Faster Low-Rank Approximation and Kernel Ridge Regression via the Block-Nyström Method

Sachin Garg SACHG@UMICH.EDU
University of Michigan

Michał Dereziński DEREZIN@UMICH.EDU
University of Michigan

Editors: Nika Haghtalab and Ankur Moitra

Abstract

The Nyström method is a popular low-rank approximation technique for large matrices that arise in kernel methods and convex optimization. Yet, when the data exhibits heavy-tailed spectral decay, the effective dimension of the problem often becomes so large that even the Nyström method may be outside of our computational budget. To address this, we propose Block-Nyström, an algorithm that injects a block-diagonal structure into the Nyström method, thereby significantly reducing its computational cost while recovering strong approximation guarantees. We show that Block-Nyström can be used to construct improved preconditioners for second-order optimization, as well as to efficiently solve kernel ridge regression for statistical learning over Hilbert spaces. Our key technical insight is that, within the same computational budget, combining several smaller Nyström approximations leads to stronger tail estimates of the input spectrum than using one larger approximation. Along the way, we provide a novel recursive preconditioning scheme for efficiently inverting the Block-Nyström matrix, and provide new statistical learning bounds for a broad class of approximate kernel ridge regression solvers.

Keywords: Randomized linear algebra, low-rank approximation, kernel methods, optimization.

1. Introduction

Fast algorithms for approximating large positive semidefinite (psd) matrices are central to many problems in computational learning. For instance, when minimizing a convex training loss, approximating the psd Hessian matrix is a key step in preconditioning first-order optimization methods (e.g., Qu et al., 2016) and designing Newton-type algorithms (e.g., Erdogdu and Montanari, 2015). Moreover, for methods such as Gaussian processes (Williams and Rasmussen, 2006) and kernel ridge regression (KRR, e.g., Bach, 2013), approximating the psd kernel matrix is an essential step for making these techniques scalable to large datasets.

The Nyström method has proven to be one of the most successful techniques for psd matrix approximation (Williams and Seeger, 2000), leading to numerous Nyström-preconditioned optimization algorithms (e.g., Avron et al., 2017; Frangella et al., 2023, 2024), as well as learning algorithms based on Nyström kernel approximations (e.g., Rudi et al., 2015, 2018; Burt et al., 2019). In the most basic version of the method, given an $n \times n$ psd matrix $\mathbf{A}$, we select a subset $S \subseteq \{1, ..., n\}$ of m coordinates (landmarks), and use them to construct $\hat{\mathbf{A}} = \mathbf{C}\mathbf{W}^{-1}\mathbf{C}^\top$, where $\mathbf{C} = \mathbf{A}_{:,S}$ is the submatrix of the columns of $\mathbf{A}$ indexed by S, whereas $\mathbf{W} = \mathbf{A}_{S,S}$ is the principal submatrix of $\mathbf{A}$ indexed by S. Given a subset S, this approximate decomposition can be obtained by evaluating nm entries of $\mathbf{A}$ and inverting the principal submatrix $\mathbf{W}$ at the cost of $O(m^3)$ operations. Essentially,

© 2025 S. Garg & M. Dereziński.

this method projects $\mathbf{A}$ onto a subspace defined by the landmarks, thus it is crucial to ensure that the subset S is small but representative of $\mathbf{A}$'s structure.

One of the key advantages of the Nyström method is that it combines good practical performance with strong theoretical guarantees. To ensure the latter, a long line of works (including Alaoui and Mahoney, 2015; Musco and Musco, 2017; Rudi et al., 2018) have developed a randomized landmark selection technique called approximate ridge leverage score sampling. Given some $\lambda > 0$, we define λ-ridge leverage scores of $\mathbf{A}$ as the diagonal entries of $\mathbf{A}(\mathbf{A}+\lambda\mathbf{I})^{-1}$, and their sum is called the λ-effective dimension, $d_\lambda(\mathbf{A}) := \operatorname{tr}(\mathbf{A}(\mathbf{A}+\lambda\mathbf{I})^{-1}) \leq n$. It is known that to obtain a strong Nyström approximation of $\mathbf{A}$ along the top part of its spectrum above the λ threshold, we must sample $m = \tilde{O}(d_\lambda(\mathbf{A}))$ landmarks according to its λ-ridge leverage scores (we use $\tilde{O}$ to hide logarithmic factors). Formally, this gives the following guarantee for $\hat{\mathbf{A}}$ as an approximation of $\mathbf{A}$:

$$(\lambda\text{-regularized } \alpha\text{-approximation}) \quad \alpha^{-1}(\mathbf{A}+\lambda\mathbf{I}) \preceq \hat{\mathbf{A}}+\lambda\mathbf{I} \preceq \mathbf{A}+\lambda\mathbf{I}, \tag{1}$$

where $\preceq$ denotes the Loewner ordering and the above sampling scheme achieves $\alpha = 2$. Such α-approximations have several downstream applications, including where α is the approximation factor in the statistical risk of Nyström-approximated KRR (Section 1.2), or where α represents the condition number after preconditioning an ℓ_2-regularized Hessian matrix (Section 1.1).

While Nyström approximations are efficient, they still carry a substantial computational cost when the λ-effective dimension is large (i.e., when the regularizer λ is small or the spectrum of $\mathbf{A}$ is heavy-tailed). Obtaining a ridge leverage score sample of size $m = \tilde{O}(d_\lambda(\mathbf{A}))$ takes up to $\tilde{O}(nm^2)$ operations (Musco and Musco, 2017), while the cost of applying and/or inverting the matrix $\hat{\mathbf{A}}+\lambda\mathbf{I}$ requires additional $\tilde{O}(nm+m^3)$ time. This leads to quadratic or even cubic dependence on the landmark sample size m, which is a challenge for learning tasks with large effective dimension, such as KRR in the unattainable setting (Lin and Cevher, 2020) and other problems with heavy-tailed spectral distributions. This leads to our main motivating question:

> *How to construct a λ-regularized psd matrix approximation when λ-effective dimension is so large that the classical Nyström approximation is outside of our time budget?*

To address this question, we propose Block-Nyström, a cost-effective alternative to the Nyström method, which, roughly speaking, sparsifies the Nyström principal submatrix $\mathbf{W} = \mathbf{A}_{S,S}$ into a block-diagonal structure. This not only reduces the inversion cost of $\mathbf{W}$, but also decreases the effective dimension associated with the method, thereby reducing the cost of ridge leverage score sampling. We show that Block-Nyström achieves a better approximation factor α than the classical Nyström under the same computational budget when the effective dimension is large and the spectral decay is heavy-tailed. We summarize Block-Nyström in the following theorem.

Theorem 1 (Block-Nyström) *Given a psd matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\lambda > 0$, and $\alpha \geq 1$, we can construct a matrix data structure $\hat{\mathbf{A}}$ of rank $m = \tilde{O}(\alpha d_{\alpha^2\lambda}(\mathbf{A}))$ in time $\tilde{O}\big((nm^2+m^3)/\alpha^2+nm\big)$ such that:*

1. *$\hat{\mathbf{A}}$ is a λ-regularized $O(\alpha)$-approximation of $\mathbf{A}$ in the sense of (1);*
2. *We can apply $\hat{\mathbf{A}}$ to a vector in $\tilde{O}(nm)$ time;*
3. *We can apply $(\hat{\mathbf{A}}+\lambda\mathbf{I})^{-1}$ to a vector in $\tilde{O}\big(nm\cdot\alpha^{o(1)}\big)$ time.*

Construction Let $S = (S_1, S_2, ..., S_q)$ be a collection of $m = qb$ coordinates drawn according to approximate $\alpha^2\lambda$-ridge leverage score sampling, split up into $q = \tilde{O}(\alpha)$ blocks of size $b = \tilde{O}(d_{\alpha^2\lambda}(\mathbf{A}))$. We define the Block-Nyström approximation as follows:

$$\hat{\mathbf{A}} = \mathbf{C}\mathbf{W}^{-1}\mathbf{C}^\top \quad \text{for} \quad \mathbf{C} = \mathbf{A}_{:,S} \quad \text{and} \quad \mathbf{W} = q\,\mathrm{diag}\left(\mathbf{A}_{S_1,S_1}, ..., \mathbf{A}_{S_q,S_q}\right). \tag{2}$$

Note that Block-Nyström can be viewed as an extension of the classical Nyström approximation (obtained by setting $q = 1$). In fact, Theorem 1 recovers the standard Nyström guarantee (including all of the time complexities up to logarithmic factors) for $\alpha = 1$. In that case, we obtain an $O(1)$-approximation at the cost of $\tilde{O}(nm^2/\alpha^2) = \tilde{O}(nd_\lambda(\mathbf{A})^2)$. As we relax the approximation guarantee by increasing α, the effective dimension decreases to $d_{\alpha^2\lambda}(\mathbf{A})$, thus reducing this dominant cost.

Can a similar relaxed approximation guarantee be achieved with the classical Nyström method? Yes, but it will not reduce the computational cost as substantially as Block-Nyström. A simple calculation shows that a Nyström approximation constructed from $\tilde{O}(d_{\alpha\lambda}(\mathbf{A}))$ landmarks drawn according to $\alpha\lambda$-ridge leverage scores also achieves an $O(\alpha)$-approximation. However, the cost will be higher than Block-Nyström, since $d_{\alpha\lambda}(\mathbf{A}) \geq d_{\alpha^2\lambda}(\mathbf{A})$, where the gap depends on the spectral profile of $\mathbf{A}$. As a concrete example, suppose that the spectrum of $\mathbf{A}$ exhibits polynomial decay, $\lambda_i(\mathbf{A}) = \Theta(i^{-1/\gamma})$, as is typical in kernel methods. Then, we have $d_{\alpha\lambda}(\mathbf{A}) = \Omega(\alpha^\gamma d_{\alpha^2\lambda}(\mathbf{A}))$, so Block-Nyström's $\tilde{O}(nd_{\alpha^2\lambda}(\mathbf{A})^2)$ runtime is faster than classical Nyström by a factor of $\tilde{\Omega}(\alpha^{2\gamma})$, which gets more significant the more heavy-tailed the spectrum of $\mathbf{A}$ is. A detailed cost comparison is deferred to the discussion of the concrete applications of Block-Nyström.

Key technical insights Why does Block-Nyström offer a computational advantage over classical Nyström? To offer some intuition, we discuss two key technical insights that unlocked our analysis.

1. *Optimizing bias-variance trade-off.* The Block-Nyström construction (2) can be equivalently formulated as an average of smaller classical Nyström approximations, $\hat{\mathbf{A}}_{[q]} = \frac{1}{q}\sum_{i=1}^q \hat{\mathbf{A}}_i$, where $\hat{\mathbf{A}}_i = \mathbf{C}_i\mathbf{W}_i^{-1}\mathbf{C}_i^\top$ with $\mathbf{C}_i = \mathbf{A}_{:,S_i}$ and $\mathbf{W}_i = \mathbf{A}_{S_i,S_i}$. Thus, $\hat{\mathbf{A}}_{[q]}$ should behave similarly to the expectation $\mathbb{E}[\hat{\mathbf{A}}_1]$. While taking expectation over the landmark samples does not substantially improve the approximation of the top part of $\mathbf{A}$'s spectrum (bias dominates variance), it dramatically improves the tail estimates (variance dominates bias). Thus, via careful matrix concentration analysis, we can optimize this bias-variance trade-off to choose the optimal amount of averaging. This also explains why Block-Nyström makes the biggest gains for heavy-tailed spectra (e.g., polynomial decay), where tail estimates matter the most.

2. *Fast inversion via recursive preconditioning.* In most applications of the Nyström method, quickly computing the regularized inverse $(\hat{\mathbf{A}} + \lambda\mathbf{I})^{-1}$ is crucial. This is usually done via the Woodbury formula: $(\mathbf{C}\mathbf{W}^{-1}\mathbf{C}^\top + \lambda\mathbf{I})^{-1} = \frac{1}{\lambda}(\mathbf{I} - \mathbf{C}(\mathbf{C}^\top\mathbf{C} + \lambda\mathbf{W})^{-1}\mathbf{C}^\top)$, which requires computing and inverting the matrix $(\mathbf{C}^\top\mathbf{C} + \lambda\mathbf{W})^{-1}$. Unfortunately, adding $\mathbf{C}^\top\mathbf{C}$ eliminates the block-diagonal structure of $\lambda\mathbf{W}$, thereby seemingly erasing the computational benefits of Block-Nyström. We propose an alternative approach, by using an iterative solver to implicitly compute the inverse $(\hat{\mathbf{A}}_{[q]} + \lambda\mathbf{I})^{-1}$. We precondition this solver by a smaller Block-Nyström matrix with increased regularization, $\hat{\mathbf{A}}_{[q/c]} + c^2\lambda\mathbf{I}$ for $c = \tilde{O}(1)$, and repeat the procedure recursively until reaching the single Nyström matrix that can be inverted via Woodbury. We show that this scheme implements inversion in nearly-linear time.

1.1. Application: Preconditioning convex optimization

One of the primary settings where the Nyström method has received considerable attention is convex optimization. Here, the task is to minimize a convex function $g : \mathbb{R}^n \to \mathbb{R}$ by utilizing its gradient and Hessian information. Prior works have used Nyström and other psd matrix approximations of the Hessian $\nabla^2 g(\mathbf{x})$ to extract second-order information from the function by constructing a preconditioner for a first-order method, both in the general convex setting (Erdogdu and Montanari, 2015; Ye et al., 2020; Frangella et al., 2024; Sun et al., 2025) as well as when g is a quadratic (Avron et al., 2017; Frangella et al., 2023; Dereziński et al., 2025). In many cases, these preconditioners are treated as a black box, so that one can simply apply our Block-Nyström method in place of existing approaches, and use guarantee (1) to ensure fast linear convergence for strongly convex objectives.

To illustrate this, we apply our Theorem 1 to a convergence result for a Nesterov-accelerated approximate Newton method (Theorem 2 in Ye et al., 2020), which implies that Block-Nyström can be effectively used to achieve fast linear convergence in a neighborhood around the solution.

Corollary 2 (Block-Nyström preconditioning) *Let $g(\mathbf{x}) = \psi(\mathbf{x}) + \lambda\|\mathbf{x}\|^2$, where $\psi : \mathbb{R}^n \to \mathbb{R}$ is convex with an L-Lipschitz-continuous Hessian, and if $L > 0$, suppose that $\mathbf{x}_0$ is in a sufficiently small neighborhood around $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} g(\mathbf{x})$. Also, let $\hat{\mathbf{A}}$ be the Block-Nyström approximation of $\nabla^2\psi(\mathbf{x}_0)$ with regularizer λ and approximation factor $\alpha \geq 1$. Then, for any $\epsilon > 0$, after $\tilde{O}(\sqrt{\alpha}\log 1/\epsilon)$ applications of $(\hat{\mathbf{A}} + \lambda\mathbf{I})^{-1}$ and of $\nabla\psi$, we can obtain $\hat{\mathbf{x}}$ such that:*

$$g(\hat{\mathbf{x}}) - g(\mathbf{x}^*) \leq \epsilon \cdot \big[g(\mathbf{x}_0) - g(\mathbf{x}^*)\big].$$

To provide a concrete computational comparison between Block-Nyström and other preconditioners, in Section 4 we study the task of minimizing a quadratic function $g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} - \mathbf{x}^\top\mathbf{b}$ under heavy-tailed spectrum. We illustrate this here on the class of input matrices that follow the spiked covariance model (Johnstone, 2001), i.e., consisting of non-isotropic low-rank signal distorted by isotropic noise. Following Dereziński and Yang (2024), we model this by allowing some top-k part of $\mathbf{A}$'s spectrum (the signal) to be arbitrarily ill-conditioned, while requiring the remaining tail (the noise) to have condition number $O(1)$. We show that a careful implementation of Block-Nyström preconditioning achieves better time complexity for solving this quadratic minimization than the previous best known preconditioner (Dereziński et al., 2025) when the number of large eigenvalues is $O(n^{0.82})$, improving the runtime from $\tilde{O}(n^{2.065})$ to $\tilde{O}(n^{2.044})$. Similar gains can be obtained for matrices with other heavy-tailed spectral decay profiles.

Corollary 3 *Given a strongly convex quadratic $g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} - \mathbf{x}^\top\mathbf{b}$ such that $\mathbf{A} \in \mathbb{R}^{n\times n}$ has at most $O(n^{0.82})$ eigenvalues larger than $O(1)$ times its smallest eigenvalue, there is an algorithm that with high probability computes $\hat{\mathbf{x}}$ such that $g(\hat{\mathbf{x}}) \leq (1+\epsilon)\cdot\min_{\mathbf{x}\in\mathbb{R}^n} g(\mathbf{x})$ in time $\tilde{O}(n^{2.044}\log 1/\epsilon)$.*

1.2. Application: Least Squares Regression over Hilbert spaces

Another significant application of the Nyström method is in speeding up kernel ridge regression for learning over Hilbert spaces. In this model, we receive n i.i.d. samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ from an unknown probability measure ρ over $\mathcal{H} \times \mathbb{R}$ where $\mathcal{H}$ is a Hilbert space, and our goal is to approximate the expected risk minimizer under square loss, $\inf_{f\in\mathcal{H}} \mathbb{E}_\rho[(f(\mathbf{x})-y)^2]$, where $f(\mathbf{x}) := \langle f, \mathbf{x}\rangle_{\mathcal{H}}$. When the minimizing hypothesis $f_{\mathcal{H}}$ lies in $\mathcal{H}$ (the attainable case), this gives rise to a very

well understood regression framework, which can be effectively learned via kernel ridge regression:

$$\hat{f} = \sum_{i=1}^{n} w_i \mathbf{x}_i, \quad \text{where} \quad \mathbf{w} = (\mathbf{K} + n\lambda^* \mathbf{I})^{-1}\mathbf{y}, \quad \mathbf{K} = \big[\langle \mathbf{x}_i, \mathbf{x}_j \rangle_{\mathcal{H}}\big]_{ij}.$$

Optimal learning rates for kernel ridge regression have been derived in the attainable case (Caponnetto and De Vito, 2007), in particular ensuring that the effective dimension under optimal regularizer $n\lambda^*$ is always bounded by $O(\sqrt{n})$. This enables using efficient Nyström-based approximations of $\mathbf{K}$ which recover the optimal rate with a low computational cost (Rudi et al., 2015, 2017, 2018).

However, the situation is different in the considerably harder unattainable case (Lin and Cevher, 2020; Lin et al., 2020), where the minimizing hypothesis does not lie in the Hilbert space, but rather, in a larger function space, as controlled by a smoothness parameter $\zeta \in (0, 1)$ (see Assumption 3). In the unattainable case (i.e., $\zeta < 1/2$), the effective dimension is often much larger (even close to n), making the Nyström method far more expensive if we wish to recover the convergence rate of full KRR. Thus, it is natural to ask how much we can reduce our computational budget by relaxing the desired convergence rate so that we can use a coarser α-approximation of $\mathbf{K}$. We do this by considering $n\lambda^*$-regularized α-approximation of $\mathbf{K}$, where $\alpha = O(n^\theta)$ for small enough $\theta > 0$.

Unfortunately, existing risk analysis for KRR applies only to projection-based kernel approximations, such as the classical Nyström method. Block-Nyström is no longer a strict projection, so to apply it, we must first extend the KRR risk analysis in the unattainable case to a broader class of kernel α-approximations. We do this in the following result, which is stated here informally for Block-Nyström, but applies much more broadly and thus should be of independent interest.

Theorem 4 (Block-Nyström KRR, informal Corollary 15) *Given Hilbert space $\mathcal{H}$ and probability measure ρ over $\mathcal{H} \times \mathbb{R}$ with smoothness $\zeta \in (0, 1/2)$ (Assumption 3) and capacity $\gamma \in (0, 1)$ (Assumption 4), let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be i.i.d. samples from ρ. Let $\hat{\mathbf{K}}$ be the Block-Nyström approximation (Theorem 1) of $\mathbf{K} = [\langle \mathbf{x}_i, \mathbf{x}_j \rangle_{\mathcal{H}}]_{i,j}$ with the optimal KRR regularizer $n\lambda^*$ and some $\alpha \geq 1$. Then, the estimate $\hat{f} = \sum_i \hat{w}_i \mathbf{x}_i$ constructed from $\hat{\mathbf{w}} = (\hat{\mathbf{K}} + n\lambda^* \mathbf{I})^{-1}\mathbf{y}$ achieves risk bound:*

$$\|\hat{f} - f_{\mathcal{H}}\|_\rho \lesssim \alpha \cdot n^{-\frac{\zeta}{\min\{1, 2\zeta+\gamma\}}}.$$

Remark 5 *The above convergence rate matches the rate achieved by full KRR (i.e., without kernel approximation) obtained by Lin et al. (2020), up to the α factor. The same guarantee can also be obtained for an α-approximation based on the classical Nyström method.*

In Section 5, we use this analysis to evaluate the computational gain of relaxing the convergence rate exponent by setting $\alpha = n^\theta$ for $\theta > 0$ small enough to still ensure convergence. We show that, similarly as in the general psd matrix approximation task, Block-Nyström achieves a better computational gain than $n\lambda^*$-regularized classical Nyström when θ is small but positive, across all values of smoothness and capacity in the unattainable setting.

1.3. Background and further related work

The Nyström method (Nyström, 1930) was first used in a machine learning context by Williams and Seeger (2000) to approximate kernel matrices arising in Gaussian process regression. Initially, uniform sampling was used for landmark selection, with other sampling schemes considered by

Kumar et al. (2009) among others, until Alaoui and Mahoney (2015) showed that sampling landmarks according to ridge leverage scores leads to strong approximation guarantees in terms of the effective dimension. Learning guarantees for Nyström-based KRR in the fixed design model were considered by Bach (2013), and then adapted to Hilbert spaces by Rudi et al. (2015, 2017). Fast ridge leverage score sampling schemes were designed for this setting by Rudi et al. (2018). Recent works have considered extensions of Nyström to the unattainable setting (e.g., Lin et al., 2020) and distributed architectures (e.g., Li et al., 2023). Other approximate KRR solvers have also been studied, including random features (e.g., Rudi and Rosasco, 2017) and stochastic optimization (e.g., Lin and Cevher, 2018; Abedsoltan et al., 2023).

The Nyström method has also been studied in randomized linear algebra as a low-rank approximation method (e.g., Halko et al., 2011), with many generalizations such as using sketching instead of landmark sampling (Gittens and Mahoney, 2013). One of the key applications in this context is preconditioning iterative algorithms for solving quadratic minimization to high precision (e.g., Avron et al., 2017; Frangella et al., 2023; Díaz et al., 2023; Abedsoltan et al., 2024; Dereziński et al., 2025), as well as for other convex optimization settings (e.g., Frangella et al., 2024).

2. Preliminaries

Let λ_i denote the i^{th} eigenvalue of $\mathbf{A}$, with $\lambda_1/\lambda_n := \kappa$ denoting the condition number of $\mathbf{A}$. We use $\preceq$ to denote the Loewner ordering and define $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^\top \mathbf{A}\mathbf{x}}$. $\mathbf{S} \in \mathbb{R}^{s\times n}$ is a sub-sampling matrix for $\{p_i\}_{i=1}^n$ if its rows are standard basis vectors $\mathbf{e}_i^\top$ sampled i.i.d. proportionally to p_i's.

Definition 6 (λ-ridge leverage scores of $\mathbf{A}$) *Given a psd matrix $\mathbf{A}$ and $\lambda > 0$, we define the ith λ-ridge leverage score $\ell_i(\mathbf{A},\lambda)$ as $[\mathbf{A}(\mathbf{A}+\lambda\mathbf{I})^{-1}]_{i,i}$. Also we define $d_\lambda(\mathbf{A}) := \mathrm{tr}(\mathbf{A}(\mathbf{A}+\lambda\mathbf{I})^{-1})$ as the λ-effective dimension. We say that $\tilde{\ell}_1, ..., \tilde{\ell}_n$ are $O(1)$-approximate λ-ridge leverage scores of $\mathbf{A}$ if $\tilde{\ell}_i \geq \ell_i(\mathbf{A},\lambda)/T$ for all i and $T = O(1)$, and $\sum_i \tilde{\ell}_i = O(d_\lambda(\mathbf{A}))$.*

In our results, we rely on two different algorithms for computing approximate ridge leverage scores.

Lemma 7 (Adapted from Theorem 7 of Musco and Musco (2017)) *There exists an algorithm that provides $\tilde{\ell}_i(\mathbf{A},\lambda)$ i.e., $O(1)$-approximate λ-ridge leverage scores of $\mathbf{A}$ in time $\tilde{O}(nd_\lambda(\mathbf{A})^2)$.*

If we additionally assume that $\mathbf{A}_{ii} = O(1)$ for all i, which is common in kernel literature, then it is possible to generate a leverage score sample even faster than what is guaranteed in Lemma 7.

Lemma 8 (Adapted from Theorem 1 of Rudi et al. (2018)) *Let $\mathbf{A}_{ii} = O(1)$ for all i. Given $\lambda > 0$, we can generate a sub-sampling matrix $\mathbf{S} \in \mathbb{R}^{s\times n}$ with $s = \tilde{O}(d_\lambda(\mathbf{A}))$, sampled i.i.d. from $O(1)$-approximate λ-ridge leverage scores of $\mathbf{A}$ in time $\tilde{O}\big(\min(\frac{1}{\lambda}, n)d_\lambda(\mathbf{A})^2\big)$.*

Ridge leverage score sampling provides the following standard Nyström approximation guarantee.

Lemma 9 (Theorem 3 of Musco and Musco (2017)) *If $\mathbf{S}$ is a sub-sampling matrix corresponding to $\tilde{O}(d_\lambda(\mathbf{A}))$ i.i.d. samples according to $O(1)$-approximate λ-ridge leverage scores of $\mathbf{A}$, then the Nyström matrix $\hat{\mathbf{A}} = \mathbf{A}\mathbf{S}^\top(\mathbf{S}\mathbf{A}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{A}$ with high probability satisfies $\|\hat{\mathbf{A}} - \mathbf{A}\| \leq \lambda$.*

3. Analysis of Block-Nyström

In this section, we give the analysis of Block-Nyström method. First, we provide the approximation guarantees by optimizing bias-variance trade-offs in Nyström, and then we describe our recursive preconditioning algorithm for inverting the regularized Block-Nyström matrix. We conclude the section with the proof of Theorem 1.

3.1. Optimizing bias-variance trade-offs

First, let us discuss how the classical Nyström guarantees relate to the λ-regularized α-approximation condition (1). Note that, given an $n \times n$ psd matrix $\mathbf{A}$ and a sub-sampling matrix $\mathbf{S} \in \mathbb{R}^{s\times n}$, the Nyström approximation can be defined as $\hat{\mathbf{A}} = \mathbf{A}^{1/2}\mathbf{P}\mathbf{A}^{1/2}$, where $\mathbf{P} = \mathbf{A}^{1/2}\mathbf{S}^\top(\mathbf{S}\mathbf{A}\mathbf{S}^\top)^\dagger\mathbf{S}\mathbf{A}^{1/2}$ is an orthogonal projection matrix onto the row-span of $\mathbf{S}\mathbf{A}^{1/2}$. This implies that $\hat{\mathbf{A}} \preceq \mathbf{A}$. Moreover, if $\mathbf{S}$ is produced by λ-ridge leverage score sampling of $s = \tilde{O}(d_\lambda(\mathbf{A}))$ landmarks, then standard guarantees (e.g., Lemma 9) give $\|\mathbf{A} - \hat{\mathbf{A}}\| \leq \lambda$. In particular, this implies that $\mathbf{A} + \lambda\mathbf{I} \preceq \hat{\mathbf{A}} + 2\lambda\mathbf{I} \preceq 2(\hat{\mathbf{A}} + \lambda\mathbf{I})$. Together, these properties yield (1) with $\alpha = 2$.

What if we wish to construct a smaller Nyström approximation with rank $\tilde{O}(d_{\lambda'}(\mathbf{A}))$ for some $\lambda' > \lambda$, and still use it with a small regularizer λ? In that case, we can trivially observe that $\mathbf{A} + \lambda\mathbf{I} \preceq (1 + \frac{\lambda'}{\lambda})(\hat{\mathbf{A}} + \lambda\mathbf{I})$, which implies guarantee (1) with $\alpha = 1 + \lambda'/\lambda$. This is not particularly satisfying, and we could not really hope to do much better, since $\hat{\mathbf{A}}$ only has $\tilde{O}(d_{\lambda'}(\mathbf{A}))$ non-zero eigenvalues, which are associated primarily with the eigenvalues of $\mathbf{A}$ above the λ' threshold. Thus, it contains little information about $\mathbf{A}$'s spectrum in the $[\lambda, \lambda']$ interval.

However, if we could use the expectation $\mathbb{E}[\hat{\mathbf{A}}] = \mathbf{A}^{1/2}\mathbb{E}[\mathbf{P}]\mathbf{A}^{1/2}$ as the approximation of $\mathbf{A}$, then the story would change. Even though $\mathbf{P}$ is a low-rank projection, its expectation $\mathbb{E}[\mathbf{P}]$ is the same rank as $\mathbf{A}$, so it carries some information about all directions in $\mathbf{A}$'s spectrum. Expected projection matrices arise in many contexts, and a number of works have studied their properties under various sketching and sub-sampling matrices $\mathbf{S}$ (e.g., Rodomanov and Kropotov, 2020; Mutny et al., 2020; Dereziński and Rebrova, 2024), including for λ'-ridge leverage score sampling (Rathore et al., 2025). From these results, we can show that if the sample size is at least $\tilde{O}(d_{\lambda'}(\mathbf{A}))$, then:

$$\text{(Lemma 17)} \qquad \mathbb{E}[\mathbf{P}] \succeq \frac{1}{2}\mathbf{A}(\mathbf{A} + \lambda'\mathbf{I})^{-1}. \tag{3}$$

This characterization suggests that $\mathbb{E}[\hat{\mathbf{A}}]$ provides meaningful estimates even for the spectral tail of $\mathbf{A}$. Indeed, we show that given a Nyström approximation $\hat{\mathbf{A}}$ based on λ'-ridge leverage score sampling, for any $\lambda \leq \lambda'$ we have:

$$(\mathbb{E}[\hat{\mathbf{A}}] + \lambda\mathbf{I}) \succeq \sqrt{\lambda/4\lambda'} \cdot (\mathbf{A} + \lambda\mathbf{I}). \tag{4}$$

So, $\mathbb{E}[\hat{\mathbf{A}}]$ is a non-trivial $O(\sqrt{\lambda'/\lambda})$-approximation of $\mathbf{A}$ for any $\lambda < \lambda'$. This is compared to the $O(\lambda'/\lambda)$-approximation achieved by $\hat{\mathbf{A}}$, which is essentially vacuous in the tail of the spectrum.

Naturally, $\mathbb{E}[\hat{\mathbf{A}}]$ is not a practical approximation strategy, but it provides the motivation for our Block-Nyström method. Our approach is to approximate $\mathbb{E}[\hat{\mathbf{A}}]$ by drawing several independent copies $\hat{\mathbf{A}}_1$, ..., $\hat{\mathbf{A}}_q$, and averaging them together, to obtain $\hat{\mathbf{A}}_{[q]} = \frac{1}{q}\sum_{i=1}^q \hat{\mathbf{A}}_i$. This reduces the inherent variance of $\hat{\mathbf{A}}$ in the tail of the spectrum and, with large enough q, should lead to an $O(\sqrt{\lambda'/\lambda})$-approximation, matching the guarantee we obtained for $\mathbb{E}[\hat{\mathbf{A}}]$.

Fortunately, via a matrix Chernoff argument, we are able to show that a relatively moderate number of Nyström copies is sufficient to recover the improved guarantee, showing that $\hat{\mathbf{A}}_{[q]}$ is with high probability an $\tilde{O}(\sqrt{\lambda'/\lambda} + \lambda'/(q\lambda))$-approximation. Here, the first term can be viewed as the contribution of the bias, whereas the second term is the variance which goes down with q.

Optimizing over this bias-variance error decomposition, we conclude that to obtain an $O(\alpha)$-approximation, one can choose $\lambda' = \alpha^2\lambda$, and then set $q = \tilde{O}(\sqrt{\lambda'/\lambda}) = \tilde{O}(\alpha)$. We summarize this in the following result, which is proven in Appendix B.

Theorem 10 (Spectral approximation with Block-Nyström) *Given $\lambda > 0$ and $\alpha \geq 1$, let $\{p_i\}_{i=1}^n$ denote the $O(1)$-approximate $\alpha^2\lambda$-ridge leverage scores of $\mathbf{A}$. Let $\hat{\mathbf{A}}_{[q]} = \frac{1}{q}\sum_{i=1}^q \hat{\mathbf{A}}_i$ where $\hat{\mathbf{A}}_i$'s are iid Nyström approximations of $\mathbf{A}$ sampled i.i.d. from $\{p_i\}_{i=1}^n$, each with $O(d_{\alpha^2\lambda}(\mathbf{A})\log^3(n/\delta))$ landmarks. If $q \geq 32\alpha\log(n/\delta)$, then with probability $1-\delta$, the matrix $\hat{\mathbf{A}}_{[q]}$ is a λ-regularized 16α-approximation of $\mathbf{A}$, i.e.,*

$$(64\alpha)^{-1}(\mathbf{A} + \lambda\mathbf{I}) \preceq \hat{\mathbf{A}}_{[q]} + \lambda\mathbf{I} \preceq \mathbf{A} + \lambda\mathbf{I}.$$

3.2. Fast inversion via recursive preconditioning

The computational benefits in the preprocessing cost of Block-Nyström are easy to verify: The cost of constructing a single Nyström approximation $\hat{\mathbf{A}} = \mathbf{C}\mathbf{W}^{-1}\mathbf{C}^\top$ scales cubically with the number of landmarks, because we must invert the matrix $\mathbf{W}$, but the cost of constructing many small Nyström approximations grows only linearly with the number of landmarks. These benefits become even more significant when we account for the cost of ridge leverage score sampling. However, it is much less straightforward to retain that speed-up if we need to also quickly invert the approximation, as is the case in kernel ridge regression, for example. Remarkably, we are able to show that this can still be done efficiently, by significantly departing from standard block-inversion arguments.

Our task is to quickly apply $(\hat{\mathbf{A}}_{[q]}+\lambda\mathbf{I})^{-1}$ to any given vector $\mathbf{v} \in \mathbb{R}^n$. A standard approach that can be applied to most low-rank approximations is to use the Woodbury inversion formula, which implies that for any matrix $\hat{\mathbf{A}} = \mathbf{C}\mathbf{W}^{-1}\mathbf{C}^\top$ we have $(\hat{\mathbf{A}}+\lambda\mathbf{I})^{-1} = \frac{1}{\lambda}(\mathbf{I} - \mathbf{C}(\mathbf{C}^\top\mathbf{C}+\lambda\mathbf{W})^{-1}\mathbf{C}^\top)$. Thus, it suffices to pre-compute the matrix $(\mathbf{C}^\top\mathbf{C} + \lambda\mathbf{W})^{-1}$ in $O(nm^2 + m^3)$ time, where m is the number of landmarks, so that we can apply the inverse to a vector $\mathbf{v}$ in $O(nm)$ operations. This can be further improved by using an iterative solver for applying the inverse $(\mathbf{C}^\top\mathbf{C} + \lambda\mathbf{W})^{-1}$, reducing the preprocessing cost to $\tilde{O}(nm + m^3)$ (Dereziński et al., 2025). In principle, this strategy could still be applied to invert the Block-Nyström approximation by relying on decomposition (2). However, unfortunately the matrix $\mathbf{C}^\top\mathbf{C} + \lambda\mathbf{W}$ does not retain the block-diagonal structure of $\mathbf{W}$, so the preprocessing cost of this approach must scale cubically with the number of landmarks, thus erasing most of the computational benefit of Block-Nyström.

Remarkably, we show that one can still nearly-match the fast $O(nm)$ inversion time for Block-Nyström without incurring any such preprocessing cost. We achieve this by recursively reducing the task of applying $(\hat{\mathbf{A}}_{[q]} + \lambda\mathbf{I})^{-1}$ to the cheaper task of applying $(\hat{\mathbf{A}}_1 + \tilde{\lambda}\mathbf{I})^{-1}$ for a carefully chosen $\tilde{\lambda}$, where $\hat{\mathbf{A}}_1$ is one of the small Nyström components in $\hat{\mathbf{A}}_{[q]}$ (proof in Appendix B).

Theorem 11 (Recursive solver for Block-Nyström) *Given $\lambda > 0$, consider some $\alpha \geq 1$ and $q = O\big(\alpha\log(n/\delta)/\theta^2\big)$, and let $\hat{\mathbf{A}}_i$ for $1 \leq i \leq q$ be Nyström approximations of $\mathbf{A}$ sampled from $O(1)$-approximate $\alpha^2\lambda$-ridge leverage scores of $\mathbf{A}$, each of rank $O(d_{\alpha^2\lambda}(\mathbf{A})\log^3(n/\delta))$. Denote*

$\hat{\mathbf{A}}_{[q]} = \frac{1}{q}\sum_{i=1}^{q} \hat{\mathbf{A}}_i$. *Then, conditioned on an event with probability* $1-\delta$, *for any* $\mathbf{v} \in \mathbb{R}^n$, $\epsilon > 0$ *and* $0 < \phi < 1$ *we can find* $\mathbf{u} \in \mathbb{R}^n$ *such that:*

$$\left\| \mathbf{u} - \left(\hat{\mathbf{A}}_{[q]} + \lambda \mathbf{I}\right)^{-1} \mathbf{v} \right\| \leq \epsilon \cdot \|\mathbf{v}\|$$

in time $\tilde{O}\big(c \cdot \alpha^{1+\phi} n d_{\alpha^2\lambda}(\mathbf{A}) \log 1/\epsilon\big)$ *with a preprocessing cost of* $\tilde{O}(\alpha(n d_{\alpha^2\lambda}(\mathbf{A}) + d_{\alpha^2\lambda}(\mathbf{A})^3))$, *for any* c *such that* $c \geq \lceil \max\big\{ \log^{2/\phi}(20c^2(1+\theta)^4), 2\big\} \rceil$.

Remark 12 *Choosing* $\theta = 1/2$, $\phi = \frac{1}{\sqrt{\log c}}$ *and sufficiently large* $c = O(\log n)$, *we get runtime* $\tilde{O}(\alpha^{1+o(1)} n d_{\alpha^2\lambda}(\mathbf{A})) = \tilde{O}(nm \cdot \alpha^{o(1)})$, *where* $m = \tilde{O}(\alpha d_{\alpha\lambda}(\mathbf{A}))$ *is the total number of landmarks across all* q *Nyström matrices. Similarly, the preprocessing cost can be written as* $\tilde{O}(nm + m^3/\alpha^2)$.

Proof Sketch. For any integer $k \geq 0$, consider $q_k = \tilde{O}(\frac{\alpha}{\theta^2 c^k})$ and $\mathbf{M}_k = \hat{\mathbf{A}}_{[q_k]} + c^{2k}\lambda \mathbf{I}$. Observe that we need to approximate $\mathbf{M}_0^{-1}\mathbf{v}$. The main idea here is to solve the linear system $\mathbf{M}_0\mathbf{u} = \mathbf{v}$ to ϵ accuracy by preconditioning $\mathbf{M}_0$ with $\mathbf{M}_1$, which in turn leads to solving the linear system $\mathbf{M}_1\mathbf{u} = \mathbf{v}$ for some $\mathbf{v} \in \mathbb{R}^n$. In fact, by solving the linear system $\mathbf{M}_1\mathbf{u} = \mathbf{v}$ to a sufficient accuracy, one can obtain an ϵ-approximate solution to $\mathbf{M}_0\mathbf{u} = \mathbf{v}$ (see Lemma 16). Now recursing on k, and letting $k = O\big(\log(\alpha)/\log(c)\big)$, one reaches $\mathbf{M}_{k+1}$ with $q_{k+1} = \tilde{O}(1)$. At that stage, we precondition $\mathbf{M}_{k+1}$ with just $\hat{\mathbf{A}}_1 + \lambda\mathbf{I}$.

The first main argument of the proof justifies preconditioning of $\mathbf{M}_j$ with $\mathbf{M}_{j+1}$ by proving that for any $0 \leq j \leq k$, we have $\kappa(\mathbf{M}_j^{-1/2}\mathbf{M}_{j+1}\mathbf{M}_j^{-1/2}) \leq c^2(1+\theta)^4$. We obtain this by relying on matrix concentration arguments to show that $\kappa\big(\mathbf{M}_j^{-1/2}\mathbb{E}\mathbf{M}_j\mathbf{M}_j^{-1/2}\big) \leq (1+\theta)^2$ and also observing that $\kappa\big((\mathbb{E}\mathbf{M}_j)^{-1/2}\mathbb{E}\mathbf{M}_{j+1}(\mathbb{E}\mathbf{M}_j)^{-1/2}\big) \leq c^2$ for each j. By design, there is a trade-off associated with choosing c, with large c leading to computational gains at the expense of quality of preconditioning and vice-a-versa. Furthermore, as each recursive step leads to approximately solving a linear system, the approximation error can accumulate multiplicatively with the depth of the recursion. Despite all that, we show that one can pick c carefully enough such that the overall runtime complexity is $\tilde{O}(c \cdot q^{1+\phi} n d_{\alpha^2\lambda}(\mathbf{A}))$ for any $\phi > 0$. We finish the proof using a recent result (Lemma 25) from Dereziński et al. (2025) stating that $(\hat{\mathbf{A}}_1 + \alpha^2\lambda\mathbf{I})^{-1}\mathbf{u} = \mathbf{v}$ can be solved appproximately in time $\tilde{O}(n d_{\alpha^2\lambda}(\mathbf{A}))$ after an $\tilde{O}(n d_{\alpha^2\lambda}(\mathbf{A}) + d_{\alpha^2\lambda}(\mathbf{A})^3)$ preprocessing cost.

3.3. Proof of Theorem 1

Let $q = \tilde{O}(\alpha)$ and $\hat{\mathbf{A}}_{[q]} + \lambda\mathbf{I} = \frac{1}{q}\sum_{i=1}^{q} \hat{\mathbf{A}}_i + \lambda\mathbf{I}$, where each $\hat{\mathbf{A}}_i$ is a Nyström approximation of $\mathbf{A}$ of rank $\tilde{O}(d_{\alpha^2\lambda}(\mathbf{A}))$ sampled from $O(1)$-approximate $\alpha^2\lambda$-ridge leverage scores of $\mathbf{A}$. Let $m = \tilde{O}(\alpha d_{\alpha^2\lambda}(\mathbf{A}))$ be the number of landmarks across all $\hat{\mathbf{A}}_i$'s. The ridge leverage scores can be approximated using recursive techniques from Musco and Musco (2017), resulting in cost $\tilde{O}(n d_{\alpha^2\lambda}(\mathbf{A})^2) = \tilde{O}(nm/\alpha^2)$. Furthermore, as each individual $\hat{\mathbf{A}}_i$ has the decomposition $\mathbf{C}_i\mathbf{W}_i^{-1}\mathbf{C}_i^\top$ where $\mathbf{C}_i \in \mathbb{R}^{n \times \tilde{O}(d_{\alpha^2\lambda}(\mathbf{A}))}$ and $\mathbf{W}_i \in \mathbb{R}^{\tilde{O}(d_{\alpha^2\lambda}(\mathbf{A})) \times \tilde{O}(d_{\alpha^2\lambda}(\mathbf{A}))}$, we compute $\mathbf{C}_i$ and $\mathbf{W}_i^{-1}$ for all q Nyström approximations costing $\tilde{O}(\alpha(n d_{\alpha^2\lambda}(\mathbf{A}) + d_{\alpha^2\lambda}(\mathbf{A})^3)) = \tilde{O}(nm + m^3/\alpha^2)$. The total construction cost is therefore $\tilde{O}\big((nm + m^3)/\alpha^2 + nm\big)$. In Theorem 10 we have shown that $\hat{\mathbf{A}}_{[q]} + \lambda\mathbf{I}$ forms an $O(\alpha)$-approximation of $\mathbf{A} + \lambda\mathbf{I}$ in the sense of (1). Secondly, for any vector $\mathbf{v} \in \mathbb{R}^n$, multiplication of $\hat{\mathbf{A}}_i$ with $\mathbf{v}$ can be carried out sequentially from right to left by exploiting the decomposition of $\hat{\mathbf{A}}_i$, resulting in total cost $q \cdot \tilde{O}(n d_{\alpha^2\lambda}(\mathbf{A})) = \tilde{O}(nm)$. Furthermore, as shown in Theorem 11, $\left(\hat{\mathbf{A}}_{[q]} + \lambda\mathbf{I}\right)^{-1}\mathbf{v}$ can be approximated to high precision in time $\tilde{O}\big(nm \cdot \alpha^{o(1)}\big)$.

4. Block-Nyström Preconditioner for Quadratic Minimization

In this section, we present an application of Block-Nyström for minimizing a quadratic function $g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{x}^\top \mathbf{b}$, given an $n \times n$ psd matrix $\mathbf{A}$ and a vector $\mathbf{b} \in \mathbb{R}^n$, specifically using Block-Nyström as a preconditioner for minimizing $g(\mathbf{x})$, concluding with the proof of Corollary 3. For a more direct comparison with recent related works, we will state the time complexities in this section using the fast matrix multiplication exponent $\omega \approx 2.372$.

We consider the setting of Corollary 3. Let $\mathbf{A}$ have at most k eigenvalues larger than $O(1)$ times its smallest eigenvalue. Our first step is to approximate $\bar{\lambda}$-ridge leverage scores of $\mathbf{A}$, where $\bar{\lambda} = \frac{1}{k}\sum_{i>k} \lambda_i(\mathbf{A})$. The recursive sampling method from Musco and Musco (2017) provides $O(1)$-approximations for all $\bar{\lambda}$-ridge leverage scores in time $\tilde{O}(nk^{\omega-1})$, however, we show that if matrix $\mathbf{A}$ has flat-tailed spectrum, then the cost of leverage score approximation can be further optimized to $\tilde{O}(\mathrm{nnz}(\mathbf{A}) + k^\omega)$.

Lemma 13 *Given an $n \times n$ psd matrix $\mathbf{A}$ with at most k eigenvalues larger than $O(1)$ times its smallest eigenvalue, consider $\bar{\lambda} = \frac{1}{k}\sum_{i>k} \lambda_i(\mathbf{A})$. Then, $d_{\bar{\lambda}(\mathbf{A})} \leq 2k$, and we can compute $O(1)$-approximations of all $\bar{\lambda}$-ridge leverage scores of $\mathbf{A}$ in time $\tilde{O}(\mathrm{nnz}(\mathbf{A}) + k^\omega)$.*

Note that the $\tilde{O}(\mathrm{nnz}(\mathbf{A}) + k^\omega)$ runtime may be preferrable to the $\tilde{O}(nk^{\omega-1})$ runtime attained by Musco and Musco (2017) when k is sufficiently large or the matrix is sufficiently sparse. The proof of Lemma 13 can be found in Appendix D. We are now ready to prove Corollary 3.

Proof of Corollary 3 We invoke the Block-Nyström method (Theorem 1) for $\lambda = \bar{\lambda}/\alpha^2$, obtaining $\hat{\mathbf{A}}$, using ridge leverage sampling based on Lemma 13. Estimating the ridge leverage scores takes time $\tilde{O}(\mathrm{nnz}(\mathbf{A}) + k^\omega)$. Obtaining the data structure $\hat{\mathbf{A}}$ requires preprocessing, where we sample each $\hat{\mathbf{A}}_i$ for $1 \leq i \leq \tilde{O}(\alpha)$ and compute $\hat{\mathbf{C}}_i$, $\mathbf{W}_i^{-1}$. This preprocessing cost adds a factor of $\tilde{O}(\alpha(nk + k^\omega))$ to the overall cost. Now, we use $(\hat{\mathbf{A}} + \lambda\mathbf{I})^{-1}$ as the preconditioner for an iterative quadratic solver such as the one given in Corollary 2. Note that Theorem 1 implies that $\hat{\mathbf{A}}$ is a λ-regularized $O(\alpha)$-approximation of $\mathbf{A}$. Therefore, condition number after preconditioning $\mathbf{A}$ with $\hat{\mathbf{A}} + \lambda\mathbf{I}$ will then be

$$\kappa\Big((\hat{\mathbf{A}} + \lambda\mathbf{I})^{-1/2}\mathbf{A}(\hat{\mathbf{A}} + \lambda\mathbf{I})^{-1/2}\Big) = O\bigg(\alpha\Big(1 + \frac{\lambda}{\lambda_{\min}}\Big)\bigg) = O\Big(\alpha + \frac{n}{k\alpha}\Big),$$

where in the last step we used that $\bar{\lambda} = O(\frac{n}{k}\lambda_{\min})$ by the assumption on $\mathbf{A}$. Setting $\alpha = \sqrt{n/k}$, we obtain that the condition number after preconditioning is $O(\sqrt{n/k})$. Every iteration of the preconditioned solver requires matrix vector products with $\mathbf{A}$ costing $O(n^2)$ and with $(\hat{\mathbf{A}} + \lambda\mathbf{I})^{-1}$, costing $\tilde{O}(\alpha^{1+o(1)}nk)$, due to Lemma 11. Furthermore, for $\alpha = O(\sqrt{n/k})$ we have $\alpha nk = n\sqrt{nk} \leq n^2$. Thus, the overall cost of the solver includes $\tilde{O}(n^2 + k^\omega\alpha)$ for preprocessing, and then $\tilde{O}((n^2 + \alpha^{1+o(1)}nk)\sqrt{\alpha})$ for solving. Note that again we have $\alpha^{1+o(1)}nk = O(n^2)$. So, overall, we get the time complexity:

$$\tilde{O}\bigg(n^2\Big(\frac{n}{k}\Big)^{1/4} + k^\omega\Big(\frac{n}{k}\Big)^{1/2}\bigg).$$

Now, observe that if the first term dominates the second term, then we can drive k up to balance them. So, by optimizing over k, we get $k = n^{\frac{2-1/4}{\omega-1/4}} \approx n^{0.82}$, and plugging this into the bound we get $\tilde{O}(n^{\frac{\omega-2}{4\omega-1}} + k^\omega\sqrt{n/k}) = \tilde{O}(n^{2.044})$ for any $k \leq n^{0.82}$. ■

Discussion: The Block-Nyström preconditioner achieves $\tilde{O}(n^{2.044} + k^{\omega}\sqrt{n/k})$ time complexity, which is better than the $\tilde{O}(n^{0.065} + k^{\omega})$ runtime of the preconditioner by Dereziński et al. (2025) for any $k \leq n^{0.82}$. Interestingly, an altogether different approach to this problem was proposed by Dereziński and Yang (2024). They used a stochastic solver instead of a preconditioner, which means that their method cannot be paired with deterministic solvers and exploit fast matrix vector products with $\mathbf{A}$, unlike preconditioner-based methods such as Block-Nyström). Their approach achieves $\tilde{O}(n^2 + nk^{\omega-1})$ runtime, so our method also improves upon this approach, for any $k \geq n^{0.76}$.

5. Least Squares Regression over Hilbert Spaces

In this section, we present an application of Block-Nyström to kernel ridge regression over Hilbert spaces in hard learning scenarios. Along the way, we give a new risk bound for a class of KRR algorithms, which should be of independent interest. We start by introducing the learning model.

Let $\mathcal{H}$ be a separable Hilbert space and $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be sampled from an unknown distribution ρ over $\mathcal{H} \times \mathbb{R}$. Let $\rho_{\mathcal{X}}$ denote the marginal distribution over $\mathcal{H}$ and $\rho(y|\mathbf{x})$ denote the conditional distribution over $\mathbf{y} \in \mathbb{R}$. We denote the inner product over $\mathcal{H}$ by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and the induced norm by $\| \cdot \|_{\mathcal{H}}$. Let $\mathcal{H}_\rho$ be the Hilbert space consisting of linear forms over $\mathcal{H}$, defined as $\mathcal{H}_\rho = \{f : \mathcal{H} \to \mathbb{R} \,|\, f(\mathbf{x}) = \langle \mathbf{x}, \omega \rangle_{\mathcal{H}} \text{ for some } \omega \in \mathcal{H}\}$. We consider expected risk minimization:

$$\inf_{f \in \mathcal{H}_\rho} \mathcal{E}(f) = \int_{\mathcal{H} \times \mathbb{R}} (f(\mathbf{x}) - y)^2 d\rho(\mathbf{x}, y). \tag{5}$$

The above formulation for nonparametric regression provides a general framework for learning over Hilbert spaces. In particular, it covers learning over reproducing kernel Hilbert spaces and has been studied extensively in literature (e.g., Cucker and Zhou, 2007; Caponnetto and De Vito, 2007; Steinwart, 2008; Bauer et al., 2007; Lin and Rosasco, 2017). Note that the minimizer of expected risk in (5) need not exist in $\mathcal{H}_\rho$, leading to hard learning scenarios, known as unattainable learning settings. We next describe the standard assumptions for this learning model.

Assumption 1 (Bounded kernel assumption) *The support of $\rho_{\mathcal{X}}$ is compact and there exists a constant $G > 0$ such that $\langle \mathbf{x}, \mathbf{x}' \rangle_{\mathcal{H}} \leq G^2 \; \forall \, \mathbf{x}, \mathbf{x}' \in \mathcal{H}$, almost surely.*

Assumption 2 (Moments assumption) *There exist constants $M, \sigma > 0$ such that for any integer $p \geq 2$, we have $\int_{\mathbb{R}} |y|^p d\rho(y|\mathbf{x}) < \frac{1}{2} p! M^{p-2} \sigma^2$ almost surely.*

Consider the Hilbert space $L^2(\mathcal{H}, \rho_{\mathcal{X}}) := \{f : \mathcal{H} \to \mathbb{R} \mid \int_{\mathcal{H}} f(\mathbf{x})^2 d\rho_{\mathcal{X}}(\mathbf{x}) < \infty\}$. The norm on $L^2(\mathcal{H}, \rho_{\mathcal{X}})$ will be denoted by $\|\cdot\|_\rho$, and for any $f \in L^2(\mathcal{H}, \rho_{\mathcal{X}})$ we have $\|f\|_\rho = (f(\mathbf{x})^2 d\rho_{\mathcal{X}}(\mathbf{x}))^{1/2}$. Let $\mathcal{S}_\rho : \mathcal{H} \to L^2(\mathcal{H}, \rho_{\mathcal{X}})$ be the map $\mathcal{S}_\rho \omega = \langle \cdot, \omega \rangle_{\mathcal{H}}$ and $\mathcal{S}_\rho^* : L^2(\mathcal{H}, \rho_{\mathcal{X}}) \to \mathcal{H}$ be the adjoint map of $\mathcal{S}_\rho$. The covariance operator $\mathcal{S}_\rho^* \mathcal{S}_\rho$ will be denoted as $\mathcal{C}$, whereas the integral operator $\mathcal{S}_\rho \mathcal{S}_\rho^*$ will be denoted as $\mathcal{L}$. Let $\mathcal{Z}_n : \mathcal{H} \to \mathbb{R}^n$ be $\mathcal{Z}_n \omega = \big(\langle \mathbf{x}_i, \omega \rangle\big)_{i=1}^n$. An important observation here is that the operator $\mathcal{Z}_n \mathcal{Z}_n^* : \mathbb{R}^n \to \mathbb{R}^n$ corresponds to the $n \times n$ psd kernel matrix $\mathbf{K} = [\langle \mathbf{x}_i, \mathbf{x}_j \rangle_{\mathcal{H}}]_{i,j}$ over the n points $\mathbf{x}_i$ for $1 \leq i \leq n$. The minimizer of $\mathcal{E}(f)$ over all functions in $L^2(\mathcal{H}, \rho_{\mathcal{X}})$ is attained by the regression function f_ρ (Cucker and Zhou, 2007) defined as $f_\rho = \int_{\mathbb{R}} y d\rho(y|\mathbf{x})$. The minimizer of (5) is given by the orthogonal projection of f_ρ onto the closure of $\mathcal{H}_\rho$ in $L^2(\mathcal{H}, \rho_{\mathcal{X}})$. The orthogonal projection of f_ρ onto the closure of $\mathcal{H}_\rho$ will be denoted by $f_{\mathcal{H}}$.

Assumption 3 (Smoothness assumption) *There exist positive constants C, R, ζ such that we have $\int_{\mathcal{H}} \big(f_{\mathcal{H}}(\mathbf{x}) - f_\rho(\mathbf{x})\big)^2 \mathbf{x} \otimes \mathbf{x} d\rho_{\mathcal{X}}(\mathbf{x}) \preceq C^2 \cdot \mathcal{C}$, and moreover, $f_{\mathcal{H}} = \mathcal{L}^\zeta g$, with $\|g\|_\rho \leq R$.*

Generally speaking, larger ζ corresponds to a more stringent assumption and implies that $f_{\mathcal{H}}$ can be well approximated by a hypothesis in $\mathcal{H}_\rho$. In particular, when $0 < \zeta < \frac{1}{2}$, the function $f_{\mathcal{H}}$ does not lie in $\mathcal{H}_\rho$ and this corresponds to the harder unattainable setting. We focus on this setting, because when $\zeta \geq 1/2$, then existing Nyström-based algorithms are essentially optimal (Rudi et al., 2018).

Assumption 4 (Capacity assumption) *Let $\lambda > 0$ and let $\mathcal{N}(\lambda)$ denote $\operatorname{tr}(\mathcal{C}(\mathcal{C}+\lambda\mathcal{I})^{-1})$, where $\mathcal{I}$ denotes the identity operator on $\mathcal{H}$. Then there exists a positive constant $0 \leq \gamma \leq 1$ and a constant c_γ such that $\mathcal{N}(\lambda) \leq c_\gamma \lambda^{-\gamma}$.*

Since $\mathcal{C}$ is a trace class operator, the above assumption is always satisfied with $\gamma = 1$ and $c_\gamma = G^2$. Moreover, if eigenvalues of $\mathcal{C}$ decay polynomially i.e. $\sigma_i = O(i^{-1/\gamma})$, then the above assumption is satisfied with $c_\gamma = O(\gamma/(\gamma-1))$. The quantity $\mathcal{N}(\lambda)$ is known as the λ-effective dimension of the covariance operator $\mathcal{C}$, and it is analogous to its matrix counterpart from Definition 6.

5.1. Expected risk of approximate KRR

Consider the λ-regularized empirical risk defined as:

$$\tilde{\mathcal{E}}(\omega, \lambda) = \frac{1}{n}\sum_{i=1}^{n} (\langle \omega, \mathbf{x}_i \rangle_{\mathcal{H}} - y_i)^2 + \lambda \|\omega\|_{\mathcal{H}}^2. \tag{6}$$

A straightforward application of the representer theorem shows that the minimizer of $\tilde{\mathcal{E}}(\omega, \lambda)$ in $\mathcal{H}$ lies in the n dimensional space $\mathcal{H}_n := \left\{ \sum_{i=1}^n w_i \mathbf{x}_i | \mathbf{w} \in \mathbb{R}^n \right\}$. Noting that $\mathcal{H}_n$ is the range of operator $\mathcal{Z}_n^*$, it is easily shown that minimizing (6) solves the linear system: $(\mathbf{K} + n\lambda\mathbf{I})\mathbf{w} = \mathbf{y}$.

The Nyström approximation of this problem can be interpreted as minimizing $\tilde{\mathcal{E}}(\omega, \lambda)$ over an m dimensional subspace of $\mathcal{H}_n$ defined by the $m \ll n$ Nyström landmarks. For a random sub-sampling matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$, the space $\mathcal{H}_m$ is given as $\left\{ \mathcal{Z}_n^* \mathbf{S}^\top \mathbf{w} | \mathbf{w} \in \mathbb{R}^m \right\}$. A simple exercise (see Lemma 28) shows that the minimizer of $\tilde{\mathcal{E}}(\omega, \lambda)$ is given by $\hat{\omega} = \hat{\mathcal{P}} \mathcal{Z}_n^* (\mathcal{Z}_n \hat{\mathcal{P}} \mathcal{Z}_n^* + n\lambda \mathbf{I})^{-1} \mathbf{y}$, where $\hat{\mathcal{P}}$ is the orthogonal projection operator onto the subspace $\mathcal{H}_m$. Finding $(\mathcal{Z}_n \hat{\mathcal{P}} \mathcal{Z}_n^* + n\lambda \mathbf{I})^{-1}\mathbf{y}$ corresponds to solving $(\hat{\mathbf{K}} + n\lambda\mathbf{I})\mathbf{w} = \mathbf{y}$, where $\hat{\mathbf{K}} = \mathcal{Z}_n \hat{\mathcal{P}} \mathcal{Z}_n^*$ is the Nyström approximation of $\mathbf{K}$. The associated statistical risk with $\hat{\omega}$ is analyzed by upper bounding $\|\mathcal{S}_\rho \hat{\omega} - f_{\mathcal{H}}\|_\rho$, since for $\hat{f} = \mathcal{S}_\rho \hat{\omega}$ one has $\mathcal{E}(\hat{f}) - \mathcal{E}(f_{\mathcal{H}}) = \|\mathcal{S}_\rho \hat{\omega} - f_{\mathcal{H}}\|_\rho^2$ (Cucker and Zhou, 2007; Steinwart, 2008).

While prior work (e.g., Lin et al., 2020) has analyzed projection-based algorithms in this model, their analysis does not cover Block-Nyström, because in this case $\hat{\mathcal{P}}$ is not an orthogonal projection. To address this, we consider the following more general family of algorithms.

Assumption 5 *Let $\hat{\mathcal{P}}$ be an operator on space $\mathcal{H}$ with properties: $0 \preceq \hat{\mathcal{P}} \preceq \mathcal{I}$ and $\mathcal{P}\hat{\mathcal{P}} = \hat{\mathcal{P}}$, where $\mathcal{P}$ denotes the orthogonal projection operator onto $\mathcal{H}_n$ and $\mathcal{I}$ denotes the identity operator. Then, for given $\lambda > 0$, we consider the hypothesis $\hat{\omega} := \hat{\mathcal{P}} \mathcal{Z}_n^* (\mathcal{Z}_n \hat{\mathcal{P}} \mathcal{Z}_n^* + n\lambda \mathbf{I})^{-1} \mathbf{y}$.*

This general learning model covers as special cases the models where $\hat{\mathcal{P}}$ is an orthogonal projection, as in Nyström KRR or sketching-based KRR methods. However, in the more general setup $\hat{\mathcal{P}}$ need not be a projection. For example, in Block-Nyström, we consider $\hat{\mathcal{P}} = \frac{1}{q}\sum_{i=1}^q \hat{\mathcal{P}}_i$ where $\hat{\mathcal{P}}_i$ are independent random orthogonal projections. In particular, $\hat{\mathcal{P}}_i = \mathcal{Z}_n^* \mathbf{S}_i^\top (\mathbf{S}_i \mathbf{K} \mathbf{S}_i^\top)^\dagger \mathbf{S}_i \mathcal{Z}_n$ where $\mathbf{S}_i \in \mathbb{R}^{m \times n}$ is a sub-sampling matrix, which yields Block-Nyström KRR, $\hat{\omega} = \mathcal{P} \mathcal{Z}_n^* (\hat{\mathbf{K}}_{[q]} + n\lambda \mathbf{I})^{-1} \mathbf{y}$ where $\hat{\mathbf{K}}_{[q]} = \frac{1}{q}\sum_{i=1}^q \hat{\mathbf{K}}_i$ and $\hat{\mathbf{K}}_i = \mathcal{Z}_n \hat{\mathcal{P}}_i \mathcal{Z}_n^*$ are iid Nyström approximations.

Theorem 14 (Expected risk of approximate KRR) *Under assumptions 1, 2, 3, 4, and 5, let* $\zeta < 1/2$, $n \geq O(G^2 \log(G^2/\delta))$ *and* $\frac{19G^2 \log(n/\delta)}{n} \leq \lambda \leq \|\mathcal{C}\|$. *If* $\hat{\mathcal{P}}$ *is constructed so that* $\hat{\mathbf{K}} = \mathcal{Z}_n^* \hat{\mathcal{P}} \mathcal{Z}_n$ *is an* $n\lambda$*-regularized* α*-approximation of* $\mathbf{K}$, *as defined in* (1), *then with high probability,*

$$\left\|\mathcal{S}_\rho \hat{\omega} - f_{\mathcal{H}}\right\|_\rho \leq \tilde{O}\Big(\alpha R \lambda^\zeta + \frac{1}{n\lambda^{1-\zeta}} + \frac{1}{\sqrt{n\lambda^\gamma}}\Big).$$

The proof of Theorem 14 (Appendix C) incorporates and extends techniques from Rudi et al. (2015), Lin et al. (2020), and Lin and Cevher (2020). While previous works require a projection-based guarantee for the approximate KRR (such as Lemma 9), our proof only requires the looser regularized approximation guarantee (1). Note that the projection-based arguments upper bound the expected risk by approximating only the top part of the spectrum of $\mathbf{K}$, while our analysis also relies on how well the tail of the spectrum is approximated, thus leveraging the benefits of Block-Nyström.

If we let $\alpha = O(1)$, then Theorem 14 recovers the optimal learning rate for least squares regression over Hilbert spaces as proved in Lin et al. (2020) for $\zeta < 1/2$. The optimal learning rate can be found by optimizing the right hand side over λ, leading to $\lambda^* \approx n^{-\frac{1}{\max\{1,2\zeta+\gamma\}}}$ [1]. This gives learning rate of $\tilde{O}(n^{-\rho^*})$, where $\rho^* = \zeta$ if $2\zeta + \gamma \leq 1$ and $\rho^* = \frac{\zeta}{2\zeta+\gamma}$ if $2\zeta + \gamma > 1$. Considering $q = 1$ and $\hat{\mathbf{K}}$ to be the Nyström approximation of $\mathbf{K}$ sampled using $n\lambda^*$-ridge leverage scores of $\mathbf{K}$, we recover the guarantee for classical Nyström KRR, as shown in Lin and Cevher (2020). In the unattainable setting, the overall time complexity is dominated by sampling Nyström landmarks according to $n\lambda^*$-ridge leverage scores of $\mathbf{K}$. Even using fast ridge leverage score sampling techniques from Rudi et al. (2018), the overall time complexity of running classical Nyström KRR could be as high as $\tilde{O}\big(n^{\frac{1+2\gamma}{\max\{1,2\zeta+\gamma\}}}\big)$. For instance, if we consider the regime when $\gamma \approx 1$ and ζ is small, then the cost could be as large as $O(n^3)$, completely erasing any computational gains.

To reduce this computational cost, suppose we allow larger α, e.g., $\alpha = n^\theta$ for a small $\theta > 0$. We show that, with the regularizer fixed to $n\lambda^*$, Block-Nyström improves upon the time complexity of classical Nyström KRR in this setting, while maintaining the excess risk factor of α.

Corollary 15 (Expected risk of Block-Nyström KRR) *Under Assumptions 1, 2, 3 and 4, suppose that* $\zeta < 1/2$ *and* $n \geq \tilde{O}(G^2 \log(G^2/\delta))$. *Also, let* $\lambda^* = \tilde{O}(n^{-\frac{1}{2\zeta+\gamma}})$ *if* $2\zeta + \gamma > 1$ *and* $\lambda^* = \tilde{O}(n^{-1})$ *if* $2\zeta + \gamma \leq 1$. *Consider Block-Nyström KRR,* $\hat{\omega}_{[q]} = \mathcal{P}_{[q]} \mathcal{Z}_n^* (\hat{\mathbf{K}}_{[q]} + n\lambda^* \mathbf{I})^{-1} \mathbf{y}$, *constructed using* $q = \tilde{O}(\alpha)$ *blocks, each with* $\tilde{O}(d_{\alpha^2\lambda^*}(\mathbf{K}))$ *leverage score samples. Then,*

$$\left\|\mathcal{S}_\rho \hat{\omega}_{[q]} - f_{\mathcal{H}}\right\|_\rho \leq \begin{cases} \alpha \cdot \tilde{O}(n^{-\frac{\zeta}{2\zeta+\gamma}}) & \text{if } 2\zeta + \gamma > 1, \\ \alpha \cdot \tilde{O}(n^{-\zeta}) & \text{if } 2\zeta + \gamma \leq 1. \end{cases}$$

5.2. Cost comparison with classical Nyström

In Lemmas 33 and 34 (Appendix C), we provide the time complexity analysis of Nyström and Block-Nyström KRR, for achieving near-optimal risk $\tilde{O}(n^{-\rho^*+\theta})$, where $\rho^* = \frac{\zeta}{\max\{1,2\zeta+\gamma\}}$ is the optimal rate and $\theta > 0$ is a small constant (i.e., $\alpha = n^\theta$ in Corollary 15). We show that when θ is sufficiently small and regularizer is $n\lambda^*$, the computational gains with Block-Nyström KRR are at the order of $n^{2\theta\gamma}$ when compared with classical Nyström KRR. In hard learning problems involving kernels with slow polynomial decay i.e., $\gamma \approx 1$, this yields a significant improvement

1. Here, $\approx$ means upto constants depending on G, $\|\mathcal{C}\|$ and $\log(n/\delta)$ factors

of $\tilde{O}(n^{2\theta})$ over the existing projection-based KRR methods. Moreover, in Appendix C, we further explore the computational gains of Block-Nyström by considering all values of θ for which we can attain convergence. In particular, we observe that the optimized time complexity exhibits a phase transition between Block-Nyström and classical Nyström, which is attained by carefully reducing the number of blocks in Block-Nyström while maintaining the same α.

6. Conclusions

We propose Block-Nyström, a low-rank approximation method for positive semidefinite matrices which is computationally faster than the Nyström method for matrices with heavy-tailed spectrum. We show that Block-Nyström speeds up preconditioning for convex optimization, and can improve the cost of near-optimal learning for kernel ridge regression over Hilbert spaces.

Acknowledgments

The authors would like to acknowledge support from NSF award CCF-2338655.

References

Amirhesam Abedsoltan, Mikhail Belkin, and Parthe Pandit. Toward large kernel models. In *International Conference on Machine Learning*, pages 61–78. PMLR, 2023.

Amirhesam Abedsoltan, Parthe Pandit, Luis Rademacher, and Mikhail Belkin. On the nyström approximation for preconditioning in kernel machines. In *International Conference on Artificial Intelligence and Statistics*, pages 3718–3726. PMLR, 2024.

Ahmed El Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 775–783, December 2015.

Haim Avron, Kenneth L Clarkson, and David P Woodruff. Faster kernel ridge regression using sketching and preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1116–1138, 2017.

Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on learning theory*, pages 185–209. PMLR, 2013.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

David Burt, Carl Edward Rasmussen, and Mark Van Der Wilk. Rates of convergence for sparse variational gaussian process regression. In *International Conference on Machine Learning*, pages 862–871. PMLR, 2019.

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

Michael B Cohen, Jelani Nelson, and David P Woodruff. Optimal approximate matrix product in terms of stable rank. In *International Colloquium on Automata, Languages, and Programming*. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, 2016.

Felipe Cucker and Ding Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

Michał Dereziński and Elizaveta Rebrova. Sharp analysis of sketch-and-project methods via a connection to randomized singular value decomposition. *SIAM Journal on Mathematics of Data Science*, 6(1):127–153, 2024.

Michał Dereziński and Jiaming Yang. Solving dense linear systems faster than via preconditioning. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1118–1129, 2024.

Michał Dereziński, Christopher Musco, and Jiaming Yang. Faster linear systems and matrix norm approximation via multi-level sketched preconditioning. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1972–2004. SIAM, 2025.

Mateo Díaz, Ethan N Epperly, Zachary Frangella, Joel A Tropp, and Robert J Webber. Robust, randomized preconditioning for kernel ridge regression. *arXiv preprint arXiv:2304.12465*, 2023.

Murat A Erdogdu and Andrea Montanari. Convergence rates of sub-sampled Newton methods. *Advances in Neural Information Processing Systems*, 28:3052–3060, 2015.

Zachary Frangella, Joel A Tropp, and Madeleine Udell. Randomized nyström preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 44(2):718–752, 2023.

Zachary Frangella, Pratik Rathore, Shipu Zhao, and Madeleine Udell. Promise: Preconditioned stochastic optimization methods by incorporating scalable curvature estimates. *Journal of Machine Learning Research*, 25(346):1–57, 2024.

Junichi Fujii, Masatoshi Fujii, Takayuki Furuta, and Ritsuo Nakamoto. Norm inequalities equivalent to heinz inequality. *Proceedings of the American Mathematical Society*, 118(3):827–830, 1993.

Alex Gittens and Michael Mahoney. Revisiting the nystrom method for improved large-scale machine learning. In *International Conference on Machine Learning*, pages 567–575. PMLR, 2013.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53 (2):217–288, 2011.

Arun Jambulapati and Aaron Sidford. Ultrasparse ultrasparsifiers and faster laplacian system solvers. *ACM Trans. Algorithms*, February 2024. ISSN 1549-6325.

Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327, 2001.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling techniques for the nystrom method. In *Artificial intelligence and statistics*, pages 304–311. PMLR, 2009.

Jian Li, Yong Liu, and Weiping Wang. Optimal convergence rates for distributed nyström approximation. *Journal of Machine Learning Research*, 24(141):1–39, 2023.

Junhong Lin and Volkan Cevher. Optimal distributed learning with multi-pass stochastic gradient methods. In *International Conference on Machine Learning*, pages 3092–3101. PMLR, 2018.

Junhong Lin and Volkan Cevher. Convergences of regularized algorithms and stochastic gradient methods with random projections. *Journal of Machine Learning Research*, 21(20):1–44, 2020.

Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

Cameron Musco and Christopher Musco. Recursive sampling for the nystrom method. *Advances in neural information processing systems*, 30, 2017.

Mojmir Mutny, Michal Derezinski, and Andreas Krause. Convergence analysis of block coordinate algorithms with determinantal sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 3110–3120, 2020.

EJ Nyström. Über die praktische auflösung von integralgleichungen mit anwendungen auf randwertaufgaben. *Acta Mathematica*, 54(1):185–204, 1930.

Zheng Qu, Peter Richtárik, Martin Takác, and Olivier Fercoq. Sdna: Stochastic dual newton ascent for empirical risk minimization. In *International Conference on Machine Learning*, pages 1823–1832. PMLR, 2016.

Pratik Rathore, Zachary Frangella, Jiaming Yang, Michał Dereziński, and Madeleine Udell. Have ASkotch: Fast cocktails for large-scale kernel ridge regression. *arXiv preprint arXiv:2407.10070*, 2025.

Anton Rodomanov and Dmitry Kropotov. A randomized coordinate descent method with volume sampling. *SIAM Journal on Optimization*, 30(3):1878–1904, 2020.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. *Advances in neural information processing systems*, 30, 2017.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in neural information processing systems*, 28, 2015.

Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. *Advances in neural information processing systems*, 30, 2017.

Alessandro Rudi, Daniele Calandriello, Luigi Carratino, and Lorenzo Rosasco. On fast leverage score sampling and optimal learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Ingo Steinwart. *Support Vector Machines*. Springer, 2008.

Jingruo Sun, Zachary Frangella, and Madeleine Udell. Sapphire: Preconditioned stochastic variance reduction for faster large-scale statistical learning. *arXiv preprint arXiv:2501.15941*, 2025.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Haishan Ye, Luo Luo, and Zhihua Zhang. Nesterov's acceleration for approximate newton. *Journal of Machine Learning Research*, 21(142):1–37, 2020.

Appendix A. Auxiliary lemmas

In this section, we mention the auxiliary results and lemmas that are used in our analysis.

Lemma 16 (Adapted from Jambulapati and Sidford (2024)) *For any pd matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ with $\mathbf{A} \preceq \mathbf{B} \preceq \kappa\mathbf{A}$ for $\kappa \geq 1$, and f that is a $\frac{1}{10\kappa}$-solver for $\mathbf{B}$, i.e., given $\mathbf{v} \in \mathbb{R}^n$, returns $f(\mathbf{v})$ such that $\|f(\mathbf{b}) - \mathbf{B}^{-1}\mathbf{v}\|_{\mathbf{B}}^2 \leq \frac{1}{10\kappa}\|\mathbf{B}^{-1}\mathbf{v}\|_{\mathbf{B}}^2$, there is an ϵ-solver for $\mathbf{A}$ that applies f and $\mathbf{A}$ at most $\lceil 4\sqrt{\kappa}\log(2/\epsilon)\rceil$ times each, and spends additional $O(n\sqrt{\kappa}\ln(1/\epsilon))$ time when run.*

Lemma 17 (Rathore et al. (2025)) *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be psd and $\lambda' > 0$. Let $\{p_i\}_{i=1}^n$ be $O(1)$-approximate λ'-ridge leverage score sampling probabilities. Let $\mathbf{S} \in \mathbb{R}^{O(d_{\lambda'}(\mathbf{A})\log^3 n)\times n}$ be a random sub-sampling matrix, drawn from probability distribution $\{p_i\}_{i=1}^n$. Consider the random projection matrix $\mathbf{P} = \mathbf{A}^{1/2}\mathbf{S}^\top(\mathbf{S}\mathbf{A}\mathbf{S}^\top)^\dagger\mathbf{S}\mathbf{A}^{1/2}$. Then $\mathbb{E}[\mathbf{P}]$ satisfies*

$$\mathbb{E}[\mathbf{P}] \succeq \frac{1}{2}\mathbf{A}\left(\mathbf{A} + \lambda'\mathbf{I}\right)^{-1}. \tag{7}$$

Lemma 18 (Cordes inequality Fujii et al. (1993)) *Let $\mathcal{A}$ and $\mathcal{B}$ be two positive bounded linear operators on a separable Hilbert space. Then for $0 \leq s \leq 1$*

$$\|\mathcal{A}^s\mathcal{B}^s\| \leq \|\mathcal{A}\mathcal{B}\|^s.$$

Lemma 19 (Lemma 5 from Rudi et al. (2015)) *If assumptions 1 and 4 hold, then for any $\delta > 0$, $n \geq O(G^2\log(G^2/\delta))$ and $\frac{19G^2\log(n/\delta)}{n} \leq \lambda \leq \|\mathcal{C}\|$, we have,*

$$\|\mathcal{C}_\lambda^{1/2}\cdot\mathcal{C}_{n\lambda}^{-1/2}\| < 2$$

with probability $1-\delta$.

The assumption on n can be removed, (see Lemma 5.3 in Lin et al. (2020)) but leads to a more complicated upper bound depending on G^2 in place of constant 2. In this work, we are interested in a large n regime and therefore assume $n \geq O(G^2\log(G^2/\delta))$ obtaining a constant upper bound.

Lemma 20 (Proposition 1 from Rudi et al. (2015)) *If assumptions 1 and 4 hold, then for any $\delta > 0$, $n = O(G^2\log(G^2/\delta))$ and $\frac{19G^2\log(n/\delta)}{n} \leq \lambda \leq \|\mathcal{C}\|$, we have,*

$$d_{n\lambda}(\mathbf{K}) \leq 3\mathcal{N}(\lambda)$$

with probability $1-\delta$.

The next two results are essentially taken from Lin et al. (2020) and upper bound the true bias and the sample variance terms in our analysis.

Lemma 21 (Adapted from Lemma 5.2 in Lin et al. (2020)) *Let $0 \le a \le \zeta$ and $\omega = \mathcal{C}_\lambda^{-1}\mathcal{S}_\rho^* f_{\mathcal{H}}$. Under assumption 3*

$$\left\|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega - f_{\mathcal{H}})\right\|_\rho \le R\lambda^{\zeta-a}.$$

Lemma 22 (Adapted from Lemma 5.6 in Lin et al. (2020)) *Let $\hat{\mathbf{y}} = \mathbf{y}/\sqrt{n}$ and $\omega = \mathcal{C}_\lambda^{-1}\mathcal{S}_\rho^* f_{\mathcal{H}}$. Then under assumptions 1, 2, 3, 4 for any $\delta > 0$*

$$\left\|\mathcal{C}_{n\lambda}^{-1/2}[\mathcal{S}_n^*\hat{\mathbf{y}} - \mathcal{C}_n\omega]\right\|_{\mathcal{H}} \le C'\Big(\lambda^\zeta + \frac{1}{n\lambda^{\max(1/2,1-\zeta)}} + \frac{1}{\sqrt{n\lambda^\gamma}}\Big)\log(1/\delta)$$

with probability $1-\delta$. The constant C' depends on $R, G, M, C, c_\gamma, \zeta$, and σ^2.

Appendix B. Spectral approximation with Block-Nyström

We start this section by proving the spectral approximation guarantee (1) provided by the Block-Nyström approximation of $\mathbf{A}$.

Theorem 23 (Spectral approximation with Block-Nyström (Restated Theorem 10)) *Let $\lambda > 0$ be fixed and $\lambda' > \lambda$. Let $\{p_i\}_{i=1}^n$ denote the $O(1)$-approximate λ'-ridge leverage scores of $\mathbf{A}$. Let $\hat{\mathbf{A}}_{[q]} = \frac{1}{q}\sum_{i=1}^q \hat{\mathbf{A}}_i$ where $\hat{\mathbf{A}}_i$'s are iid Nyström approximations of $\mathbf{A}$ sampled independently from $\{p_i\}_{i=1}^n$, each with $\tilde{O}(d_{\lambda'}(\mathbf{A})\log(n/\delta))$ landmarks and $\mathbf{A}_\lambda = \mathbf{A} + \lambda\mathbf{I}$. Then with probability $1-\delta$,*

$$\|\mathbf{A}_\lambda^{1/2}(\hat{\mathbf{A}}_{[q]} + \lambda\mathbf{I})^{-1}\mathbf{A}_\lambda^{1/2}\| \le 16\sqrt{\frac{\lambda'}{\lambda}}$$

if $q > 64\sqrt{\frac{\lambda'}{\lambda}}\log(n/\delta)$.

Proof Define random matrices $\mathbf{Z}_i = \mathbf{A}_\lambda^{-1/2}(\hat{\mathbf{A}}_i + \lambda\mathbf{I})\mathbf{A}_\lambda^{-1/2}$, where $\hat{\mathbf{A}}_i = \mathbf{A}\mathbf{S}_i^\top(\mathbf{S}_i\mathbf{A}\mathbf{S}_i^\top)^\dagger\mathbf{S}_i\mathbf{A}$ and $\mathbf{A}_\lambda = \mathbf{A} + \lambda\mathbf{I}$. Here each $\mathbf{S}_i$ is a sub-sampling matrix corresponding to a set drawn independently from $\{1, 2, \cdots, n\}$ with probability distribution $\{p_i\}_{i=1}^n$. As $\hat{\mathbf{A}}_i \preceq \mathbf{A}$ for all i, we have $\mathbf{Z}_i \preceq \mathbf{I}$ for all i. Therefore $\|\mathbf{Z}_i\| \le 1 := R$. Let $\mu_{\min}$ denote the minimum eigenvalue of $\mathbb{E}\mathbf{Z}_i$. We have,

$$\begin{aligned}
\mu_{\min} &= \lambda_{\min}(\mathbf{A}_\lambda^{-1/2}\mathbb{E}[\hat{\mathbf{A}}_i + \lambda\mathbf{I}]\mathbf{A}_\lambda^{-1/2}) \\
&\ge \frac{1}{2}\lambda_{\min}\Big(\mathbf{A}_\lambda^{-1/2}\big(\mathbf{A}\mathbf{A}_{\lambda'}^{-1}\mathbf{A} + \lambda\mathbf{I}\big)\mathbf{A}_\lambda^{-1/2}\Big) \\
&= \frac{1}{2}\min_j\Big(\frac{1}{\lambda_j+\lambda}\cdot\big(\frac{\lambda_j^2}{\lambda_j+\lambda'} + \lambda\big)\Big) \\
&= \frac{1}{2}\min_j\Big(\frac{\lambda_j^2 + \lambda\lambda_j + \lambda\lambda'}{\lambda_j^2 + \lambda\lambda_j + \lambda'\lambda_j + \lambda\lambda'}\Big)
\end{aligned}$$

The second inequality holds due to Lemma 17 as $\mathbb{E}\hat{\mathbf{A}}_i \succeq \frac{1}{2}\mathbf{A}(\mathbf{A}+\lambda'\mathbf{I})^{-1}\mathbf{A}$. Considering λ_j as a variable x, a routine calculus exercise shows that the minimum is attained when $x = \sqrt{\lambda\lambda'}$. Therefore, we get,

$$\mu_{\min} \geq \frac{1}{2}\frac{\lambda\lambda' + \lambda\sqrt{\lambda\lambda'} + \lambda\lambda'}{\lambda\lambda' + \lambda\sqrt{\lambda\lambda'} + \lambda'\sqrt{\lambda\lambda'} + \lambda\lambda'} \geq \frac{1}{2}\frac{\lambda\lambda'}{4\lambda'\sqrt{\lambda\lambda'}} = \frac{1}{8}\sqrt{\frac{\lambda}{\lambda'}}.$$

Applying matrix Chernoff concentration inequality we get

$$\Pr\Big(\lambda_{\min}\big(\frac{1}{q}\sum_{i=1}^{q}\mathbf{Z}_i\big) \leq (1-\epsilon)\mu_{\min}\Big) \leq n\cdot\exp\Big(-\frac{\epsilon^2 q\mu_{\min}}{2R}\Big).$$

Substitute $R = 1$ and $\epsilon = \frac{1}{2}$, we get if $q > 64\sqrt{\frac{\lambda'}{\lambda}}\log(n/\delta)$, then with probability $1-\delta$ $\lambda_{\min}\big(\frac{1}{q}\sum_{i=1}^{q}\mathbf{Z}_i\big) \geq \frac{\mu_{\min}}{2}$. This gives us

$$\frac{1}{16}\sqrt{\frac{\lambda}{\lambda'}}\mathbf{A}_\lambda \preceq \frac{1}{q}\sum_{i=1}^{q}\hat{\mathbf{A}}_i + \lambda\mathbf{I} \preceq \mathbf{A}_\lambda,$$

and finishes the proof. ■

The following result is needed in the proof of our recursive solver for inverting Block-Nyström.

Theorem 24 (Matrix concentration for Block-Nyström) *Let $\lambda' > 0$ and for $1 \leq i \leq q$, let $\hat{\mathbf{A}}_i$ be Nyström approximations of $\mathbf{A}$ sampled from $O(1)$-approximate λ'-ridge leverage scores of $\mathbf{A}$, each with $\tilde{O}(d_{\lambda'}(\mathbf{A})\log(n/\delta))$ landmarks. Then for any $\lambda > 0$ satisfying $\lambda \leq \lambda'$ and for any $0 < \theta < 1$, $q > \frac{200\sqrt{\lambda'/\lambda}\log(2n/\delta)}{\theta^2}$, with probability $1-\delta$ we get*

$$(1-\theta/2)(\mathbb{E}\hat{\mathbf{A}}+\lambda\mathbf{I}) \preceq \frac{1}{q}\sum_{i=1}^{q}\hat{\mathbf{A}}_i + \lambda\mathbf{I} \preceq (1+\theta/2)(\mathbb{E}\hat{\mathbf{A}}+\lambda\mathbf{I}).$$

Proof First note that due to Lemma 17, $\big(\mathbb{E}\hat{\mathbf{A}}_i+\lambda\mathbf{I}\big)^{-1} \preceq 2\big(\mathbf{A}\mathbf{A}_{\lambda'}^{-1}\mathbf{A}+\lambda\mathbf{I}\big)^{-1}$. Define $\mathbf{Z}_i = (\mathbb{E}\hat{\mathbf{A}}+\lambda\mathbf{I})^{-1/2}(\hat{\mathbf{A}}_i+\lambda\mathbf{I})(\mathbb{E}\hat{\mathbf{A}}+\lambda\mathbf{I})^{-1/2}$. We have $\mathbb{E}[\mathbf{Z}_i] = \mathbf{I}$ and therefore $\mu_{\max}(\mathbb{E}\mathbf{Z}_i) = \mu_{\min}(\mathbb{E}\mathbf{Z}_i) = 1$, where $\mu_{\max}$ and $\mu_{\min}$ denotes the maximum and minimum eigenvalues of $\mathbb{E}\mathbf{Z}_i$. Now we upper bound $\|\mathbf{Z}_i\|$. Let $\mathbf{A}_\lambda = \mathbf{A}+\lambda\mathbf{I}$.

$$\begin{aligned}
\|\mathbf{Z}_i\| &\leq \|(\mathbb{E}\hat{\mathbf{A}}+\lambda\mathbf{I})^{-1/2}(\mathbf{A}+\lambda\mathbf{I})(\mathbb{E}\hat{\mathbf{A}}+\lambda\mathbf{I})^{-1/2}\| \\
&\leq 2\lambda_{\max}\Big(\mathbf{A}_\lambda^{1/2}(\mathbb{E}\hat{\mathbf{A}}+\lambda\mathbf{I})^{-1}\mathbf{A}_\lambda^{1/2}\Big) \\
&\leq 2\max_i\Big(\frac{\lambda_i+\lambda}{\frac{\lambda_i^2}{\lambda_i+\lambda'}+\lambda}\Big) \\
&= 2\max_i\Big(\frac{1+\frac{\lambda}{\lambda_i}}{\frac{\lambda_i}{\lambda_i+\lambda'}+\frac{\lambda}{\lambda_i}}\Big) := \xi_i.
\end{aligned}$$

Note that if $\lambda_i \geq \lambda'$, then $\xi_i < 3$. In case $\lambda_i < \lambda'$, we write $\lambda_i = c_i\lambda_n$ for some $c_i \geq 1$. We have

$$\xi_i < \frac{1+\frac{\lambda}{\lambda_i}}{\frac{\lambda_i}{2\lambda'}+\frac{\lambda}{\lambda_i}} = \frac{1+\frac{\lambda}{c_i\lambda_n}}{\frac{c_i\lambda_n}{2\lambda'}+\frac{\lambda}{c_i\lambda_n}}$$

and consider the following two cases:

1. *Case 1:* $\frac{c_i\lambda_n}{2\lambda'} \geq \frac{\lambda}{c_i\lambda_n}$ or $c_i \geq \frac{\sqrt{2\lambda\lambda'}}{\lambda_n}$. Then
$$\xi_i < \frac{1+\frac{\lambda}{c_i\lambda_n}}{\frac{c_i\lambda_n}{2\lambda'}} < 1 + \frac{2\lambda'}{c_i\lambda_n} \leq 1 + \sqrt{\frac{2\lambda'}{\lambda}}.$$

2. *Case 2:* $\frac{c_i\lambda_n}{2\lambda'} < \frac{\lambda}{c_i\lambda_n}$ or $c_i < \frac{\sqrt{2\lambda\lambda'}}{\lambda_n}$. Then
$$\xi_i < \frac{1+\frac{\lambda}{c_i\lambda_n}}{\frac{\lambda}{c_i\lambda_n}} < 1 + \sqrt{\frac{2\lambda'}{\lambda}}.$$

Therefore $\|\mathbf{Z}_i\| < 2\max\left\{3, 1+\sqrt{\frac{2\lambda'}{\lambda}}\right\} < 20\sqrt{\frac{\lambda'}{\lambda}} := R$. Using matrix Chernoff inequality, for any $\theta > 0$
$$\Pr\left(\lambda_{\min}\left(\frac{1}{q}\sum_{i=1}^{q}\mathbf{Z}_i\right) \leq 1-\theta/2\right) \leq n.\exp\left(-\frac{q\theta^2}{8R}\right),$$
and,
$$\Pr\left(\lambda_{\max}\left(\frac{1}{q}\sum_{i=1}^{q}\mathbf{Z}_i\right) \geq 1+\theta/2\right) \leq n.\exp\left(-\frac{q\theta^2}{(8+2\theta)R}\right).$$
For any $\theta < 1$ and $q > 200\sqrt{\lambda'/\lambda}\log(2n/\delta)/\theta^2$, with probability $1-\delta$
$$-\theta/2\mathbf{I} \preceq \frac{1}{q}\sum_{i=1}^{q}\mathbf{Z}_i - \mathbf{I} \preceq \theta/2\mathbf{I}.$$
and finally with probability $1-\delta$
$$(1-\theta/2)(\mathbb{E}\hat{\mathbf{A}}+\lambda\mathbf{I}) \preceq \frac{1}{q}\sum_{i=1}^{q}\hat{\mathbf{A}}_i + \lambda\mathbf{I} \preceq (1+\theta/2)(\mathbb{E}\hat{\mathbf{A}}+\lambda\mathbf{I}).$$

■

We use the following result from Dereziński et al. (2025) as a black box in our recursive preconditioning scheme for solving the Block-Nyström linear system approximately. The following lemma provides an algorithm to solve the resulting linear system at the final depth of our recursive scheme.

Lemma 25 (Based on Lemma 4.3 from Dereziński et al. (2025)) *Given matrix $\mathbf{A} \succ 0$, let $\hat{\mathbf{A}} = \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{\top}$ be its Nyström approximation where $\mathbf{C} = \mathbf{A}\mathbf{S}^{\top} \in \mathbb{R}^{n\times m}$ and $\mathbf{W} = \mathbf{S}\mathbf{A}\mathbf{S}^{\top}$. For $\psi > 0$, denote $\mathbf{M} = \hat{\mathbf{A}} + \psi\mathbf{I}$ as the Nyström preconditioner and assume that $\mathbf{M} \approx_{O(\kappa)} \mathbf{A} + \psi\mathbf{I}$. Given vector $\mathbf{v} \in \mathbb{R}^n$ and $\epsilon > 0$, there exists an algorithm that provides $\hat{\mathbf{w}}$: an approximate solution to the linear system $\mathbf{M}\mathbf{w} = \mathbf{v}$ satisfying*
$$\|\hat{\mathbf{w}} - \mathbf{M}^{-1}\mathbf{v}\|_{\mathbf{M}} \leq \epsilon\|\mathbf{M}^{-1}\mathbf{v}\|_{\mathbf{M}}.$$
in time $O((nm+m^3)\log(m/\delta) + (nm+m^2)\log(\kappa/\epsilon))$, where κ is the condition number of $\mathbf{A}$.

The following lemma is a computational result we use in our proof of Theorem 27.

Lemma 26 *For any $x, \phi, \theta > 0$, $c > 1$ and $p = c(1+\theta)^2$*

$$(4(1+\theta)^2)^{\frac{\log(x)}{\log(c)}} \leq x^{\phi/2}$$
$$\log(20p^2)^{\frac{\log(x)}{\log(c)}} \leq (x)^{\phi/2}$$

if and only if $c > \max\left((4(1+\theta)^2)^{2/\phi}, \log^{2/\phi}(20p^2), 1\right)$.
Proof

$$(4(1+\theta)^2)^{\frac{\log(x)}{\log(c)}} = x^{\frac{\log(4(1+\theta)^2)}{\log(c)}} \leq x^{\phi/2}$$

if and only if $c \geq (4(1+\theta)^2)^{2/\phi}$. Furthermore for $p = c(1+\theta)^2$,

$$\log(20p^2)^{\frac{\log(x)}{\log(c)}} \leq (x)^{\phi/2}$$

if and only if $c \geq \log^{2/\phi}(20p^2)$. ■

The following theorem is an extended version of Theorem 11. The statement of Theorem 11 can be recovered by setting $\tilde{\lambda} = \lambda$. We provide a recursive preconditioning scheme for solving the Block-Nyström linear system having near-linear dependence on the number of blocks q.

Theorem 27 (Recursive solver for Block-Nyström) *Given $\lambda, \tilde{\lambda}$ and λ' satisfying $\lambda \leq \tilde{\lambda} \leq \lambda'$, for $1 \leq i \leq q$ let $\hat{\mathbf{A}}_i$ are Nyström approximations of $\mathbf{A}$ sampled from $O(1)$-approximate λ'-ridge leverage scores of $\mathbf{A}$, each with $m = O(d_{\lambda'}(\mathbf{A})\log(n/\delta))$ landmarks and let $q = O\big(\sqrt{\frac{\lambda'}{\tilde{\lambda}}}\log(n/\delta)/\theta^2\big)$. Denote $\hat{\mathbf{A}}_{[q]} = \frac{1}{q}\sum_{i=1}^{q}\hat{\mathbf{A}}_i$. There exists an event with probability $1-\delta$ for any $\delta > 0$ such that for any $\mathbf{v} \in \mathbb{R}^n$, $\epsilon > 0$ and $0 < \phi < 1$ we can find $\mathbf{u} \in \mathbb{R}^n$ such that*

$$\|\mathbf{u} - \big(\hat{\mathbf{A}}_{[q]} + \lambda\mathbf{I}\big)^{-1}\mathbf{v}\|^2_{\hat{\mathbf{A}}_{[q]}+\lambda\mathbf{I}} \leq \epsilon \cdot \|\big(\hat{\mathbf{A}}_{[q]} + \lambda\mathbf{I}\big)^{-1}\mathbf{v}\|^2_{\hat{\mathbf{A}}_{[q]}+\lambda\mathbf{I}}$$

in time $O\Big(\sqrt{\tilde{\lambda}/\lambda}\cdot nmq^{1+\phi}\cdot cs_1s_2k\log(1/\epsilon)\Big)$, where $k = \max\Big\{1, \Big\lceil \log\big(\sqrt{\lambda'/\tilde{\lambda}}\big)/\log(c)\Big\rceil\Big\}$, $s_1 = \max\big(1, \log\big(\tilde{\lambda}/\lambda\big)\big)$, $s_2 = \log(m\kappa/\delta)\log^{3/2}(n/\delta)\log(20c^2(1+\theta)^4)$, and c is chosen so that $c \geq \lceil\max\Big((4(1+\theta)^2)^{2/\phi}, \log^{2/\phi}(20c^2(1+\theta)^4\Big)\Big\rceil$.

Proof Let $\epsilon > 0$, $\theta > 0$ be fixed and some $c > 1$ (to be determined later). Let $r_k = \lambda'/(c^{2k}\tilde{\lambda})$ where $k \geq 0$ is an integer. Furthermore let $q_k = O(\sqrt{r_k}\log(n/\delta)/\theta^2)$. We have $\hat{\mathbf{A}}_{[q]} + \lambda\mathbf{I} = \frac{1}{q_0}\sum_{i=1}^{q_0}\hat{\mathbf{A}}_i + \lambda\mathbf{I}$. Let $\mathbf{M}_k = \frac{1}{q_k}\sum_{i=1}^{q_k}\hat{\mathbf{A}}_i + \frac{\lambda'}{r_k}\mathbf{I}$. Note that

$$(\lambda/\tilde{\lambda})\mathbf{M}_0 \preceq \hat{\mathbf{A}}_{[q]} + \lambda\mathbf{I} \preceq \mathbf{M}_0.$$

Using Lemma 16, for any given $\mathbf{v}$ we can find $\mathbf{u} \in \mathbb{R}^n$ such that $\|\mathbf{u} - \big(\hat{\mathbf{A}}_{[q]} + \lambda\mathbf{I}\big)^{-1}\mathbf{v}\|^2_{\hat{\mathbf{A}}_{[q]}+\lambda\mathbf{I}} \leq \epsilon\|\big(\hat{\mathbf{A}}_{[q]} + \lambda\mathbf{I}\big)^{-1}\mathbf{v}\|^2_{\hat{\mathbf{A}}_{[q]}+\lambda\mathbf{I}}$ in time $4\sqrt{\tilde{\lambda}/\lambda}\big(T(\mathbf{M}_0) + q_0nm\big)\log(2/\epsilon) + n\sqrt{\tilde{\lambda}/\lambda}\log(1/\epsilon)$, where

$T(\mathbf{M}_0)$ is time complexity to find a $\mathbf{u}$ given $\mathbf{v}$ such that $\|\mathbf{u} - \mathbf{M}_0^{-1}\mathbf{v}\|_{\mathbf{M}_0}^2 \leq \frac{\lambda}{10\tilde{\lambda}}\|\mathbf{M}_0^{-1}\mathbf{v}\|_{\mathbf{M}_0}^2$. The total time complexity is given as

$$O\Big(\sqrt{\tilde{\lambda}/\lambda}\big(T(\mathbf{M}_0) + q_0 nm\big)\log(1/\epsilon)\Big) \tag{8}$$

Now note that according to Theorem 24 and our choices of q_0 and q_1, with probability $1 - \delta/n$

$$\mathbf{M}_0 \preceq (1+\theta)\Big(\mathbb{E}\hat{\mathbf{A}} + \frac{\lambda'}{r_0}\mathbf{I}\Big) \preceq (1+\theta)\Big(\mathbb{E}\hat{\mathbf{A}} + \frac{\lambda'}{r_1}\mathbf{I}\Big) \preceq (1+\theta)^2\mathbf{M}_1.$$

Furthermore,

$$\mathbf{M}_0 \succeq \frac{1}{1+\theta}\Big(\mathbb{E}\hat{\mathbf{A}} + \frac{\lambda'}{r_0}\mathbf{I}\Big) \succeq \frac{1}{c^2(1+\theta)}\Big(\mathbb{E}\hat{\mathbf{A}} + \frac{\lambda'}{r_1}\mathbf{I}\Big) \succeq \frac{1}{c^2(1+\theta)^2}\mathbf{M}_1.$$

Again using Lemma 16 we get

$$\begin{aligned} T(\mathbf{M}_0) &\leq 4c(1+\theta)^2\big(T(\mathbf{M}_1) + q_0 nm\big)\log(20\tilde{\lambda}/\lambda) + nc(1+\theta)^2\log(10\tilde{\lambda}/\lambda) \\ &= O\Big((T(\mathbf{M}_1) + q_0 nm)cs_1\Big). \end{aligned} \tag{9}$$

where $s_1 = \log(20\tilde{\lambda}/\lambda)$. Here $T(\mathbf{M}_1)$ is time complexity to find a $\mathbf{u}$ given $\mathbf{v}$ such that $\|\mathbf{u} - \mathbf{M}_1^{-1}\mathbf{v}\|_{\mathbf{M}_1}^2 \leq \frac{1}{10c^2(1+\theta)^4}\|\mathbf{M}_1^{-1}\mathbf{v}\|_{\mathbf{M}_1}^2$. Similar as above we can show that

$$\frac{1}{c^2(1+\theta)^2}\mathbf{M}_2 \preceq \mathbf{M}_1 \preceq (1+\theta)^2\mathbf{M}_2$$

upper bounding $T(\mathbf{M}_1)$ as

$$T(\mathbf{M}_1) \leq 4c(1+\theta)^2\big(T(\mathbf{M}_2) + q_1 nm\big)\log(20c^2(1+\theta)^4) + nc(1+\theta)^2\log(10c^2(1+\theta)^4).$$

Now recursively continuing in this manner by conditioning over high probability events $\frac{1}{1+\theta}\mathbb{E}[\mathbf{M}_k] \preceq \mathbf{M}_k \preceq (1+\theta)\mathbb{E}[\mathbf{M}_k]$ according to Theorem 24, we get

$$T(\mathbf{M}_k) \leq 4c(1+\theta)^2\big(T(\mathbf{M}_{k+1}) + q_k nm\big)\log(20c^2(1+\theta)^4) + nc(1+\theta)^2\log(10c^2(1+\theta)^4).$$

In fact, we can show that for $p = c(1+\theta)^2$

$$\begin{aligned} T(\mathbf{M}_1) &\leq 4^k p^k \log^k(20p^2)T(\mathbf{M}_{k+1}) + nm\sum_{i=1}^{k} 4^i q_i p^i \log^i(20p^2) + \frac{n}{4}\sum_{i=1}^{k} 4^i p^i \log^i(20p^2) \\ &\leq 4^k p^k \log^k(20p^2)T(\mathbf{M}_{k+1}) + 2nmq_0\sum_{i=1}^{k}\Big(\frac{4p\log(20p^2)}{c}\Big)^i \\ &\leq 4^k p^k \log^k(20p^2)T(\mathbf{M}_{k+1}) + 2knmq_0 \cdot \big(4(1+\theta)^2\big)^k \cdot \big(\log(20p^2)\big)^k. \end{aligned}$$

In second to last inequality we used $q_i \geq 1$ for all i and the formula $q_i = q_0/c^i$. Choose k such that $r_{k+1} < 2$ i.e., $k = \left\lceil \frac{\log\sqrt{\lambda'/\tilde{\lambda}}}{\log(c)} \right\rceil$ is sufficient. Now, let $c \geq \max\Big\{(4(1+\theta)^2)^{2/\phi}, \log^{2/\phi}(20p^2)\Big\}$.

Using Lemma 26 we get

$$\begin{aligned} T(\mathbf{M}_1) &\leq 4^k p^k \log^k(20p^2) T(\mathbf{M}_{k+1}) + 2knmq_0 \cdot \Big(\sqrt{\frac{\lambda'}{\tilde{\lambda}}}\Big)^{\phi} \\ &= \frac{q_0}{q_k} \cdot \big(4(1+\theta)^2\big)^k \cdot \big(\log(20p^2)\big)^k \cdot T(\mathbf{M}_{k+1}) + 2knmq_0 \cdot \Big(\sqrt{\frac{\lambda'}{\tilde{\lambda}}}\Big)^{\phi} \\ &\leq \Big(T(\mathbf{M}_{k+1}) + 2knm\Big) \cdot q_0^{1+\phi} \end{aligned} \tag{10}$$

where in the last inequality we used $\sqrt{\frac{\lambda'}{\tilde{\lambda}}} < q_0$. Now we precondition $\mathbf{M}_{k+1}$ with $\hat{\mathbf{A}}_1 + \frac{\lambda'}{r_{k+1}}\mathbf{I}$. As we have

$$\frac{1}{q_{k+1}}\Big(\hat{\mathbf{A}}_1 + \frac{\lambda'}{r_{k+1}}\mathbf{I}\Big) \preceq \mathbf{M}_{k+1} \preceq \mathbf{A} + \frac{\lambda'}{r_{k+1}}\mathbf{I} \preceq \mathbf{A} + \lambda'\mathbf{I} \preceq 2(\hat{\mathbf{A}}_1 + \lambda'\mathbf{I}) \preceq 4\Big(\hat{\mathbf{A}}_1 + \frac{\lambda'}{r_{k+1}}\mathbf{I}\Big).$$

where the second last inequality holds with probability $1 - \delta/n$, due to Lemma 9. Using Lemma 16 we get

$$T(\mathbf{M}_{k+1}) \leq 8\sqrt{q_{k+1}}\big(T(\hat{\mathbf{A}}_1) + q_{k+1}nm\big)\log(20p^2) + 2n\sqrt{q_{k+1}}\log(10p^2)$$

where $T(\hat{\mathbf{A}}_1)$ is the time complexity to find a $\mathbf{u}$ given $\mathbf{v}$ such that $\Big\|\mathbf{u} - \Big(\hat{\mathbf{A}}_1 + \frac{\lambda'}{r_{k+1}}\mathbf{I}\Big)^{-1}\mathbf{v}\Big\|^2_{\hat{\mathbf{A}}_1 + \lambda'/r_{k+1}\mathbf{I}} \leq \frac{1}{40q_{k+1}}\Big\|\Big(\hat{\mathbf{A}}_1 + \frac{\lambda'}{r_{k+1}}\mathbf{I}\Big)^{-1}\mathbf{v}\Big\|^2_{\hat{\mathbf{A}}_1 + \lambda'/r_{k+1}\mathbf{I}}$. It only remains to upper bound $T(\hat{\mathbf{A}}_1)$. We apply Lemma 25 to upper bound $T(\hat{\mathbf{A}}_1)$. Let $\psi = \lambda'/r_k$ and apply Lemma 25 with $\mathbf{M} = \hat{\mathbf{A}}_1 + \psi\mathbf{I}$, we get that $\mathbf{M}^{-1}\mathbf{x}$ can be solved approximately in time $O\Big(nm\log(m\kappa/\delta)\Big)$, thus upper bounding $T(\hat{\mathbf{A}}_1)$. Substituting this in expression for $T(\mathbf{M}_{k+1})$ we get

$$\begin{aligned} T(\mathbf{M}_{k+1}) &= O\Big((q_{k+1}nm)\log(m\kappa/\delta)\sqrt{q_{k+1}}\log(20p^2)\Big) \\ &= O\Big(nm\log(m\kappa/\delta)\log^{3/2}(n/\delta)\log(20p^2)\Big) \end{aligned}$$

Denoting $s_2 = \log(m\kappa/\delta)\log^{3/2}(n/\delta)\log(20p^2)$ and substituting the upper bound for $T(\mathbf{M}_{k+1})$ in (10) we get

$$T(\mathbf{M}_1) = O\Big(\big(nms_2 + 2knm\big) \cdot q_0^{1+\phi}\Big).$$

Back substituting the above in (9) we get

$$T(\mathbf{M}_0) = O\Big(\big(nms_2 + 2knm\big) \cdot q_0^{1+\phi} \cdot cs_1 + nmq_0 \cdot cs_1\Big)$$

Finally substituting the above in expression (8), the overall time complexity is given as

$$\begin{aligned} &O\Big(\sqrt{\tilde{\lambda}/\lambda}\big(\big(nms_2 + 2knm\big) \cdot q_0^{1+\phi} \cdot cs_1 + nmq_0 \cdot cs_1 + q_0 nm\big)\log(1/\epsilon)\Big) \\ =&O\Big(\sqrt{\tilde{\lambda}/\lambda} \cdot nmq_0^{1+\phi} \cdot cs_1 s_2 k \log(1/\epsilon)\Big). \end{aligned}$$

Taking the union bound over all the considered high probability events finishes the proof. ■

Appendix C. Statistical risk of KRR in random design setting

Consider the Hilbert space $L^2(\mathcal{H}, \rho_{\mathcal{X}})$ consisting of square integrable functions from $\mathcal{H}$ to $\mathbb{R}$, i.e., $L^2(\mathcal{H}, \rho_{\mathcal{X}}) = \{f : \mathcal{H} \to \mathbb{R} \mid \int_{\mathcal{H}} f(\mathbf{x})^2 d\rho_{\mathcal{X}}(\mathbf{x}) < \infty\}$. The norm on $L^2(\mathcal{H}, \rho_{\mathcal{X}})$ will be denoted by $\|\cdot\|_\rho$, and for any $f \in L^2(\mathcal{H}, \rho_{\mathcal{X}})$ we have $\|f\|_\rho = \Big(f(\mathbf{x})^2 d\rho_{\mathcal{X}}(\mathbf{x})\Big)^{1/2}$. Let $\mathcal{S}_\rho : \mathcal{H} \to L^2(\mathcal{H}, \rho_{\mathcal{X}})$ be the map $\mathcal{S}_\rho \omega = \langle \cdot, \omega \rangle_{\mathcal{H}}$ and $\mathcal{S}_\rho^* : L^2(\mathcal{H}, \rho_{\mathcal{X}}) \to \mathcal{H}$ be the adjoint map of $\mathcal{S}_\rho$. The covariance operator $\mathcal{S}_\rho^* \mathcal{S}_\rho : \mathcal{H} \to \mathcal{H}$ will be denoted as $\mathcal{C}$, whereas the integral operator $\mathcal{S}_\rho \mathcal{S}_\rho^* : L^2(\mathcal{H}, \rho_{\mathcal{X}}) \to L^2(\mathcal{H}, \rho_{\mathcal{X}})$ will be denoted as $\mathcal{L}$. Under assumption 1, $\mathcal{C}$ and $\mathcal{L}$ are positive, self-adjoint, and bounded linear operators. Let $\mathcal{Z}_n : \mathcal{H} \to \mathbb{R}^n$ as $\mathcal{Z}_n \omega = \big(\langle \mathbf{x}_i, \omega \rangle\big)_{i=1}^n$ and $\mathcal{S}_n = \mathcal{Z}_n / \sqrt{n}$. The operator $\mathcal{S}_n^* \mathcal{S}_n : \mathcal{H} \to \mathcal{H}$ is known as the empirical covariance operator and will be denoted by $\mathcal{C}_n$. Note that $\mathcal{Z}_n^* \mathcal{Z}_n = n\mathcal{C}_n$. An important observation here is that the operator $\mathcal{Z}_n \mathcal{Z}_n^* : \mathbb{R}^n \to \mathbb{R}^n$ is the linear map $\alpha \to \mathbf{K}\alpha$ where $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the kernel matrix over the n points $\mathbf{x}_i$ for $1 \le i \le n$. For $\lambda > 0$, the regularized operators $\mathcal{C} + \lambda \mathcal{I}$, $\mathcal{C}_n + \lambda \mathcal{I}$ will be denoted by $\mathcal{C}_\lambda$ and $\mathcal{C}_{n\lambda}$ respectively. The following provides further intuition on the various operators considered in the analysis of statistical risk of KRR:

- For any $g \in L^2(\mathcal{H}, \rho_{\mathcal{X}})$, $\mathcal{S}_\rho^* g = \int_{\mathcal{H}} \mathbf{x} g(\mathbf{x}) d\rho_{\mathcal{X}}(\mathbf{x})$, $\mathcal{L} g = \int_{\mathcal{H}} \langle \cdot, \mathbf{x} \rangle_{\mathcal{H}} g(\mathbf{x}) d\rho_{\mathcal{X}}(\mathbf{x})$.
- For any $\omega \in \mathcal{H}$, $\mathcal{S}_\rho^* \mathcal{S}_\rho \omega = \int_{\mathcal{H}} \langle \cdot, \omega \rangle_{\mathcal{H}} \mathbf{x} d\rho_{\mathcal{X}}(\mathbf{x})$, $\mathcal{S}_n^* \mathcal{S}_n \omega = \frac{1}{n} \sum_{i=1}^n \langle \mathbf{x}_i, \omega \rangle_{\mathcal{H}} \mathbf{x}_i$.
- For any $\alpha \in \mathbb{R}^n$, $\mathcal{Z}_n^* \alpha = \sum_{i=1}^n \alpha_i \mathbf{x}_i$ and $\mathcal{Z}_n \mathcal{Z}_n^* \alpha = \mathbf{K}\alpha$.

The following lemma provide two different ways to represent the hypothesis $\hat{\omega}$ learned by the classical Nyström KRR.

Lemma 28 (Representations of learned hypothesis) *Let $\hat{\mathbf{K}}$ be an m-rank Nyström approximation of $\mathbf{K}$. Let $\hat{\mathcal{P}}$ denote the projection operator onto the corresponding subspace $\mathcal{H}_m$. Then the hypothesis learned by classical Nyström KRR is given as*

$$\begin{aligned}
\hat{\omega} &= \hat{\mathcal{P}} \mathcal{Z}_n^* \big(\hat{\mathbf{K}} + n\lambda \mathbf{I}\big)^{-1} \mathbf{y} \\
&= \hat{\mathcal{P}} \mathcal{Z}_n^* (\mathcal{Z}_n \hat{\mathcal{P}} \mathcal{Z}_n^* + n\lambda \mathbf{I})^{-1} \mathbf{y} \\
&= \hat{\mathcal{P}} (\hat{\mathcal{P}} \mathcal{Z}_n^* \mathcal{Z}_n \hat{\mathcal{P}} + n\lambda \mathcal{I})^{-1} \hat{\mathcal{P}} \mathcal{Z}_n^* \mathbf{y}.
\end{aligned}$$

Proof Recall that

$$\tilde{\mathcal{E}}(\omega, \lambda) = \frac{1}{n} \sum_{i=1}^n (\langle \omega, \mathbf{x}_i \rangle_{\mathcal{H}} - y_i)^2 + \lambda \|\omega\|_{\mathcal{H}}^2.$$

Let $\hat{\omega}$ be the minimizer of $\tilde{\mathcal{E}}(\omega, \lambda)$. First note $\mathcal{Z}_n \hat{\omega} = \big(\langle \hat{\omega}, \mathbf{x}_i \rangle_{\mathcal{H}}\big)_{i=1}^n$. Substituting $\hat{\omega} = \mathcal{Z}_n^* \mathbf{S}^\top \hat{\mathbf{w}}$ where $\mathbf{S} \in \mathbb{R}^{m \times n}$ is a random sketching matrix, we get $\mathcal{Z}_n \mathcal{Z}_n^* \mathbf{S}^\top \hat{\mathbf{w}} = \big(\langle \hat{\omega}, \mathbf{x}_i \rangle_{\mathcal{H}}\big)_{i=1}^n$. Therefore $\sum_{i=1}^n (\langle \hat{\omega}, \mathbf{x}_i \rangle_{\mathcal{H}} - y_i)^2 = \|\mathbf{K}\mathbf{S}^\top \hat{\mathbf{w}} - \mathbf{y}\|^2$, by noting $\mathcal{Z}_n \mathcal{Z}_n^* = \mathbf{K}$. In particular $\hat{\mathbf{w}}$ is the solution of the system $(\mathbf{S}\mathbf{K}^2\mathbf{S}^\top + n\lambda \mathbf{S}\mathbf{K}\mathbf{S}^\top)\hat{\mathbf{w}} = \mathbf{S}\mathbf{K}\mathbf{y}$, implying,

$$\hat{\mathbf{w}} = (\mathbf{S}\mathbf{K}^2\mathbf{S}^\top + n\lambda \mathbf{S}\mathbf{K}\mathbf{S}^\top)^\dagger \mathbf{S}\mathbf{K}\mathbf{y}.$$

The learned $\hat{\omega}$ is given as

$$\begin{aligned}
\hat{\omega} &= \mathcal{Z}_n^* \mathbf{S}^\top (\mathbf{S}\mathbf{K}^2\mathbf{S}^\top + n\lambda \mathbf{S}\mathbf{K}\mathbf{S}^\top)^\dagger \mathbf{S}\mathbf{K}\mathbf{y} \\
&= \mathcal{Z}_n^* \mathbf{S}^\top (\mathbf{S}\mathcal{Z}_n(\mathcal{Z}_n^*\mathcal{Z}_n + n\lambda\mathcal{I})\mathcal{Z}_n^*\mathbf{S}^\top)^\dagger \mathbf{S}\mathcal{Z}_n\mathcal{Z}_n^*\mathbf{y} \\
&= \mathcal{Z}_n^* \mathbf{S}^\top (\mathbf{S}\mathcal{Z}_n(\mathbf{S}\mathcal{Z}_n^\dagger)\mathbf{S}\mathcal{Z}_n(\mathcal{Z}_n^*\mathcal{Z}_n + n\lambda\mathcal{I})\mathcal{Z}_n^*\mathbf{S}^\top(\mathcal{Z}_n^*\mathbf{S}^\top)^\dagger\mathcal{Z}_n^*\mathbf{S}^\top)^\dagger \mathbf{S}\mathcal{Z}_n\mathcal{Z}_n^*\mathbf{y} \\
&= \mathcal{Z}_n^* \mathbf{S}^\top (\mathbf{S}\mathcal{Z}_n(\hat{\mathcal{P}}\mathcal{Z}_n^*\mathcal{Z}_n\hat{\mathcal{P}} + n\lambda\mathcal{I})\mathcal{Z}_n^*\mathbf{S}^\top)^\dagger \mathbf{S}\mathcal{Z}_n\mathcal{Z}_n^*\mathbf{y} \\
&= \hat{\mathcal{P}}(\hat{\mathcal{P}}\mathcal{Z}_n^*\mathcal{Z}_n\hat{\mathcal{P}} + n\lambda\mathcal{I})^{-1}\hat{\mathcal{P}}\mathcal{Z}_n^*\mathbf{y}.
\end{aligned}$$

The second last equality holds because $\mathcal{Z}_n^*\mathbf{S}^\top(\mathcal{Z}_n^*\mathbf{S}^\top)^\dagger$ and $(\mathbf{S}\mathcal{Z}_n)^\dagger\mathbf{S}\mathcal{Z}_n$ are orthogonal projections on $\mathcal{H}_m$. Furthermore using the push-through identity we have $(\hat{\mathcal{P}}\mathcal{Z}_n^*\mathcal{Z}_n\hat{\mathcal{P}} + n\lambda\mathcal{I})^{-1}\hat{\mathcal{P}}\mathcal{Z}_n^* = \hat{\mathcal{P}}\mathcal{Z}_n^*(\mathcal{Z}_n^*\hat{\mathcal{P}}\mathcal{Z}_n + n\lambda\mathbf{I})^{-1}\mathbf{y}$ leading to

$$\begin{aligned}
\hat{\omega} &= \hat{\mathcal{P}}\mathcal{Z}_n^*(\mathcal{Z}_n\hat{\mathcal{P}}\mathcal{Z}_n^* + n\lambda\mathbf{I})^{-1}\mathbf{y} \\
&= \hat{\mathcal{P}}\mathcal{Z}_n^*\big(\hat{\mathbf{K}} + n\lambda\mathbf{I}\big)^{-1}\mathbf{y}.
\end{aligned}$$

The last equality holds because $\hat{\mathbf{K}} = \mathcal{Z}_n\hat{\mathcal{P}}\mathcal{Z}_n^*$. ■

A similar representation like Lemma 28 holds for Block-Nyström KRR as shown before the proof of Theorem 30. In the next result, we first show that for any given $\mathbf{x}$ in $\mathcal{H}$, the output y can be predicted in time linear in number of blocks, provided the Block Nyström hypothesis has been precomputed.

Lemma 29 (Prediction with Block-Nyström) *For some $q \geq 1$, consider $\mathcal{P}_{[q]} = \frac{1}{q}\sum_{i=1}^q \hat{\mathcal{P}}_i$ where $\hat{\mathcal{P}}_i$ denotes an orthogonal projection onto some m dimensional subspace of $\mathcal{H}_n$. In particular, $\hat{\mathcal{P}}_i = \mathcal{Z}_n^*\mathbf{S}_i^\top(\mathbf{S}_i\mathbf{K}\mathbf{S}_i^\top)^\dagger\mathbf{S}_i\mathcal{Z}_n$ where $\mathbf{S}_i \in \mathbb{R}^{m\times n}$ is a sub-sampling matrix. Let $\hat{\omega}_{[q]} = \mathcal{P}_{[q]}\mathcal{Z}_n^*(\hat{\mathbf{K}}_{[q]} + n\lambda\mathbf{I})^{-1}\mathbf{y}$ where $\hat{\mathbf{K}}_{[q]} = \frac{1}{q}\sum_{i=1}^q \hat{\mathbf{K}}_i + n\lambda\mathbf{I}$ and $\hat{\mathbf{K}}_i = \mathcal{Z}_n\hat{\mathcal{P}}_i\mathcal{Z}_n^*$ are iid Nyström approximations of $\mathbf{K}$. Then for any $\mathbf{x} \in \mathcal{H}$, $\langle\hat{\omega}_{[q]}, \mathbf{x}\rangle_{\mathcal{H}}$ can be computed in time $O(qm)$ after a preprocessing cost of $\tilde{O}\Big(qm^3 + nmq^{1+o(1)}\Big)$, assuming $(\hat{\mathbf{K}}_{[q]} + n\lambda\mathbf{I})^{-1}\mathbf{y}$ can be computed in time $\tilde{O}(nmq^{1+o(1)})$.*

Proof Given any $\mathbf{x} \in \mathcal{H}$, we can predict the response y for $\mathbf{x}$ by computing $\langle\hat{\omega}_{[q]}, \mathbf{x}\rangle_{\mathcal{H}}$ as follows.

$$\begin{aligned}
\langle\omega_{[q]}, \mathbf{x}\rangle_{\mathcal{H}} &= \frac{1}{q}\sum_{i=1}^q \langle\hat{\mathcal{P}}_i\mathcal{Z}_n^*(\hat{\mathbf{K}}_{[q]} + n\lambda\mathbf{I})^{-1}\mathbf{y}, \mathbf{x}\rangle_{\mathcal{H}} \\
&= \frac{1}{q}\sum_{i=1}^q \langle\mathcal{Z}_n^*\mathbf{S}_i^\top(\mathbf{S}_i\mathbf{K}\mathbf{S}_i^\top)^\dagger\mathbf{S}_i\mathcal{Z}_n\mathcal{Z}_n^*(\hat{\mathbf{K}}_{[q]} + n\lambda\mathbf{I})^{-1}\mathbf{y}, \mathbf{x}\rangle_{\mathcal{H}} \\
&= \frac{1}{q}\sum_{i=1}^q \langle(\mathbf{S}_i\mathbf{K}\mathbf{S}_i^\top)^\dagger\mathbf{S}_i\mathcal{Z}_n\mathcal{Z}_n^*(\hat{\mathbf{K}}_{[q]} + n\lambda\mathbf{I})^{-1}\mathbf{y}, \mathbf{S}_i\mathcal{Z}_n\mathbf{x}\rangle.
\end{aligned}$$

For any fixed i, $(\mathbf{S}_i\mathbf{K}\mathbf{S}_i^\top)^\dagger$ can be computed in time $O(m^3)$, $\mathbf{S}_i\mathcal{Z}_n\mathcal{Z}_n^*$ in time $O(nm)$[2], and $\mathbf{S}_i\mathcal{Z}_n\mathcal{Z}_n^*(\hat{\mathbf{K}}_{[q]} + n\lambda\mathbf{I})^{-1}\mathbf{y}$ in time $O(nm)$ given $\mathbf{S}_i\mathcal{Z}_n\mathcal{Z}_n^*$ and $(\hat{\mathbf{K}}_{[q]} + n\lambda\mathbf{I})^{-1}\mathbf{y}$ have been precomputed. For any $\mathbf{x} \in \mathcal{H}$, $\mathbf{S}_i\mathcal{Z}_n\mathbf{x}$ can be computed in time $O(m)$. The overall preprocessing cost to find $\langle\hat{\omega}_{[q]}, \mathbf{x}\rangle_{\mathcal{H}}$ is $\tilde{O}\Big(qm^3 + nmq^{1+o(1)}\Big)$. ■

2. We are assuming that $\langle\mathbf{x}_i, \mathbf{x}_j\rangle_{\mathcal{H}}$ takes times $O(1)$ for any $1 \leq i, j \leq n$.

In Theorem 27, we provide a recursive algorithm to approximate $(\hat{\mathbf{K}}_{[q]}+n\lambda\mathbf{I})^{-1}\mathbf{y}$ in time $\tilde{O}(nmq^{1+o(1)})$ with high precision, for an appropriately chosen q depending on λ and sampling distribution for Nyström landmarks. We now provide the proof of statistical risk of Block-Nyström KRR. For $q \geq 1$ and $1 \leq i \leq q$, consider $\hat{\mathcal{P}}_i$ to denote orthogonal projections on m dimensional subspaces of $\mathcal{H}_n$ sampled in an iid manner and $\hat{\mathbf{K}}_i = \mathcal{Z}_n\hat{\mathcal{P}}_i\mathcal{Z}_n^*$. Let $\hat{\mathcal{P}}_{[q]}$ denote the average of q projections i.e., $\hat{\mathcal{P}}_{[q]} = \frac{1}{q}\sum_{i=1}^q \hat{\mathcal{P}}_i$ and $\hat{\mathbf{K}}_{[q]} = \frac{1}{q}\sum_{i=1}^q \hat{\mathbf{K}}_i = \mathcal{Z}_n\hat{\mathcal{P}}_{[q]}\mathcal{Z}_n^*$. In Block-Nyström, we consider the following hypothesis in $\mathcal{H}$

$$\begin{aligned}\omega_{[q]} &= \hat{\mathcal{P}}_{[q]}\mathcal{Z}_n^*(\hat{\mathbf{K}}_{[q]} + n\lambda\mathbf{I})^{-1}\mathbf{y}\\ &= \hat{\mathcal{P}}_{[q]}\mathcal{Z}_n^*(\mathcal{Z}_n\hat{\mathcal{P}}_{[q]}\mathcal{Z}_n^* + n\lambda\mathbf{I})^{-1}\mathbf{y}\\ &= \hat{\mathcal{P}}_{[q]}^{1/2}\big(\hat{\mathcal{P}}_{[q]}^{1/2}\mathcal{Z}_n^*\mathcal{Z}_n\hat{\mathcal{P}}_{[q]}^{1/2} + n\lambda\mathcal{I}\big)^{-1}\hat{\mathcal{P}}_{[q]}^{1/2}\mathcal{Z}_n^*\mathbf{y}.\end{aligned}$$

Recall f_ρ denotes the regression function and $f_{\mathcal{H}}$ denotes the projection of f_ρ onto $\bar{\mathcal{H}}_\rho$. We define $\omega = \mathcal{C}_\lambda^{-1}\mathcal{S}_\rho^* f_{\mathcal{H}}$.

The following theorem provides a more general version of Theorem 14, bounding the generalization error in terms of a range of norms parameterized by a, following prior works (Lin and Cevher, 2020; Lin et al., 2020). The bounds in terms of the expected risk stated in Theorem 14 can be recovered by taking $a = 0$.

Theorem 30 (Expected risk of approximate KRR, extended Theorem 14) *Under assumptions 1, 2, 3, 4, 5,* $0 \leq a \leq \zeta$, $\zeta < \frac{1}{2}$, $n \geq O(G^2\log(G^2/\delta))$ *and* $\frac{19G^2\log(n/\delta)}{n} \leq \lambda \leq \|\mathcal{C}\|$.

$$\big\|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_{[q]} - f_{\mathcal{H}})\big\|_\rho \leq \lambda^{-a}\cdot\tilde{O}\Big(R\lambda^\zeta\big\|\mathbf{K}_{n\lambda}^{1/2}(\hat{\mathbf{K}}_{[q]} + n\lambda\mathbf{I})^{-1}\mathbf{K}_{n\lambda}^{1/2}\big\| + \frac{1}{n\lambda^{1-\zeta}} + \frac{1}{\sqrt{n\lambda^\gamma}}\Big)$$

where $\mathbf{K}_{n\lambda} = \mathbf{K} + n\lambda\mathbf{I}$.

Proof For simplicity of notation we will denote $(\hat{\mathbf{K}}_{[q]} + n\lambda\mathbf{I}) = \bar{\mathbf{K}}_{n\lambda}$, $\hat{\mathcal{P}}_{[q]} = \bar{\mathcal{P}}$. We start with the following decomposition:

$$\begin{aligned}\big\|\mathcal{L}^{-a}\big(\mathcal{S}_\rho\omega_{[q]} - f_{\mathcal{H}}\big)\big\|_\rho &\leq \big\|\mathcal{L}^{-a}\mathcal{S}_\rho\big(\omega_{[q]} - \omega\big)\big\|_\rho + \underbrace{\big\|\mathcal{L}^{-a}\big(\mathcal{S}_\rho\omega - f_{\mathcal{H}}\big)\big\|_\rho}_{\text{True Bias}}\\ &\leq \big\|\mathcal{L}^{-a}\mathcal{S}_\rho\big(\omega_{[q]} - \omega\big)\big\|_\rho + R\lambda^{\zeta-a} \qquad (11)\end{aligned}$$

where in last inequality we used Lemma 21 to upper bound the true bias term. Consider the first term now:

$$\begin{aligned}\big\|\mathcal{L}^{-a}\mathcal{S}_\rho\big(\omega_{[q]} - \omega\big)\big\|_\rho &= \big\|\mathcal{L}^{-a}\mathcal{S}_\rho\mathcal{C}^{a-1/2}\mathcal{C}^{1/2-a}\big(\omega_{[q]} - \omega\big)\big\|_\rho\\ &\leq \big\|\mathcal{L}^{-a}\mathcal{S}_\rho\mathcal{C}^{a-1/2}\big\|\cdot\big\|\mathcal{C}^{1/2-a}\big(\omega_{[q]} - \omega\big)\big\|_{\mathcal{H}}\\ &\leq \big\|\mathcal{C}^{1/2-a}\big(\omega_{[q]} - \omega\big)\big\|_{\mathcal{H}}. \qquad (12)\end{aligned}$$

The last inequality is obtained after upperbounding $\left\|\mathcal{L}^{-a}\mathcal{S}_\rho\mathcal{C}^{a-1/2}\right\|$ by 1. Now we substitute for $\omega_{[q]}$. Continuing we have,

$$\begin{aligned}\left\|\mathcal{C}^{1/2-a}(\omega_{[q]}-\omega)\right\|_{\mathcal{H}} &= \left\|\mathcal{C}^{1/2-a}\left[\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathbf{y}-\omega\right]\right\|_{\mathcal{H}}\\ &= \left\|\mathcal{C}^{1/2-a}\left[\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}(\mathbf{y}-\mathcal{Z}_n\omega+\mathcal{Z}_n\omega)-\omega\right]\right\|_{\mathcal{H}}\\ &\le \underbrace{\left\|\mathcal{C}^{1/2-a}\left[\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}(\mathbf{y}-\mathcal{Z}_n\omega)\right]\right\|_{\mathcal{H}}}_{\mathcal{T}_v\text{: Variance}} + \underbrace{\left\|\mathcal{C}^{1/2-a}\left[\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n\omega-\omega\right]\right\|_{\mathcal{H}}}_{\mathcal{T}_b\text{: Bias}}.\end{aligned} \tag{13}$$

We upper bound the bias and variance terms separately, considering the variance term first:

$$\begin{aligned}\mathcal{T}_v &= \left\|\mathcal{C}^{1/2-a}\left[\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}(\mathbf{y}-\mathcal{Z}_n\omega)\right]\right\|_{\mathcal{H}}\\ &= \left\|\mathcal{C}^{1/2-a}\mathcal{C}_\lambda^{a-1/2}\mathcal{C}_\lambda^{1/2-a}\mathcal{C}_{n\lambda}^{a-1/2}\mathcal{C}_{n\lambda}^{1/2-a}\left[\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}(\mathbf{y}-\mathcal{Z}_n\omega)\right]\right\|_{\mathcal{H}}\\ &\le \left\|\mathcal{C}^{1/2-a}\mathcal{C}_\lambda^{a-1/2}\right\|\cdot\left\|\mathcal{C}_\lambda^{1/2-a}\mathcal{C}_{n\lambda}^{a-1/2}\right\|\cdot\left\|\mathcal{C}_{n\lambda}^{-a}\right\|\cdot\left\|\mathcal{C}_{n\lambda}^{1/2}\left[\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}(\mathbf{y}-\mathcal{Z}_n\omega)\right]\right\|_{\mathcal{H}}.\end{aligned}$$

We have $\left\|\mathcal{C}^{1/2-a}\mathcal{C}_\lambda^{a-1/2}\right\|\le 1$, and $\left\|\mathcal{C}_\lambda^{1/2-a}\mathcal{C}_{n\lambda}^{a-1/2}\right\|\le\left\|\mathcal{C}_\lambda^{1/2}\mathcal{C}_{n\lambda}^{-1/2}\right\|^{1-2a}\le 2$ by using Lemma 18 and Lemma 19. Therefore we get,

$$\begin{aligned}\mathcal{T}_v &\le 2\lambda^{-a}\left\|\mathcal{C}_{n\lambda}^{1/2}\left[\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}(\mathbf{y}-\mathcal{Z}_n\omega)\right]\right\|_{\mathcal{H}}\\ &= 2\lambda^{-a}\left\|\mathcal{C}_{n\lambda}^{1/2}\left[\bar{\mathcal{P}}^{1/2}\bar{\mathcal{P}}^{1/2}\mathcal{Z}_n^*(\mathcal{Z}_n\bar{\mathcal{P}}^{1/2}\bar{\mathcal{P}}^{1/2}\mathcal{Z}_n^*+n\lambda\mathbf{I})^{-1}(\mathbf{y}-\mathcal{Z}_n\omega)\right]\right\|_{\mathcal{H}}.\end{aligned}$$

In the last equality we substituted $\bar{\mathbf{K}}_{n\lambda}^{-1}=(\mathcal{Z}_n\bar{\mathcal{P}}\mathcal{Z}_n^*+n\lambda\mathbf{I})^{-1}$. Using push-through identity we have $\bar{\mathcal{P}}^{1/2}\mathcal{Z}_n^*(\mathcal{Z}_n\bar{\mathcal{P}}^{1/2}\bar{\mathcal{P}}^{1/2}\mathcal{Z}_n^*+n\lambda\mathbf{I})^{-1}=(\bar{\mathcal{P}}^{1/2}\mathcal{Z}_n^*\mathcal{Z}_n\bar{\mathcal{P}}^{1/2}+n\lambda\mathcal{I})^{-1}\bar{\mathcal{P}}^{1/2}\mathcal{Z}_n^*$. Substituting this in the last equality we have,

$$\begin{aligned}\mathcal{T}_v &\le 2\lambda^{-a}\left\|\mathcal{C}_{n\lambda}^{1/2}\bar{\mathcal{P}}^{1/2}(\bar{\mathcal{P}}^{1/2}\mathcal{Z}_n^*\mathcal{Z}_n\bar{\mathcal{P}}^{1/2}+n\lambda\mathcal{I})^{-1}\mathcal{P}^{1/2}\left[\mathcal{Z}_n^*\mathbf{y}-\mathcal{Z}_n^*\mathcal{Z}_n\omega\right]\right\|_{\mathcal{H}}\\ &= 2\lambda^{-a}\left\|\mathcal{C}_{n\lambda}^{1/2}\bar{\mathcal{P}}^{1/2}(\bar{\mathcal{P}}^{1/2}\mathcal{C}_n\bar{\mathcal{P}}^{1/2}+\lambda\mathcal{I})^{-1}\mathcal{P}^{1/2}\left[\mathcal{S}_n^*\hat{\mathbf{y}}-\mathcal{S}_n\omega\right]\right\|_{\mathcal{H}}\\ &\le 2\lambda^{-a}\left\|\mathcal{C}_{n\lambda}^{1/2}\bar{\mathcal{P}}^{1/2}(\bar{\mathcal{P}}^{1/2}\mathcal{C}_n\bar{\mathcal{P}}^{1/2}+\lambda\mathcal{I})^{-1}\mathcal{P}^{1/2}\mathcal{C}_{n\lambda}^{1/2}\right\|\cdot\underbrace{\left\|\mathcal{C}_{n\lambda}^{-1/2}\left[\mathcal{S}_n^*\hat{\mathbf{y}}-\mathcal{C}_n\omega\right]\right\|_{\mathcal{H}}}_{\text{sample variance}}.\end{aligned}$$

The second equality holds because $\mathcal{Z}_n=\sqrt{n}\mathcal{S}_n$ and $\mathcal{S}_n^*\mathcal{S}_n=\mathcal{C}_n$ and $\hat{\mathbf{y}}=\mathbf{y}/\sqrt{n}$. We use Lemma 22 to upper bound the term $\left\|\mathcal{C}_{n\lambda}^{-1/2}\left[\mathcal{S}_n^*\hat{\mathbf{y}}-\mathcal{C}_n\omega^\lambda\right]\right\|_{\mathcal{H}}$. The remaining multiplicative term can be upper bounded by 1 as

$$\begin{aligned}&\bar{\mathcal{P}}^{1/2}\mathcal{C}_n\bar{\mathcal{P}}^{1/2}+\lambda\mathcal{I}\succeq\bar{\mathcal{P}}^{1/2}\mathcal{C}_{n\lambda}\bar{\mathcal{P}}^{1/2}=\bar{\mathcal{P}}^{1/2}\mathcal{C}_{n\lambda}^{1/2}\mathcal{C}_{n\lambda}^{1/2}\bar{\mathcal{P}}^{1/2}\\ \Rightarrow\ &(\bar{\mathcal{P}}^{1/2}\mathcal{C}_n\bar{\mathcal{P}}^{1/2}+\lambda\mathcal{I})^{-1}\preceq(\bar{\mathcal{P}}^{1/2}\mathcal{C}_{n\lambda}^{1/2}\mathcal{C}_{n\lambda}^{1/2}\bar{\mathcal{P}}^{1/2})^\dagger=(\mathcal{C}_{n\lambda}^{1/2}\bar{\mathcal{P}}^{1/2})^\dagger(\bar{\mathcal{P}}^{1/2}\mathcal{C}_{n\lambda}^{1/2})^\dagger.\end{aligned}$$

Therefore,

$$\left\|\mathcal{C}_{n\lambda}^{1/2}\bar{\mathcal{P}}^{1/2}(\bar{\mathcal{P}}^{1/2}\mathcal{C}_n\bar{\mathcal{P}}^{1/2}+\lambda\mathcal{I})^{-1}\bar{\mathcal{P}}^{1/2}\mathcal{C}_{n\lambda}^{1/2}\right\|\le\left\|\mathcal{C}_{n\lambda}^{1/2}\bar{\mathcal{P}}^{1/2}(\mathcal{C}_{n\lambda}^{1/2}\bar{\mathcal{P}}^{1/2})^\dagger(\bar{\mathcal{P}}^{1/2}\mathcal{C}_{n\lambda}^{1/2})^\dagger\bar{\mathcal{P}}^{1/2}\mathcal{C}_{n\lambda}^{1/2}\right\|\le 1.$$

The upperbound on the variance term becomes

$$\mathcal{T}_v\le\lambda^{-a}\cdot O\left(\lambda^\zeta+\frac{1}{n\lambda^{1-\zeta}}+\frac{1}{\sqrt{n\lambda^\gamma}}\right)\log(1/\delta). \tag{14}$$

We now upper bound the bias term:

$$\begin{aligned}\mathcal{T}_b &= \big\|\mathcal{C}^{1/2-a}\big(\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n - \mathcal{I}\big)\omega\big\|_{\mathcal{H}} \\ &= \big\|\mathcal{C}^{1/2-a}\big(\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n - \mathcal{I}\big)\mathcal{C}_\lambda^{-1}\mathcal{S}_\rho^* f_{\mathcal{H}}\big\|_{\mathcal{H}} \\ &= \big\|\mathcal{C}^{1/2-a}\big(\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n - \mathcal{I}\big)\mathcal{C}_\lambda^{-1}\mathcal{S}_\rho^*\mathcal{L}^\zeta g\big\|_{\mathcal{H}}.\end{aligned}$$

The last equality holds due to the source assumption 4. Using $\mathcal{S}_\rho^*\mathcal{L}^\zeta = \mathcal{S}_\rho^*(\mathcal{S}_\rho\mathcal{S}_\rho^*)^\zeta = (\mathcal{S}_\rho^*\mathcal{S}_\rho)^\zeta\mathcal{S}_\rho^* = \mathcal{C}^\zeta\mathcal{S}_\rho^* g$, we get

$$\begin{aligned}\mathcal{T}_b &= \big\|\mathcal{C}^{1/2-a}\big(\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n - \mathcal{I}\big)\mathcal{C}_\lambda^{-1}\mathcal{C}^\zeta\mathcal{S}_\rho^* g\big\|_{\mathcal{H}} \\ &\le \big\|\mathcal{C}^{1/2-a}\big(\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n - \mathcal{I}\big)\mathcal{C}_\lambda^{-1}\mathcal{C}^\zeta\mathcal{S}_\rho^*\big\| \cdot \big\|g\big\|_\rho \\ &\le R\big\|\mathcal{C}^{1/2-a}\big(\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n - \mathcal{I}\big)\mathcal{C}_\lambda^{-1}\mathcal{C}^\zeta\mathcal{C}^{1/2}\big\| \\ &\le R\big\|\mathcal{C}_\lambda^{1/2-a}\big(\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n - \mathcal{I}\big)\mathcal{C}_\lambda^{\zeta-1/2}\big\|.\end{aligned}$$

In the second to last inequality we used $\big\|g\big\|_\rho \le R$ and in the last inequality we used $\mathcal{C}^{\zeta+1/2} \preceq \mathcal{C}_\lambda^{\zeta+1/2}$, and $\mathcal{C}^{1/2-a} \preceq \mathcal{C}_\lambda^{1/2-a}$ as $a \le \zeta < \frac{1}{2}$. Continuing we get,

$$\begin{aligned}\mathcal{T}_b &\le R\big\|\mathcal{C}_\lambda^{-a}\big\| \cdot \big\|\mathcal{C}_\lambda^{1/2}\mathcal{C}_{n\lambda}^{-1/2}\big\| \cdot \big\|\mathcal{C}_{n\lambda}^{-1/2}\big\| \cdot \big\|\mathcal{C}_{n\lambda}(\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n - \mathcal{I})\big\| \cdot \|\mathcal{C}_\lambda^{\zeta-1/2}\| \\ &\le \frac{2R\lambda^{-a}\lambda^\zeta}{\lambda}\big\|\mathcal{C}_{n\lambda}(\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n - \mathcal{I})\big\| \\ &= \frac{2R\lambda^{-a}\lambda^\zeta}{n\lambda}\big\|(\mathcal{Z}_n^*\mathcal{Z}_n + n\lambda\mathbf{I})(\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n - \mathcal{I})\big\| \\ &= \frac{2R\lambda^{-a}\lambda^\zeta}{n\lambda}\big\|(\mathcal{Z}_n^*\mathcal{Z}_n + n\lambda\mathbf{I})(\mathcal{Z}_n^\dagger\mathcal{Z}_n\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n - \mathcal{I})\big\| \\ &= \frac{2R\lambda^{-a}\lambda^\zeta}{n\lambda}\big\|(\mathcal{Z}_n^*\mathcal{Z}_n + n\lambda\mathcal{I})(\mathcal{Z}_n^\dagger\bar{\mathbf{K}}\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n - \mathcal{I})\big\|,\end{aligned}$$

where we used the observation that $\mathcal{Z}_n^\dagger\mathcal{Z}_n$ is a projection on $\mathcal{H}_n$, implying $\mathcal{Z}_n^\dagger\mathcal{Z}_n\bar{\mathcal{P}} = \bar{\mathcal{P}}$. Also we substituted $\mathcal{Z}_n\bar{\mathcal{P}}\mathcal{Z}_n^* = \bar{\mathbf{K}}$. Multiplying the terms we get,

$$\begin{aligned}\mathcal{T}_b &\le \frac{2R\lambda^{-a}\lambda^\zeta}{n\lambda}\big\|\mathcal{Z}_n^*\bar{\mathbf{K}}\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n + n\lambda\mathcal{Z}_n^\dagger\bar{\mathbf{K}}\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n - \mathcal{Z}_n^*\mathcal{Z}_n - n\lambda\mathcal{I}\big\| \\ &\le \frac{2R\lambda^{-a}\lambda^\zeta}{n\lambda}\Big(\|\mathcal{Z}_n^*(\bar{\mathbf{K}}\bar{\mathbf{K}}_{n\lambda}^{-1} - \mathbf{I})\mathcal{Z}_n + n\lambda\mathcal{Z}_n^\dagger\bar{\mathbf{K}}\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n - n\lambda\mathcal{I}\|\Big).\end{aligned}$$

The terms $\big\|\mathcal{Z}_n^*(\bar{\mathbf{K}}\bar{\mathbf{K}}_{n\lambda}^{-1} - \mathbf{I})\mathcal{Z}_n\big\|$ and $\big\|\mathcal{Z}_n^\dagger\bar{\mathbf{K}}\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n\big\|$ can be upper bounded separately as

$$\big\|\mathcal{Z}_n^*(\bar{\mathbf{K}}\bar{\mathbf{K}}_{n\lambda}^{-1} - \mathbf{I})\mathcal{Z}_n\big\| = n\lambda\|\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n\| = n\lambda\|\mathbf{K}^{1/2}\bar{\mathbf{K}}_{n\lambda}^{-1}\mathbf{K}^{1/2}\| \le n\lambda\|\mathbf{K}_{n\lambda}^{1/2}\bar{\mathbf{K}}_{n\lambda}^{-1}\mathbf{K}_{n\lambda}^{1/2}\|$$

where we used $\mathcal{Z}_n\mathcal{Z}_n^* = \mathbf{K}$. In second term substitute $\bar{\mathbf{K}} = \mathcal{Z}_n\bar{\mathcal{P}}\mathcal{Z}_n^*$ to get:

$$\begin{aligned}\|n\lambda\mathcal{Z}_n^\dagger\bar{\mathbf{K}}\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n\| &= n\lambda\|\mathcal{Z}_n^\dagger\mathcal{Z}_n\bar{\mathcal{P}}\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n\| \le n\lambda\|\mathcal{Z}_n^\dagger\mathcal{Z}_n\| \cdot \|\bar{\mathcal{P}}\| \cdot \|\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n\| \\ &\le n\lambda\|\mathcal{Z}_n^*\bar{\mathbf{K}}_{n\lambda}^{-1}\mathcal{Z}_n\| \le n\lambda\|\mathbf{K}_{n\lambda}^{1/2}\bar{\mathbf{K}}_{n\lambda}^{-1}\bar{\mathbf{K}}_{n\lambda}^{1/2}\|.\end{aligned}$$

Combining these upperbounds we get,

$$\mathcal{T}_b \leq 2R\lambda^{-a}\lambda^{\zeta}\big(1 + \|\mathbf{K}_{n\lambda}^{1/2}\bar{\mathbf{K}}_{n\lambda}^{-1}\mathbf{K}_{n\lambda}^{1/2}\|\big). \tag{15}$$

Combining (15), (14), (13), (12), and (11), we conclude

$$\big\|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_{[q]} - f_{\mathcal{H}})\big\|_\rho \leq \lambda^{-a}\cdot O\Big(\lambda^{\zeta}\big\|\bar{\mathbf{K}}_{n\lambda}^{1/2}\bar{\mathbf{K}}_{n\lambda}^{-1}\bar{\mathbf{K}}_{n\lambda}^{1/2}\big\| + \frac{1}{n\lambda^{1-\zeta}} + \frac{1}{\sqrt{n\lambda^{\gamma}}}\Big)\log(1/\delta).$$

■

Using Theorem 30 we prove the following two corollaries for classical Nyström and Block Nyström KRR respectively, recovering the optimal generalization error up to a multiplicative factor of α.

For the following corollary, consider $\lambda' = \alpha\lambda$ and construct a Nyström approximation to $\mathbf{K}$ by sampling landmarks from $n\lambda'$-ridge leverage scores of $\mathbf{K}$, but still adding $n\lambda$ as the regularizer. We refer to this as classical Nyström.

Corollary 31 (Statistical risk of KRR with classical Nyström) *Let $0 \leq a \leq \zeta$, $\zeta < \frac{1}{2}$, $n \geq O(G^2\log(G^2/\delta))$, $\lambda^* = \tilde{O}(n^{-\frac{1}{2\zeta+\gamma}})$ if $2\zeta+\gamma > 1$, and $\lambda^* = \tilde{O}(n^{-1})$ if $2\zeta+\gamma \leq 1$. Let $\lambda' = \alpha\lambda^* < 1$ for some $\alpha > 1$. Then under assumptions 1-4 and for any $\delta > 0$*

$$\big\|\mathcal{L}^{-a}(\mathcal{S}_\rho\hat{\omega} - f_{\mathcal{H}})\big\|_\rho \leq \begin{cases}\alpha\cdot\tilde{O}(n^{-\frac{\zeta-a}{2\zeta+\gamma}}) \text{ if } 2\zeta+\gamma > 1\\ \alpha\cdot\tilde{O}(n^{-(\zeta-a)}) \text{ if } 2\zeta+\gamma \leq 1\end{cases}$$

with probability $1-\delta$. Here $\hat{\omega}$ denotes the hypothesis $\hat{\mathcal{P}}\mathcal{Z}_n^*(\hat{\mathbf{K}} + n\lambda^*\mathbf{I})^{-1}\mathbf{y}$, $\hat{\mathcal{P}}$ is orthogonal projection onto the m dimensional subspace spanned by Nyström landmarks sampled using $O(1)$-approximate $n\lambda'$-ridge leverage scores of $\mathbf{K}$ and $\hat{\mathbf{K}}$ denotes the corresponding Nyström approximation to $\mathbf{K}$.

Proof Let λ^* be the optimal regularizer depending on the case if $2\zeta+\gamma > 1$ or $2\zeta+\gamma \leq 1$. Consider $\lambda' = \alpha\lambda^*$. We have with probability $1-\delta$

$$\mathbf{K} + n\lambda^*\mathbf{I} \preceq \mathbf{K} + n\lambda'\mathbf{I} \preceq 2(\hat{\mathbf{K}} + n\lambda'\mathbf{I}) \preceq \frac{2\lambda'}{\lambda^*}(\hat{\mathbf{K}} + n\lambda^*\mathbf{I}) \preceq 2\alpha(\mathbf{K} + n\lambda^*\mathbf{I})$$

implying $\|(\mathbf{K} + n\lambda^*\mathbf{I})^{1/2}\big(\hat{\mathbf{K}} + n\lambda^*\mathbf{I}\big)^{-1}(\mathbf{K} + n\lambda^*\mathbf{I})^{1/2}\| \leq 2\alpha$. Substituting this in the statement of Theorem 30 finishes the proof. ■

We provide an extended version of Corollary 15. The following extension is useful when the excess risk factor α is potentially quite large and the resulting linear system $(\hat{\mathbf{K}}_{[q]} + n\lambda^*\mathbf{I})\mathbf{w} = \mathbf{y}$ is more expensive to solve than sampling Nyström landmarks from $\alpha^2\lambda^*$-ridge leverage scores of $\mathbf{K}$, due to large number of blocks as $q = \tilde{O}(\alpha)$. To recover corollary 15 one can consider $\beta = \alpha$ in the following result.

Corollary 32 (Expected risk of Block-Nyström-KRR) *Under Assumptions 1, 2, 3 and 4, suppose that $\zeta < 1/2$ and $n \geq \tilde{O}(G^2\log(G^2/\delta))$. Also, let $\lambda^* = \tilde{O}(n^{-\frac{1}{2\zeta+\gamma}})$ if $2\zeta+\gamma > 1$ and*

$\lambda^* = \tilde{O}(n^{-1})$ if $2\zeta + \gamma \leq 1$. Consider Block-Nyström-KRR, $\hat{\omega}_{[q]} = \mathcal{P}_{[q]}\mathcal{Z}_n^*(\hat{\mathbf{K}}_{[q]} + n\lambda^*\mathbf{I})^{-1}\mathbf{y}$, constructed using $q = \tilde{O}(\beta)$ blocks, each with $\tilde{O}(d_{n\lambda'}(\mathbf{K}))$ leverage score samples where $1 \leq \beta \leq \alpha$ and $\lambda' = \beta\alpha\lambda^*$. Then,

$$\left\|\mathcal{S}_\rho\hat{\omega}_{[q]} - f_{\mathcal{H}}\right\|_\rho \leq \begin{cases} \alpha \cdot \tilde{O}(n^{-\frac{\zeta}{2\zeta+\gamma}}) & \text{if } 2\zeta+\gamma > 1, \\ \alpha \cdot \tilde{O}(n^{-\zeta}) & \text{if } 2\zeta+\gamma \leq 1. \end{cases}$$

Proof of Corollary 32. Let λ^* be the optimal regularizer depending on the case if $2\zeta + \gamma > 1$ or $2\zeta + \gamma \leq 1$. Consider $\tilde{\lambda} = \frac{\alpha}{\beta}\lambda^*$ and $\lambda' = \beta\alpha\lambda^*$. We have with probability $1 - \delta$,

$$\mathbf{K} + n\lambda^*\mathbf{I} \preceq \mathbf{K} + n\tilde{\lambda}\mathbf{I} \preceq 8\sqrt{\frac{\lambda'}{\tilde{\lambda}}}\Big(\frac{1}{q}\sum_{i=1}^{q}\hat{\mathbf{K}}_i + n\tilde{\lambda}\mathbf{I}\Big) \preceq 8\sqrt{\frac{\lambda'}{\tilde{\lambda}}}\cdot\frac{\tilde{\lambda}}{\lambda^*}\Big(\frac{1}{q}\sum_{i=1}^{q}\hat{\mathbf{K}}_i + n\lambda^*\mathbf{I}\Big) \preceq 8\alpha(\mathbf{K} + n\lambda^*\mathbf{I})$$

For $q > O(\beta\log(n/\delta))$. In the second inequality, we used Theorem 23. Therefore,

$$(\mathbf{K} + n\lambda^*\mathbf{I})^{-1} \preceq \Big(\frac{1}{q}\sum_{i=1}^{q}\hat{\mathbf{K}}_i + n\lambda^*\mathbf{I}\Big)^{-1} \preceq 8\alpha(\mathbf{K} + n\lambda^*\mathbf{I})^{-1}$$

implying $\Big\|(\mathbf{K} + n\lambda^*\mathbf{I})^{1/2}\Big(\frac{1}{q}\sum_{i=1}^{q}\hat{\mathbf{K}}_i + n\lambda^*\mathbf{I}\Big)^{-1}(\mathbf{K} + n\lambda^*\mathbf{I})^{1/2}\Big\| < 8\alpha$. Substituting this in the statement of Theorem 30 finishes the proof. ■

In the following lemma, we derive explicit expressions for the time complexity of classical Nyström KRR for any $\alpha \geq 1$. In particular, we sample landmarks using $n\lambda'$-ridge leverage scores of $\mathbf{K}$ where $\lambda' = \alpha\lambda^*$, obtaining the Nyström approximation $\hat{\mathbf{K}}$. We then add $n\lambda^*\mathbf{I}$ to $\hat{\mathbf{K}}$, obtaining an $n\lambda^*$-regularized α-approximation to $\mathbf{K}$. For better intuition, it is helpful to think of α as n^θ for small $\theta > 0$. In particular, the form of the exponent θ changes as the regime changes from $2\zeta + \gamma > 1$ to $2\zeta + \gamma \leq 1$. We capture this regime change by considering $\theta = \mu/\max\{1, 2\zeta + \gamma\}$. We consider $a = 0$ to obtain the time complexity for obtaining α multiplicative bound for the expected risk.

Lemma 33 (Cost analysis of KRR with classical Nyström) *Let $a = 0$ and $\zeta < \frac{1}{2}$. Let $\lambda' = \alpha\lambda^*$, where $\lambda^* = \tilde{O}(n^{-\frac{1}{2\zeta+\gamma}})$ and $\alpha = n^{\frac{\mu}{2\zeta+\gamma}}$ if $2\zeta + \gamma > 1$, and $\lambda^* = \tilde{O}(n^{-1})$ and $\alpha = n^\mu$ if $2\zeta + \gamma < 1$. Furthermore, let $\mu \leq \zeta$. Then the total time complexity of classical Nyström KRR in regime of $2\zeta + \gamma > 1$ is given as:*

$$\mathcal{T}_{nys} = \begin{cases} \tilde{O}\Big(n^{\frac{(1+2\gamma)(1-\mu)}{2\zeta+\gamma}}\Big) & \text{if } \mu \leq \frac{1-2\zeta}{1+\gamma} \\ \tilde{O}\Big(n^{\frac{2\zeta+\gamma(2-\mu)}{2\zeta+\gamma}}\Big) & \text{otherwise.} \end{cases}$$

and if $2\zeta + \gamma \leq 1$,

$$\mathcal{T}_{nys} = \begin{cases} \tilde{O}\Big(n^{(1+2\gamma)(1-\mu)}\Big) \text{if } \mu \leq \frac{\gamma}{1+\gamma}, \\ \tilde{O}\Big(n^{1+\gamma(1-\mu)}\Big) \text{otherwise.} \end{cases}$$

Proof In the regime of $2\zeta + \gamma > 1$, the optimal λ^* is $\tilde{O}(n^{-\frac{1}{2\zeta+\gamma}})$. Let $\lambda' = \alpha\lambda^*$ where $\alpha = n^{\frac{\mu}{2\zeta+\gamma}}$. Let $m = d_{n\lambda'}(\mathbf{K})$. The overall time complexity of classical Nyström KRR is the sum of three costs as follows:

- Landmarks sampling using Lemma 8 : $\tilde{O}(m^2/\lambda') = \tilde{O}(\frac{1}{(\lambda')^{1+2\gamma}}) = \tilde{O}(\frac{1}{\alpha^{1+2\gamma}} \cdot n^{\frac{1+2\gamma}{2\zeta+\gamma}})$.
- Cost of inverting $\tilde{O}(m) \times \tilde{O}(m)$ submatrix of $\mathbf{K}$: $\tilde{O}(m^3) = \tilde{O}(\frac{1}{\alpha^{3\gamma}} \cdot n^{\frac{3\gamma}{2\zeta+\gamma}})$.
- Cost of solving linear system $(\hat{\mathbf{K}} + n\lambda^*\mathbf{I})\mathbf{w} = \mathbf{y}$ using preconditioned conjugate gradient: $\tilde{O}(nm) = \tilde{O}(\frac{1}{\alpha^{\gamma}} \cdot n^{\frac{2\zeta+2\gamma}{2\zeta+\gamma}})$.

Under the assumption that $\mu \leq \zeta$, it is easy to show that cost of sampling landmarks is larger than the cost of inverting the principal submatrix. Now comparing the first and third costs it is easy to see that cost of landmarks sampling dominates the cost of solving the linear system as long as $\mu < \frac{1-2\zeta}{1+\gamma}$. Therefore overall time complexity is given as

$$\begin{cases} \tilde{O}\left(n^{\frac{(1+2\gamma)(1-\mu)}{2\zeta+\gamma}}\right) & \text{if } \mu \leq \frac{1-2\zeta}{1+\gamma} \\ \tilde{O}\left(n^{\frac{2\zeta+\gamma(2-\mu)}{2\zeta+\gamma}}\right) & \text{otherwise.} \end{cases}$$

Note that we always assume $\mu \leq \zeta$. Now consider the regime $2\zeta+\gamma < 1$, the optimal $\lambda^* = \tilde{O}(1/n)$ and the corresponding learning rate $\tilde{O}(n^{-\zeta})$. Let $\lambda' = \alpha\lambda^*$ where now $\alpha = n^{\mu}$ for $0 \leq \mu \leq \zeta$. We have

- Nyström landmarks sampling using Lemma 8: $\tilde{O}(m^2/\lambda') = \tilde{O}\left(n^{(1+2\gamma)(1-\mu)}\right)$.
- Cost of inverting $\tilde{O}(m) \times \tilde{O}(m)$ matrix : $\tilde{O}(m^3) = \tilde{O}(n^{3\gamma(1-\mu)})$
- Cost of solving linear system $(\hat{\mathbf{K}} + n\lambda^*\mathbf{I})\mathbf{w} = \mathbf{y}$ using preconditioned conjugate gradient: $\tilde{O}(n^{1+\gamma(1-\mu)})$

The overall cost is then given as

$$\begin{cases} \tilde{O}\left(n^{(1+2\gamma)(1-\mu)}\right) \text{if } \mu \leq \frac{\gamma}{1+\gamma}, \\ \tilde{O}\left(n^{1+\gamma(1-\mu)}\right) \text{otherwise.} \end{cases}$$

Here also additional constraint is $\mu \leq \zeta$. ■

We now provide explicit expressions for the time complexity of Block-Nyström KRR.

Lemma 34 (Cost analysis of KRR with Block-Nyström) *Let $a = 0$, $\zeta < \frac{1}{2}$ and λ^* be the optimal regularizer depending on the regime if $2\zeta+\gamma > 1$ or $2\zeta+\gamma \leq 1$. Let $\lambda' = \beta\alpha\lambda^*$ for some $1 \leq \beta \leq \alpha$ and define q as follows*

$$q = \begin{cases} \lceil 200\beta \log(n)/\delta \rceil \text{ if } \beta > 1 \\ 1 \text{ if } \beta = 1. \end{cases}$$

If $2\zeta + \gamma > 1$, let $\lambda^* = \tilde{O}(n^{-\frac{1}{2\zeta+\gamma}})$, $\alpha = n^{\frac{\mu}{2\zeta+\gamma}}$ and $\beta = n^{\frac{\nu}{2\zeta+\gamma}}$ where $0 \leq \mu \leq \zeta$ and $\nu = \max\left\{0, \min\left\{\mu, \frac{1-2\zeta-\mu(1.5+\gamma)}{\gamma+\phi+0.5}\right\}\right\}$. Then the total time complexity of Block-Nyström KRR is given

as:

$$\mathcal{T}_{Blk} = \begin{cases} \tilde{O}\left(n^{\frac{(1+2\gamma)(1-\mu)-2\gamma\nu}{2\zeta+\gamma}}\right) \text{ if } \mu < \frac{1-2\zeta}{\gamma+1.5}, \\ \tilde{O}\left(n^{\frac{(1+2\gamma)(1-\mu)}{2\zeta+\gamma}}\right) \text{ if } \frac{1-2\zeta}{\gamma+1.5} \leq \mu \leq \frac{1-2\zeta}{\gamma+1}, \\ \tilde{O}\left(n^{\frac{2\zeta+\gamma(2-\mu)}{2\zeta+\gamma}}\right) \text{otherwise}. \end{cases}$$

If $2\zeta+\gamma \leq 1$, *let* $\lambda^* = \tilde{O}(n^{-1})$, $\alpha = n^{\mu}$, $\beta = n^{\nu}$ *where* $0 \leq \mu \leq \zeta$ *and* $\nu = \max\left\{0, \min\left\{\mu, \frac{\gamma-\mu(\gamma+1.5)}{\gamma+\phi+0.5}\right\}\right\}$. *Then the overall cost of Block-Nyström is given as*

$$\mathcal{T}_{Blk} = \begin{cases} \tilde{O}\left(n^{(1+2\gamma)(1-\mu)-2\gamma\nu}\right) \text{ if } \mu < \frac{\gamma}{\gamma+1.5} \\ \tilde{O}\left(n^{(1+2\gamma)(1-\mu)}\right) \text{ if } \frac{\gamma}{\gamma+1.5} \leq \mu \leq \frac{\gamma}{1+\gamma} \\ \tilde{O}\left(n^{1+\gamma(1-\mu)}\right) \text{otherwise} \end{cases}$$

Proof We consider $\tilde{\lambda} = \frac{\alpha}{\beta}\lambda^*$ and $\lambda' = \beta\alpha\lambda^*$. For sampling Nyström landmarks we use Lemma 8 q times leading to cost of $\tilde{O}(q \cdot m^2/\lambda')$, where $m = d_{n\lambda'}(\mathbf{K})$. First consider the regime $2\zeta + \gamma > 1$. Similar to Lemma 33 we break down the cost in three components:

- Nyström landmarks sampling : $\tilde{O}(qm^2/\lambda') = \tilde{O}(\beta^{-2\gamma}\alpha^{-(1+2\gamma)}n^{\frac{1+2\gamma}{2\zeta+\gamma}})$.
- Cost of inverting q $\tilde{O}(m)\times\tilde{O}(m)$ principal submatrices of $\mathbf{K}$: $\tilde{O}(qm^3) = \tilde{O}\left(\beta^{1-3\gamma}\alpha^{-3\gamma}n^{\frac{3\gamma}{2\zeta+\gamma}}\right)$.
- Cost of solving linear system $(\hat{\mathbf{K}}_{[q]} + n\lambda^*\mathbf{I})\mathbf{w} = \mathbf{y}$ using techniques from Lemma 27: $\tilde{O}\left(\sqrt{\tilde{\lambda}/\lambda^*} \cdot q^{1+\phi}nm\right) = \tilde{O}\left(\sqrt{\alpha/\beta} \cdot \beta^{1+\phi-\gamma}\alpha^{-\gamma}n^{\frac{2\zeta+2\gamma}{2\zeta+\gamma}}\right) = \tilde{O}\left(\beta^{1/2+\phi-\gamma}\alpha^{1/2-\gamma}n^{\frac{2\zeta+2\gamma}{2\zeta+\gamma}}\right)$.

Under the assumption that $\mu \leq \zeta$, one can show that the cost of landmarks sampling is larger than the cost of inverting the principal submatrices. Comparing the first and the third costs we get

$$\begin{aligned} &\beta^{-2\gamma}\alpha^{-(1+2\gamma)}n^{\frac{1+2\gamma}{2\zeta+\gamma}} \geq \beta^{1/2+\phi-\gamma}\alpha^{1/2-\gamma}n^{\frac{2\zeta+2\gamma}{2\zeta+\gamma}} \\ \Rightarrow &n^{\frac{1-2\zeta}{2\zeta+\gamma}} \geq \beta^{1/2+\phi+\gamma}\alpha^{3/2+\gamma} \\ \Rightarrow &1-2\zeta \geq \mu(1.5+\gamma) + \nu(0.5+\phi+\gamma) \\ \Rightarrow &\nu \leq \frac{1-2\zeta-\mu(1.5+\gamma)}{\gamma+\phi+0.5}. \end{aligned}$$

We set $\nu = \max\left\{0, \min\left\{\mu, \frac{1-2\zeta-\mu(1.5+\gamma)}{\gamma+\phi+0.5}\right\}\right\}$. If $\nu > 0$, then we run Block-Nyström KRR and if $\nu = 0$ i.e. $\beta = 1$ and consequently $q = 1$, we run classical Nyström KRR. The overall time complexity is given as

$$\begin{cases} \tilde{O}\left(n^{\frac{(1+2\gamma)(1-\mu)-2\gamma\nu}{2\zeta+\gamma}}\right) \text{ if } \mu < \frac{1-2\zeta}{\gamma+1.5}, \\ \tilde{O}\left(n^{\frac{(1+2\gamma)(1-\mu)}{2\zeta+\gamma}}\right) \text{ if } \frac{1-2\zeta}{\gamma+1.5} \leq \mu \leq \frac{1-2\zeta}{\gamma+1}, \\ \tilde{O}\left(n^{\frac{2\zeta+\gamma(2-\mu)}{2\zeta+\gamma}}\right) \text{ otherwise.} \end{cases}$$

Now let $2\zeta + \gamma \leq 1$. The cost of Block-Nyström KRR is given as

- Landmarks sampling: $\tilde{O}(qm^2/\lambda') = \tilde{O}\Big(\beta^{-2\gamma}\alpha^{-(1+2\gamma)}n^{1+2\gamma}\Big)$.
- Cost of inverting q $\tilde{O}(m)\times\tilde{O}(m)$ principal submatrices of $\mathbf{K}$: $\tilde{O}(qm^3) : \tilde{O}\Big(\beta^{1-3\gamma}\alpha^{-3\gamma}n^{3\gamma}\Big)$
- Cost of solving linear system $(\hat{\mathbf{K}}_{[q]} + n\lambda^*\mathbf{I})\mathbf{w} = \mathbf{y}$ using techniques from Lemma 27: $\tilde{O}\Big(\sqrt{\tilde{\lambda}/\lambda^*}\cdot q^{1+\phi}nm\Big) = \tilde{O}\Big(\sqrt{\alpha/\beta}\cdot\beta^{1+\phi-\gamma}\alpha^{-\gamma}n^{1+\gamma}\Big) = \tilde{O}\Big(\beta^{1/2+\phi-\gamma}\alpha^{1/2-\gamma}n^{1+\gamma}\Big)$

Again, one can show that the cost of landmarks sampling is larger than the cost of inverting the principal submatrices. Comparing the first and third costs we get

$$\begin{aligned}
\beta^{-2\gamma}\alpha^{-(1+2\gamma)}n^{1+2\gamma} &\geq \beta^{1/2+\phi-\gamma}\alpha^{1/2-\gamma}n^{1+\gamma} \\
\Rightarrow n^{\gamma} &\geq \beta^{1/2+\phi+\gamma}\alpha^{3/2+\gamma} \\
\Rightarrow \gamma &\geq \mu(\gamma+3/2)+\nu(\gamma+\phi+1/2) \\
\Rightarrow \nu &\leq \frac{\gamma-\mu(\gamma+1.5)}{\gamma+\phi+0.5}
\end{aligned}$$

As $0 \leq \nu \leq \mu$ we set $\nu = \max\Big\{0, \min\big\{\mu, \frac{\gamma-\mu(\gamma+1.5)}{\gamma+\phi+0.5}\big\}\Big\}$. Again, $\nu > 0$ corresponds to Block-Nyström and if $\nu = 0$ we run classical Nyström KRR. The overall cost is therefore given as

$$\begin{cases}
\tilde{O}\Big(n^{(1+2\gamma)(1-\mu)-2\gamma\nu}\Big) \text{ if } \mu < \frac{\gamma}{\gamma+1.5} \\
\tilde{O}\Big(n^{(1+2\gamma)(1-\mu)}\Big) \quad \text{if } \frac{\gamma}{\gamma+1.5} \leq \mu \leq \frac{\gamma}{1+\gamma} \\
\tilde{O}\Big(n^{1+\gamma(1-\mu)}\Big) \text{ otherwise.}
\end{cases}$$

■

Detailed cost comparison between classical Nyström and Block Nyström KRR. As shown in Lemma 34 there is a regime in which Block-Nyström provides computational gains over classical Nyström KRR. In particular, we have proven that as long as $\alpha = n^{\theta}$ for $\theta < \frac{1}{\gamma+1.5}\cdot\frac{\min\{\gamma,1-2\zeta\}}{\max\{1,2\zeta+\gamma\}}$, Block-Nyström enjoys faster time complexity and achieves the same excess risk factor of α as obtained by classical Nyström. Furthermore, within this range of values for θ, the computational gains enjoyed by Block-Nyström change their trend. As θ increases from 0 to $\frac{1}{2\gamma+2+o(1)}\cdot\frac{\min\{\gamma,1-2\zeta\}}{\max\{1,2\zeta+\gamma\}}$, the ratio of time complexity of two methods, $\mathcal{T}_{Blk}/\mathcal{T}_{nys}$, scales as $1/n^{2\theta\gamma}$, whereas when θ increases from $\frac{1}{2\gamma+2+o(1)}\cdot\frac{\min\{\gamma,1-2\zeta\}}{\max\{1,2\zeta+\gamma\}}$ to $\frac{1}{\gamma+1.5}\cdot\frac{\min\{\gamma,1-2\zeta\}}{\max\{1,2\zeta+\gamma\}}$, we use a slightly different averaging scheme where we reduce the number of blocks to avoid computational overhead resulting due to potentially having to solve the linear system $(\hat{\mathbf{K}}_{[q]} + n\lambda^*\mathbf{I})\mathbf{w} = \mathbf{y}$ with $q = \tilde{O}(\alpha)$ number of blocks. In this case we consider $q = \tilde{O}(\beta)$ with β diminishing from α to 1 as θ increases in this later regime. This idea of reduced averaging leads $\mathcal{T}_{Blk}/\mathcal{T}_{nys}$ to scale as $\frac{1}{n^{\xi}}$ where $\xi = \frac{1}{\gamma+o(1)+0.5}\Big(\frac{\min\{\gamma,1-2\zeta\}}{\max\{1,2\zeta+\gamma\}} - (\gamma+1.5)\theta\Big)$, meaning larger θ has diminishing computational gains. At the critical point of $\theta = \frac{1}{\gamma+1.5}\frac{\min\{\gamma,1-2\zeta\}}{\max\{1,2\zeta+\gamma\}}$, Block Nyström KRR reduces to classical Nyström KRR as β becomes 1.

Appendix D. Proof of Lemma 13

In this section we provide proof of Lemma 13. We start by restating the lemma, replacing the notation $\bar{\lambda}$ with λ for notational convenience. For psd matrices, we use $\mathbf{A} \approx_c \mathbf{B}$ to denote $c^{-1}\mathbf{B} \preceq \mathbf{A} \preceq c\mathbf{B}$, and $\omega \approx 2.372$ denotes the fast matrix multiplication exponent.

Lemma 35 *Given an $n \times n$ psd matrix $\mathbf{A}$ with at most k eigenvalues larger than $O(1)$ times its smallest eigenvalue, consider $\lambda = \frac{1}{k}\sum_{i>k}\lambda_i(\mathbf{A})$. Then, $d_\lambda(\mathbf{A}) \leq 2k$, and we can compute $O(1)$-approximations of all λ-ridge leverage scores of $\mathbf{A}$ in time $\tilde{O}(\mathrm{nnz}(\mathbf{A}) + k^\omega)$.*

Proof The fact that $d_\lambda(\mathbf{A}) \leq 2k$ is shown in Lemma 2.1 of Dereziński et al. (2025). Next, let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ be the eigenvalues of $\mathbf{A}$. We will show that it suffices to approximate the ridge leverage scores of $\mathbf{A}^2$ instead of $\mathbf{A}$. Let $\tilde{\lambda} = \frac{1}{k}\sum_{i>k}\lambda_i^2$, so that $d_{\tilde{\lambda}}(\mathbf{A}^2) \leq 2k$. Also, define $\bar{\kappa}_k(\mathbf{A}) = \frac{1}{n-k}\sum_{i>k}\frac{\lambda_i}{\lambda_n}$. We're going to show that:

$$\bar{\kappa}_k(\mathbf{A}) \cdot \mathbf{A}(\mathbf{A}+\lambda\mathbf{I})^{-1} \preceq \bar{\kappa}_k(\mathbf{A}^2) \cdot \mathbf{A}^2(\mathbf{A}^2+\tilde{\lambda}\mathbf{I})^{-1}. \tag{16}$$

Since the matrices commute, it is enough to show this for each eigenvalue. Observe that we have:

$$\frac{\bar{\kappa}_k(\mathbf{A}^2)}{\bar{\kappa}_k(\mathbf{A})} = \frac{\sum_{j>k}\lambda_j^2}{\sum_{j>k}\lambda_n\lambda_j} \in [1, \lambda_{k+1}/\lambda_n].$$

This means that

$$\begin{aligned}\frac{\lambda_i}{\lambda_i + \frac{1}{k}\sum_{j>k}\lambda_j} &= \frac{\lambda_i^2}{\lambda_i^2 + \frac{1}{k}\sum_{j>k}\lambda_i\lambda_j} \leq \frac{\lambda_i^2}{\lambda_i^2 + \frac{1}{k}\sum_{j>k}\lambda_n\lambda_j} \\ &= \frac{\lambda_i^2}{\lambda_i^2 + \frac{\bar{\kappa}_k(\mathbf{A})}{\bar{\kappa}_k(\mathbf{A}^2)}\frac{1}{k}\sum_{j>k}\lambda_j^2} \leq \frac{\bar{\kappa}_k(\mathbf{A}^2)}{\bar{\kappa}_k(\mathbf{A})}\frac{\lambda_i^2}{\lambda_i^2 + \frac{1}{k}\sum_{j>k}\lambda_j^2},\end{aligned}$$

which implies (16). Since λ-ridge leverage scores of $\mathbf{A}$ are the diagonal entries of $\mathbf{A}(\mathbf{A}+\lambda\mathbf{I})^{-1}$, this shows that for any i:

$$\begin{aligned}\ell_i(\mathbf{A},\lambda) &= \mathbf{e}_i^\top \mathbf{A}(\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{e}_i \\ &\leq \frac{\bar{\kappa}_k(\mathbf{A}^2)}{\bar{\kappa}_k(\mathbf{A})}\mathbf{e}_i^\top \mathbf{A}^2(\mathbf{A}^2+\tilde{\lambda}\mathbf{I})^{-1}\mathbf{e}_i \\ &= O(1)\cdot\ell_i(\mathbf{A}^2,\tilde{\lambda}),\end{aligned}$$

where in the last step we used that $\bar{\kappa}_k(\mathbf{A}^2) = O(1)$ because all of the tail eigenvalues are within a constant of each other. This, together with the fact that $\sum_i \ell_i(\mathbf{A}^2,\tilde{\lambda}) = O(k)$, implies that it suffices to perform approximate ridge leverage score sampling with respect to $\ell_i(\mathbf{A}^2,\tilde{\lambda})$.

Next, we show how to construct approximations of $\ell_i(\mathbf{A}^2,\tilde{\lambda})$ in $\tilde{O}(\mathrm{nnz}(\mathbf{A}) + k^\omega)$ time. First, notice that

$$\ell_i(\mathbf{A}^2,\tilde{\lambda}) = \mathbf{a}_i^\top(\mathbf{A}^2+\tilde{\lambda}\mathbf{I})^{-1}\mathbf{a}_i,$$

where $\mathbf{a}_i$ is the ith row of $\mathbf{A}$. Next, we approximate the inner matrix by using a $\tilde{O}(k) \times n$ sparse sketching matrix Π (Cohen et al., 2016), computing $\mathbf{M} = \Pi\mathbf{A}$ in $\tilde{O}(\mathrm{nnz}(\mathbf{A}))$ time, so that with high probability $\mathbf{M}^\top\mathbf{M} + \tilde{\lambda}\mathbf{I} \approx_2 \mathbf{A}^2 + \tilde{\lambda}\mathbf{I}$. Thus, it suffices to approximate:

$$\begin{aligned}\mathbf{a}_i^\top(\mathbf{M}^\top\mathbf{M} + \tilde{\lambda}\mathbf{I})^{-1}\mathbf{a}_i &= \|(\mathbf{M}^\top\mathbf{M} + \tilde{\lambda}\mathbf{I})^{-1/2}\mathbf{a}_i\|^2 \\ &= \|\mathbf{B}_{\tilde{\lambda}}(\mathbf{M}^\top\mathbf{M} + \tilde{\lambda}\mathbf{I})^{-1}\mathbf{a}_i\|^2, \qquad \text{where} \quad \mathbf{B}_{\tilde{\lambda}} = \begin{bmatrix}\mathbf{M} \\ \sqrt{\tilde{\lambda}}\mathbf{I}\end{bmatrix}.\end{aligned}$$

Note that if $\mathbf{S}$ is an $O(\log n) \times 2n$ Johnson-Lindenstrauss embedding, then we can approximate the above norms to within a constant factor with $\tilde{\ell}_i = \|\mathbf{S}\mathbf{B}_{\tilde{\lambda}}(\mathbf{M}^\top\mathbf{M} + \tilde{\lambda}\mathbf{I})^{-1}\mathbf{a}_i\|^2$. In order to compute these estimates, we first pre-compute the $O(\log n) \times d$ matrix $\mathbf{S}\mathbf{B}_{\tilde{\lambda}}$, and then compute $\mathbf{R} = \mathbf{S}\mathbf{B}_{\tilde{\lambda}} \cdot (\mathbf{M}^\top\mathbf{M} + \tilde{\lambda}\mathbf{I})^{-1}$ by solving $O(\log n)$ linear systems with $\mathbf{M}^\top\mathbf{M} + \tilde{\lambda}\mathbf{I}$. The linear systems can be solved by following the preconditioning strategy of Dereziński et al. (2025), observing that:

$$(\mathbf{M}^\top\mathbf{M} + \tilde{\lambda}\mathbf{I})^{-1} = \frac{1}{\tilde{\lambda}}\Big(\mathbf{I} - \mathbf{M}^\top(\mathbf{M}\mathbf{M}^\top + \tilde{\lambda}\mathbf{I})^{-1}\mathbf{M}\Big).$$

Thus, it suffices to solve a linear system with $\mathbf{M}\mathbf{M}^\top + \tilde{\lambda}\mathbf{I}$. To do this, we construct a $k \times \tilde{O}(k)$ sketch $\tilde{\mathbf{M}} = \mathbf{M}\mathbf{S}_2^\top$ such that $\tilde{\mathbf{M}}\tilde{\mathbf{M}}^\top \approx_{O(1)} \mathbf{M}\mathbf{M}^\top$ via a subspace embedding (Cohen et al., 2016), and precondition that linear system with $\mathbf{B}^{-1} = (\tilde{\mathbf{M}}\tilde{\mathbf{M}}^\top + \tilde{\lambda}\mathbf{I})^{-1} \approx_{O(1)} (\mathbf{M}\mathbf{M}^\top + \tilde{\lambda}\mathbf{I})^{-1}$. Thus, after precomputing $\mathbf{B}^{-1}$ at the cost of $\tilde{O}(\mathrm{nnz}(\mathbf{M}) + k^\omega)$, we can solve the linear system $(\mathbf{M}^\top\mathbf{M} + \tilde{\lambda}\mathbf{I})^{-1}\mathbf{v}$ in time $\tilde{O}(\mathrm{nnz}(\mathbf{A}) + k^2)$.

Finally, we compute the $O(\log n) \times n$ matrix $\mathbf{R}\mathbf{A}$ and then let $\tilde{\ell}_i$ be the squared row norms of this matrix. The overall costs are $\tilde{O}(\mathrm{nnz}(\mathbf{A}) + k^\omega)$ for precomputing $\mathbf{R}$ and then $\tilde{O}(\mathrm{nnz}(\mathbf{A}))$ for computing the squared row norms.

■