Gain with no Pain: Efficiency of Kernel-PCA by Nyström Sampling

Nicholas Sterge Penn State, USA
Bharath Sriperumbudur Penn State, USA
Lorenzo Rosasco Univ. of Genoa, Italy
Alessandro Rudi INRIA/ENS, France

Abstract

In this paper, we analyze a Nyström based approach to efficient large scale kernel principal component analysis (PCA). The latter is a natural nonlinear extension of classical PCA based on considering a nonlinear feature map or the corresponding kernel. Like other kernel approaches, kernel PCA enjoys good mathematical and statistical properties but, numerically, it scales poorly with the sample size. Our analysis shows that Nyström sampling greatly improves computational efficiency without incurring any loss of statistical accuracy. While similar effects have been observed in supervised learning, this is the first such result for PCA. Our theoretical findings are based on a combination of analytic and concentration of measure techniques. Our study is more broadly motivated by the question of understanding the interplay between statistical and computational requirements for learning.

1 Introduction

Achieving good statistical accuracy under budgeted computational resources is a central theme in modern machine learning (Bottou and Bousquet, 2008). Indeed, the problem of understanding the interplay and trade-offs between statistical and computational requirements has recently received much attention. Nonparametric learning, and in particular kernel methods, have provided a natural framework to pursue these questions, see e.g., Musco and Musco (2017); Rudi et al. (2015); Alaoui and Mahoney (2014); Bach (2013); Calandriello et al. (2018); Orabona et al. (2008). On the one hand, these methods are developed in a sound mathematical setting and their statistical properties are well studied. On the other hand, from a numerical point of view, they scale poorly to large scale problems, and hence improved computational efficiency is of particular interest.

While initial studies have mostly focused on approximating kernel matrices (Drineas and Mahoney, 2005; Gittens and Mahoney, 2013; Jin et al., 2013; Zhang et al., 2008), recent results have highlighted the importance of considering downstream learning tasks, with a focus on the interplay between statistical accuracy and computational complexity. In particular, results in supervised learning have shown there are regimes where computational gains can be achieved with no loss of statistical accuracy (Rudi et al., 2015; Rudi and Rosasco, 2017). A basic intuition is that approximate computations provide a form of implicit regularization, hence memory and time requirements can be tailored to statistical accuracy allowed by the data (Rudi et al., 2015). To which extent similar effects can be proved beyond supervised learning is largely unexplored. Indeed, the only result we are aware of in this direction was recently shown for kernel k-means in (Calandriello et al., 2018).

In this paper, we analyze one of the most basic unsupervised approaches, namely PCA, or rather its nonlinear version, that is kernel PCA (Schölkopf et al., 1998) and its approximate version through Nyström sampling (Williams and Seeger, 2001). The empirical behavior of approximate kernel PCA using Nyström sampling is well understood (Zhang et al., 2008), where it has been shown to lead to significant performance gains; however, theoretical results to this end are quite limited. It is well known that Nyström kernel PCA (NY-KPCA) with m subsamples achieves a time complexity of $O(nm^2 + m^3)$ and a space complexity of $O(m^2)$, in contrast to

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

$O(n^3)$ and $O(n^2)$ time and space complexities of KPCA, where n is the sample size, implying that the NY-KPCA has a better computational and memory requirement when $m < n$. However, to the best of our knowledge, no results are known on the statistical behavior of NY-KPCA that answers the question of whether the computational gain is achieved at the expense of statistical efficiency or not. The main contribution of the paper is a rigorous statistical analysis of NY-KPCA in terms of finite sample bounds on the error in reconstructing a kernel function based on its projections onto an ℓ-dimensional eigenspace associated with a certain covariance operator (see Theorem 2 and related Corollaries 3 and 4). In particular, we show that NY-KPCA can achieve the same *error* of KPCA with $m < n$, thereby demonstrating computational gains at no statistical loss. Moreover, we show that adaptive sampling using leverage scores (Alaoui and Mahoney, 2014) can lead to further gains. More precisely, we show that the requirement on the number of sub-samples, m varies between $(\log n)^2$ and $n^\theta \log n$ $(\theta < 1)$ depending on ℓ (dimension of the eigenspace), type of sub-sampling (uniform or adaptive) and the smoothness of the RKHS controlled by the rate of decay of eigenvalues of the covariance operator.

We note that some recent papers, e.g., (Sriperumbudur and Sterge, 2018; Ullah et al., 2018), have focused on approximate kernel PCA using random features (Rahimi and Recht, 2008). The notion of reconstruction error considered in these works is different from that of KPCA (Shawe-Taylor et al., 2005; Blanchard et al., 2007). The reason for this are certain technicalities that arise in random feature approximation (for more details, see the discussion following Corollary 4). As a consequence, these results are not directly comparable to our current work and KPCA. In contrast, our results based on Nyström approximation are directly comparable to that of KPCA, wherein we show that the proposed NY-KPCA has similar statistical behavior but better computational complexity than KPCA.

Definitions and Notation For $\mathbf{a} := (a_1, \ldots, a_d) \in \mathbb{R}^d$ and $\mathbf{b} := (b_1, \ldots, b_d) \in \mathbb{R}^d$ define $\|\mathbf{a}\|_2 := \sqrt{\sum_{i=1}^d a_i^2}$ and $\langle \mathbf{a}, \mathbf{b} \rangle_2 := \sum_{i=1}^d a_i b_i$. $\mathbf{a} \otimes_2 \mathbf{b} := \mathbf{a}\mathbf{b}^\top$ denotes the tensor product of $\mathbf{a}$ and $\mathbf{b}$. $\mathbf{I}_n$ denotes an $n \times n$ identity matrix. $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$. $[n] := \{1, \ldots, n\}$ for $n \in \mathbb{N}$. For constants a and b, $a \lesssim b$ (*resp.* $a \gtrsim b$) denotes that there exists a positive constant c (*resp.* c') such that $a \leq cb$ (*resp.* $a \geq c'b$). For a random variable A with law P and a constant b, $A \lesssim_p b$ denotes that for any $\delta > 0$, there exists a positive constant $c_\delta < \infty$ such that $P(A \leq c_\delta b) \geq \delta$.
For $x, y \in H$, a Hilbert space, $x \otimes_H y$ is an element of the tensor product space $H \otimes H$ which can also be seen as an operator from H to H as $(x \otimes_H y)z = x\langle y, z \rangle_H$ for any $z \in H$. $\alpha \in \mathbb{R}$ is called an *eigenvalue* of a bounded self-adjoint operator S if there exists an $x \neq 0$ such that $Sx = \alpha x$ and such an x is called the *eigenvector / eigenfunction* of S and α. An eigenvalue is said to be *simple* if it has multiplicity one. For an operator $S : H \to H$, $\|S\|_{\mathcal{L}^1(H)}$, $\|S\|_{\mathcal{L}^2(H)}$ and $\|S\|_{\mathcal{L}^\infty(H)}$ denote the trace, Hilbert-Schmidt and operator norms of S, respectively.

2 Kernel PCA by Nyström Sampling

In this section, we review kernel principal component analysis (KPCA) (Schölkopf et al., 1998) in population and empirical settings and introduce approximate kernel PCA using Nyström approximation. We assume the following for the rest of the paper:

Assumption 1. *$\mathcal{X}$ is a separable topological space and $(\mathcal{H}, k)$ is a separable RKHS of real-valued functions on $\mathcal{X}$ with a bounded, continuous, strictly positive definite kernel k satisfying* $\sup_{x \in \mathcal{X}} k(x, x) =: \kappa < \infty$.

2.1 KPCA and empirical KPCA

Let X be a zero-mean random variable with law $\mathbb{P}$ defined on $\mathcal{X}$. When $\mathcal{X} = \mathbb{R}^d$, classical PCA (Jolliffe, 1986) finds $\mathbf{a} \in \mathbb{R}^d$ such that $\mathrm{Var}\left[\langle \mathbf{a}, X \rangle_2\right]$ is maximized, with the constraint $\|\mathbf{a}\|_2 = 1$. Defining $C := \mathbb{E}_{X \sim \mathbb{P}}[XX^\top]$, the solution is simply the unit eigenvector of C corresponding to its largest eigenvalue. In practice, PCA is computed by replacing C with an empirical approximation $C_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ based on a sample $X_1, \ldots, X_n$. Kernel PCA extends this idea to an RKHS, $\mathcal{H}$ defined on $\mathcal{X}$, by finding $f \in \mathcal{H}$ with unit norm such that $\mathrm{Var}[f(X)]$ is maximized. Since $\mathrm{Var}[f(X)] = \langle f, Cf \rangle_{\mathcal{H}}$ assuming $\mathbb{E}[f(X)] = 0$ for all $f \in \mathcal{H}$, we have $f^* = \arg\max\{\langle f, Cf \rangle_{\mathcal{H}} : \|f\|_{\mathcal{H}} = 1\}$ where C is the (uncentered) covariance operator on $\mathcal{H}$ defined as

$$C := \int_{\mathcal{X}} k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x)\, d\mathbb{P}(x).$$

The boundedness of k in Assumption 1 ensures that C is trace class and thus compact. Since C is positive and self-adjoint, the spectral theorem (Reed and

Simon, 1980) gives

$$C = \sum_{i \in I} \lambda_i \phi_i \otimes_{\mathcal{H}} \phi_i,$$

where $(\lambda_i)_{i\in I} \subset \mathbb{R}^+$ are the eigenvalues and $(\phi_i)_{i\in I}$ is the orthonormal system of eigenfunctions that span $\overline{\mathcal{R}(C)}$ with index set I being either finite or countable (in this case $\lambda_i \to 0$ as $i \to \infty$). The solution to the KPCA problem is thus the eigenfunction of C corresponding to its largest eigenvalue. We make the following simplifying assumption for the ease of presentation.

Assumption 2. *The eigenvalues $(\lambda_i)_{i\in I}$ of C are simple, positive, and w.l.o.g. satisfy a decreasing rearrangement, i.e., $\lambda_1 > \lambda_2 \ldots$.*

Assumption 2 ensures that $(\phi_i)_{i\in I}$ form an orthonormal basis and the eigenspace corresponding to each λ_i is one-dimensional. This means the orthogonal projection operator onto the ℓ-eigenspace of C, i.e. $\text{span}\{(\phi_i)_{i=1}^{\ell}\}$, is given by

$$P^{\ell}(C) = \sum_{i=1}^{\ell} \phi_i \otimes_{\mathcal{H}} \phi_i. \tag{1}$$

The above construction corresponds to population version of KPCA when the data distribution $\mathbb{P}$ is known. If $\mathbb{P}$ is unknown and the knowledge of $\mathbb{P}$ is available only through the training set $(X_i)_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$, then KPCA cannot be carried out as C depends on $\mathbb{P}$. Therefore, an approximation to C is used to perform KPCA. Most commonly, this approximation is chosen to be the empirical estimator of C defined as

$$C_n = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i) \otimes_{\mathcal{H}} k(\cdot, X_i)$$

resulting in empirical kernel PCA (EKPCA). Note that C_n is a finite rank, positive, and self-adjoint operator. Thus the spectral theorem (Reed and Simon, 1980) yields

$$C_n = \sum_{i=1}^{n} \hat{\lambda}_i \hat{\phi}_i \otimes_{\mathcal{H}} \hat{\phi}_i,$$

where $(\hat{\lambda}_i)_{i=1}^n \subset \mathbb{R}^+$ and $(\hat{\phi}_i)_{i=1}^n \subset \mathcal{H}$ are the eigenvalues and eigenfunctions of C_n. Similar to Assumption 2, we assume the following:

Assumption 3. $\text{rank}(C_n) = n$. *The eigenvalues $(\hat{\lambda}_i)_{i=1}^n$ of C_n are simple and w.l.o.g. satisfy a decreasing rearrangement, i.e., $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots$.*

The eigensystem $(\hat{\lambda}_i, \hat{\phi}_i)_{i=1}^n$ of C_n can be obtained by solving an n-dimensional system involving the eigendecomposition of the Gram matrix $\mathbf{K} = [k(X_i, X_j)]_{i,j\in[n]}$, which scales as $O(n^3)$ (Schölkopf et al., 1998). In particular, the eigenvalues of $\mathbf{K}$ are related to those of C_n as $\lambda_i(\mathbf{K}) = n\hat{\lambda}_i$. Moreover, if $\mathbf{u}_i$ is an orthonormal eigenvector of $\mathbf{K}$ corresponding to the eigenvalue $\lambda_i(\mathbf{K})$, then it holds for all $x \in \mathcal{X}$,

$$\phi_i(x) = \frac{1}{\sqrt{n\hat{\lambda}_i}} \sum_{j=1}^{n} k(x, x_j) u_{i,j}. \tag{2}$$

The above result proven in (Schölkopf et al., 2001) can be seen as a representer theorem (Kimeldorf and Wahba, 1971) for KPCA. Finally, note that, for some $\ell \leq n$, the orthogonal projection operator onto $\text{span}\{(\hat{\phi}_i)_{i=1}^{\ell}\}$ is given by

$$P^{\ell}(C_n) = \sum_{i=1}^{\ell} \hat{\phi}_i \otimes_{\mathcal{H}} \hat{\phi}_i. \tag{3}$$

2.2 Approximate kernel PCA using Nyström method

For large sample sizes, since performing KPCA is computationally intensive, various approximation schemes that has been explored in the kernel machine literature can be deployed to speed up EKPCA. Recently, one such approximation involving random Fourier features has been studied by Sriperumbudur and Sterge (2018) and Ullah et al. (2018) to speed EKPCA while maintaining its statistical performance. In this paper, we explore the popular Nyström approximation (Williams and Seeger, 2001; Drineas and Mahoney, 2005; Zhang et al., 2008) to speed up EKPCA and study the trade-offs between computational gains and statistical accuracy. The general idea in Nyström method is to obtain a low-rank approximation to the Gram matrix $\mathbf{K}$, and replace $\mathbf{K}$ by this approximation in kernel algorithms, resulting in computational speedup. Since $\mathbf{K}$ is related to C_n (as discussed in Section 2.1), Nyström method can also be seen as obtaining a low rank approximation to C_n, which is what we exploit in obtaining a Nyström approximate KPCA. It follows from (2) that the eigenfunctions of C_n lie in the space

$$\mathcal{H}_n = \left\{ f \in \mathcal{H} \,\middle|\, f = \sum_{i=1}^{n} \alpha_i k(\cdot, X_i), \alpha_1, ..., \alpha_n \in \mathbb{R} \right\}.$$

Therefore, it can be seen that EKPCA is a solution to the following problem

$$\arg\max \left\{ \langle f, C_n f \rangle_{\mathcal{H}} : f \in \mathcal{H}_n,\ \|f\|_{\mathcal{H}} = 1 \right\},$$

assuming $\mathbf{K}$ is invertible[1]. Extending this representation, suppose for fixed $m < n$ points $\{\tilde{X}_1, \ldots, \tilde{X}_m\}$ are sampled uniformly without replacement from $\{X_1, \ldots, X_n\}$, yielding the following low-dimensional subspace of $\mathcal{H}_n$,

$$\mathcal{H}_m = \left\{ f \in \mathcal{H} \,\middle|\, f = \sum_{i=1}^{m} \alpha_i k(\cdot, \tilde{X}_i), \alpha_1, ..., \alpha_m \in \mathbb{R} \right\}.$$

We propose Nyström KPCA (NY-KPCA) as a solution to the following problem:

$$\arg\max \{ \langle f, C_n f \rangle_{\mathcal{H}} : f \in \mathcal{H}_m, \, \|f\|_{\mathcal{H}} = 1 \}, \quad (4)$$

where the maximum is taken over functions in $\mathcal{H}_m$, or equivalently, over $\alpha \in \mathbb{R}^m$. Basically, we are considering a plain Nyström approximation where the m centers, $\{\tilde{X}_1, \ldots, \tilde{X}_m\}$, are sampled uniformly without replacement from the training set; however, other subsampling methods are possible, see Section 2.2.1. The following result shows that the solution to (4) is obtained by solving a finite dimensional linear system, which has better computational complexity than that of EKPCA. To this end, we first introduce some notation, $\mathbf{K}_{mm} = [k(\tilde{X}_i, \tilde{X}_j)]_{i,j\in[m]}$, $\mathbf{K}_{nm} = [k(X_i, \tilde{X}_j)]_{i\in[n],j\in[m]} \in \mathbb{R}^{n\times m}$, $\mathbf{K}_{mn} = \mathbf{K}_{nm}^\top$.

Proposition 1. *Define the $m \times m$ matrix $\mathbf{M} = \mathbf{K}_{mm}^{-1/2}\mathbf{K}_{mn}\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1/2}$. The solution to (4) is given by*

$$\hat{\phi}_{1,m} = \tilde{Z}_m^* \mathbf{K}_{mm}^{-1/2} \mathbf{u}_{1,m},$$

where $\mathbf{u}_{1,m}$ is the eigenvector of $\frac{1}{n}\mathbf{M}$ corresponding to its largest eigenvalue and $\tilde{Z}_m^* : \mathbb{R}^m \to \mathcal{H}$, $\alpha \mapsto \sum_{i=1}^m \alpha_i k(\cdot, \tilde{X}_i)$.

Proof. Define the following operators on $\mathcal{H}$

$$Z_n : \mathcal{H} \to \mathbb{R}^n, \quad f \mapsto (f(X_1), \ldots, f(X_n))^\top, \text{ and}$$
$$\tilde{Z}_m : \mathcal{H} \to \mathbb{R}^m, \quad f \mapsto (f(\tilde{X}_1), \ldots, f(\tilde{X}_m))^\top.$$

The adjoint of $\tilde{Z}_m$ (Smale and Zhou, 2007) is given by

$$\tilde{Z}_m^* : \mathbb{R}^m \to \mathcal{H}, \, \alpha \mapsto \sum_{i=1}^m \alpha_i k(\cdot, \tilde{X}_i).$$

Thus, any $f \in \mathcal{H}_m$ may be written as $\tilde{Z}_m^*\alpha$, for some $\alpha \in \mathbb{R}^m$ and so $\langle f, C_n f\rangle_{\mathcal{H}} = \frac{1}{n}\left\langle \tilde{Z}_m^*\alpha, Z_n^* Z_n \tilde{Z}_m^* \alpha \right\rangle_{\mathcal{H}} = \frac{1}{n}\alpha^\top \tilde{Z}_m Z_n^* Z_n \tilde{Z}_m^* \alpha$, where we used $Z_n^* Z_n = \frac{1}{n} C_n$. It is easy to verify that $Z_n \tilde{Z}_m^* = \mathbf{K}_{nm}$ and $\tilde{Z}_m Z_n^* = \mathbf{K}_{mn}$. Therefore, (4) can be written as

$$\arg\max \left\{ \frac{1}{n}\alpha^\top \mathbf{K}_{mn}\mathbf{K}_{nm}\alpha : \alpha^\top \mathbf{K}_{mm}\alpha = 1 \right\}. \quad (5)$$

[1]The existence of $\mathbf{K}^{-1}$ is guaranteed by strict positive definiteness of k, provided all X_i in the training set are unique.

Letting $\mathbf{u} = \mathbf{K}_{mm}^{1/2}\alpha$ simplifies the constraint in (5) to $\mathbf{u}^\top\mathbf{u} = 1$, and we write (5) as

$$\arg\max \left\{ \frac{1}{n}\mathbf{u}^\top \mathbf{K}_{mm}^{-1/2}\mathbf{K}_{mn}\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1/2}\mathbf{u} : \mathbf{u}^\top\mathbf{u} = 1 \right\}.$$

The solution to the above problem is the unit eigenvector of $\frac{1}{n}\mathbf{K}_{mm}^{-1/2}\mathbf{K}_{mn}\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1/2}$ corresponding to its largest eigenvalue. Denoting this eigenvector as $\mathbf{u}_{1,m}$, we obtain a function $\hat{\phi}_{1,m} \in \mathcal{H}$ solving the NY-KPCA problem in (4) via $\hat{\phi}_{1,m} = \tilde{Z}_m^* \mathbf{K}_{mm}^{-1/2}\mathbf{u}_{1,m}$. □

The cost of computing $\mathbf{M}$ is $O(nm^2 + m^3)$ and the cost of computing its eigendecomposition is $O(m^3)$. Thus, for $m < n$, the cost of NY-KPCA scales as $O(nm^2)$, which is lower than the $O(n^3)$ cost of EKPCA. Define $\tilde{\mathbf{K}} := \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}$, which is usually called the Nyström approximation (Williams and Seeger, 2001; Drineas and Mahoney, 2005) to the Gram matrix $\mathbf{K}$. It is easy to verify that $\mathbf{M}$ and $\tilde{\mathbf{K}}$ have same eigenvalues since $\mathbf{M} = \mathbf{K}_{mm}^{-1/2}\mathbf{K}_{mn}\left(\mathbf{K}_{mm}^{-1/2}\mathbf{K}_{mn}\right)^\top$ and $\tilde{\mathbf{K}} = \left(\mathbf{K}_{mm}^{-1/2}\mathbf{K}_{mn}\right)^\top \mathbf{K}_{mm}^{-1/2}\mathbf{K}_{mn}$, and $\operatorname{rank}(\mathbf{M}) = \operatorname{rank}(\tilde{\mathbf{K}})$. Therefore we work with $\tilde{\mathbf{K}}$ and make the following assumption on its eigenvalues.

Assumption 4. $\operatorname{rank}(\tilde{\mathbf{K}}) = m$. *The eigenvalues $(\hat{\lambda}_{i,m})_{i=1}^m$ of $\frac{1}{n}\tilde{\mathbf{K}}$ are simple and w.l.o.g. satisfy a decreasing rearrangement, i.e., $\hat{\lambda}_{1,m} > \hat{\lambda}_{2,m} > \ldots$.*

The symmetry of $\mathbf{M}$ guarantees orthonormality of $(\mathbf{u}_{i,m})_i$, and the orthonormality of $(\hat{\phi}_{i,m})_i$ follows. For some $\ell \leq m$, the orthogonal projector onto $\operatorname{span}\{\hat{\phi}_{i,m}\}_{i=1}^\ell$ is given by

$$P_m^\ell(C_n) = \sum_{i=1}^{\ell} \hat{\phi}_{i,m} \otimes_{\mathcal{H}} \hat{\phi}_{i,m}. \quad (6)$$

One may ask if $\hat{\phi}_{i,m}$ are eigenfunctions of some operator on $\mathcal{H}$. Denote P_m as the orthogonal projector onto $\mathcal{H}_m$. It is simple to verify (Rudi et al., 2015, Theorem 2) that $P_m = \tilde{Z}_m^* \mathbf{K}_{mm}^{-1} \tilde{Z}_m$ and that $(\hat{\lambda}_{i,m}, \hat{\phi}_{i,m})$ are the orthonormal eigenfunctions of $P_m C_n P_m$, i.e., $P_m C_n P_m \hat{\phi}_{i,m} = \hat{\lambda}_{i,m}\hat{\phi}_{i,m}$ for all $i \in [m]$. Therefore, we may think of $P_m C_n P_m$ as a low-rank approximation to C_n.

2.2.1 Approximate leverage scores

In the above discussion on Nyström KPCA, $\tilde{\mathbf{X}} := \{\tilde{X}_1, \ldots, \tilde{X}_m\}$ is a subset of the training set $\mathbf{X} := \{X_1, \ldots, X_n\}$ with the entries of $\tilde{\mathbf{X}}$ being sampled

uniformly without repetition from $\mathbf{X}$. As an alternative to uniform sampling, $\tilde{\mathbf{X}}$ can be sampled according to the leverage score distribution (Alaoui and Mahoney, 2015; Drineas et al., 2012; Cohen et al., 2015). For any $s > 0$, the leverage scores associated with the training data $\mathbf{X}$ are defined as

$$(l_i(s))_{i=1}^n, \quad l_i(s) = [\mathbf{K}(\mathbf{K} + ns\mathbf{I}_n)^{-1}]_{ii}, i \in [n]$$

with the leverage score distribution being $p_i(s) = \frac{l_i(s)}{\sum_{i=1}^n l_i(s)}$ according to which $\mathbf{X}$ can be sampled independently with replacement to achieve $\tilde{\mathbf{X}}$. Since the leverage scores are computationally intensive to compute, usually, they are approximated and one such approximation is T-approximate leverage scores (Rudi et al., 2018; Cohen et al., 2015).

Definition 1. *(T-approximate leverage scores) For a given $s > 0$, let $(l_i(s))_{i=1}^n$ be the leverage scores associated with the training data $\{X_1, ..., X_n\}$. Let $\delta > 0$, $s_0 > 0$, and $T \geq 1$. $(\hat{l}_i(s))_{i=1}^n$ are T-approximate leverage scores, with confidence δ, if the following holds with probability at least $1 - \delta$:*

$$T^{-1} l_i(s) \leq \hat{l}_i(s) \leq T l_i(s), \quad \forall i \in [n], \quad s > s_0.$$

Given T-approximate leverage scores for $s > s_0$, $\tilde{\mathbf{X}}$ can be obtained by sampling $\mathbf{X}$ with replacement according to the sampling distribution $\hat{p}_i(s) = \hat{l}_i(s) / \sum_{i=1}^n \hat{l}_i(s)$. Having obtained $\tilde{\mathbf{X}}$, (4) can be solved exactly as in Proposition 1. We refer to this method as approximate leverage score (ALS) Nyström subsampling.

3 Computational vs. Statistical Trade-Off: Main Results

As shown in the earlier section, Nyström kernel PCA approximates the solution to empirical kernel PCA with less computational expense. In this section, we explore whether this computational saving is obtained at the expense of statistical performance. As in Sriperumbudur and Sterge (2018), we measure the statistical performance of KPCA, EKPCA, and NY-KPCA in terms of reconstruction error. In linear PCA, the reconstruction error, given by

$$\mathbb{E}_{X \sim \mathbb{P}} \left\| \left(I - P^\ell(C)\right) X \right\|_2^2, \tag{7}$$

which is the error involved in reconstructing a random variable X by projecting it onto the ℓ-eigenspace (i.e., span of the top-ℓ eigenvectors) associated with its covariance matrix, $C = \mathbb{E}[XX^\top]$ through the orthogonal projection operator $P^\ell(C)$. Clearly, the error is zero when $\ell = d$. The analog of the reconstruction error in KPCA, as well as EKPCA and NY-KPCA, can be similarly stated in terms of their projection operators, (1), (3), and (6) as follows. For any orthogonal projection operator $P : \mathcal{H} \to \mathcal{H}$, define the reconstruction error as

$$R(P) := \mathbb{E}_{X \sim \mathbb{P}} \left\| (I - P)\, k(\cdot, X) \right\|_{\mathcal{H}}^2.$$

For the linear kernel, this is exactly the reconstruction error of PCA. The following lemma, proved in Section A.1, presents an alternate expression for $R(P)$ based on which, we define the reconstruction error in KPCA, EKPCA and NY-KPCA as

$$R_{C,\ell} := R(P^\ell(C)), \; R_{C_n,\ell} := R(P^\ell(C_n)),$$
$$\text{and } R^{nys}_{C_n,\ell} := R(P^\ell_m(C_n)), \tag{8}$$

respectively.

Lemma 1. $R(P) = \|(I - P)C^{1/2}\|^2_{\mathcal{L}^2(\mathcal{H})}$.

The following theorem provides finite-sample bounds on the reconstruction error associated with NY-KPCA, under both uniform and approximate leverage score subsampling, from which convergence rates may be obtained.

Theorem 2. *Suppose Assumptions 1-4 hold. For any $t > 0$, define $\mathcal{N}_C(t) = \operatorname{tr}((C + tI)^{-1}C)$ and $\mathcal{N}_{C,\infty}(t) = \sup_{x \in \mathcal{X}} \langle k(\cdot, x), (C + tI)^{-1} k(\cdot, x) \rangle_{\mathcal{H}}$. Then the following hold:*

(i) Suppose $n > 3$, $0 < \delta < 1$, $\frac{9\kappa}{n} \log \frac{n}{\delta} \leq t \leq \lambda_1$, and $m \geq (67 \vee 5\mathcal{N}_{C,\infty}(t)) \log \frac{4\kappa}{t\delta}$. Then, for plain Nyström subsampling:

$$\mathbb{P}^n \left\{ (X_i)_{i=1}^n : R^{nys}_{C_n,\ell} \leq \mathcal{N}_C(t) \left(6\lambda_\ell + 42t\right) \right\} \geq 1 - 2\delta. \tag{9}$$

(ii) For $0 < \delta < 1$, suppose there exists $T \geq 1$ such that $(\hat{l}_i(t))_{i=1}^n$ are $T-$approximate leverage scores with confidence δ for any $\frac{19\kappa}{n} \log \frac{2n}{\delta} \leq t \leq \lambda_1$. Assume approximate leverage score Nyström subsampling is used with $m \geq (78T^2\mathcal{N}_C(t) \vee 334) \log \frac{8n}{\delta}$, and $n \geq 1655\kappa + 223\kappa \log \frac{2\kappa}{\delta}$. Then

$$\mathbb{P}^n \left\{ (X_i)_{i=1}^n : R^{nys}_{C_n,\ell} \leq \mathcal{N}_C(t) \left(6\lambda_\ell + 42t\right) \right\} \geq 1 - 3\delta. \tag{10}$$

Proof. (i) For ease of notation, we will let $C_{n,t} = C_n + tI$, $\|\cdot\|_{\mathcal{L}^2(\mathcal{H})} = \|\cdot\|_2$, and $\|\cdot\|_{\mathcal{L}^\infty(\mathcal{H})} = \|\cdot\|_\infty$. For $t > 0$, we have

$$\begin{aligned} R^{nys}_{C_n,\ell} &= \left\| (I - P^\ell_m(C_n))C^{1/2} \right\|_2^2 \\ &= \left\| (I - P^\ell_m(C_n))C^{1/2}_{n,t} C^{-1/2}_{n,t} C^{1/2} \right\|_2^2 \\ &\leq A \cdot B, \end{aligned}$$

where $A = \left\|(I - P_m^\ell(C_n))C_{n,t}^{1/2}\right\|_\infty^2$ and $B = \left\|C_{n,t}^{-1/2}C^{1/2}\right\|_2^2$. First, we have

$$\begin{aligned}
B &= \left\|C_{n,t}^{-1/2}C^{1/2}\right\|_2^2 \\
&= \left\|C_{n,t}^{-1/2}(C+tI)^{1/2}(C+tI)^{-1/2}C^{1/2}\right\|_2^2 \\
&\le \left\|C_{n,t}^{-1/2}(C+tI)^{1/2}\right\|_\infty^2 \left\|(C+tI)^{-1/2}C^{1/2}\right\|_2^2 \\
&= \left\|C_{n,t}^{-1/2}(C+tI)^{1/2}\right\|_\infty^2 \mathcal{N}_C(t), \qquad (11)
\end{aligned}$$

where we used the fact $\left\|(C+tI)^{-1/2}C^{1/2}\right\|_2^2 = \operatorname{tr}(C^{1/2}(C+tI)^{-1}C^{1/2}) = \operatorname{tr}((C+tI)^{-1}C) =: \mathcal{N}_C(t)$. Next, we have

$$\begin{aligned}
A = \left\|(I - P_m^\ell(C_n))C_{n,t}^{1/2}\right\|_\infty^2 &\le 2\underbrace{\left\|(I-P_m)C_{n,t}^{1/2}\right\|_\infty^2}_{A_1} \\
&+2\underbrace{\left\|(P_m - P_m^\ell(C_n))C_{n,t}^{1/2}\right\|_\infty^2}_{A_2}, \qquad (12)
\end{aligned}$$

where $P_m = Z_m^*(\mathbf{K}_{mm})^{-1}Z_m$ is the orthogonal projector onto $\mathcal{H}_m$ (see Section 2.2). A_1 can be bounded as

$$A_1 \le D_1 \cdot D_2, \qquad (13)$$

where $D_1 = \left\|(I-P_m)(C+tI)^{1/2}\right\|_\infty^2$ and $D_2 = \left\|(C+tI)^{-1/2}C_{n,t}^{1/2}\right\|_\infty^2$. A_2 is bounded as

$$\begin{aligned}
A_2 &\overset{(*)}{=} \left\|(I - P_m^\ell(C_n))P_m C_{n,t}^{1/2}\right\|_\infty^2 \\
&= \left\|(I - P_m^\ell(C_n))P_m C_{n,t} P_m (I - P_m^\ell(C_n))\right\|_\infty \\
&\le \left\|(I - P_m^\ell(C_n))P_m C_n P_m (I - P_m^\ell(C_n))\right\|_\infty \\
&\quad + t\left\|(I - P_m^\ell(C_n))P_m (I - P_m^\ell(C_n))\right\|_\infty, \\
&\overset{(**)}{\le} \hat{\lambda}_{\ell+1,m} + t, \qquad (14)
\end{aligned}$$

where we used the facts that $\mathcal{R}(P_m^\ell(C_n)) \subset \mathcal{R}(P_m)$ in $(*)$ and $P_m^\ell(C_n)$ projects onto the ℓ-eigenspace of $P_m C_n P_m$ in $(**)$. Here $\mathcal{R}(A)$ denotes the range of operator A. $\hat{\lambda}_{\ell+1,m}$ can be bounded as

$$\begin{aligned}
\hat{\lambda}_{\ell+1,m} &\le |\hat{\lambda}_{\ell+1,m} - \hat{\lambda}_{\ell+1}| + \hat{\lambda}_{\ell+1} \\
&\overset{(\dagger)}{\le} \frac{1}{n}\left\|\tilde{\mathbf{K}} - \mathbf{K}\right\|_{\mathcal{L}^\infty(\mathbb{R}^n)} + \hat{\lambda}_\ell, \qquad (15)
\end{aligned}$$

where $(\dagger)$ follows from the Hoffman-Wiedlandt inequality (R. Bhatia, 1994). We may rewrite (15) as

$$\begin{aligned}
&\frac{1}{n}\left\|\tilde{\mathbf{K}} - \mathbf{K}\right\|_{\mathcal{L}^\infty(\mathbb{R}^n)} = \frac{1}{n}\left\|Z_n(I-P_m)Z_n^*\right\|_{\mathcal{L}^\infty(\mathbb{R}^n)} \\
&= \left\|(I-P_m)C_n(I-P_m)\right\|_\infty \\
&= \left\|C_n^{1/2}(I-P_m)C_n^{1/2}\right\|_\infty \\
&\le \left\|C_n^{1/2}(C+tI)^{-1/2}\right\|_\infty^2 \left\|(C+tI)^{1/2}(I-P_m)\right\|_\infty^2 \\
&\overset{(\ddagger)}{\le} \left\|C_{n,t}^{1/2}(C+tI)^{-1/2}\right\|_\infty^2 \left\|(C+tI)^{1/2}(I-P_m)\right\|_\infty^2, \qquad (16)
\end{aligned}$$

where we used

$$\begin{aligned}
&\left\|C_n^{1/2}(C+tI)^{-1/2}\right\|_\infty^2 \\
&\quad\le \left\|C_n^{1/2}C_{n,t}^{-1/2}\right\|_\infty^2 \left\|C_{n,t}^{1/2}(C+tI)^{-1/2}\right\|_\infty^2
\end{aligned}$$

and $\left\|C_n^{1/2}C_{n,t}^{-1/2}\right\|_\infty^2 \le 1$ in $(\ddagger)$. The result follows by combining (11)–(16) and employing Lemmas A.1 and A.2.

(ii) The proof follows exactly as in *(i)*; however, we bound $\left\|(I-P_m)(C+tI)^{1/2}\right\|_{\mathcal{L}^\infty(\mathcal{H})}^2$ with Lemma A.3 and $t = \frac{19\kappa}{n}\log\frac{2n}{\delta}$. □

Remark 1. Theorem 2 can be presented in terms of the empirical effective dimension $d_{\text{eff}} = \operatorname{tr}\left(\mathbf{K}(\mathbf{K}+nt\mathbf{I}_n)^{-1}\right)$ rather than its population counterpart, $\mathcal{N}_C(t)$, as in Alaoui and Mahoney (2015), by noting that $\mathcal{N}_C(t) \lesssim_{\mathbb{P}^n} d_{\text{eff}} + \frac{1}{nt}$. However, it is beneficial to present in terms of $\mathcal{N}_C(t)$ as it fosters easy comparison of the empirical and approximate reconstruction errors to the population error.

To understand the significance of Theorem 2, we have to compare it to the behavior of the reconstruction error associated with EKPCA, i.e., $R_{C_n,\ell}$. Rudi et al. (2015, Theorem 3.1) showed that for $n > 3$, $0 < \delta < 1$ and $\frac{9\kappa}{n}\log\frac{n}{\delta} \le t \le \lambda_1$,

$$\mathbb{P}^n\left\{(X_i)_{i=1}^n : R_{C_n,\ell} \le 9\mathcal{N}_C(t)\,(\lambda_\ell + t)\right\} \ge 1 - \delta. \qquad (17)$$

Comparing (9) and (10) to (17), it is clear that NY-KPCA has a statistical behavior similar to that EKPCA, with the bounds differing only up to constants. However, it is not obvious whether such a behavior is achieved for $m < n$, i.e., the order of dependence of m on n is not clear. To clarify this, in the following, we present two corollaries to Theorem 2, which compare the asymptotic convergence rates of $R_{C,\ell}$, $R_{C_n,\ell}$ and $R_{C_n,\ell}^{nys}$ under an additional assumption on the decay rate of eigenvalues of C.

Corollary 3 (Polynomial decay of eigenvalues)**.** *Suppose* $\underline{A}i^{-\alpha} \leq \lambda_i \leq \bar{A}i^{-\alpha}$ *for* $\alpha > 1$ *and* $\underline{A}, \bar{A} \in (0,\infty)$*. Let* $\ell = n^{\frac{\theta}{\alpha}}$*,* $\theta > 0$*. Then the following hold:*

(i)
$$n^{-\theta(1-\frac{1}{\alpha})} \lesssim R_{C,\ell} \lesssim n^{-\theta(1-\frac{1}{\alpha})};$$

(ii)
$$R_{C_n,\ell} \lesssim_{\mathbb{P}^n} \begin{cases} n^{-\theta(1-\frac{1}{\alpha})}, & \theta < 1 \\ \left(\frac{\log n}{n}\right)^{1-\frac{1}{\alpha}}, & \theta \geq 1 \end{cases};$$

(iii) *For plain Nyström subsampling:*

$$R^{nys}_{C_n,\ell} \lesssim_{\mathbb{P}^n} \begin{cases} n^{-\theta(1-\frac{1}{\alpha})},\ \theta < 1,\ m \gtrsim n^{\theta}\log n \\ \left(\frac{\log n}{n}\right)^{1-\frac{1}{\alpha}},\ \theta \geq 1,\ m \gtrsim \frac{n}{\log n}\log\frac{n}{\log n} \end{cases};$$

(iv) *For approximate leverage score Nyström subsampling:*

$$R^{nys}_{C_n,\ell} \lesssim_{\mathbb{P}^n} \begin{cases} n^{-\theta(1-\frac{1}{\alpha})},\ \theta < 1,\ m \gtrsim n^{\frac{\theta}{\alpha}}\log n \\ \left(\frac{\log n}{n}\right)^{1-\frac{1}{\alpha}},\ \theta \geq 1,\ m \gtrsim \frac{n^{\frac{1}{\alpha}}}{(\log n)^{\frac{1}{\alpha}-1}} \end{cases}.$$

Proof. (i) From Theorem 2 (i) we have $R_{C,\ell} = \sum_{i>\ell}\lambda_i \lesssim \sum_{i>\ell} i^{-\alpha} \lesssim \int_{\ell}^{\infty} x^{-\alpha}dx \lesssim \ell^{1-\alpha} = n^{-\theta(1-\frac{1}{\alpha})}$. Similarly, $R_{C,\ell} = \sum_{i>\ell}\lambda_i \gtrsim \sum_{i>\ell} i^{-\alpha} \gtrsim \int_{\ell}^{\infty} x^{-\alpha}dx \gtrsim \ell^{1-\alpha} = n^{-\theta(1-\frac{1}{\alpha})}$.

(ii) This is Theorem 3.2 of (Rudi et al., 2015) with $\alpha = \frac{1}{2}$, $r = \alpha$, $p = 2$, and $\ell = n^{\frac{\theta}{\alpha}}$.

(iii) Theorem 2 (iii) and Proposition B.1 yield

$$R^{nys}_{C_n,\ell} \lesssim_{\mathbb{P}^n} t^{-\frac{1}{\alpha}}n^{-\theta} + t^{1-\frac{1}{\alpha}} \leq \begin{cases} t^{1-\frac{1}{\alpha}}, t \geq n^{-\theta} \\ t^{-\frac{1}{\alpha}}n^{-\theta}, t \leq n^{-\theta} \end{cases},$$

where $\frac{\log n}{n} \lesssim t \leq \lambda_1$ and $m \gtrsim \mathcal{N}_{C,\infty}(t)\log\frac{1}{t}$ with $\mathcal{N}_{C,\infty}(t) = \sup_{x\in\mathcal{X}}\left\langle k(\cdot,x), (C+tI)^{-1}k(\cdot,x)\right\rangle_{\mathcal{H}} \lesssim \frac{1}{t}$.

First, consider the case when $t \geq n^{-\theta}$. This means

$$R^{nys}_{C_n,\ell} \lesssim \inf\left\{t^{1-\frac{1}{\alpha}} : t \gtrsim \frac{\log n}{n} \vee n^{-\theta}, m \gtrsim \frac{1}{t}\log\frac{1}{t}\right\}.$$

For $\theta < 1$, we obtain $R^{nys}_{C_n,\ell} \lesssim \inf\left\{t^{1-\frac{1}{\alpha}} : t \gtrsim n^{-\theta}, m \gtrsim \frac{1}{t}\log\frac{1}{t}\right\} \leq n^{-\theta(1-\frac{1}{\alpha})}$ if $m \gtrsim n^{\theta}\log n$. For $\theta \geq 1$, we obtain $R^{nys}_{C_n,\ell} \lesssim \inf\left\{t^{1-\frac{1}{\alpha}} : t \gtrsim \frac{\log n}{n}, m \gtrsim \frac{1}{t}\log\frac{1}{t}\right\} \leq \left(\frac{\log n}{n}\right)^{(1-\frac{1}{\alpha})}$ if $m \gtrsim \frac{n}{\log n}\log\frac{n}{\log n}$.

Next, consider the case when $t \leq n^{-\theta}$ which means $R^{nys}_{C_n,\ell} \lesssim \inf\left\{t^{-\frac{1}{\alpha}}n^{-\theta} : \frac{\log n}{n} \lesssim t \lesssim n^{-\theta}, m \gtrsim \frac{1}{t}\log\frac{1}{t}\right\} \leq n^{-\theta(1-\frac{1}{\alpha})}$ when $\theta < 1$ and $m \gtrsim n^{\theta}\log n$.

(iv) Theorem 2(iv) and Proposition B.1 yield

$$R^{nys}_{C_n,\ell} \lesssim_{\mathbb{P}^n} t^{-\frac{1}{\alpha}}n^{-\theta} + t^{1-\frac{1}{\alpha}} \leq \begin{cases} t^{1-\frac{1}{\alpha}}, t \geq n^{-\theta} \\ t^{-\frac{1}{\alpha}}n^{-\theta}, t \leq n^{-\theta} \end{cases},$$

where $\frac{\log n}{n} \lesssim t \leq \lambda_1$ and $m \gtrsim \mathcal{N}_C(t)\log n \gtrsim t^{-\frac{1}{\alpha}}\log n$. The result follows by carrying out the analysis as in (iii) for $\theta < 1$ and $\theta \geq 1$. □

Remark 2. (i) The above result shows that the reconstruction errors associated with KPCA and EKPCA have similar asymptotic behavior as long as ℓ does not grow to infinity too fast, i.e., $\theta < 1$. On the other hand, for $\theta \geq 1$, the reconstruction error of EKPCA has slower asymptotic convergence to zero than that of KPCA. If ℓ grows to infinity faster with the rate controlled by θ, then the variance term dominates the bias resulting in a slower convergence rate compared to that of KPCA.

(ii) Comparing (ii) and (iii) in the above result, we note that EKPCA and NY-KPCA have similar convergence behavior as long as m is large enough where the size of m is controlled by the growth of ℓ through θ. For the case of $\theta \geq 1$ in (iii), we require $m \gtrsim \frac{n}{\log n}\log\frac{n}{\log n}$ which means asymptotically m should be of the same order as n. On the other hand, the approximate leverage score Nyström subsampling gives same convergence rates as that of EKPCA but requiring far fewer samples than that for NY-KPCA with plain Nyström subsampling. These results show that for the interesting case of $\theta < 1$ where EKPCA performance matches with that of KPCA, NY-KPCA also achieves similar performance, albeit with lower computational requirement.

Corollary 4 (Exponential decay of eigenvalues)**.** *Suppose* $\underline{B}e^{-\tau i} \leq \lambda_i \leq \bar{B}e^{-\tau i}$ *for* $\tau > 0$ *and* $\underline{B}, \bar{B} \in (0,\infty)$*. Let* $\ell = \frac{1}{\tau}\log n^{\theta}$ *for* $\theta > 0$*. Then the following hold:*

(i)
$$n^{-\theta} \lesssim R_{C,\ell} \lesssim n^{-\theta};$$

(ii)
$$R_{C_n,\ell} \lesssim_{\mathbb{P}^n} \begin{cases} \frac{\log n}{n^{\theta}}, & \theta < 1 \\ \frac{(\log n)^2}{n}, & \theta \geq 1 \end{cases};$$

(iii) *For plain Nyström subsampling:*

$$R^{nys}_{C_n,\ell} \lesssim_{\mathbb{P}^n} \begin{cases} \frac{\log n}{n^{\theta}},\ \theta < 1,\ m \gtrsim n^{\theta}\log n \\ \frac{(\log n)^2}{n},\ \theta \geq 1,\ m \gtrsim \frac{n}{\log n}\log\frac{n}{\log n} \end{cases};$$

(*iv*) *For approximate leverage score Nyström subsampling:*

$$R^{nys}_{C_n,\ell} \lesssim_{\mathbb{P}^n} \begin{cases} \frac{\log n}{n^\theta}, \ \theta < 1, \ m \gtrsim (\log n)^2 \\ \frac{(\log n)^2}{n}, \ \theta \geq 1, \ m \gtrsim \log n \log \frac{n}{\log n} \end{cases} .$$

Proof. (*i*) From Theorem 2 (*i*) we have $R_{C,\ell} = \sum_{i>\ell} \lambda_i \lesssim \sum_{i>\ell} e^{-\tau i} \lesssim \int_\ell^\infty e^{-\tau x} dx \lesssim e^{-\tau \ell} = n^{-\theta}$ and $R_{C,\ell} = \sum_{i>\ell} \lambda_i \gtrsim \sum_{i>\ell} e^{-\tau i} \gtrsim \int_{\ell+1}^\infty e^{-\tau x} dx \gtrsim e^{-\tau(\ell+1)} = e^{-\tau} n^{-\theta}$.

(*ii*) Theorem 2 (*ii*) and Proposition B.2 yield

$$R_{C_n,\ell} \lesssim_{\mathbb{P}^n} \left(n^{-\theta} + t\right) \log \frac{1}{t} \leq \begin{cases} n^{-\theta} \log \frac{1}{t}, t \leq n^{-\theta} \\ t \log \frac{1}{t}, t \geq n^{-\theta} \end{cases} ,$$

where $\frac{\log n}{n} \lesssim t \leq \lambda_1$.

For the case of $t \leq n^{-\theta}$, we obtain $R_{C_n,\ell} \lesssim \inf \left\{ n^{-\theta} \log \frac{1}{t} : \frac{\log n}{n} \lesssim t \leq n^{-\theta} \right\} = n^{-\theta} \log n$, where the constraint is only valid for $\theta < 1$.

On the other hand, for $t \geq n^{-\theta}$, we obtain $R_{C_n,\ell} \lesssim \inf \left\{ t \log \frac{1}{t} : t \gtrsim \frac{\log n}{n} \vee n^{-\theta} \right\} = \frac{\log n}{n} \log \left(\frac{n}{\log n} \right) \leq \frac{(\log n)^2}{n}$, which holds for $\theta \geq 1$.

(*iii*) Arguing similarly as in (*ii*), it follows that for $\theta < 1$ and $m \gtrsim n^\theta \log n$, we obtain a rate of $n^{-\theta} \log n$ for $R^{nys}_{C_n,\ell}$. Similarly for $\theta \geq 1$ and $m \geq \frac{n}{\log n} \log \left(\frac{n}{\log n} \right)$, we obtain a rate of $n^{-1} (\log n)^2$.

(*iv*) Arguing as in (*ii*) and enforcing the restriction $m \gtrsim \log n \log \frac{1}{t}$ imposed by Theorem 2 (*ii*) yields the result. □

Corollary 4 shares similar behavior to that Corollary 3 as discussed in Remark 2 but just that it yields faster rates since the RKHS is smooth as determined by the rate of decay of eigenvalues. In addition, the approximate leverage score Nyström subsampling based KPCA requires only $(\log n)^2$ subsamples to match the performance of EKPCA resulting in substantial computational savings without any loss in statistical accuracy.

Comparison to random feature approximation. Sriperumbudur and Sterge (2018); Ullah et al. (2018) studied the question of computational vs. statistical tradeoff in kernel PCA using random feature approximation (Rahimi and Recht, 2008). However, our results for Nyström approximation are not directly comparable to theirs. These works considered a reconstruction error defined in (8) through Lemma 1, however, in the $L^2(\mathbb{P})$ norm, which is weaker than the RKHS norm. Using the $L^2(\mathbb{P})$ norm is needed for random feature approximation based KPCA as the random features in general do not lie in $\mathcal{H}$, while KPCA and EKPCA generate eigenfunctions in $\mathcal{H}$, which makes comparison infeasible. This was addressed by embedding all of them in $L^2(\mathbb{P})$ and comparing their behavior in $L^2(\mathbb{P})$ norm—for classical PCA this would correspond to considering the error $\mathbb{E}[(X^\top (I - P^\ell(C))X)^2]$ rather than (7). On the other hand, the eigenfunctions of NY-KPCA lie in $\mathcal{H}$ making comparison to KPCA and EKPCA feasible in $\mathcal{H}$ norm, which is what did in this work. For a direct comparison with random features, we have study the reconstruction errors in (8) in $L^2(\mathbb{P})$, which will be a focus of our future work.

4 Conclusions & Further Work

In this paper, we considered the problem of analyzing the approximation of kernel PCA using Nyström method. The Nyström approximation avoids some of the technical difficulties associated with random features and allows to derive statistical error estimates that are directly comparable to those of KPCA. Our results indicate there are regimes where computational gains can be achieved while preserving statistical accuracy. These results parallel recent findings in supervised learning and extend them to unsupervised learning.

Our study opens a number of possible questions. For example, still for KPCA, it would be interesting to understand the properties of Nyström sampling in combination with iterative eigensolvers, both batch (e.g., the power method) and stochastic (e.g., Oja's rule). The application of our approach to other spectral methods, such as those used in graph and manifold learning, would be interesting. Beyond PCA and spectral methods, our study naturally yields the question of which other learning problems can have analogous statistical and computational trade-offs, e.g., independence tests based on covariance and cross-covariance operators (Gretton et al., 2008), or mean embeddings (Sriperumbudur et al., 2010).

References

Alaoui, A. and Mahoney, M. (2015). Fast randomized kernel ridge regression with statistical guarantees. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 775–783. Curran Associates, Inc.

Alaoui, A. and Mahoney, M. W. (2014). Fast ran-

domized kernel methods with statistical guarantees. *arXiv*.

Bach, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In Shalev-Shwartz, S. and Steinwart, I., editors, *Proc. of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 185–209. PMLR.

Blanchard, G., Bousquet, O., and Zwald, L. (2007). Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2):259–294.

Bottou, L. and Bousquet, O. (2008). The tradeoffs of large scale learning. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20, pages 161–168. Curran Associates, Inc.

Calandriello, D., Lazaric, A., and Valko, M. (2018). Distributed adaptive sampling for kernel matrix approximation. *CoRR*, abs/1803.10172.

Cohen, M. B., lee, Y. T., Musco, C., Musco, C., Peng, R., and Sidford, A. (2015). Uniform sampling for matrix approximation. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 181–190. ACM.

Drineas, P., Magdon-Ismail, M., Mahoney, M. W., and Woodruff, D. P. (2012). Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506.

Drineas, P. and Mahoney, M. W. (2005). On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175.

Gittens, A. and Mahoney, M. (2013). Revisiting the Nyström method for improved large-scale machine learning. In Dasgupta, S. and McAllester, D., editors, *Proc. of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 567–575. PMLR.

Gohberg, I., Goldberg, S., and Kaashoek, R. (2003). *Basic Classes of Linear Operators*. Birkhauser, Basel, Switzerland.

Gretton, A., Fukumizu, K., Teo, C.-H., Song, L., Schölkopf, B., and Smola, A. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592. MIT Press.

Jin, R., Yang, T., Mahdavi, M., Li, Y.-F., and Zhou, Z.-H. (2013). Improved bounds for the Nyström method with application to kernel classification. *IEEE Transactions on Information Theory*, 59(10):6939–6949.

Jolliffe, I. (1986). *Principal Component Analysis*. Springer-Verlag, New York, USA.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95.

Musco, C. and Musco, C. (2017). Recursive sampling for the Nyström method. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 3833–3845. Curran Associates, Inc.

Orabona, F., Keshet, J., and Caputo, B. (2008). The projectron: A bounded kernel-based perceptron. In Cohen, W. W., McCallum, A., and Roweis, S. T., editors, *ICML*, volume 307 of *ACM International Conference Proceeding Series*, pages 720–727. ACM.

R. Bhatia, L. E. (1994). The Hoffman-Wielandt inequality in infinite dimensions. In *Proceedings of the Indian Academy of Sciences - Mathematical Sciences*, volume 104, pages 483–494.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc.

Reed, M. and Simon, B. (1980). *Methods of Modern Mathematical Physics: Functional Analysis I*. Academic Press, New York.

Rudi, A., Calandriello, D., Carratino, L., and Rosasco, L. (2018). On fast leverage score sampling and optimal learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 5672–5682. Curran Associates, Inc.

Rudi, A., Camoriano, R., and Rosasco, L. (2015). Less is more: Nyström computational regularization. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 1657–1665. Curran Associates, Inc.

Rudi, A., Canas, G. D., and Rosasco, L. (2013). On the sample complexity of subspace learning. In *Advances in Neural Information Processing Systems 26*, pages 2067–2075.

Rudi, A. and Rosasco, L. (2017). Generalization properties of learning with random features. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 3215–3225. Curran Associates, Inc.

Schölkopf, B., Herbrich, R., and Smola, A. (2001). A generalized representer theorem. In *Proc. of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory*, pages 416–426, London, UK. Springer-Verlag.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319.

Shawe-Taylor, J., Williams, C., Christianini, N., and Kandola, J. (2005). On the eigenspectrum of the Gram matrix and the generalisation error of kernel PCA. *EEE Transactions on Information Theory*, 51(7):2510–2522.

Smale, S. and Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. G. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561.

Sriperumbudur, B. K. and Sterge, N. (2018). Approximate kernel PCA using random features: Computational vs. statistical trade-off. *arXiv*.

Ullah, M. E., Mianjy, P., Marinov, T. V., and Arora, R. (2018). Streaming kernel PCA with $\tilde{O}(\sqrt{n})$ random features. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 7322–7332. Curran Associates, Inc.

Williams, C. and Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In T. K. Leen, T. G. Diettrich, V. T., editor, *Advances in Neural Information Processing Systems 13*, pages 682–688, Cambridge, MA. MIT Press.

Zhang, K., Tsang, I. W., and Kwok, J. T. (2008). Improved Nyström low-rank approximation and error analysis. In *Proceedings of the 25th international conference on Machine learning*, pages 1232–1239. ACM.