Spectral methods are of fundamental importance in statistics and machine learning, because they underlie algorithms from classical principal components analysis to more recent approaches that exploit manifold structure. In most cases, the core technical problem can be reduced to computing a low-rank approximation to a positive-definite kernel. For the growing number of applications dealing with very large or high-dimensional datasets, however, the optimal approximation afforded by an exact spectral decomposition is too costly, because its complexity scales as the cube of either the number of training examples or their dimensionality. Motivated by such circumstances, we present new algorithms for the approximation of positive-semidefinite kernels, together with error bounds that improve on results in the literature. We approach this problem by seeking to determine, in an efficient manner, the most informative subset of our data relative to the kernel approximation task at hand. This leads to two new strategies based on the Nyström method that are directly applicable to massive datasets. The first of these—based on sampling—leads to a randomized algorithm whereupon the kernel induces a probability distribution on its set of partitions, whereas the latter approach—based on sorting—provides for the selection of a partition in a deterministic way. We detail their numerical implementation and provide simulation results for a variety of representative problems in statistical data analysis, each of which demonstrates the improved performance of our approach relative to existing methods.

Spectral Methods in Machine Learning

Before describing our main results, we briefly survey the different spectral methods used in machine learning, and show how our results can be applied to a variety of classical and more contemporary algorithms. Let \( \{x_1, \ldots, x_n\} \) be a collection of data points in \( \mathbb{R}^m \). Spectral methods can be classified according to whether they rely on:

**Outer characteristics of the point cloud.** These are methods such as PCA or Fisher discriminant analysis. They require the spectral analysis of a positive-definite kernel of dimension \( m \), the extrinsic dimensionality of the data.

**Inner characteristics of the point cloud.** These are methods such as MDS, along with recent extensions that rely on it (more or less) to perform an embedding of the data points. They require the spectral analysis of a kernel of dimension \( n \), the cardinality of the point cloud.

In turn, the requisite spectral analysis task becomes prohibitive as the (intrinsic or extrinsic) size of the dataset becomes large. For methods such as PCA and MDS, the analysis task consists of finding the best rank-\( k \) approximation to a symmetric, positive-semidefinite (PSD) matrix—a problem whose efficient solution is the main focus of our article. Many other methods (e.g., refs. 1–5) are reduced by only a few adjustments to this same core problem of kernel approximation.

In particular, techniques such as Fisher discriminant analysis or Laplacian eigenmaps require the solution of a generalized eigenvalue problem via Lanczos iteration. More recently, these have been extended to the task of computing a data-specific metric (e.g., refs. 6 and 7) or embedding the data on a manifold (e.g., refs. 8 and 9).

The goal of this article is twofold. First, we aim to demonstrate quantifiable performance-complexity trade-offs for spectral methods in machine learning, by exploiting the distinction between the amount of data to be analyzed and the amount of information those data represent relative to the kernel approximation task at hand. Second, and equally important, we seek to provide practitioners with new strategies for very large datasets that perform well in practice. Our approach depends on the Nyström extension, a kernel approximation technique for integral equations whose potential as a heuristic for machine learning problems has been previously noted (7, 8). We make this notion precise by revealing the power of the Nyström method and giving quantitative bounds on its performance.

Our main results yield two efficient algorithms—one randomized, the other deterministic—that determine a way of sampling a dataset prior to application of the Nyström method. The former computes a simple rank statistic of the data, and the latter involves sampling from an induced probability distribution. Each of these approaches yields easily implementable numerical schemes, for which we provide empirical evidence of improved performance in simulation relative to existing methods for low-rank kernel approximation.

**Author contributions:** M.-A.B. and P.J.W. designed research, performed research, and wrote the paper.

The authors declare no conflict of interest.

**Communicated by David L. Donoho, Stanford University, Stanford, CA, October 21, 2008 (received for review December 19, 2007)**

**References**

1. To whom correspondence should be addressed. E-mail: patrick@seas.harvard.edu.

© 2009 by The National Academy of Sciences of the USA
The essence of the method is hence to use only partial information about the kernel to first solve a simpler eigenvalue problem, and then to extend the eigenvectors obtained thereby by using complete knowledge of the kernel. The same idea may in turn be applied to extend the solution of a reduced matrix eigenvalue problem to approximate the eigenvectors of an SPSD matrix G (8).

Specifically, one may approximate k eigenvectors of G by decomposing and then extending a k \times k principal submatrix of G. First, let G be partitioned as

$$G = \begin{bmatrix} A & B^T \\ B & C \end{bmatrix},$$  \tag{1}

with A \in \mathbb{R}^{k \times k}; we say that this partition corresponds to the multi-index I = \{1, 2, \ldots, k\}. Now define spectral decompositions G = UAU^T and A = U_A \Lambda_A U_A^T; the Nyström extension then provides an approximation for k eigenvectors in U as

$$\tilde{U} := U_A \Lambda_A^{-1} A = U_A \Lambda_A U_A^T.$$  \tag{2}

In turn, the approximations \tilde{U} \approx U and \Lambda_A \approx \Lambda may be composed to yield an approximation G \approx G according to

$$\tilde{G} := \tilde{U} \Lambda_A \tilde{U}^T = \begin{bmatrix} A & B^T \\ B & C \end{bmatrix}.$$  \tag{3}

We call \tilde{G} the Nyström approximation to G corresponding to I = \{1, 2, \ldots, k\}; the extension of this definition to arbitrary multi-index I will be made formal below. We see from Eq. 2 that the main computational burden now takes place on a principal submatrix A of dimension k \times n, and hence the Nyström extension provides a practical means of scaling up spectral methods in machine learning to very large kernels. From Eqs. 1 and 3 we deduce the resultant approximation error to be

$$\|G - \tilde{G}\| = \|C - BA^{-1}B^T\|,$$  \tag{4}

where Sc(A) := C - BA^{-1}B^T is known as the Schur complement of A in G (9). The characterization of Eq. 4 ties the quality of the Nyström approximation explicitly to the partitioning of G; intuitively, this error reflects the loss of information that results from discarding submatrix C while retaining A and B.

**Main Results**

The Nyström method yields a means of approximating G conditioned on a particular choice of partition, hence shifting the computational load to determining that partition. To this end, we provide two algorithms for efficiently selecting from among all \binom{n}{k} possible partitions of G while controlling the approximation error of Eq. 4.

We first generalize the partitioning introduced above as follows. Let I,J ⊆ \{1, 2, \ldots, n\} be multi-indices of respective cardinalities k and l that contain pairwise distinct elements in \{1, 2, \ldots, n\}. We write I = \{i_1, \ldots, i_k\}, J = \{j_1, \ldots, j_l\}, and denote by \tilde{I} the complement of I in \{1, 2, \ldots, n\}. In order to characterize the Nyström approximation error induced by an arbitrary partition, we write G_{I\cup\tilde{J}} for the k \times l matrix whose (p, q)-th entry is given by (G_{I\cup\tilde{J}})_{pq} = G_{ij}, and abbreviate G_I for G_{I\cup\tilde{J}}.

Determining an optimal partition of G is thus seen to be equivalent to selecting a multi-index I such that the error

$$\|G - \tilde{G}\| = \|G_I - G_{I\cup\tilde{J}}G_I^{-1}G_{I\cup\tilde{J}}\| = \|Sc(G_I)\|$$  \tag{5}

induced by the Nyström approximation \tilde{G} corresponding to I is minimized. This naturally leads us to the algorithmic question of how to select the multi-index I in an efficient yet effective manner. In the sequel we propose both a randomized and a deterministic algorithm for accomplishing this task, and derive the resultant average or worst-case approximation error. To understand the
power of this approach, however, it is helpful to first consider conditions under which the Nyström method is capable of providing perfect reconstruction of \( G \).

Of course, if we take for \( I \) the entire set \( \{1, 2, \ldots, n\} \), then the Nyström extension yields \( G = \tilde{G} \) trivially. However, note that if \( G \) is of rank \( k < n \), then there exist multi-indices of cardinality \( k \) such that the Nyström method provides an exact reconstruction: exactly those such that \( \text{rank}(G_I) = \text{rank}(G) = k \), since this implies

\[
S_C(G_I) = G_I - G_{1,\ldots,l}G_{l,\ldots,1}G_{1,\ldots,l} = 0. \tag{6}
\]

We verify Eq. 6 presently, but the intuition behind it is as follows. If \( G \) is SPSD and of rank \( k \), then it can be expressed as a Gram matrix whose entries comprise the inner products of a set of \( n \) vectors in \( \mathbb{R}^k \). Knowing the correlation of these \( n \) vectors with a subset of \( k \) linearly independent vectors in turn allows us to reconstruct them exactly. Hence, in this case, the information contained in \( G_I \) is sufficient to reconstruct \( G \), and the Nyström method performs the reconstruction.

Before introducing our two algorithms for efficient partition selection and bounding their performance, we require the following result, which gives an explicit characterization of the Schur complement in terms of ratios of determinants.

**Lemma 1** [Crabtree–Haynsworth (10)]. Let \( G_I \) be a nonsingular principal submatrix of some SPSD matrix \( G \). Then the Schur complement of \( G_I \) in \( G \) is given element-wise by

\[
(S_C(G_I))_{ij} = \frac{\det(G_{ij}I + G_{\bar{i}j})}{\det(G_I)}. \tag{7}
\]

We may use the Crabtree–Haynsworth characterization of Lemma 1 to deduce Eq. 6 as follows. First, notice that if \( \text{rank}(G_I) = k = |I| \), then Eq. 7 implies that the diagonal of \( S_C(G_I) \) is zero. To wit, we have \( S_C(G_I))_{ii} = \det(G_{iI})/\det(G_I) \), with the numerator equal to the determinant of the \((k + 1)\)-dimensional principal submatrix of a positive-definite matrix of rank \( k \), and hence zero. However, it is known that positive definiteness of \( G \) implies positive definiteness of \( S_C(G_I) \) for any multi-index \( I \) (9), allowing us to conclude that \( S_C(G_I) \) is identical zero if \( \text{rank}(G_I) = \text{rank}(G) = k \).

**Randomized Multi-index Selection by Weighted Sampling.** Our first algorithm for selecting a multi-index \( I \) rests on the observation that since \( G \) is positive definite, it induces a probability distribution on the set of all \( I : |I| = k \) as follows:

\[
p_{G,k}(I) := Z^{-1} \det(G_I), \tag{8}
\]

where \( Z = \sum_{|I| = k} \det(G_I) \) is a normalizing constant.

Our corresponding algorithm for low-rank kernel approximation consists of first selecting \( I \) by sampling \( I \sim p_{G,k}(I) \) according to Eq. 8, and then implementing the Nyström extension to obtain \( \tilde{G} \) from \( G_I \) and \( G_{1,\ldots,l} \) in analogy to Eqs. 2 and 3. This algorithm is well behaved in the sense that if \( G \) is of rank \( k \) and we seek a rank-\( k \) approximant \( \tilde{G} \), then \( \tilde{G} = G \) and we realize the potential for perfect reconstruction afforded by the Nyström extension. Indeed, \( \det(G_I) \neq 0 \) implies that \( \text{rank}(G_I) = k \), and so Eq. 6 in turn implies that \( \|G_I - G_{1,\ldots,l}G_{l,\ldots,1}G_{1,\ldots,l}\| = 0 \) when \( \text{rank}(G) = k \).

For the general case whereupon \( \text{rank}(G) \geq k \), we have the following error bound in expectation:

**Theorem 1.** Let \( G \) be a real, \( n \times n \), positive quadratic form with eigenvalues \( \lambda_1 \geq \cdots \geq \lambda_n \). Let \( \tilde{G} \) be the Nyström approximation to \( G \) corresponding to \( I \), with \( I \sim p_{G,k}(I) \). Then

\[
\mathbb{E}\|G - \tilde{G}\| \leq (k + 1) \sum_{l=k+1}^n \lambda_l, \tag{9}
\]

**Proof:** By Eq. 5, we seek to bound

\[
\mathbb{E}\|G - \tilde{G}\| = \sum_{l:|I|=k} \frac{1}{\det(G_I)} \sum_{l:|I|=k} \det(G_I)\|S_C(G_I)\|.
\]

Denote the eigenvalues of \( S_C(G_I) \) as \( \mu_j \), positive definiteness and subadditivity of the square root imply that

\[
\|S_C(G_I)\| \leq \sum_j \mu_j^2 \leq \sum_j \tilde{\mu}_j = \text{tr}(S_C(G_I)). \tag{10}
\]

The Crabtree–Haynsworth characterization of Lemma 1 yields

\[
\text{tr}(S_C(G_I)) = \sum_{i\in l} \det(G_{i|I}) \det(G^\dagger_{i|I}),
\]

and thus

\[
\mathbb{E}\|G - \tilde{G}\| \leq \frac{1}{Z} \sum_{l:|I|=k} \sum_{l:|I|=k} \det(G_{i|I}) \tag{11}
\]

where we recall that \( Z = \sum_{|I| = k} \det(G_I) \).

Every multi-index of cardinality \( k + 1 \) appears exactly \( k + 1 \) times in the double sum of 11, whence

\[
\mathbb{E}\|G - \tilde{G}\| \leq \frac{(k + 1)}{Z} \sum_{l:|I|=k} \det(G_I). \tag{12}
\]

As \( G \) is an SPSD matrix, the Cauchy–Binet Theorem tells us that the sum of its principal \((k + 1)\)-minors can be expressed as the sum of \((k + 1)\)-fold products of its ordered eigenvalues:

\[
\sum_{l:|I|=k} \det(G_I) = \sum_{l:|I|=k} \lambda_1 \lambda_2 \cdots \lambda_{k+1}.
\]

It thus follows that

\[
\sum_{l:|I|=k} \det(G_I) \leq \sum_{l:|I|=k} \lambda_1 \lambda_2 \cdots \lambda_{k+1} \sum_{l:|I|=k} \lambda_l = \sum_{l:|I|=k} \det(G_I) \sum_{l:|I|=k} \lambda_l.
\]

Combining the above relation with 12, we obtain

\[
\mathbb{E}\|G - \tilde{G}\| \leq \frac{(k + 1)}{Z} \sum_{l:|I|=k} \det(G_I) \sum_{l:|I|=k} \lambda_l = (k + 1) \sum_{l:|I|=k} \lambda_l,
\]

which concludes the proof.

**Deterministic Multi-index Selection by Sorting.** Theorem 1 provides for an SPSD approximant \( \tilde{G} \) such that \( \mathbb{E}\|G - \tilde{G}\| \leq (k + 1) \sum_{l=k+1}^n \lambda_l \) in the Frobenius norm, compared with the optimal deterministic result \( \|G - \tilde{G}\| = \sqrt{(k + 1) \lambda_k} \) afforded by the full spectral decomposition. However, this probabilistic bound raises two practical algorithmic issues. First of all, sampling from the probability distribution \( p_{G,k}(I) \propto \det(G_I) \), whose support has cardinality \( \binom{n}{k} \), does not necessarily offer any computational savings over an exact spectral decomposition—a consideration we address in detail later, through the introduction of approximate sampling methods.

Moreover, in certain situations, practitioners may require a greater level of confidence in the approximation than is given by a bound in expectation. Although we cannot necessarily hope to preserve the quality of the bound of Theorem 1, we may sacrifice its power to obtain corresponding gains in the deterministic nature of the result and in computational efficiency. To this end, our deterministic algorithm for low-rank kernel approximation consists of letting \( I \) contain the indices of the \( k \) largest diagonal elements of \( G \) and then implementing the Nyström extension analogously.
to Eqs. 2 and 3. The following theorem bounds the corresponding worst-case error:

**Theorem 2.** Let $G$ be a real positive-definite kernel, let $I$ contain the indices of its $k$ largest diagonal elements, and let $\tilde{G}$ be the corresponding Nyström approximation. Then

$$
\|G - \tilde{G}\| \leq \sum_{i \notin I} G_{ii}.
$$

[13]

The proof of Theorem 2 is straightforward, once we have the following generalization of the Hadamard inequality (9):

**Lemma 2 [Fischer’s Lemma].** If $G$ is a positive-definite matrix and $G_I$ a nonsingular principal submatrix then

$$
\det(G_{I\setminus I}) < \det(G_I)G_{ii}.
$$

**Proof of Theorem 2:** We have from Eq. 10 that $\|G - \tilde{G}\| \leq \text{tr}(S_C(G_I))$; applying Lemma 1 in turn gives

$$
\|G - \tilde{G}\| \leq \frac{1}{\det(G_I)} \sum_{i \notin I} \det(G_{I\setminus I(i)}).
$$

after which Lemma 2 yields the final result.

While yielding only a worst-case error bound, this algorithm is easily implemented and appears promising in the context of array signal processing (11). Beginning with the case $k = 1$, it may be seen through repeated application of Theorem 2 to constitute a simple stepwise-greedy approximation to optimal multi-index selection.

**Remarks and Discussion.** The Nyström extension, in conjunction with efficient techniques for multi-index selection, hence provides a means of approximate spectral analysis in situations where the exact eigendecomposition of a positive-definite kernel is prohibitively expensive. As a strategy for dealing with very large, high-dimensional datasets in the context of both the classical and contemporary statistical analysis techniques described earlier, this approach lends itself easily to a straightforward implementation in practical settings, and also carries with it the accompanying performance guarantees of Theorems 1 and 2 through the two algorithms presented above.

In considering the performance and complexity of these 2 algorithms, we first compare them with the only other result known to us for explicitly quantifying the approximation error of an SPSD matrix using the Nyström extension (12). This algorithm consists of choosing row/column subsets by sampling, independently and with replacement, indices in proportion to elements of $(G^2_{ii})^m_{i=1}$, the squares of the main diagonal entries of $G$. The resultant probabilistic bound is written to include the possibility of sampling $c \geq k$ indices to obtain a rank-$k$ approximation obeying (in Frobenius norm)

$$
\mathbb{E}\|G - \tilde{G}\| \leq \|G - G_k\| + \frac{2\sqrt{2}}{\sqrt{c/k}} \sum_{i=1}^{n} G_{ii}^2,
$$

[14]

an additive error bound relative to that of the optimal rank-$k$ approximation $G_k$ obtained via exact spectral decomposition.

Two important points follow from a comparison of the bounds of our Theorems 1 and 2 with that of 14. First, inspection of 13 (Theorem 2) and 14 reveals that a conservative sufficient condition for the former to improve upon the latter when $c = k$ is that $\text{tr}(G) \geq n$ (also bearing in mind that the 13 is deterministic, whereas 14 holds only in expectation). A comparison of 9 (Theorem 1) and 14 reveals the more desirable relative form of the former, which involves only the $(n - k)$ smallest eigenvalues of $G$ and avoids an additive error term. Recall that Eq. 9 also guarantees zero error for an approximation whose rank $k$ equals the rank of $G$.

A direct implementation of Theorem 1, however, requires sampling from $p_{G_k}(I)$, which may be computationally infeasible. In the sequel we demonstrate that an approximate sampling is sufficient to outperform other algorithms for SPSD kernel approximation. Moreover, a sharp decrease in error is observed in simulations when $k$ meets or exceeds the effective rank of $G$. This feature is especially desirable for modern spectral methods such as those described in the introduction, which yield very large matrices of low effective rank; whereas the number of data points $n$ determines the dimensionality of the kernel matrix $G$, its effective rank is given by the number of components of the manifold $M$ from which the data are sampled plus $\dim(M)$, a sum typically much smaller than $n$.

We also remark on similarities and differences between our strategies and ongoing work in the theoretical computer science community to derive complexity-class results for randomized low-rank approximation of arbitrary $m \times n$ matrices. Though our goals and corresponding algorithms are quite different in their approach and scope of application, it is of interest to note that our Theorem 1 can in fact be viewed as a kernel-level version of a theorem of ref. 13, where a related notion termed *volume sampling* is employed for column selection. However, in ref. 13, as in the seminal work of ref. 6 and others building upon it, approximations are obtained by applying linear projections to the approximand; although different algorithms define different projections, they do...
not in general guarantee the return of an SPSD approximant when applied to an SPSD matrix. The same holds true for approaches motivated by numerical analysis; in recent work, the authors of ref. 14 apply the method of ref. 15 to obtain a low-rank approximation termed an interpolative decomposition, and focus on its use in obtaining accurate and stable approximations to matrices with low numerical rank.

With reference to these various lines of work, we remark that a projection method applied to a matrix $A$ can naturally be related to the Nyström extension applied to $AA^T$, though in our application setting it is of specific interest to work directly with the kernel in question. In particular, our results indicate how, by restricting to quadratic forms, one is able to exploit more specialized results from linear algebra than in the case of arbitrary rectangular matrices. We refer the reader to ref. 12 for an extended discussion of the various differences between projection-based approaches and the Nyström extension.

We conclude these remarks with a discussion of the computational complexity of the above algorithms for spectral decomposition. Recall that an exact spectral decomposition requires $O(n^3)$ operations, with algorithms specialized for sparse matrices running in time $O(n^2)$ (4). The deterministic algorithm of Theorem 2 requires finding the $k$ largest diagonal elements of $G$, which can be done in $O(n \log k)$ steps. In analogy to Eq. 2, the subsequent spectral decomposition of $G_k = U_k \Lambda_k U_k^T$ can be done in $O(k^2)$, and the final step of calculating $G_{k+1} = U_k \Lambda_k^{-1}$ requires time $O((n-k)k^2 + k^3)$, as $\Lambda_k$ is diagonal. The total running time of this deterministic algorithm is hence $O(n \log k + kn^2 + (n-k)k^3)$, which compares favorably with previously known methods when $k$ is small. The algorithm of Theorem 1 selects multi-index $I$ at random, and thus the sorting complexity $O(n \log k)$ is replaced by the complexity of sampling from $p_{G_k}(I) \propto \det(G_k)$. Below we describe an approximate sampling technique based on stochastic simulation whose complexity is $O(k^2)$, owing to the computation of determinants, with a multiplicative constant depending on the precise simulation method employed.

Numerical Implementation and Simulation Results

We now detail the implementation of our algorithms, and present simulation results for cases of practical interest that are representative of recent and more classical methods in spectral machine learning. Though simulations imply the adoption of a measure on the input space of SPSD matrices, our results hold for every SPSD matrix.

We first describe an approximate sampling technique adopted as an alternative to sampling directly from $p_{G_k}(I)$ according to Eq. 8. Among several standard approaches (16), we chose to employ the Metropolis algorithm to simulate an ergodic Markov chain that admits $p_{G_k}(I)$ as its equilibrium distribution, via a traversal of the state space $\{I : |I| = k\}$ according to a straightforward proposal step that seeks to exchange one element of $I$ with one of $I$ at each iteration. We made no attempt to optimize this choice, as its performance in practice was observed to be satisfactory, with distance to $p_{G_k}(\cdot)$ in total variation norm typically observed.
to be small after on the order of 50\(|I|\) iterations of the chain. This approximate sampling technique yields a complete algorithm for low-complexity spectral analysis, as described in the algorithm above and implemented in subsequent experiments.†

Our first experiment was designed to evaluate the relative approximation error \(20 \log_{10} \|G - \hat{G}\|/\|G\|\) incurred by the Nystrom extension for the randomized algorithms of Theorem 1 and ref. 12. To do so we simulated \(G\) from the ensemble of Wishart matrices‡ according to \(G = G_1 + 5 \times 10^{-2} G_2\), where \(G_1 \sim \mathcal{W}_1(I, n)\) and \(G_2 \sim \mathcal{W}_1(I, n)\); all generated matrices \(G\) were thus SPSD and of full rank, but with their \(k\) principal eigenvalues significantly larger than the remainder. We set \(n = 500\) and \(k = 50\), and averaged over 10,000 matrices drawn at random, with outputs averaged over 100 trials for each realization of \(G\). A third algorithm indicating the Nystrom extension’s baseline performance was provided by selecting a multi-index of cardinality \(k\) uniformly at random. Fig. 1 shows the comparative results of these three algorithms, with that of Theorem 1 (implemented using the algorithm described above) outperforming that of (12), whose sampling in proportion to \(\hat{G}_i\) fails to yield an improvement over the baseline method of sampling uniformly over the set of all multi-indices. Additionally, for approximants of rank 50 or higher, we observe a marked decline in approximation error for the algorithm of Theorem 1, as expected according to Eq. 6.

In a second experiment, we compared the performance of these 3 algorithms in the context of nonlinear embeddings. To do so we sampled 500 points uniformly at random from the unit circle, and then computed the approximate spectral decomposition of the 500-dimensional matrix required by the diffusion maps algorithm of ref. 5. Corresponding kernel approximation errors in the Frobenius norm were measured for each of the randomized algorithms used the Nyström extension to transfer the main computational burden from one of kernel eigen-analysis to a combinatorial task of partition selection, thereby rendering the overall approximation problem more amenable to quantifiable complexity-precision trade-offs. We then presented 2 new algorithms to determine a partition of the kernel prior to Nyström approximation, with one employing a randomized approach to multi-index selection and the other a rank statistic. For the former, we gave a relative error bound in expectation for positive-definite kernel approximation; for the latter, we bounded its deterministic worst-case error. We also detailed a practical implementation of our algorithms and verified via simulations the improvements in performance yielded by our approach. In cases where optimal approaches rely on an exact spectral decomposition, our results yield strategies for very large datasets, and come with accompanying performance guarantees. In this way they provide practitioners with direct access to spectral methods for large-scale machine learning and statistical data analysis tasks.

**Summary**

In this article we have introduced two alternative strategies for the approximate spectral decomposition of large kernels, and demonstrated their applicability to machine learning tasks. We used the Nyström extension to transfer the main computational burden from one of kernel eigen-analysis to a combinatorial task of partition selection, thereby rendering the overall approximation problem more amenable to quantifiable complexity-precision trade-offs. We then presented 2 new algorithms to determine a partition of the kernel prior to Nyström approximation, with one employing a randomized approach to multi-index selection and the other a rank statistic. For the former, we gave a relative error bound in expectation for positive-definite kernel approximation; for the latter, we bounded its deterministic worst-case error. We also detailed a practical implementation of our algorithms and verified via simulations the improvements in performance yielded by our approach. In cases where optimal approaches rely on an exact spectral decomposition, our results yield strategies for very large datasets, and come with accompanying performance guarantees. In this way they provide practitioners with direct access to spectral methods for large-scale machine learning and statistical data analysis tasks.

**ACKNOWLEDGMENTS.** M.-A.B. thanks Roger Brockett for valuable discussions. This work is supported in part by Defense Advanced Research Projects Agency Grant HR0011-07-1-0007 and by National Science Foundation Grants DMS-0631636 and CBET-0730389.


