Nyström Method with Kernel K-means++ Samples as Landmarks

Dino Oglic[1 2] **Thomas Gärtner**[2]

Abstract

We investigate, theoretically and empirically, the effectiveness of kernel K-means++ samples as landmarks in the Nyström method for low-rank approximation of kernel matrices. Previous empirical studies (Zhang et al., 2008; Kumar et al., 2012) observe that the landmarks obtained using (kernel) K-means clustering define a good low-rank approximation of kernel matrices. However, the existing work does not provide a theoretical guarantee on the approximation error for this approach to landmark selection. We close this gap and provide the first bound on the approximation error of the Nyström method with kernel K-means++ samples as landmarks. Moreover, for the frequently used Gaussian kernel we provide a theoretically sound motivation for performing Lloyd refinements of kernel K-means++ landmarks in the instance space. We substantiate our theoretical results empirically by comparing the approach to several state-of-the-art algorithms.

1. Introduction

We consider the problem of finding a good low-rank approximation for a given symmetric and positive definite matrix. Such matrices arise in kernel methods (Schölkopf & Smola, 2001) where the data is often first transformed to a symmetric and positive definite matrix and then an off-the-shelf matrix-based algorithm is used for solving classification and regression problems, clustering, anomaly detection, and dimensionality reduction (Bach & Jordan, 2005). These learning problems can often be posed as convex optimization problems for which the representer theorem (Wahba, 1990) guarantees that the optimal solution can be found in the subspace of the kernel feature space spanned by the instances. To find the optimal solution in a problem with n instances, it is often required to perform a matrix inversion or eigendecomposition which scale as $\mathcal{O}\left(n^3\right)$. To overcome this computational shortcoming and scale kernel methods to large scale datasets, Williams & Seeger (2001) have proposed to use a variant of the Nyström method (Nyström, 1930) for low-rank approximation of kernel matrices. The approach is motivated by the fact that frequently used kernels have a fast decaying spectrum and that small eigenvalues can be removed without a significant effect on the precision (Schölkopf & Smola, 2001). For a given sub-set of l landmarks, the Nyström method finds a low-rank approximation in time $\mathcal{O}\left(l^2 n + l^3\right)$ and kernel methods with the low-rank approximation in place of the kernel matrix scale as $\mathcal{O}\left(l^3\right)$. In practice, $l \ll n$ and the approach can scale kernel methods to millions of instances.

The crucial step in the Nyström approximation of a symmetric and positive definite matrix is the choice of landmarks and an optimal choice is a difficult discrete/combinatorial problem directly influencing the goodness of the approximation (Section 2). A large part of the existing work has, therefore, focused on providing approximation guarantees for different landmark selection strategies. Following this line of research, we propose to select landmarks using the kernel K-means++ sampling scheme (Arthur & Vassilvitskii, 2007) and provide the first bound on the relative approximation error in the Frobenius norm for this strategy (Section 3). An important part of our theoretical contribution is the first complete proof of a claim by Ding & He (2004) on the relation between the subspace spanned by optimal K-means centroids and left singular vectors of the feature space (Proposition 1). While our proof covers the general case, that of Ding & He (2004) is restricted to data matrices with piecewise constant right singular vectors.

Having given a bound on the approximation error for the proposed landmark selection strategy, we provide a brief overview of the existing landmark selection algorithms and discuss our work in relation to approaches directly comparable to ours (Section 4). For the frequently used Gaussian kernel, we also theoretically motivate the instance space Lloyd refinements (Lloyd, 1982) of kernel K-means++ landmarks. The results of our empirical study are presented in Section 5 and indicate a superior performance of the proposed approach over competing methods. This is in agreement with the previous studies on K-means centroids as landmarks by Zhang et al. (2008) and Kumar et al. (2012).

[1] Institut für Informatik III, Universität Bonn, Germany [2] School of Computer Science, The University of Nottingham, United Kingdom. Correspondence to: Dino Oglic <dino.oglic@uni-bonn.de>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

2. Nyström Method

In this section, we review the Nyström method for low-rank approximation of kernel matrices. The method was originally proposed for the approximation of integral eigenfunctions (Nyström, 1930) and later adopted to low-rank approximation of kernel matrices by Williams & Seeger (2001). We present it here in a slightly different light by following the approach to subspace approximations by Smola & Schölkopf (2000, 2001).

Let $\mathcal{X}$ be an instance space and $X = \{x_1, x_2, \cdots, x_n\}$ an independent sample from a Borel probability measure defined on $\mathcal{X}$. Let $\mathcal{H}$ be a reproducing kernel Hilbert space with a positive definite kernel $h\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Given a set of landmark points $Z = \{z_1, \cdots z_m\}$ (not necessarily a subset of the sample) the goal is to approximate kernel functions $h(x_i, \cdot)$ for all $i = \overline{1, n}$ using linear combinations of the landmarks. This goal can be formally stated as

$$\min_{\alpha \in \mathbb{R}^{m \times n}} \sum_{i=1}^{n} \left\| h(x_i, \cdot) - \sum_{j=1}^{m} \alpha_{j,i} h(z_j, \cdot) \right\|_{\mathcal{H}}^{2} . \qquad (1)$$

Let H denote the kernel matrix over all samples and landmarks and let H_Z denote the block in this matrix corresponding to the kernel values between the landmarks. Additionally, let h_x denote a vector with entries corresponding to the kernel values between an instance x and the landmarks. After expanding the norm, the problem is transformed into

$$\min_{\alpha \in \mathbb{R}^{m \times n}} \sum_{i=1}^{n} H_{ii} - 2h_{x_i}^{\top} \alpha_i + \alpha_i^{\top} H_Z \alpha_i , \qquad (2)$$

where α_i denotes the ith column of α. Each summand in the optimization objective is a convex function depending only on one column of α. Hence, the optimal solution is

$$\alpha = H_Z^{-1} H_{Z \times X} .$$

From here it then follows that the optimal approximation $\tilde{H}_{X|Z}$ of the matrix H_X using landmarks Z is given by

$$\tilde{H}_{X|Z} = H_{X \times Z} H_Z^{-1} H_{Z \times X} .$$

While the problem of computing the optimal projections of instances to a subspace spanned by the landmarks is convex and solvable in closed form (see above), the problem of choosing the best set of landmarks is a combinatorial problem that is difficult to solve. To evaluate the effectiveness of the subspace spanned by a given set of landmarks it is standard to use the Schatten matrix norms (Weidmann, 1980). The *Schatten p-norm* of a symmetric and positive definite matrix H is defined as $\|H\|_p = \left(\sum_{i=1}^{n} \lambda_i^p\right)^{1/p}$, where $\lambda_i \geq 0$ are eigenvalues of H and $p \geq 1$. For $p = \infty$ the Schatten p-norm is equal to the operator norm and for $p = 2$ it is equal to the Frobenius norm. The three most frequently used Schatten norms are $p = 1, 2, \infty$ and for these norms the following inequalities hold:

$$\|H\|_{\infty} = \max_i \lambda_i \leq \sqrt{\sum_i \lambda_i^2} = \sqrt{\operatorname{tr}\left(H^{\top} H\right)} = \|H\|_2$$
$$\leq \sum_i \lambda_i = \operatorname{tr}(H) = \|H\|_1 .$$

From Eq. (1) and (2) it follows that for a subspace spanned by a given set of landmarks Z, the 1-norm approximation error of the optimal projections onto this space is given by

$$L(\alpha^*) = \operatorname{tr}(H_X) - \operatorname{tr}(\tilde{H}_{X|Z}) = \left\| H_X - \tilde{H}_{X|Z} \right\|_1 .$$

The latter equation follows from the properties of trace and the fact that $\Xi = H_X - \tilde{H}_{X|Z}$ is a symmetric and positive definite matrix with $\Xi_{ij} = \langle \xi(x_i, \cdot), \xi(x_j, \cdot) \rangle_{\mathcal{H}}$ and $\xi(x_i, \cdot) = h(x_i, \cdot) - \sum_{k=1}^{m} \alpha_{k,i}^{*} h(z_k, \cdot)$.

For a good Nyström approximation of a kernel matrix it is crucial to select the landmarks to reduce the error in one of the frequently used Schatten p-norms, i.e.,

$$Z^* = \operatorname*{arg\,min}_{Z \subset \operatorname{span}(X),\ |Z| = K} \left\| H_X - \tilde{H}_{X|Z} \right\|_p .$$

Let us denote with V_K and Λ_K the top K eigenvectors and eigenvalues of the kernel matrix H_X. Then, at the low-rank approximation $\tilde{H}_{X|Z}^{*} = V_K \Lambda_K V_K^{\top}$, the Schatten p-norm error attains its minimal value (Golub & van Loan, 1996).

3. Kernel K-means++ Samples as Landmarks

We start with a review of K-means clustering (Lloyd, 1982) and then give the first complete proof of a claim stated in Ding & He (2004) and Xu et al. (2015) on the relation between the subspace spanned by the top $(K-1)$ left singular vectors of the data matrix and that spanned by optimal K-means centroids. Building on a result by Arthur & Vassilvitskii (2007) we then bound the relative approximation error in the Frobenius norm of the Nyström method with kernel K-means++ samples as landmarks.

Let the instance space $\mathcal{X} \subset \mathbb{R}^d$ and let K denote the number of clusters. In K-means clustering the goal is to choose a set of centers $C = \{c_1, \cdots, c_K\}$ minimizing the potential

$$\phi(C) = \sum_{x \in X} \min_{c \in C} \|x - c\|^2 = \sum_{k=1}^{K} \sum_{x \in \mathcal{P}_k} \|x - c_k\|^2 ,$$

where $\mathcal{P}_k = \{x \in X \mid \mathcal{P}(x) = c_k\}$ is a clustering cell and $\mathcal{P}\colon \mathcal{X} \to C$ denotes the centroid assignment function. For a clustering cell $\mathcal{P}_k$ the centroid is computed as $\frac{1}{|\mathcal{P}_k|} \sum_{x \in \mathcal{P}_k} x$. In the remainder of the section, we denote with $P \in \mathbb{R}^{n \times K}$ the *cluster indicator matrix* of the clustering C such that $p_{ij} = 1/\sqrt{n_j}$ when instance x_i is assigned to centroid c_j, and $p_{ij} = 0$ otherwise. Here n_j denotes the number of instances assigned to centroid c_j. Without loss of generality, we assume that the columns of the data matrix $X \in \mathbb{R}^{d \times n}$ are centered instances (i.e., $\sum_{i=1}^{n} x_i/n = 0$).

Now, using the introduced notation we can write the clustering potential as (Ding & He, 2004; Boutsidis et al., 2009)

$$\phi(C) = \left\| X - XPP^\top \right\|_2^2 .$$

Denoting with p_i the ith column in P we have that it holds $p_i^\top p_j = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and otherwise $\delta_{ij} = 0$. Hence, it holds that $P^\top P = \mathbb{I}_K$ and P is an orthogonal projection matrix with rank K. Let $\mathcal{C}$ denote the family of all possible clustering indicator matrices of rank K. Then, the K-means optimization problem is equivalent to the constrained low-rank approximation problem

$$P^* = \underset{P \in \mathcal{C}}{\arg\min} \ \left\| X - XPP^\top \right\|_2^2 .$$

From here, using the relation between the squared Schatten 2-norm and the matrix trace we obtain

$$P^* = \underset{P \in \mathcal{C}}{\arg\min} \ \operatorname{tr}\left(X^\top X\right) - \operatorname{tr}\left(P^\top X^\top X P\right) . \tag{3}$$

In the remainder of the section, we refer to the constrained optimization objective from Eq. (3) as the *discrete* problem. For this problem, Ding & He (2004) observe that the set of vectors $\{p_1, \cdots, p_K, \mathbf{e}/\sqrt{n}\}$ is linearly dependent ($\mathbf{e}$ is a vector of ones) and that the rank of the optimization problem can be reduced. As $\sum_{i=1}^{K} \sqrt{n_i} p_i = \mathbf{e}$, there exists a linear orthonormal transformation of the subspace basis given by the columns of P such that one of the vectors in the new basis of the subspace spanned by P is $\mathbf{e}/\sqrt{n}$. Such transformations are equivalent to a rotation of the subspace. Let $R \in \mathbb{R}^{K \times K}$ denote an orthonormal transformation matrix such that the vectors $\{p_i\}_{i=1}^{K}$ map to $\{q_i\}_{i=1}^{K}$ with $q_K = \frac{1}{\sqrt{n}}\mathbf{e}$. This is equivalent to requiring that the Kth column in R is $r_K = \left(\sqrt{\frac{n_1}{n}}, \cdots, \sqrt{\frac{n_K}{n}}\right)^\top$ and $q_i^\top \mathbf{e} = 0$ for $i = \overline{1, K-1}$. Moreover, from $Q = PR$ and $R^\top R = \mathbb{I}_K$ it follows that

$$Q^\top Q = R^\top P^\top P R = R^\top R = \mathbb{I}_K .$$

Hence, if we denote with Q_{K-1} the matrix-block with the first $(K-1)$ columns of Q then the problem from Eq. (3) can be written as (Ding & He, 2004; Xu et al., 2015)

$$\begin{aligned} Q_{K-1}^* = & \underset{Q_{K-1} \in \mathbb{R}^{n \times (K-1)}}{\arg\max} && \operatorname{tr}\left(Q_{K-1}^\top X^\top X Q_{K-1}\right) \\ s.t. & && Q_{K-1}^\top Q_{K-1} = \mathbb{I}_{K-1} \\ & && Q = PR \ \wedge \ q_K = \frac{1}{\sqrt{n}}\mathbf{e}. \end{aligned}$$

While P is an orthonormal indicator/sparse matrix of rank K, Q is a piecewise constant and in general non-sparse orthonormal matrix of the same rank. The latter optimization problem can be relaxed by not adding the structural constraints $Q = PR$ and $q_K = \mathbf{e}/\sqrt{n}$. The resulting optimization problem is known as the Rayleigh–Ritz quotient (e.g., see Lütkepohl, 1997) and in the remainder of the section we refer to it as the *continuous* problem. The optimal solution to the continuous problem is (up to a rotation of the basis) defined by the top $(K-1)$ eigenvectors from the eigendecomposition of the positive definite matrix $X^\top X$ and the optimal value of the relaxed optimization objective is the sum of the eigenvalues corresponding to this solution. As the continuous solution is (in general) not sparse, the discrete problem is better described with non-sparse piecewise constant matrix Q than with sparse indicator matrix P.

Ding & He (2004) and Xu et al. (2015) have formulated a theorem which claims that the subspace spanned by optimal K-centroids is in fact the subspace spanned by the top $(K-1)$ left singular vectors of X. The proofs provided in these works are, however, restricted to the case when the discrete and continuous/relaxed version of the optimization problem match. We address here this claim without that restriction and amend their formulation accordingly. For this purpose, let $C^* = \{c_1, \cdots, c_K\}$ be K centroids specifying an optimal K-means clustering (i.e., minimizing the potential). The between cluster scatter matrix $S = \sum_{i=1}^{K} n_i c_i c_i^\top$ projects any vector $x \in \mathcal{X}$ to a subspace spanned by the centroid vectors, i.e., $Sx = \sum_{i=1}^{K} n_i \left(c_i^\top x\right) c_i \in \operatorname{span}\{c_1, \cdots, c_K\}$. Let also λ_K denote the Kth eigenvalue of $H = X^\top X$ and assume the eigenvalues are listed in descending order. A proof of the following proposition is provided in Appendix A.

Proposition 1. *Suppose that the subspace spanned by optimal K-means centroids has a basis that consists of left singular vectors of X. If the gap between the eigenvalues λ_{K-1} and λ_K is sufficiently large (see the proof for explicit definition), then the optimal K-means centroids and the top $(K-1)$ left singular vectors of X span the same subspace.*

Proposition 2. *In contrast to the claim by Ding & He (2004) and Xu et al. (2015), it is possible that no basis of the subspace spanned by optimal K-means centroids consists of left singular vectors of X. In that case, the subspace spanned by the top $(K-1)$ left singular vectors is different from that spanned by optimal K-means centroids.*

Let $X = U\Sigma V^\top$ be an SVD decomposition of X and denote with U_K the top K left singular vectors from this decomposition. Let also $U_K^\perp$ denote the dual matrix of U_K and $\phi\left(C^* \mid U_K\right)$ the clustering potential given by the projections of X and C^* onto the subspace U_K.

Proposition 3. *Let H_K denote the optimal rank K approximation of the Gram matrix $H = X^\top X$ and let C^* be an* *optimal K-means clustering of X. Then, it holds*

$$\phi\left(C^*\right) \leq \left\|H - H_{K-1}\right\|_1 + \phi\left(C^* \mid U_{K-1}\right).$$

Let us now relate Proposition 1 to the result from Section 2 where we were interested in finding a set of landmarks spanning the subspace that preserves most of the variance of the data in the kernel feature space. Assuming that the conditions from Proposition 1 are satisfied, the Nyström approximation using optimal kernel K-means centroids as landmarks projects the data to a subspace with the highest possible variance. Hence, under these conditions optimal kernel K-means landmarks provide the optimal rank $(K-1)$ reconstruction of the kernel matrix. However, for a kernel K-means centroid there does not necessarily exist a point in the instance space that maps to it. To account for this and the hardness of the kernel K-means clustering problem (Aloise et al., 2009), we propose to approximate the centroids with kernel K-means++ samples. This sampling strategy iteratively builds up a set of landmarks such that in each iteration an instance is selected with probability proportional to its contribution to the clustering potential in which previously selected instances act as cluster centers. For a problem with n instances and dimension d, the strategy selects K landmarks in time $\mathcal{O}\left(Knd\right)$.

Before we give a bound on the Nyström approximation with kernel K-means++ samples as landmarks, we provide a result by Arthur & Vassilvitskii (2007) on the approximation error of the optimal clustering using this sampling scheme.

Theorem 4. *[Arthur & Vassilvitskii (2007)] If a clustering C is constructed using the K-means++ sampling scheme then the corresponding clustering potential $\phi\left(C\right)$ satisfies*

$$\mathbb{E}\left[\phi\left(C\right)\right] \leq 8\left(\ln K + 2\right)\phi\left(C^*\right),$$

where C^* is an optimal clustering and the expectation is taken with respect to the sampling distribution.

Having presented all the relevant results, we now give a bound on the approximation error of the Nyström method with kernel K-means++ samples as landmarks. A proof of the following theorem is provided in Appendix A.

Theorem 5. *Let H be a kernel matrix with a finite rank factorization $H = \Phi\left(X\right)^\top \Phi\left(X\right)$. Denote with H_K the optimal rank K approximation of H and let $\tilde{H}_K$ be the Nyström approximation of the same rank obtained using kernel K-means++ samples as landmarks. Then, it holds*

$$\mathbb{E}\left[\frac{\|H-\tilde{H}_K\|_2}{\|H-H_K\|_2}\right] \leq 8(\ln(K+1)+2)(\sqrt{n-K}+\Theta_K),$$

with $\Theta_K = {}^{\phi(C^* \mid U_K)}/_{\|H-H_K\|_2}$, where U_K denotes the top K left singular vectors of $\Phi\left(X\right)$ and C^* an optimal kernel K-means clustering with $(K+1)$ clusters.

Corollary 6. *When $\phi\left(C^* \mid U_K\right) \leq \sqrt{n-K}\left\|H - H_K\right\|_2$, then the additive term $\Theta_K \leq \sqrt{n-K}$ and*

$$\mathbb{E}\left[\frac{\|H-\tilde{H}_K\|_2}{\|H-H_K\|_2}\right] \in \mathcal{O}\left(\ln K\sqrt{n-K}\right). \qquad (4)$$

The given bound for low-rank approximation of symmetric and positive definite matrices holds for the Nyström method with kernel K-means++ samples as landmarks *without any Lloyd iterations* (Lloyd, 1982). To obtain even better landmarks, it is possible to first sample candidates using the kernel K-means++ sampling scheme and then attempt a Lloyd refinement in the instance space (motivation for this is provided in Section 4.3). If the clustering potential is decreased as a result of this, the iteration is considered successful and the landmarks are updated. Otherwise, the refinement is rejected and current candidates are selected as landmarks. This is one of the landmark selection strategies we analyze in our experiments (e.g., see Appendix C).

Let us now discuss the properties of our bound with respect to the rank of the approximation. From Corollary 6 it follows that the bound on the relative approximation error increases initially (for small K) with $\ln K$ and then decreases as K approaches n. This is to be expected as a larger K means we are trying to find a higher dimensional subspace and initially this results in having to solve a more difficult problem. The bound on the low-rank approximation error is, on the other hand, obtained by multiplying with $\left\|H - H_K\right\|_2$ which depends on the spectrum of the kernel matrix and decreases with K. In order to be able to generalize at all, one has to assume that the spectrum falls rather sharply and typical assumptions are $\lambda_i \in \mathcal{O}\left(i^{-a}\right)$ with $a > 1$ or $\lambda_i \in \mathcal{O}\left(e^{-bi}\right)$ with $b > 0$ (e.g., see Section 4.3, Bach, 2013). It is simple to show that for $a \geq 2$, $K > 1$, and $\lambda_i \in \mathcal{O}\left(i^{-a}\right)$ such falls are sharper than $\ln K$ (Corollary 7, Appendix A). Thus, our bound on the low-rank approximation error decreases with K for sensible choices of the kernel function. Note that a similar state-of-the-art bound on the relative approximation error by Li et al. (2016) exhibits worse behavior and grows linearly with K.

4. Discussion

We start with a brief overview of alternative approaches to landmark selection in the Nyström method for low-rank approximation of kernel matrices. Following this, we focus on a bound that is the most similar to ours, that of K-DPP-Nyström (Li et al., 2016). Then, for the frequently used Gaussian kernel, we provide a theoretically sound motivation for performing Lloyd refinements of kernel K-means++ landmarks in the instance space instead of the kernel feature space. These refinements are computationally cheaper than those performed in the kernel feature space and can only improve the positioning of the landmarks.

4.1. Related Approaches

As pointed in Sections 1 and 2, the choice of landmarks is instrumental for the goodness of the Nyström low-rank approximations. For this reason, the existing work on the Nyström method has focused mainly on landmark selection techniques with theoretical guarantees. These approaches can be divided into four groups: *i*) random sampling, *ii*) greedy methods, *iii*) methods based on the Cholesky decomposition, *iv*) vector quantization (e.g., K-means clustering).

The simplest strategy for choosing the landmarks is by uniformly sampling them from a given set of instances. This was the strategy that was proposed by Williams & Seeger (2001) in the first paper on the Nyström method for low-rank approximation of kernel matrices. Following this, more sophisticated non-uniform sampling schemes were proposed. The schemes that received a lot of attention over the past years are the selection of landmarks by sampling proportional to column norms of the kernel matrix (Drineas et al., 2006), diagonal entries of the kernel matrix Drineas & Mahoney (2005), approximate leverage scores (Alaoui & Mahoney, 2015; Gittens & Mahoney, 2016), and submatrix determinants (Belabbas & Wolfe, 2009; Li et al., 2016). From this group of methods, the approximate leverage score sampling and the K-DPP Nyström method (see Section 4.2) are considered state-of-the-art methods in low-rank approximation of kernel matrices.

The second group of landmark selection techniques are greedy methods. A well-performing representative from this group is a method for sparse approximations proposed by Smola & Schölkopf (2000) for which it was later independently established (Kumar et al., 2012) that it performs very well in practice—second only to K-means clustering.

The third group of methods relies on the incomplete Cholesky decomposition to construct a low-rank approximation of a kernel matrix (Fine & Scheinberg, 2002; Bach & Jordan, 2005; Kulis et al., 2006). An interesting aspect of the work by Bach & Jordan (2005) and that of Kulis et al. (2006) is the incorporation of side information/labels into the process of finding a good low-rank approximations of a given kernel matrix.

Beside these approaches, an influential ensemble method for low-rank approximation of kernel matrices was proposed by Kumar et al. (2012). This work also contains an empirical study with a number of approaches to landmark selection. Kumar et al. (2012) also note that the landmarks obtained using instance space K-means clustering perform the best among non-ensemble methods.

4.2. K-DPP Nyström Method

The first bound on the Nyström approximation with landmarks sampled proportional to submatrix determinants was given by Belabbas & Wolfe (2009). Li et al. (2016) recognize this sampling scheme as a determinantal point process and extend the bound to account for the case when l landmarks are selected to make an approximation of rank $K \leq l$. That bound can be formally specified as (Li et al., 2016)

$$\mathbb{E}\left[\frac{\|H - \tilde{H}_K\|_2}{\|H - H_K\|_2}\right] \leq \frac{l+1}{l+1-K}\sqrt{n-K}. \qquad (5)$$

For $l = K$, the bound can be derived from that of Belabbas & Wolfe (Theorem 1, 2009) by applying the inequalities between the corresponding Schatten p-norms.

The bounds obtained by Belabbas & Wolfe (2009) and Li et al. (2016) can be directly compared to the bound from Corollary 6. From Eq. (5), for $l = K + 1$, we get that the expected relative approximation error of the K-DPP Nyström method scales like $\mathcal{O}\left(K\sqrt{n-K}\right)$. For a good worst case guarantee on the generalization error of learning with Nyström approximations (see, e.g., Yang et al., 2012), the parameter K scales as $\sqrt{n}$. Plugging this parameter estimate into Eq. (4), we see that the upper bound on the expected error with kernel K-means++ landmarks scales like $\mathcal{O}\left(\sqrt{n}\ln n\right)$ and that with K-DPP landmarks like $\mathcal{O}\left(n\right)$.

Having compared our bound to that of K-DPP landmark selection, we now discuss some specifics of the empirical study performed by Li et al. (2016). The crucial step of that landmark selection strategy is the ability to efficiently sample from a K-DPP. To achieve this, the authors have proposed to use a Markov chain with a worst case mixing time linear in the number of instances. The mixing bound holds provided that a data-dependent parameter satisfies a condition which is computationally difficult to verify (Section 5, Li et al., 2016). Moreover, there are cases when this condition is not satisfied and for which the mixing bound does not hold. In their empirical evaluation of the K-DPP Nyström method, Li et al. (2016) have chosen the initial state of the Markov chain by sampling it using the K-means++ scheme and then run the chain for 100-300 iterations. While the choice of the initial state is not discussed by the authors, one reason that this could be a good choice is because it starts the chain from a high density region. To verify this hypothesis, we simulate the K-DPP Nyström method by choosing the initial state uniformly at random and run the chain for 1 000 and 10 000 steps (Section 5). Our empirical results indicate that starting the K-DPP chain with K-means++ samples is instrumental for performing well with this method in terms of runtime and accuracy (Figure 6, Li et al., 2016). Moreover, for the case when the initial state is sampled uniformly at random, our study indicates that the chain might need at least one pass through the data to reach a region with good landmarks. The latter is computationally inefficient already on datasets with more than 10 000 instances.

4.3. Instance Space K-means Centroids as Landmarks

We first address the approach to landmark selection based on K-means clustering in the instance space (Zhang et al., 2008) and then give a theoretically sound motivation for why these landmarks work well with the frequently used Gaussian kernel. The outlined reasoning motivates the instance space Lloyd refinements of kernel K-means++ samples and it can be extended to other kernel feature spaces by following the derivations from Burges (1999).

The only existing bound for instance space K-means landmarks was provided by Zhang et al. (2008). However, this bound only works for kernel functions that satisfy

$$\left(h\left(a,b\right)-h\left(c,d\right)\right)^2 \leq \eta\left(h,\mathcal{X}\right)\left(\left\|a-c\right\|^2-\left\|b-d\right\|^2\right),$$

for all $a,b,c,d \in \mathcal{X}$ and a data- and kernel-dependent constant $\eta\left(h,\mathcal{X}\right)$. In contrast to this, our bound holds for all kernels over Euclidean spaces. The bound given by Zhang et al. (2008) is also a worst case bound, while ours is a bound in the expectation. The type of the error itself is also different, as we bound the relative error and Zhang et al. (2008) bound the error in the Frobenius norm. The disadvantage of the latter is in the sensitivity to scaling and such bounds become loose even if a single entry of the matrix is large (Li et al., 2016). Having established the difference in the type of the bounds, it cannot be claimed that one is sharper than the other. However, it is important to note that the bound by Zhang et al. (Proposition 3, 2008) contains the full clustering potential $\phi\left(C^*\right)$ multiplied by ${}^{n\sqrt{n}}/_{K}$ as a term and this is significantly larger than the rank dependent term from our bound (e.g., see Theorem 5).

Burges (1999) has investigated the geometry of kernel feature spaces and a part of that work refers to the Gaussian kernel. We review the results related to this kernel feature space and give an intuition for why K-means clustering in the instance space provides a good set of landmarks for the Nyström approximation of the Gaussian kernel matrix. The reasoning can be extended to other kernel feature spaces as long as the manifold onto which the data is projected in the kernel feature space is a flat Riemannian manifold with the geodesic distance between the points expressed in terms of the Euclidean distance between instances (e.g., see Riemmannian metric tensors in Burges, 1999).

The frequently used Gaussian kernel is given by

$$h\left(x,y\right)=\left\langle\Phi\left(x\right),\Phi\left(y\right)\right\rangle=\exp\left({}^{\|x-y\|^2}/_{2\sigma^2}\right),$$

where the feature map $\Phi\left(x\right)$ is infinite dimensional and for a subset X of the instance space $\mathcal{X} \in \mathbb{R}^d$ also infinitely continuously differentiable on X. As in Burges (1999) we denote with $\mathcal{S}$ the image of X in the reproducing kernel Hilbert space of h. The image $\mathcal{S}$ is a $r \leq d$ dimensional surface in this Hilbert space. As noted by Burges (1999) the image $\mathcal{S}$ is a Hausdorff space (Hilbert space is a metric space and, thus, a Hausdorff space) and has a countable basis of open sets (the reproducing kernel Hilbert space of the Gaussian kernel is separable). So, for $\mathcal{S}$ to be a differentiable manifold (Boothby, 1986) the image $\mathcal{S}$ needs to be locally Euclidean of dimension $r \leq d$.

We assume that our set of instances X is mapped to a differentiable manifold in the reproducing kernel Hilbert space $\mathcal{H}$. On this manifold a Riemannian metric can be defined and, thus, the set X is mapped to a Riemannian manifold $\mathcal{S}$. Burges (1999) has showed that the Riemannian metric tensor induced by this kernel feature map is $g_{ij}=\frac{\delta_{ij}}{\sigma^2}$, where $\delta_{ij}=1$ if $i=j$ and zero otherwise ($1 \leq i,j \leq d$). This form of the tensor implies that the manifold is flat.

From the obtained metric tensor, it follows that the squared geodesic distance between two points $\Phi\left(x\right)$ and $\Phi\left(y\right)$ on $\mathcal{S}$ is equal to the σ-scaled Euclidean distance between x and y in the instance space, i.e., $d_{\mathcal{S}}\left(\Phi\left(x\right),\Phi\left(y\right)\right)^2={}^{\|x-y\|^2}/_{\sigma^2}$. For a cluster $\mathcal{P}_k$, the geodesic centroid is a point on $\mathcal{S}$ that minimizes the distance to other cluster points (centroid in the K-means sense). For our instance space, we have that

$$c_k^*=\arg\min_{c\in\mathbb{R}^d}\sum_{x\in\mathcal{P}_k}\left\|x-c\right\|^2 \Rightarrow c_k^*=\frac{1}{|\mathcal{P}_k|}\sum_{x\in\mathcal{P}_k}x.$$

Thus, by doing K-means clustering in the instance space we are performing approximate geodesic clustering on the manifold onto which the data is embedded in the Gaussian kernel feature space. It is important to note here that a centroid from the instance space is only an approximation to the geodesic centroid from the kernel feature space – the preimage of the kernel feature space centroid does not necessarily exist. As the manifold is flat, geodesic centroids are 'good' approximations to kernel K-means centroids. Hence, by selecting centroids obtained using K-means clustering in the instance space we are making a good estimate of the kernel K-means centroids. For the latter centroids, we know that under the conditions of Proposition 1 they span the same subspace as the top $(K-1)$ left singular vectors of a finite rank factorization of the kernel matrix and, thus, define a good low-rank approximation of the kernel matrix.

5. Experiments

Having reviewed the state-of-the-art methods in selecting landmarks for the Nyström low-rank approximation of kernel matrices, we perform a series of experiments to demonstrate the effectiveness of the proposed approach and substantiate our claims from Sections 3 and 4. We achieve this by comparing our approach to the state-of-the-art in landmark selection – approximate leverage score sampling (Gittens & Mahoney, 2016) and the K-DPP Nyström method (Belabbas & Wolfe, 2009; Li et al., 2016).

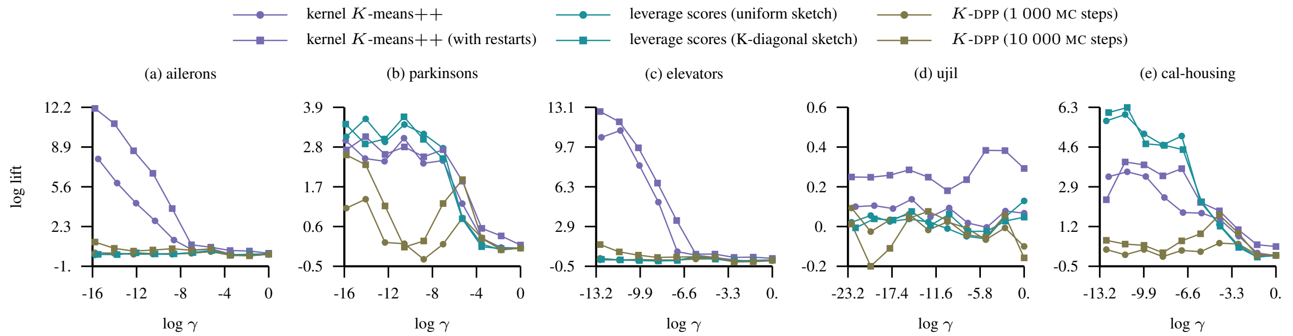

Figure 1. The figure shows the lift of the approximation error in the Frobenius norm as the bandwidth parameter of the Gaussian kernel varies and the approximation rank is fixed to $K = 100$. The lift of a landmark selection strategy indicates how much better it is to approximate the kernel matrix with landmarks obtained using that strategy compared to the uniformly sampled ones.

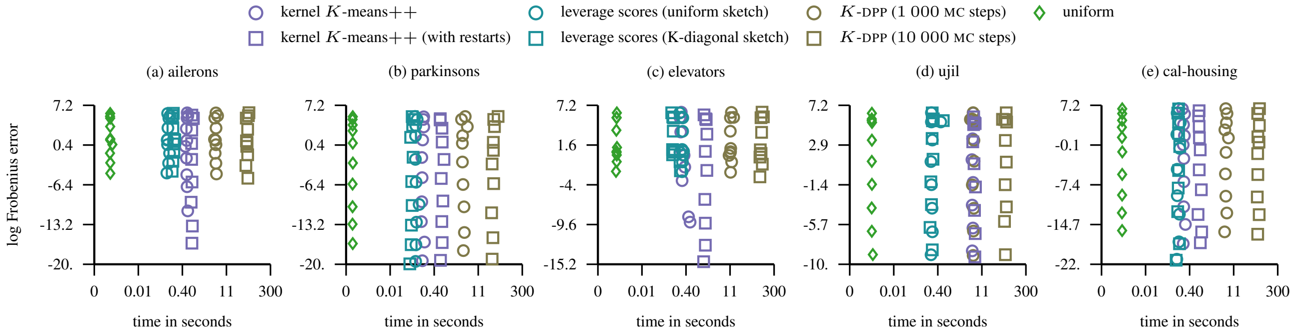

Figure 2. The figure shows the time it takes to select landmarks via different schemes together with the corresponding error in the Frobenius norm while the bandwidth of the Gaussian kernel varies and the approximation rank is fixed to $K = 100$.

Before we present and discuss our empirical results, we provide a brief summary of the experimental setup. The experiments were performed on 13 real-world datasets available at the UCI and LIACC repositories. Each of the selected datasets consists of more than 5 000 instances. Prior to running the experiments, the datasets were standardized to have zero mean and unit variance. We measure the goodness of a landmark selection strategy with the lift of the approximation error in the Frobenius norm and the time needed to select the landmarks. The lift of the approximation error of a given strategy is computed by dividing the error obtained by sampling landmarks uniformly without replacement (Williams & Seeger, 2001) with the error of the given strategy. In contrast to the empirical study by Li et al. (2016), we do not perform any sub-sampling of the datasets with less than 25 000 instances and compute the Frobenius norm error using full kernel matrices. On one larger dataset with more than 25 000 instances the memory requirements were hindering our parallel implementation and we, therefore, subsampled it to 25 000 instances (*ct-slice* dataset, Appendix C). By performing our empirical study on full datasets, we are avoiding a potentially negative influence of the sub-sampling on the effectiveness of the compared landmark selection strategies, time consumed, and the accuracy of the approximation error. Following previous empirical studies (Drineas & Mahoney, 2005; Kumar et al., 2012; Li et al., 2016), we evaluate the goodness of landmark selection strategies using the Gaussian kernel and repeat all experiments 10 times to account for their non-deterministic nature. We refer to $\gamma = {}^1\!/\!\sigma^2$ as the bandwidth of the Gaussian kernel and in order to determine the bandwidth interval we sample 5 000 instances and compute their squared pairwise distances. From these distances we take the inverse of 1 and 99 percentile values as the right and left endpoints. To force the kernel matrix to have a large number of significant spectral components (i.e., the Gaussian kernel matrix approaches to the identity matrix), we require the right bandwidth endpoint to be at least 1. From the logspace of the determined interval we choose 10 evenly spaced values as bandwidth parameters. In the remainder of the section, we summarize our findings with 5 datasets and provide the complete empirical study in Appendix C.

In the first set of experiments, we fix the approximation rank and evaluate the performance of the landmark selection strategies while varying the bandwidth of the Gaussian kernel. Similar to Kumar et al. (2012), we observe that for most datasets at a standard choice of bandwidth – inverse median squared pairwise distance between instances – the princi-

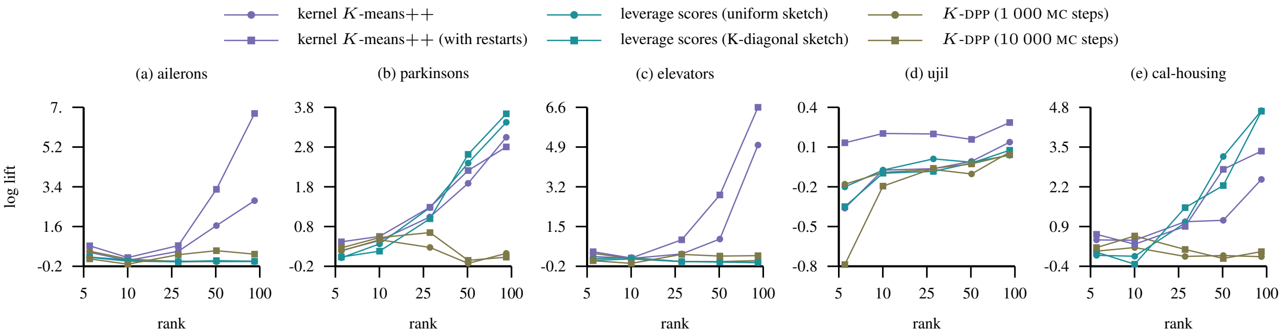

Figure 3. The figure shows the improvement in the lift of the approximation error measured in the Frobenius norm that comes as a result of the increase in the rank of the approximation. The bandwidth parameter of the Gaussian kernel is set to the inverse of the squared median pairwise distance between the samples.

pal part of the spectral mass is concentrated at the top 100 eigenvalues and we set the approximation rank $K = 100$. Figure 1 demonstrates the effectiveness of evaluated selection strategies as the bandwidth varies. More precisely, as the log value of the bandwidth parameter approaches to zero the kernel matrix is close to being the identity matrix, thus, hindering low-rank approximations. In contrast to this, as the bandwidth value gets smaller the spectrum mass becomes concentrated in a small number of eigenvalues and low-rank approximations are more accurate. Overall, the kernel K-means++ sampling scheme performs the best across all 13 datasets. It is the best performing method on 10 of the considered datasets and a competitive alternative on the remaining ones. The improvement over alternative approaches is especially evident on datasets *ailerons* and *elevators*. The approximate leverage score sampling is on most datasets competitive and achieves a significantly better approximation than alternatives on the dataset *cal-housing*. The approximations for the K-DPP Nyström method with $10\ 000$ MC steps are more accurate than that with $1\ 000$ steps. The low lift values for that method seem to indicate that the approach moves rather slowly away from the initial state sampled uniformly at random. This choice of the initial state is the main difference in the experimental setup compared to the study by Li et al. (2016) where the K-DPP chain was initialized with K-means++ sampling scheme.

Figure 2 depicts the runtime costs incurred by each of the sampling schemes. It is evident that compared to other methods the cost of running the K-DPP chain with uniformly chosen initial state for more than $1\ 000$ steps results in a huge runtime cost without an appropriate reward in the accuracy. From this figure it is also evident that the approximate leverage score and kernel K-means++ sampling are efficient and run in approximately the same time apart from the dataset *ujil* (see also *ct-slice*, Appendix C). This dataset has more than 500 attributes and it is time consuming for the kernel K-means++ sampling scheme (our implementation does not cache/pre-compute the kernel matrix). While on such large dimensional datasets the kernel K-means++ sampling scheme is not as fast as the approximate leverage score sampling, it is still the best performing landmark selection technique in terms of the accuracy.

In Figure 3 we summarize the results of the second experiment where we compare the improvement in the approximation achieved by each of the methods as the rank of the approximation is increased from 5 to 100. The results indicate that the kernel K-means++ sampling achieves the best increase in the lift of the approximation error. On most of the datasets the approximate leverage score sampling is competitive. That method also performs much better than the K-DPP Nyström approach initialized via uniform sampling.

As the landmark subspace captured by our approach depends on the gap between the eigenvalues and that of the approximate leverage score sampling on the size of the sketch matrix, we also evaluate the strategies in a setting where l landmarks are selected in order to make a rank $K < l$ approximation of the kernel matrix. Similar to the first experiment, we fix the rank to $K = 100$ and in addition to the already discussed case with $l = K$ we consider cases with $l = K \ln n$ and $l = K \ln K$. Due to space restrictions, the details of this experiment are provided in Appendix C. The results indicate that there is barely any difference between the lift curves for the kernel K-means++ sampling with $l = K \ln K$ and $l = K \ln n$ landmarks. In their empirical study, Gittens & Mahoney (2016) have observed that for uniformly selected landmarks, $\epsilon \in [0, 1]$, and $l \in \mathcal{O}\left(K \ln n\right)$, the average rank K approximation errors are within $(1 + \epsilon)$ of the optimal rank K approximation errors. Thus, based on that and our empirical results it seems sufficient to take $K \ln K$ landmarks for an accurate rank K approximation of the kernel matrix. Moreover, the gain in the accuracy for our approach with $l = K \ln K$ landmarks comes with only a slight increase in the time taken to select the landmarks. Across all datasets, the proposed sampling scheme is the best performing landmark selection technique in this setting.

Acknowledgment: We are grateful for access to the University of Nottingham High Performance Computing Facility.

References

Alaoui, Ahmed E. and Mahoney, Michael W. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems 28*, 2015.

Aloise, Daniel, Deshpande, Amit, Hansen, Pierre, and Popat, Preyas. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 2009.

Arthur, David and Vassilvitskii, Sergei. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.

Bach, Francis R. Sharp analysis of low-rank kernel matrix approximations. In *Proceedings of the 26th Annual Conference on Learning Theory*, 2013.

Bach, Francis R. and Jordan, Michael I. Predictive low-rank decomposition for kernel methods. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

Belabbas, Mohamed A. and Wolfe, Patrick J. Spectral methods in machine learning: New strategies for very large datasets. *Proceedings of the National Academy of Sciences of the USA*, 2009.

Boothby, William M. *An introduction to differentiable manifolds and Riemannian geometry*. Academic Press, 1986.

Boutsidis, Christos, Drineas, Petros, and Mahoney, Michael W. Unsupervised feature selection for the K-means clustering problem. In *Advances in Neural Information Processing Systems 22*, 2009.

Burges, Christopher J. C. Geometry and invariance in kernel based methods. In *Advances in Kernel Methods*. MIT Press, 1999.

Ding, Chris and He, Xiaofeng. K-means clustering via principal component analysis. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

Drineas, Petros and Mahoney, Michael W. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 2005.

Drineas, Petros, Kannan, Ravi, and Mahoney, Michael W. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 2006.

Fine, Shai and Scheinberg, Katya. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2002.

Gittens, Alex and Mahoney, Michael W. Revisiting the Nyström method for improved large-scale machine learning. *Journal Machine Learning Research*, 2016.

Golub, Gene H. and van Loan, Charles F. *Matrix Computations*. Johns Hopkins University Press, 1996.

Kanungo, Tapas, Mount, David M., Netanyahu, Nathan S., Piatko, Christine D., Silverman, Ruth, and Wu, Angela Y. A local search approximation algorithm for K-means clustering. In *Proceedings of the Eighteenth Annual Symposium on Computational Geometry*, 2002.

Kulis, Brian, Sustik, Mátyás, and Dhillon, Inderjit. Learning low-rank kernel matrices. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

Kumar, Sanjiv, Mohri, Mehryar, and Talwalkar, Ameet. Sampling methods for the Nyström method. *Journal of Machine Learning Research*, 2012.

Li, Chengtao, Jegelka, Stefanie, and Sra, Suvrit. Fast DPP sampling for Nyström with application to kernel methods. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Lloyd, Stuart. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 1982.

Lütkepohl, Helmut. *Handbook of Matrices*. Wiley, 1997.

Nyström, Evert J. Über die praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben. *Acta Mathematica*, 1930.

Schölkopf, Bernhard and Smola, Alexander J. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press, 2001.

Smola, Alexander J. and Schölkopf, Bernhard. Sparse greedy matrix approximation for machine learning. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

Wahba, Grace. *Spline models for observational data*. SIAM, 1990.

Weidmann, Joachim. *Linear operators in Hilbert spaces*. Springer-Verlag, 1980.

Williams, Christopher K. I. and Seeger, Matthias. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*. 2001.

Xu, Qin, Ding, Chris, Liu, Jinpei, and Luo, Bin. PCA-guided search for K-means. *Pattern Recognition Letters*, 2015.

Yang, Tianbao, Li, Yu-feng, Mahdavi, Mehrdad, Jin, Rong, and Zhou, Zhi-Hua. Nyström method vs random Fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems 25*. 2012.

Zhang, Kai, Tsang, Ivor W., and Kwok, James T. Improved Nyström low-rank approximation and error analysis. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.