An Analysis Framework for Landmark Selection and Nyström Kernel Approximation: Bounds, Algorithms, and Applications

Anonymous CVPR submission

Paper ID 465

Abstract

In recent years, the spectral analysis of appropriately defined kernels has emerged as a principled way to extract the nonlinear structure inherent in vision tasks as diverse as segmentation and recognition. For computational reasons, a landmark selection process is often employed to choose a partial kernel, followed by a Nyström extension that provides its minimally positive definite completion. The literature remains open on the question of optimal landmark selection, however, and here we develop an analysis framework for this problem that subsumes previous approaches. We first show how it leads to quantitative performance bounds for both existing and new algorithms. We then discuss a range of methods for optimizing the landmark selection process through a stochastic procedure, and finally demonstrate their effectiveness in two key applications: the low-level vision task of image segmentation, and the high-level task of nonlinear dimensionality reduction from video data.

1. Introduction

Algorithmic aspects, both of space and time complexity, have historically had a high impact on the efficacy of computer vision solutions. A fundamental tool in many such tasks is the singular value decomposition, which lies at the heart of spectral methods both classical and new for extracting relevant information from data. Canonical examples range from principal components analysis and its brethren, which act on the scatter matrix of a set of points and hence depend on their extrinsic dimensionality, to methods such as multidimensional scaling and its modern extensions [16], which scale instead as the cardinality of the point cloud.

Applications in vision and related areas as diverse as image segmentation [8], image matting [15], spectral mesh processing [17], and object recognition through the use of appearance manifolds [14] all rely in turn on the eigendecomposition of a suitably defined kernel. However, at a complexity of $O(n^3)$ the full spectral analysis of real-world data sets is often prohibitively costly—requiring in practice an approximation to the exact spectral decomposition. Indeed the aforementioned tasks typically fall into this category, and several share the common feature that their kernel approximations are obtained in exactly the same way—via the process of selecting a subset of examples or coordinates termed landmarks to serve as a basis for computation.

Since their introduction, and furthermore as data sets continue to increase in size and dimensionality, landmark methods have seen wide use by the learning community, particularly in vision [26, and references above]. These methods exploit the high level of redundancy in typical massive data sets by seeking a small (in relative terms) number of important examples or coordinates that summarize most relevant information in the data; this amounts in effect to an adaptive compression scheme. Separate from this subset selection problem is the actual reconstruction of the corresponding spectral analysis task—and this in turn is accomplished via the so-called Nyström extension [23, 27].

While the Nyström reconstruction admits the unique property of providing, conditioned upon a set of selected landmarks, the minimal kernel completion with respect to the partial ordering of positive semi-definiteness, the literature is currently open on the question of optimal landmark selection. Choosing the most appropriate set of landmarks for a specific data set is a fundamental task if spectral methods are to successfully “scale up” to the order of the large data sets currently seen in contemporary vision applications and expected in the future. Improvements will in turn translate directly to either a more efficient compression of the input (i.e., fewer landmarks will be needed) or a more accurate approximation for a given compression size. When choosing landmarks in a data-adaptive way can clearly offer improvement over approaches such as selecting them uniformly at random [1, 6], this latter approach remains by far the most popular with practitioners [8, 9, 25, 26].

While it is clear that data-dependent landmark selection methods offer the potential of at least some improvement over non-adaptive methods such as uniform sampling [17],
bounds on performance as a function of computation have not been rigorously addressed in the literature to date. One important reason for this has been the lack of a unifying framework to understand the problems of landmark selection and sampling, and to provide for approximation bounds and quantitative performance guarantees. In this paper we put forward an analysis framework for landmark selection that places previous approaches in context, and show how it leads to quantitative performance bounds on Nyström kernel approximation. We discuss a range of methods for optimizing the landmark selection process through sampling, and demonstrate their effectiveness on two diverse applications: the low-level vision task of image segmentation, and the high-level task of extracting latent dynamic processes from video data for purposes of motion recognition.

2. Spectral methods & the Nyström completion

2.1. Spectral methods in brief

Spectral methods rely on low-rank approximations of appropriately defined positive semi-definite kernels. To this end, let \( Q \geq 0 \) denote a kernel of dimension \( n \), with \( \geq \) denoting the corresponding partial order (i.e., \( Q \geq Q' \Rightarrow Q - Q' \geq 0 \)). Such a kernel can in turn be expressed in spectral coordinates as \( Q = U \Lambda U^T \), where \( U \) is an orthogonal matrix such that \( U^T U = I \) (the identity), and \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n) \) contains the (nonnegative) eigenvalues of \( Q \), assumed sorted in non-increasing order.

To measure the error in approximating a kernel \( Q \geq 0 \), we require the following notion of unitary invariance.

**Definition 1** (Unitary Invariance). A matrix norm \( \| \cdot \| \) is termed unitarily invariant if, for all unitary matrices \( U, V \), we have that \( \| UQV^T \| = \| Q \| \) for every (square) matrix \( Q \).

A unitarily invariant norm therefore depends only on the spectrum of its argument, and for any such norm the optimal rank-\( k \) approximation to \( Q \) is given by \( Q_k := U \Lambda_k U^T \) where \( \Lambda_k = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k, 0, \ldots, 0) \). When a given kernel \( Q \) is expressed in spectral coordinates, evaluating the quality of any low-rank approximation \( Q \) is a trivial task, requiring only an ordering of the eigenvalues. The cost of obtaining the spectral coordinates, however, is \( \mathcal{O}(n^3) \), which is often too costly to be computed in practice.

To this end, methods that rely on either the extrinsic dimensionality of a point cloud (e.g., principal components), or on the intrinsic dimensionality of a set of training examples (e.g., multidimensional scaling) via its cardinality, pose a computational issue for vision problems. To illustrate, let \( x_1, x_2, \ldots, x_N \in \mathbb{R}^n \) comprise the data of interest. “Outer” methods of the former category employ a rank-\( k \) approximation of the matrix \( Q := \sum_{i=1}^N x_i x_i^T \), which is of dimension \( n \). Alternatively, “inner” methods will introduce an additional (conditionally) positive-definite function \( q(x_i, x_j) \), such as \( \langle x_i, x_j \rangle \) or \( \exp(-\|x_i - x_j\|^2/\sigma^2) \), and obtain a \( k \)-dimensional embedding of the data via the affinity matrix \( Q_{ij} := q(x_i, x_j) \), which is of dimension \( N \).

2.2. The Nyström method and landmark selection

As detailed in Section 1, the Nyström method has found many applications in machine learning in general and computer vision in particular, as a means of obtaining an approximate spectral analysis of the kernel of interest \( Q \). In brief, the method solves a matrix completion problem in a way that preserves positive semi-definiteness as follows.

**Definition 2** (Nyström Extension). Fix a subset \( J \subset \{1, 2, \ldots, n\} \) of cardinality \( k < n \), and let \( Q_J \) denote the corresponding principal submatrix of an \( n \)-dimensional kernel \( Q \geq 0 \). Set \( J = \{1, 2, \ldots, k\} \) without loss of generality and partition \( Q \) as follows:

\[
Q = \begin{bmatrix} Q_J & Y \\ Y^T & Z \end{bmatrix}.
\]

The Nyström extension then approximates \( Q \) by

\[
\tilde{Q} = \begin{bmatrix} Q_J & Y \\ Y^T & Y Q_J^{-1} Y \end{bmatrix} \geq 0.
\]

Here \( Q_J \in \mathbb{R}^{k \times k} \) and \( Z \in \mathbb{R}^{(n-k) \times (n-k)} \) are always positive semi-definite, being principal submatrices, and \( Y \in \mathbb{R}^{k \times (n-k)} \).

Denoting \( Q_J = U_J \Lambda_J U_J^T \), this corresponds to approximating the eigenvectors and eigenvalues of \( Q \) by

\[
\tilde{\Lambda} = \Lambda_J, \quad \tilde{U} = \begin{bmatrix} U_J \\ Y^T U_J \Lambda_J^{-1} \end{bmatrix}.
\]

We have that \( \text{rank}(\tilde{Q}) \leq |J| = k \), and (assuming \( k \ll n \)) the complexity of reconstruction is of order \( \mathcal{O}(n^2 k) \). Approximate eigenvectors \( \tilde{U} \) can be obtained in \( \mathcal{O}(nk^2) \), and can be orthogonalized by an additional projection.

The Nyström method thus serves as a means of completing a partial kernel, conditioned upon a selected subset \( J \) of rows and columns of \( Q \). The landmark selection problem becomes the one of choosing the subset \( J \) of fixed cardinality \( k \) such that \( \| Q - \tilde{Q} \| \) is minimized for some unitarily invariant norm, with a lower bound given by \( \| Q - Q_k \| \), where \( Q_k \) is the optimal rank-\( k \) approximation obtained by setting the \( (n-k) \) smallest eigenvalues of \( Q \) to zero.

According to the difference between (1) and (2), the approximation error \( \| Q - \tilde{Q} \| \) can in general be expressed in terms of the Schur complement of \( Q_J \) in \( Q \), defined as \( Z - Y^T Q_J^{-1} Y \) according to the conformal partition of \( Q \) in (1) above, and correspondingly for an appropriate permutation of rows and columns in the general case.

With reference to Definition 2, we thus have the optimal landmark selection problem as follows.
Problem 1 (Optimal Landmark Selection). Choose \( J, |J| = k \) such that \( \|Q - \bar{Q}\| = \|Z - Y^T Q^{-1} Y\| \) is minimized.

It remains an open question as to whether, for any unitarily invariant norm, this subset selection problem can be solved in less than \( O(n^3) \) operations, the threshold above which the exact spectral decomposition becomes the best option. In fact, there are no known exact algorithms other than \( O(n^k) \) brute-force enumeration in the general case.

3. Analysis framework for landmark selection

Attempts to solve the landmark selection problem can be divided into two categories: deterministic methods that typically minimize some objective function in an iterative or stepwise greedy fashion \([17, 22, 25, 28]\), and randomized algorithms that instead proceed by sampling \([1, 6, 8, 27]\).

As we show in this section, the most canonical of these approaches can all be recovered within a generalized stochastic sampling framework to be introduced below.

3.1. Nyström error characterization

It is instructive to first consider Problem 1 in more detail, in order that we may better characterize properties of the Nyström approximation error. To this end, we adopt the trace norm \( \| \cdot \|_{tr} \) as our unitarily invariant norm of interest.

Definition 3 (Trace Norm). Fix an arbitrary matrix \( Q \in \mathbb{R}^{m \times n} \) and let \( \sigma_i(Q) \) denote its \( i \)th singular value. Then the trace norm of \( Q \) is defined as

\[
\|Q\|_{tr} = \text{tr}(\sqrt{Q^T Q}) = \sum_{i = 1}^{\min(m, n)} \sigma_i(Q) \tag{4}
\]

\[
\|Q\|_{tr} = \text{tr}(Q) \text{ for } Q \succeq 0. \tag{5}
\]

Since any positive semi-definite kernel \( Q \succeq 0 \) admits the Gram decomposition \( Q = X^T X \), this implies the following relation in Frobenius norm \( \| \cdot \|_F \), to be revisited shortly:

\[
\|Q\|_{tr} = \|X^T X\|_{tr} = \text{tr}(X^T X) = \|X\|^2_F. \tag{6}
\]

The key property of this norm for our purposes follows from a well-known linear-algebraic notion.

Proposition 1 (Dominance of Trace Norm). Amongst all unitarily invariant norms \( \| \cdot \| \), we have that \( \| \cdot \|_{tr} \geq \| \cdot \| \).

Adopting this norm for Problem 1 therefore allows us to provide minimax arguments, and its unitary invariance implies the natural property that results depend only on the spectrum of the kernel \( Q \succeq 0 \) under consideration, just as in the case of the optimal rank-\( k \) approximant \( Q_k \).

To this end, note that any Schur complement is itself positive semi-definite. Recalling from Definition 2 that the error incurred by the Nyström approximation is the norm of the corresponding Schur complement, and applying the definition of the trace norm as per (5), we obtain the following characterization of Problem 1 under trace norm.

Proposition 2 (Nyström Error in Trace Norm). Fix a subset \( J \subset \{1, 2, \ldots, n\} \) of cardinality \( k < n \), and denote by \( \bar{Q} \) its complement in \( \{1, 2, \ldots, n\} \). Then the error in trace norm induced by the Nyström approximation of an \( n \)-dimensional kernel \( Q \succeq 0 \) according to Definition 2, conditioned on the choice of subset \( J \), may be expressed as follows:

\[
\|Q - \bar{Q}\|_{tr} = \text{tr}(Q_{J \times J}) - \text{tr}(Y_{J \times J}^T Q_{J \times J}^{-1} Y_{J \times J}), \tag{7}
\]

where \( J \times \bar{J} \) denotes rows indexed by \( J \) and columns by \( \bar{J} \).

Remark 1. While each term in the expression of Proposition 2 depends on the selected subset \( J \), if all elements of the diagonal of \( Q \) are the identity, then the term \( \text{tr}(Q_{J \times J}) \) is constant. This has motivated approaches to Problem 1 based on minimizing exclusively the latter term \([25, 28]\).

We conclude with an illuminating proposition that follows self-evidently from the Gram decomposition of (6).

Proposition 3 (Nyström Error as Regression Residual). Let \( Q \succeq 0 \) have the Gram decomposition \( Q = X^T X \), and let \( X \) be partitioned as \([X_J, \ X_{\bar{J}}]\) in accordance with Proposition 2. Then the Nyström error in trace norm of (7) is the error sum-of-squares obtained by projecting columns of \( X_J \) onto the closed linear span of columns of \( X_{\bar{J}} \).

3.2. Annealed determinantal distributions

With this error characterization in hand, we may now define and introduce the notion of annealed determinantal distributions, a key component of our analysis framework.

Definition 4 (Annealed Determinantal Distributions). Let \( Q \succeq 0 \) be a positive semi-definite kernel of dimension \( n \), and fix an exponent \( s \geq 0 \). Then for fixed \( k \leq n \), \( Q \) admits a family of probability distributions defined on the set of all \( J \subset \{1, 2, \ldots, n\} : |J| = k \) as follows:

\[
p^s(J) \propto \det(Q_J)^s; \quad s \geq 0, \ |J| = k. \tag{8}
\]

This distribution is well defined owing to the property that all principal submatrices of a positive semi-definite matrix are themselves positive semi-definite, and hence have nonnegative determinant. The term annealing is suggestive of its use in stochastic computation and search, where a probability distribution or energy function is gradually raised to some nonnegative power over the course of an iterative sampling or optimization procedure.

Indeed, for \( 0 < s < 1 \) the determinantal annealing of Definition 4 amounts to a flattening of the distribution of \( \det(Q_J) \), whereas for \( 1 < s < \infty \) it becomes more and more peaked. In the limiting cases we recover, of course,
the uniform distribution on the range of $\det(Q_J)$, and respectively mass concentrated on its maximal element(s).

It is instructive to consider these limiting cases in more detail. Taking $s = 0$, we observe that the method of uniform sampling typically favored by practitioners [8, 9, 25, 26] is trivially recovered, with negligible associated computational cost. By extending a result of [1], we may bound the induced error as follows.

**Proposition 4 (Uniform Sampling).** Let $Q \succeq 0$ have the Nyström extension $\tilde{Q}$, where subset $J : |J| = k$ is chosen uniformly at random. Then

$$\mathbb{E}\|Q - \tilde{Q}\|_{tr} \leq \frac{n - k}{n} \operatorname{tr}(Q). \quad (9)$$

Note that by averaging the effects of all eigenvalues of $Q$, this approach fails to place zero probability of selection on singular subsets, implying that perfect reconstruction of rank-$k$ kernels cannot be guaranteed. However, the following proposition is easily verified [1].

**Proposition 5 (Perfect Reconstruction).** For all $s > 0$, sampling on subsets of cardinality $k$ according to the annealed determinantal distribution of Definition 4 is guaranteed to yield perfect reconstruction whenever $\operatorname{rank}(Q) \leq k$.

Considering the limiting case as $s \to \infty$, we equivalently recover the problem of maximizing the determinant, which is well known to be NP-hard. This special case admits the following interesting correspondence, since if $x$ is a vector-valued random variable with covariance $Q$, then the Schur complement of $Q_J$ in $Q$ represents the conditional covariance of $x_J$.

**Proposition 6 (Minimax Relative Entropy).** Fix an $n$-dimensional kernel $Q \succeq 0$ as the covariance of a random vector $x \in \mathbb{R}^n$ and an integer $k < n$. Minimizing the maximum relative entropy of $x_J$, conditional upon having observed coordinates $x_J$, corresponds by Schur’s Theorem to selecting $J$ such that $\det(Q_J)$ is maximized.

To this end the bound in [11] extends to the case of the trace norm as follows.

**Proposition 7 (Volume Maximization).** Let $\tilde{Q}$ denote the Nyström completion of a kernel $Q \succeq 0$ via subset $J = \arg\max_{|J| = k} \det(Q_J)$. Then

$$\|Q - \tilde{Q}\|_{tr} \leq (k + 1)(n - k)\lambda_{k+1}(Q), \quad (10)$$

where $\lambda_{k+1}(Q)$ is the $(k + 1)$th largest eigenvalue of $Q$.

We conclude with a recent result [1] bounding the expected error for the case $s = 1$, that in turn improves upon the additive error bound of [6] for sampling according to the squared diagonal elements of $Q$.

**Proposition 8 (Volume Sampling).** Let $Q \succeq 0$ have the Nyström extension $\tilde{Q}$, where subset $J : |J| = k$ is chosen according to the annealed determinantal distribution of (8) with $s = 1$. Then

$$\mathbb{E}\|Q - \tilde{Q}\|_{tr} \leq (k + 1) \sum_{i=k+1}^{n} \lambda_i(Q), \quad (11)$$

with $\lambda_{k+1}(Q)$ the $(k + 1)$th largest eigenvalue of $Q$.

This approach can also be interpreted in terms of the volume sampling approach proposed by [5], applied to the Gram matrix $X_J^T X_J$ of an “arbitrary” matrix $X_J$, as $\det(Q_J) = \det(X_J^T X_J) = \det(X_J)^2$. By this same argument, Deshpande et al. [5] show the result of Proposition 11 to be essentially the the best possible.

We conclude by noting that for most values of $s$, sampling from the distribution $p^s(J)$ presents a combinatorial problem, owing to the $\binom{n}{k}$ distinct $k$-subsets $J$ associated with an $n$-dimensional kernel $Q$. To this end, a simple Markov chain Monte Carlo method has been proposed in [11] and shown to be effective for sampling according to the determinantal distribution on $k$-subsets induced by $Q$. This Metropolis algorithm can easily be extended to the cases covered by Definition 4 for all $s \geq 0$. We also note that tridiagonal approximations to $\det(Q_J)$ can be computed in $O(k^3)$ operations and hence offer an alternative to the $O(k^5)$ cost of exact determinant computation.

### 4. Image segmentation

In the past decade, spectral methods have emerged as a major component of many image segmentation algorithms [21, 24]. Typical algorithms rely on the eigendecomposition of a kernel $Q = U \Lambda U^T$, representing normalized affinities between image pixels, in order to obtain an embedding in which pixels belonging to the same segment are close together in the Euclidean metric.

For typical pixel counts, however, computing the spectral decomposition of $Q$, or even the normalizing matrix, may be infeasible. This problem can be made tractable by using the Nyström method to approximate a matrix $W$ of unnormalized affinities, as explored in the series of papers [2, 8, 9], and also in other works such as [28]. In these papers, the subsets $J$ for the Nyström extension were selected by uniform sampling.

We show that in this context, the subset selection methods introduced in Section 3 can provide better performances than sampling uniformly. Figure 1 shows a noisy synthetic image composed of 3 rectangles, similar to one in [8] that was used to illustrate the Nyström method. For this image, uniformly sampling of $J$ when $|J|$ is small typically results in a poor approximation to $W$, as large sections of the image will not have high affinity with any of the selected pixels in
5. Nonlinear dimensionality reduction

Video data sets are often generated by a dynamical process evolving on a low-dimensional manifold, for example a line in case of a translation or a circle in case of a rotation. Extracting this low-dimensional space has applications in object recognition though appearance manifolds [13], motion recognition [3], pose estimation [7], and others.

In this context, nonlinear dimensionality reduction algorithms [16] are the key ingredient mapping the video stream to a lower-dimensional space. The vast majority of these algorithms require one to obtain the eigenvectors of a positive definite kernel \( K \) of size equal to the number of frames in the video stream, which quickly becomes prohibitive and entails the use of approximations to spectral analysis of \( K \).

We have tested the efficacy of the Nyström extension coupled with the subset selection procedures given in Section 3 on different video data sets. In Figure 4 we show the exact embedding in three dimensions, using the diffusion maps algorithm [4], of a video from the Honda/UCSD [12, 13] data set, as well as some selected frames. In this video, the subject rotates his head in front of the camera in several directions, with each motion starting from the resting position of looking straight at the camera. We observe that to each of these motions is associated a circular path, and that they all originate from the same area (the lower-front-right area) of the graph, which corresponds to the resting position.

In Figure 5, we evaluated the average approximation error to the diffusion map kernel corresponding to this video for an approximation rank between 2 and 20. The results are averaged over 2000 trials. The sampling from the determinant distribution is done via a Monte-Carlo algorithm similar to [1] and the determinant maximization is obtained
by keeping the subset $J$ with the largest corresponding determinant $Q_J$ over a random choice of 2500 subsets. For this setting, sampling according to the determinant distribution yields the best results uniformly across the range of approximations. We observe that keeping the subset with maximal determinant does not give a good approximation at low ranks. A further analysis showed that in this case the chosen landmarks tend to concentrate around the lower-front-right region of the graph, which yields a good approximation locally in this part of the space but fails to recover other regions properly. This behavior illustrates the appeal of randomized methods, which avoid such pitfalls.

In another example, a subject moves slowly in front of a camera at an uneven speed. The resulting embedding, given again by the diffusion map algorithm, is a non-uniformly sampled straight line. In this case, we can thus evaluate by visual inspection the effect of an approximation of the diffusion map kernel on the quality of the embedding. This is shown in Figure 6 where typical results from different subset selection methods are displayed. We see that sampling according to the determinant recovers the linear structure dynamical process, up to an affine transformation, whereas sampling uniformly yields some folding of the curve over itself at the extremeties and center.

In Figure 7, we show the approximation error of the kernel associated with this video averaged over 2000 trials, similarly to the previous example. In this case, maximizing the determinant yields the best overall performance. We observe that sampling according to the determinant easily outperforms choosing the subset uniformly at random.

**References**


