{E}(f_\mathcal{H}) = \inf_{f \in \mathcal{H}} \mathcal{E}(f)$. Moreover, we assume that $\rho(y|x)$ has sub-exponential tails, i.e. in terms of moments of y: there exist σ, b satisfying $0 \leq \sigma \leq b$, such that, for any $x \in X$, the following holds

$$\mathbb{E}\left[|y - f_\mathcal{H}(x)|^p \mid x\right] \leq \frac{1}{2}p!\sigma^2 b^{p-2}, \quad \forall p \geq 2. \tag{39}$$

Note that the assumption above is satisfied, when y is supported in an interval or when it has sub-gaussian or sub-exponential tails. The last assumption is known as *source condition* [17]. There exist $r \in [1/2, 1]$ and $g \in \mathcal{H}$, such that

$$f_\mathcal{H} = C^{r-1/2}g, \tag{40}$$

where C is the correlation operator defined in Def. 5. Finally define $R \geq 1$ such that $R \geq \|g\|_\mathcal{H}$. Note that assuming the existence of $f_\mathcal{H}$, the source condition is always satisfied with $r = 1/2, g = f_\mathcal{H}$ and $R = \max(1, \|f_\mathcal{H}\|_\mathcal{H})$, however if it is satisfied with larger r it leads to faster learning rates.

Theorem 8. *Let $\delta \in (0, 1]$. Let n, λ, M satisfy $n \geq 1655\kappa^2 + 223\kappa^2 \log \frac{24\kappa^2}{\delta}$, $M \geq 334 \log \frac{192n}{\delta}$ and $\frac{19\kappa^2}{n} \log \frac{24n}{\delta} \leq \lambda \leq \|C\|$. Let $\widehat{f}_{\lambda,M,t}$ be the FALKON estimator in Def. 4, after $t \in \mathbb{N}$ iterations. Under the assumptions in Eq. (39), (40), the following holds with probability at least $1 - \delta$,*

$$\mathcal{R}(\widehat{f}_{\lambda,M,t})^{1/2} \leq 6R \left(\frac{b\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{\sigma^2 \mathcal{N}(\lambda)}{n}} \right) \log \frac{24}{\delta} + 7R\lambda^r, \tag{41}$$

1. *either, when the Nyström points are selected uniformly sampled (see Sect. A) and*

$$M \geq 70 \Big[1 + \mathcal{N}_\infty(\lambda) \Big] \log \frac{48\kappa^2}{\lambda\delta}, \qquad t \geq 2 \log \frac{8(b + \kappa\|f_{\mathcal{H}}\|_{\mathcal{H}})}{R\lambda^r}, \tag{42}$$

2. *or, when the Nyström points are selected by means of (q, λ_0, δ)-approximate leverage scores (see Sect. A), with $q \geq 1$, $\lambda_0 = \frac{19\kappa^2}{n} \log \frac{48n}{\delta}$ and*

$$M \geq 215 \Big[1 + q^2 \mathcal{N}(\lambda) \Big] \log \frac{192\kappa^2 n}{\lambda\delta}, \qquad t \geq 2 \log \frac{8(b + \kappa\|f_{\mathcal{H}}\|_{\mathcal{H}})}{R\lambda^r}. \tag{43}$$

Proof. Let $\mu = \delta/4$. By Proposition 2 of [13], under the assumptions in Eq. 39 and Eq. 40, when $n \geq 1655\kappa^2 + 223\kappa^2 \log \frac{6\kappa^2}{\mu}$, $M \geq 334 \log \frac{48n}{\mu}$, and $\frac{19\kappa^2}{n} \log \frac{6n}{\mu} \leq \lambda \leq \|C\|$, we have with probability $1 - \mu$

$$\mathcal{R}(\widetilde{f}_{\lambda,M})^{1/2} \leq 6R \left(\frac{b\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{\sigma^2 \mathcal{N}(\lambda)}{n}} \right) \log \frac{6}{\mu} + 3R\mathcal{C}(M)^r + 3R\lambda^r,$$

where

$$\mathcal{C}(M) = \min \left\{ t > 0 \;\middle|\; (67 + 5\mathcal{N}_\infty(t)) \log \frac{12\kappa^2}{t\mu} \leq M \right\},$$

when the Nyström centers are selected with uniform sampling, otherwise

$$\mathcal{C}(M) = \min \left\{ \lambda_0 \leq t \leq \|C\| \;\middle|\; 78q^2 \mathcal{N}(t) \log \frac{48n}{\mu} \leq M \right\},$$

when the Nyström centers are selected via approximate sampling, with $\lambda_0 = \frac{19\kappa^2}{n} \log \frac{12n}{\mu}$. In particular, note that $\mathcal{C}(M) \leq \lambda$, in both cases, when M satisfies Eq. (42) for uniform sampling, or Eq. (43) for approximate leverage scores. Now, by applying the computational oracle inequality in Thm. 6, for uniform sampling, or Thm. 7, for approximate leverage scores, the following holds with probability $1 - 2\mu$

$$R(\widehat{f}_{\lambda,M,t})^{1/2} \leq \mathcal{R}(\widetilde{f}_{\lambda,M})^{1/2} + 4\widehat{v} \, e^{-\frac{t}{2}} \sqrt{1 + \frac{9\kappa^2}{\lambda n} \log \frac{n}{\mu}},$$

with $\widehat{v}^2 := \frac{1}{n} \sum_{i=1}^n y_i^2$. In particular, note that, since we require $\lambda \geq \frac{19\kappa^2}{n} \log \frac{12n}{\mu}$, we have

$$4\sqrt{1 + \frac{9\kappa^2}{\lambda n} \log \frac{n}{\mu}} \leq 5.$$

Now, we choose t such that $5\widehat{v}e^{-t/2} \leq R\lambda^r$, that is $t \geq 2 \log \frac{5\widehat{v}}{R\lambda^r}$. The last step consists in bounding $\widehat{v}$ in probability. Since it depends on the random variables $y_1, \ldots, y_n$ we bound it in the following way. By recalling that $|f_{\mathcal{H}}(x)| = |\langle K_x, f_{\mathcal{H}} \rangle_{\mathcal{H}}| \leq \|K_x\|_{\mathcal{H}} \|f_{\mathcal{H}}\|_{\mathcal{H}} \leq \kappa \|f_{\mathcal{H}}\|_{\mathcal{H}}$ for any $x \in X$, we have

$$\widehat{v} = \frac{1}{\sqrt{n}} \|\widehat{y}\| \leq \sqrt{\sum_{i=1}^n \frac{(y_i - f_{\mathcal{H}}(x_i))^2}{n}} + \sqrt{\sum_{i=1}^n \frac{f_{\mathcal{H}}(x_i)^2}{n}} \leq \sqrt{\sum_{i=1}^n \frac{(y_i - f_{\mathcal{H}}(x_i))^2}{n}} + \kappa \|f_{\mathcal{H}}\|_{\mathcal{H}}.$$

Since the training set examples $(x_i, y_i)_{i=1}^n$ are i.i.d. with probability ρ we can apply the Bernstein inequality [43] to the random variables $z_i = (y_i - f_{\mathcal{H}}(x_i))^2 - s$, with $s = \mathbb{E}(y_i - f_{\mathcal{H}}(x_i))^2$ (since x_i, y_i are i.i.d. each z_i has the same distribution and so the same expected value s). In particular, we need to bound the moments of z_i's. By the assumption in Eq. 39, z_i are zero mean and

$$\mathbb{E}|z_i|^{2p} \leq \frac{1}{2}(2p)!\sigma^2 b^{2p-2} \leq \frac{1}{2}p!(4\sigma b)^2 (4b^2)^{p-2}, \qquad p \geq 2$$

and so, by applying the Bernstein inequality, the following holds with probability $1 - \mu$

$$\left| \sum_{i=1}^n \frac{z_i}{n} \right| \leq \frac{8b^2 \log \frac{2}{\mu}}{3n} + \sqrt{\frac{8\sigma^2 b^2 \log \frac{2}{\mu}}{n}} \leq \frac{1}{4}b^2,$$

27

where the last step is due to the fact that we require $n \geq 223\kappa^2 \log \frac{6}{\mu}$, that $b \geq \sigma$ and that $\kappa \geq 1$ by definition. So, by noting that $s \leq \sigma^2 \leq b^2$ (see Eq. 39), we have

$$\widehat{v} \; \leq \; \kappa \|f_\mathcal{H}\|_\mathcal{H} + \sqrt{s + \sum_{i=1}^n \frac{z_i}{n}} \; \leq \; \kappa \|f_\mathcal{H}\|_\mathcal{H} + \sqrt{s} + \frac{1}{2}b \; \leq \; \frac{3}{2}b + \kappa \|f_\mathcal{H}\|_\mathcal{H},$$

with probability at least $1 - \mu$. Now by taking the intersection of the three events, the following holds with probability at least $1 - 4\mu$

$$\mathcal{R}(\widehat{f}_{\lambda,M,t})^{1/2} \leq 6R \left(\frac{b\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{\sigma^2 \mathcal{N}(\lambda)}{n}} \right) \log \frac{6}{\mu} + 7R\lambda^r.$$

$\square$

Now we provide the generalization error bounds for the setting where we only assume the existence of $f_\mathcal{H}$.

Theorem. 3 *Let $\delta \in (0, 1]$. Let the outputs y be bounded in $[-\frac{a}{2}, \frac{a}{2}]$, almost surely, with $a > 0$. For any $n \geq \max(\frac{1}{\|C\|}, 82\kappa^2 \log \frac{373\kappa^2}{\sqrt{\delta}})^2$ the following holds. When*

$$\lambda = \frac{1}{\sqrt{n}}, \qquad M \geq 5(67 + 20\sqrt{n}) \log \frac{48\kappa^2 n}{\delta}, \qquad t \geq \frac{1}{2} \log(n) + 5 + 2\log(a + 3\kappa),$$

then with probability $1 - \delta$,

$$\mathcal{R}(\widehat{f}_{\lambda,M,t}) \leq \frac{c_0 \log^2 \frac{24}{\delta}}{\sqrt{n}},$$

where $\widehat{f}_{\lambda,M,t}$ is the FALKON estimator in Def. 4 (see also Sect. 3 Alg. 1) with Nyström uniform sampling, and the constant $c_0 = 49\|f_\mathcal{H}\|_\mathcal{H}^2(1 + a\kappa + 2\kappa^2\|f_\mathcal{H}\|_\mathcal{H})^2$.

Proof. Here we assume $y \in [-\frac{a}{2}, \frac{a}{2}]$ a.s., so Eq. 39 is satisfied with $\sigma = b = a + 2\kappa\|f_\mathcal{H}\|_\mathcal{H}$, indeed

$$\mathbb{E}[|y - f_\mathcal{H}(x)|^p \mid x] \; \leq \; \mathbb{E}[2^{p-1}|y|^p \mid x] + 2^{p-1}|f_\mathcal{H}(x)|^p \; \leq \; \frac{1}{2}(a^p + 2^p\kappa^p\|f_\mathcal{H}\|_\mathcal{H}^p) \leq \frac{1}{2}p!(a + 2\kappa\|f_\mathcal{H}\|_\mathcal{H})^p,$$

where we used the fact that $|f_\mathcal{H}(x)| = |\langle K_x, f_\mathcal{H}\rangle_\mathcal{H}| \leq \|K_x\|\|f_\mathcal{H}\| \leq \kappa\|f_\mathcal{H}\|_\mathcal{H}$. Moreover, Eq. (40) is satisfied with $r = 1/2$ and $g = f_\mathcal{H}$, while $R = \max(1, \|f_\mathcal{H}\|_\mathcal{H})$.

To complete the proof we show that the assumptions on λ, M, n satisfy the condition required by Thm. 8, then we apply it and derive the final bound. Set $\lambda = n^{-1/2}$ and define $n_0 = \max(\|C\|^{-1}, 82\kappa^2 \log \frac{373\kappa^2}{\sqrt{\delta}})^2$. The condition $n \geq n_0$, satisfies the condition on n required by Thm. 8. Moreover both $\lambda = n^{-1/2}$ and $M \geq 75\sqrt{n} \log \frac{48\kappa^2 n}{\delta}$ satisfy respectively the conditions on λ, M required by Thm. 8, when $n \geq n_0$. Finally note that the condition on t implies the condition required by Thm. 8, indeed, since $R = \max(1, \|f\|_\mathcal{H})$, we have $a/R \leq a$ and $\|f_\mathcal{H}\|_\mathcal{H}/R \leq 1$, so

$$2\log \frac{8(a + \kappa\|f_\mathcal{H}\|_\mathcal{H})}{R\lambda^r} \; = \; \log \left[64 \left(\frac{a}{R} + \frac{3\kappa\|f_\mathcal{H}\|_\mathcal{H}}{R} \right)^2 \sqrt{n} \right]$$

$$\leq \; \log(64(a + 3\kappa)^2\sqrt{n}) \; \leq \; \log 64 + \frac{1}{2}\log n + 2\log(a + 3\kappa).$$

So, by applying Thm. 8 with R, r defined as above and recalling that $\mathcal{N}(\lambda) \leq \mathcal{N}_\infty(\lambda) \leq \frac{\kappa^2}{\lambda}$, we have

$$\mathcal{R}(\widehat{f}_{\lambda,M,t})^{1/2} \leq 6R \left(\frac{b\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{\sigma^2 \mathcal{N}(\lambda)}{n}} \right) \log \frac{24}{\delta} + 7R\lambda^r \leq 6R \left(\frac{b\kappa}{\sqrt{\lambda}n} + \frac{\sigma\kappa}{\sqrt{\lambda n}} \right) \log \frac{24}{\delta} + 7R\lambda^{\frac{1}{2}}$$

$$= 6Rb\kappa(1 + n^{-1/2})n^{-1/4} \log \frac{24}{\delta} + 7Rn^{-1/4} \leq \frac{7R(b\kappa + 1) \log \frac{24}{\delta}}{n^{1/4}}.$$

with probability $1 - \delta$. For the last step we used the fact that $b = \sigma$, that $6(1 + n^{-1/2}) \leq 7$, since $n \geq n_0$, and that $\log \frac{24}{\delta} > 1$. $\square$

To state the result for fast rates, we need to define explicitly the *capacity condition* on the intrinsic dimension. There exists $Q > 0, \gamma \in (0, 1]$ such that

$$\mathcal{N}(\lambda) \leq Q^2 \lambda^{-\gamma}, \quad \forall \lambda \geq 0. \tag{44}$$

Note that, by definition of $\mathcal{N}(\lambda)$, the assumption above is always satisfied with $Q = \kappa$ and $\gamma = 1$.

Theorem 5 *Let $\delta \in (0,1]$. Let the outputs y be bounded in $[-\frac{a}{2}, \frac{a}{2}]$, almost surely, with $a > 0$. Under the assumptions in Eq. 40, 44 and $n \geq \|C\|^{-s} \vee \left(\frac{102\kappa^2 s}{s-1} \log \frac{912}{\delta}\right)^{\frac{s}{s-1}}$, with $s = 2r + \gamma$, the following holds. When*

$$\lambda = n^{-\frac{1}{2r+\gamma}}, \qquad t \geq \log(n) + 5 + 2\log(a + 3\kappa^2),$$

1. *and either Nyström uniform sampling is used with*

$$M \geq 70\left[1 + \mathcal{N}_\infty(\lambda)\right] \log \frac{8\kappa^2}{\lambda\delta}, \tag{45}$$

2. *and or Nyström (q, λ_0, δ)-approximate leverage scores (Def. 1), with $q \geq 1$, $\lambda_0 = \frac{19\kappa^2}{n} \log \frac{48n}{\delta}$ and*

$$M \geq 215\left[1 + q^2 \mathcal{N}(\lambda)\right] \log \frac{8\kappa^2}{\lambda\delta}, \tag{46}$$

then with probability $1 - \delta$,

$$\mathcal{R}(\widehat{f}_{\lambda,M,t}) \leq c_0 \log^2 \frac{24}{\delta} \ n^{-\frac{2r}{2r+\gamma}},$$

where $\widehat{f}_{\lambda,M,t}$ is the FALKON estimator in Sect. 3 (Alg. 1). In particular n_0, c_0 do not depend on λ, M, n and c_0 do not depend on δ.

Proof. The proof is similar to the one for the slow learning rate (Thm. 3), here we take into account the additional assumption in Eq. (40), 44 and the fact that r may be bigger than $1/2$. Moreover we assume $y \in [-\frac{a}{2}, \frac{a}{2}]$ a.s., so Eq. 39 is satisfied with $\sigma = b = a + 2\kappa\|f_\mathcal{H}\|_\mathcal{H}$, indeed

$$\mathbb{E}[|y - f_\mathcal{H}(x)|^p \mid x] \leq \mathbb{E}[2^{p-1}|y|^p \mid x] + 2^{p-1}|f_\mathcal{H}(x)|^p \leq \frac{1}{2}(a^p + 2^p \kappa^p \|f_\mathcal{H}\|_\mathcal{H}^p) \leq \frac{1}{2}p!(a + 2\kappa\|f_\mathcal{H}\|_\mathcal{H})^p,$$

where we used the fact that $|f_\mathcal{H}(x)| = |\langle K_x, f_\mathcal{H}\rangle_\mathcal{H}| \leq \|K_x\|\|f_\mathcal{H}\| \leq \kappa\|f_\mathcal{H}\|_\mathcal{H}$.

To complete the proof we show that the assumptions on λ, M, n satisfy the required conditions to apply Thm. 8. Then we apply it and derive the final bound. Set $\lambda = n^{-1/(2r+\gamma)}$ and define $n_0 = \|C\|^{-s} \vee \left(\frac{102\kappa^2 s}{s-1} \log \frac{912}{\delta}\right)^{\frac{s}{s-1}}$, with $s = 2r + \gamma$. Since $1 < s \leq 3$, the condition $n \geq n_0$, satisfies the condition on n required to apply Thm. 8. Moreover, for any $n \geq n_0$, both $\lambda = n^{-1/(2r+\gamma)}$ and M satisfying Eq. (45) for Nyström uniform sampling, and Eq. 46 for Nyström leverage scores, satisfy respectively the conditions on λ, M required to apply Thm. 8. Finally note that the condition on t implies the condition required by Thm. 8, indeed, since $2r/(2r + \gamma) \leq 1$,

$$2\log \frac{8(b + \kappa\|f_\mathcal{H}\|_\mathcal{H})}{R\lambda^r} = \log\left[64\left(\frac{a}{R} + \frac{3\kappa\|f_\mathcal{H}\|_\mathcal{H}}{R}\right)^2 n^{\frac{2r}{2r+\gamma}}\right]$$

$$\leq \log 64 + 2\log \frac{a + 3\kappa\|f_\mathcal{H}\|_\mathcal{H}}{R} + \frac{2r}{2r+\gamma}\log n$$

$$\leq \log 64 + 2\log \frac{a + 3\kappa\|f_\mathcal{H}\|_\mathcal{H}}{R} + \log n,$$

$$\leq \log 64 + 2\log(a + 3\kappa^2) + \log n.$$

where the last step is due to the fact that $a/R \leq 1$ and $\|f_h\|_\mathcal{H}/R \leq \|C^{r-1/2}\| \leq \|C\|^{1/2} \leq \kappa$, since $R := \max(1, \|g\|_\mathcal{H})$, and $\|f_\mathcal{H}\|_\mathcal{H} \leq \|C^{r-1/2}\|\|g\|_\mathcal{H}$, by definition. So, by applying Thm. 8 with R, r defined as above and recalling that $\mathcal{N}_\infty(\lambda) \leq \frac{\kappa^2}{\lambda}$ by construction and that $\mathcal{N}(\lambda) \leq Q^2\lambda^{-\gamma}$ by the capacity condition in Eq. (44), we have

$$\mathcal{R}(\widehat{f}_{\lambda,M,t})^{1/2} \leq 6R\left(\frac{b\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{\sigma^2 \mathcal{N}(\lambda)}{n}}\right)\log \frac{24}{\delta} + 7R\lambda^r \leq 6R\left(\frac{b\kappa}{\sqrt{\lambda}n} + \frac{Q\sigma}{\sqrt{\lambda^\gamma n}}\right)\log \frac{24}{\delta} + 7R\lambda^r$$

$$= 6Rb\left(\kappa n^{-\frac{r+\gamma-1/2}{2r+\gamma}} + Q\right)n^{-\frac{r}{2r+\gamma}}\log \frac{24}{\delta} + 7Rn^{-\frac{r}{2r+\gamma}}$$

$$\leq 7R(b(\kappa + Q) + 1)\log \frac{24}{\delta} \ n^{-\frac{r}{2r+\gamma}}.$$

with probability $1 - \delta$. For the last step we used the fact that $b = \sigma$, that $r + \gamma - 1/2 \geq 0$, since $r \geq 1/2$ by definition, and that $\log \frac{24}{\delta} > 1$. $\qquad\square$

F   Longer comparison with previous works

In the literature of KRR there are some papers that propose to solve Eq. 5 with iterative preconditioned methods [3, 4, 26, 5, 6]. In particular the one of [3] is based, essentially, on an incomplete singular value decomposition of the kernel matrix. Similarly, the ones proposed by [5, 6] are based on singular value decomposition obtained via randomized linear algebra approaches. The first covers the linear case, while the second deals with the kernel case. [4, 26] use a preconditioner based on the solution of a randomized projection problem based respectively on random features and Nyström.

While such preconditioners are suitable in the case of KRR, their computational cost becomes too expensive when applied to the random projection case. Indeed, performing an incomplete svd of the matrix K_{nM} even via randomized linear algebra approaches would require $O(nMk)$ where k is the number of singular values to compute. To achieve a good preconditioning level (and so having $t \approx \log n$) we should choose k such that $\sigma_k(K_{nM}) \approx \lambda$. When the kernel function is bounded, without further assumptions on the eigenvalue decay of the kernel matrix, we need $k \approx \lambda^{-1}$ [17, 13]. Since randomized projection requires $\lambda = n^{-1/2}$, $M = O(\sqrt{n})$ to achieve optimal generalization bounds, we have $k \approx \sqrt{n}$ and so the total cost of the incomplete svd preconditioner is $O(n^2)$. On the same lines, applying the preconditioner proposed by [4, 26] requires $O(nM^2)$ to be computed and there is no natural way to find a similar sketched preconditioner as the one in Eq. (10) in the case of [4], with reduced computational cost. In the case of [26], the preconditioner they use is exactly the matrix H^{-1}, whose computation amounts to solve the original problem in Eq. (8) with direct methods and requires $O(nM^2)$.

A similar reasoning hold for methods that solve the Nyström linear system (8) with iterative approaches [22, 23, 24]. Indeed on the positive side, they have a computational cost of $O(nMt)$. However they are affected by the poor conditioning of the linear system in Eq. 8. Indeed, even if H or K_{MM} in Eq. 8 are invertible, their condition number can be arbitrarily large (while in the KRR case it is bounded by λ^{-1}), and so many iterations are often needed to achieve optimal generalization (E.g. by using early stopping in [23] they need $t \approx \lambda^{-1}$).

G   MATLAB Code for FALKON

Algorithm 2 Complete `MATLAB` code for FALKON. It requires $O(nMt + M^3)$ in time and $O(M^2)$ in memory. See Sect. 3 for more details, and Sect. 4 for theoretical properties.

Input: Dataset $X = (x_i)_{i=1}^n \in \mathbb{R}^{n \times D}, \hat{y} = (y_i)_{i=1}^n \in \mathbb{R}^n, M \in \mathbb{N}$ numbers of Nyström centers to select, lev_scores $\in \mathbb{R}^n$ approximate leverage scores (set lev_scores = [ ] for selecting Nyström centers via uniform sampling), function KernelMatrix computing the Kernel matrix of two sets of points, regularization parameter λ, number of iterations t.
Output: Nyström coefficients α.

```matlab
function alpha = FALKON(X, Y, lev_scores, M, KernelMatrix, lambda, t)
    n = size(X,1);
    [C, D] = selectNystromCenters(X, lev_scores, M, n);

    KMM = KernelMatrix(C,C);
    T = chol(D*KMM*D + eps*M*eye(M));
    A = chol(T*T'/M + lambda*eye(M));

        function w = KnMtimesVector(u, v)
            w = zeros(M,1); ms = ceil(linspace(0, n, ceil(n/M)+1));
            for i=1:ceil(n/M)
                Kr = KernelMatrix( X(ms(i)+1:ms(i+1),:), C );
                w = w + Kr'*(Kr*u + v(ms(i)+1:ms(i+1),:));
            end
        end

        function w = BHB(u)
            w = A'\(T'\(KnMtimesVector(T\(A\u), zeros(n,1))/n) + lambda*(A\u));
        end

    r = A'\(T'\KnMtimesVector(zeros(M,1), Y/n));

    beta = conjgrad(@BHB, r, t);
    alpha = T\(A\beta);
end

function beta = conjgrad(funA, r, tmax)
    p = r; rsold = r'*r; beta = zeros(size(r,1), 1);

    for i = 1:tmax
        Ap = funA(p);
        a = rsold/(p'*Ap);
        beta = beta + a*p;
        r = r - a*Ap; rsnew = r'*r;
        p = r + (rsnew/rsold)*p;
        rsold = rsnew;
    end
end

function [C, D] = selectNystromCenters(X, lev_scores, M, n)
    if isempty(lev_scores) %Uniform Nystrom
        D = eye(M);
        C = X(randperm(n,M),:);
    else % Appr. Lev. Scores Nystrom
        prob = lev_scores(:)./sum(lev_scores(:));
        [count, ind] = discrete_prob_sample(M, prob);
        D = diag(1./sqrt(n*prob(ind).*count));
        C = X(ind,:);
    end
end

function [count, ind] = discrete_prob_sample(M, prob)
    bins = histcounts(rand(M,1), [0; cumsum(prob(:))]);
    ind = find(bins > 0);
    count = bins(ind);
end
```