On the Approximation of Quadratic Forms and Sparse Matrix Products
New bounds and algorithms for high-dimensional data sets

Patrick J. Wolfe (joint work with Mohamed-Ali Belabbas)

Statistics and Information Sciences Laboratory
School of Engineering and Applied Sciences
Department of Statistics, Harvard University

Isaac Newton Institute for Mathematical Sciences
Cambridge University, 03 June 2008
1 Introduction
   - Matrix products
   - Spectral methods
   - The Nyström extension

2 Matrix Products and Low-Rank Approximations
   - Optimal rescaling of product terms
   - Subset selection methods
   - Non-adaptive random projections

3 Numerical Results and Algorithmic Implementation
   - Experimental results
   - Conclusions
Outline

Approximation of matrix products

1. Introduction
   - Matrix products
   - Spectral methods
   - The Nyström extension

2. Matrix Products and Low-Rank Approximations
   - Optimal rescaling of product terms
   - Subset selection methods
   - Non-adaptive random projections

3. Numerical Results and Algorithmic Implementation
   - Experimental results
   - Conclusions
Outline
Approximation of matrix products

1 Introduction
- Matrix products
- Spectral methods
- The Nyström extension

2 Matrix Products and Low-Rank Approximations
- Optimal rescaling of product terms
- Subset selection methods
- Non-adaptive random projections

3 Numerical Results and Algorithmic Implementation
- Experimental results
- Conclusions
Matrix Products
Matrix products and low-rank approximation of positive kernels

- Evaluating products of large matrices is a very common operation:
  - The naive algorithm (taking inner products of rows and columns) has complexity $O(n^3)$
  - The best known algorithm has complexity $O(n^{2.38})$
  - The existence of $O(n^2)$ algorithms is an open question

- Here we address the approximation of such products by sparse sampling of the matrices involved:
  - We first reduce the problem to one of approximating a positive definite kernel derived from a Hadamard product
  - We then give a randomized and a deterministic algorithm together with error bounds and experimental results
Consider matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, with $A_i$ denoting the $i^{\text{th}}$ column of $A$ and $B_i^T$ the $i^{\text{th}}$ row of $B$.

Then the matrix product $AB$ can be written as a sum of rank-one operators as follows:

$$AB = \sum_{i=1}^{n} A_i B_i^T.$$ 

An approximation is given by keeping (and possibly rescaling) only a few terms in the sum.

We will show that optimal row/column selection and rescaling both depend on approximating an induced quadratic form

$$Q := (A^T A) \odot (BB^T),$$

with $\odot$ the Hadamard (i.e., entrywise) product.
Why care about approximating large matrices and/or their products? Some examples from statistical signal processing:

- Number of interferers in radar processing, image segmentation, low-dimensional embeddings, . . .

Key question: What is the “relevant” information contained in the (potentially very large) data set or signal?

- Spectral methods reduce this question to finding a low-rank approximation to a Symmetric Positive Definite ( SPD) matrix

Kernel-based methods continue to see wide use:

- Older methods: principal components analysis (1901), multidimensional scaling (1958), . . .
- Newer methods: isomap, Laplacian eigenmaps, Hessian eigenmaps, diffusion maps, . . .
An SPD matrix $Q$ can be written in *spectral coordinates*

$$Q = U\Lambda U^T,$$

where $U$ is orthogonal and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ is diagonal.

The $\lambda_i$’s are the eigenvalues of $Q$, ordered such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$, and the $u_i$’s are the eigenvectors.

For any unitarily invariant norm $\| \cdot \|$, we have that

$$\arg\min_{\tilde{Q}: \text{rank}(\tilde{Q})=k} \| Q - \tilde{Q} \| = U\Lambda_k U^T =: Q_k,$$

where $\Lambda_k = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k, 0, \ldots, 0)$. 

Finding a low-rank approximation is easy...
How to Approximate an SPD Matrix, in Practice?
Finding a low-rank approximation is hard!

Changing to spectral coordinates is done using the Singular Value Decomposition of $Q$, which requires $O(n^3)$ operations.

- On a Pentium IV 3GHZ desktop PC, with 1GB RAM, 512k Cache:
- Extrapolating to $n = 10^6$, factoring $Q$ takes more than 4 months.
- When $n$ increases, $Q$ quickly becomes too large to be stored in memory.

![Graph showing the elapsed time vs. size of the SPD Matrix](image_url)
This presents a practical problem for large data sets!

- A commonly used “trick” is to **sparsify** the kernel.
  - Fix $\varepsilon > 0$. If $Q_{ij} \leq \varepsilon$, set $Q_{ij} = 0$
  - Questions: How to choose $\varepsilon$? How accurate is the result?

- Alternative approach: discard some of the data.
  - How to construct a low-rank approximation using just some of the data? The **Nyström extension** provides an answer

- The idea is as follows:
  - Write $Q = X^T X$, so that $Q$ is a **Gram matrix** for vectors $X_1, \ldots, X_n$.
  - Choose a subset $J$ of vectors $X_i$ and their correlations with all the other vectors to find a low-rank approximation $\tilde{Q}$.
Historically, the Nyström extension was introduced to obtain numerical solutions to integral equations.

Let $k : [0, 1] \times [0, 1] \to \mathbb{R}$ be an SPD kernel and $(f_i, \lambda_i)$, $i \in \mathbb{N}$, denote its pairs of eigenfunctions and eigenvalues:

$$\int_0^1 k(x, y)f_i(y) \, dy = \lambda_i f_i(x), \quad i \in \mathbb{N}.$$ 

The Nyström extension approximates the eigenvectors of $k(x, y)$ based on a discretization of the interval $[0, 1]$.

- Define the $M + 1$ points $x_i$ by $x_m = x_{m-1} + 1/M$ with $x_0 = 0$, so that the $x_i$’s are evenly spaced along the interval $[0, 1]$.
- Form the Gram matrix $K_{mn} := k(x_m, x_n)$ and decompose it.
The Nyström Extension

Extend the solution

- We now solve a finite-dimensional problem

$$\frac{1}{M+1} \sum_n K_{mn} v_i(n) = \lambda_i^v v_i(m), \quad i = 0, 1, \ldots, M,$$

where \((v_i, \lambda_i^v)\) represent the eigenvector-eigenvalues pairs associated with \(K\).

- What do we do with these eigenvectors? We extend them to give an estimate \(\hat{f}_i\) of the \(i^{th}\) eigenfunction as follows:

$$\hat{f}_i(x) = \frac{1}{(M+1)\lambda_i^v} \sum_m k(x, x_m) v_i(m).$$

- In essence: only use partial information about the kernel to solve a simpler eigenvalue problem, and then to extend the solution using complete knowledge of the kernel.
The Nyström Extension
In finite dimensions

How do we translate this to a finite dimensional setting?

We approximate $k$ eigenvectors of an SPD matrix $Q$ by decomposing and extending a $k \times k$ principal submatrix of $Q$.

We partition $Q$ as follows:

$$ Q = \begin{bmatrix} Q_J & Y \\ Y^T & Z \end{bmatrix}, $$

with $Q_J \in \mathbb{R}^{k \times k}$; we say that this partition corresponds to the multi-index $J = \{1, 2, \ldots, k\}$.

Define spectral decompositions $Q = U \Lambda U^T$, $Q_J = U_J \Lambda_J U_J^T$. 
The Nyström Extension

The approximation error

- The Nyström extension then provides an approximation for $k$ eigenvectors in $U$ as

$$\tilde{U} := \begin{bmatrix} U_J & \Lambda_J^{1/2} \\ Y^T U_J \Lambda_J^{1/2} \end{bmatrix}; \quad Q_J = U_J \Lambda_J U_J^T.$$

- In turn, the approximations $\tilde{U} \approx U$ and $\Lambda_J \approx \Lambda$ may be composed to yield an approximation $\tilde{Q} \approx Q$ according to

$$\tilde{Q} := \tilde{U} \Lambda_J \tilde{U}^T = \begin{bmatrix} Q_J & Y \\ Y^T & Y^T Q_J^{-1} Y \end{bmatrix}.$$

- The resultant approximation error is

$$\| Q - \tilde{Q} \| = \| Z - Y^T Q_J^{-1} Y \|,$$

the norm of the Schur complement of $Q_J$ in $Q$. 
Outline

Approximate Linear Algebra

1. Introduction
   - Matrix products
   - Spectral methods
   - The Nyström extension

2. Matrix Products and Low-Rank Approximations
   - Optimal rescaling of product terms
   - Subset selection methods
   - Non-adaptive random projections

3. Numerical Results and Algorithmic Implementation
   - Experimental results
   - Conclusions
How are matrix products and the Nyström extension related?

- Consider again $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}$, with $A_i$ denoting the $i^{th}$ column of $A$ and $B^i$ the $i^{th}$ row of $B$.

- Then the matrix product $AB$ can be written as a sum of rank-one operators: $AB = \sum_i A_i B^i$. An approximation is given by keeping (and rescaling) only a few terms in the sum:

$$AB \approx \sum_{i \in J} w_i A_i B^i.$$

- This leads to two immediate issues:
  - Which terms to keep? (How to choose $J$?)
  - How to rescale them? (What is $w$?)
For fixed $J$, we may characterize the optimal rescaling as follows:

**Theorem (Belabbas & W, 2008)**

Let the $n \times n$ SPD matrix $Q = \begin{bmatrix} Q_J & Y \\ Y^T & Z \end{bmatrix}$ be defined element-wise as $Q_{ij} = \langle A_i, A_j \rangle \langle B^i, B^i \rangle$, where $J = \{1, 2, \ldots, k\}$ without loss of generality, and $Q_J$ is the corresponding principal submatrix.

Then the best product approximation using terms $\{A_i B^i\}_{i \in J}$ is given by

$$AB \approx \tilde{AB} := \sum_{i \in J} w_i A_i B^i,$$

with $w := Q_J^{-1} r$, and $r_i := \sum_{j=1}^n \langle A_i, A_j \rangle \langle B^i, B^j \rangle$.

Moreover, if $E$ is the $(n-k) \times (n-k)$ matrix with all entries 1, then

$$\|AB - \tilde{AB}\|^2 = \text{tr}(S_C(Q_J)E) \quad \text{with} \quad S_C(Q_J) := Z - Y^T Q_J^{-1} Y.$$
Approximating the Optimal Subset Selection Procedure
Randomized and Deterministic Approaches

- This result tells us how well we can approximate the product $AB$ granted that we know only a few rows/columns of $A$ and $B$, and their correlations with the remaining rows & columns.

- It also allows us to characterize the best subset $J$ of size $k$ as that which minimizes $\text{tr}(S_C(Q_J)E)$.

- We therefore seek a subset $J$ whose associated Schur complement $S_C(Q_J)$ has the lowest possible power along the one-dimensional subspace of $\mathbb{R}^{n-k}$ spanned by $[1 \ 1 \ \cdots \ 1]$.

- We now present two approximations to this optimal subset selection problem: one based on a random choice of subsets, and an alternative “greedy” approach which yields a worst-case error bound.
Recall that, conditioned upon having chosen a subset $J \subset \{1, \ldots, n\}$, our squared approximation error is given by

$$\|AB - \tilde{AB}\|^2 = \text{tr}(S_C(Q_J)E).$$

Since $S_C(Q_J)$ is positive definite, we have that $\text{tr}(S_C(Q_J))$ is larger than the largest eigenvalue of $S_C(Q_J)$, and we can bound this squared error term as follows:

$$\text{tr}(S_C(Q_J)E) = [1 \ 1 \ \cdots \ 1] S_C(Q_J) [1 \ 1 \ \cdots \ 1]^T \leq (n - k) \|S_C(Q_J)\|_2 \leq (n - k) \text{tr}(S_C(Q_J)).$$

Note that equality is obtained when $S_C(Q_J) \propto E$, and hence this bound is tight.
Recall that we defined $Q_{ij} = \langle A_i, A_j \rangle \langle B^i, B^j \rangle$, such that $Q$ is the Hadamard product of SPD matrices $A^T A$ and $BB^T$.

By the Schur Theorem, $Q$ is also SPD, implying:

- There exists a matrix $X \in \mathbb{R}^{n \times n}$ such that $Q = X^T X$;
- All the principal minors $\det(Q_J)$ of $Q$ are positive.

We may thus define a probability distribution over multi-indices $J \subset \{1, \ldots, n\}, \vert J \vert = k$ according to

$$p_{Q,k}(J) := \frac{1}{K} \det(Q_J),$$

where $K = \sum_{J, \vert J \vert = k} \det(Q_J)$ is a normalizing constant.

Terms in $p_{Q,k}(J)$ are hence proportional to diagonal elements of the so-called $k$th compound matrix of $Q$, arising from the skew-symmetric tensor product in multilinear algebra.
The Cauchy-Binet Theorem for diagonalizing compound matrices, along with the Crabtree-Haynsworth Lemma characterizing Schur complements, then yields the following result:

**Theorem (Belabbas & W, 2007)**

Let $Q \in \mathbb{R}^{n \times n}$ be a positive quadratic form with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$. If multi-index $J$ of cardinality $k$ is chosen with probability proportional to $\det(Q_J)$, then we have that

$$
\mathbb{E} \operatorname{tr}(S_C(Q_J)) \leq (k + 1) \sum_{i=k+1}^{n} \lambda_i.
$$

This result appears to be the first relative-error bound for the approximation of SPD kernels.
Via Jensen’s inequality and our previous result, this theorem leads to an upper bound on the average error in terms of $X : Q = X^T X$ and its optimal rank-$k$ approximant $X_k$:

$$\mathbb{E} \|AB - \tilde{A}B\| \leq \sqrt{(n - k)(k + 1)} \|X - X_k\|.$$ 

The term $\|X - X_k\|$ may be further characterized by majorization results for SPD matrices as follows:

$$\|X - X_k\|^2 \leq \min (\sigma_1^2(A)\|B\|^2, \sigma_1^2(B)\|A\|^2) - \|X_k\|^2.$$ 

While sampling directly in proportion to the determinants of submatrices of cardinality $k$ is clearly infeasible, approximate sampling is possible via Markov chain Monte Carlo methods.
By bounding the maximal rather than average error, we obtain a deterministic result and corresponding “greedy” algorithm.

Since the error term is the sum of all the terms in the Schur complement, we can look to bound its largest element:

Lemma (Belabbas & W, 2008)

1. The largest entry in $S_C(Q_J) := Z - Y^T Q_J^{-1} Y$ is smaller than the largest diagonal element of $Z$.

2. Choosing $J$ to contain indices of the $k$ largest diagonal terms $Q_{ii} := \langle A_i, A_i \rangle \langle B^i, B^i \rangle$ yields the approximation error bound

$$\|AB - \widetilde{AB}\| \leq \sqrt{(n - k) \sum_{i \notin J} (\|A_i\|^2 \|B^i\|^2)}.$$

Wolfe (Harvard University)

Quadratic Forms and Matrix Products

Mar 2008 21 / 28
Non-Adaptive Random Projections
Assuming nothing is known about $A$ and $B$

- Under our previous assumptions, simple statistics about $A$ and $B$ were used to select a good partition. If we assume no such knowledge, however, the Johnson-Lindenstrauss Lemma provides us with a simple algorithm:

- If $W$ is a $k \times n$ matrix with independent unit Normal entries and $x, y \in \mathbb{R}^n$, then for $0 < \epsilon < 1$ we have that

$$\Pr \left( \left| \langle x, y \rangle - k^{-1} \langle Wx, Wy \rangle \right| \leq \epsilon \|x\| \|y\| \right) \geq 1 - 4e^{-\frac{k}{2} \left( \frac{\epsilon^2}{2} - \frac{\epsilon^3}{3} \right)}.$$

- If $A^i$ is the $i$th row of $A$, and $B_j$ the $j$th column of $B$, we have

$$(AB)_{ij} - k^{-1}(AW^T WB)_{ij} = \langle A^i, B_j \rangle - k^{-1} \langle WA^i, WB_j \rangle,$$

and thus we see that approximating $AB$ by $k^{-1}AW^T WB$ yields a good approximation with high probability.
Outline

Approximate Linear Algebra

1. Introduction
   - Matrix products
   - Spectral methods
   - The Nyström extension

2. Matrix Products and Low-Rank Approximations
   - Optimal rescaling of product terms
   - Subset selection methods
   - Non-adaptive random projections

3. Numerical Results and Algorithmic Implementation
   - Experimental results
   - Conclusions
We performed three sets of experiments, in which the performance of four subset selection methods was compared:

- A baseline uniform sampling on $k$-subsets
- Sampling according the row powers (Drineas et al., 2006)
- Sampling in proportion to the $k$-principal minors of $Q$
- Selecting $k$-subsets greedily ($O(m(n + k^2) + k^3)$)

We also compared the choice of reweighting following subset selection, in one case applying the optimal reweighting, and in the other simply reweighting according to the row powers:

$$\tilde{AB} = \sum_{i \in J} \frac{1}{\sqrt{|J||A_i||B_j|^2}} A_i B_j.$$

Wolfe (Harvard University)
Experimental Results
Subset selection methods, each with row power reweighting

- We drew 200 random matrices $A \in \mathbb{R}^{60 \times 15}$ and $B \in \mathbb{R}^{15 \times 90}$ in total, each having independent unit Normal entries.

- We then averaged the error of the randomized algorithm over 20 trials per matrix product, with relative error reported as

\[ 20 \log_{10} \left( \frac{\|AB - \tilde{AB}\|}{\|A\|\|B\|} \right) \]
Experimental Results
Subset selection methods, each with optimal reweighting

We followed the same procedure in a second set of experiments, but applied the optimal rescaling to selected row/column products.

Performance in this case is (as expected) better overall, but with the ordering of the methods unchanged.

Figure: Error using optimal reweighting
As a final experiment, we compared greedy subset selection followed by either rescaling with two non-adaptive methods:

- Choosing subsets uniformly at random and rescaling by $n/k$
- Computing random projections à la Johnson-Lindenstrauss

**Figure:** Adaptive vs. non-adaptive methods
Conclusions

Summary and future work

Analysis can establish trade-offs between complexity and precision for basic linear algebra operations:

- Conditioned on a choice of subset, the optimal rescaling term for approximating matrix products as the sum of rank-one operators was given.
- Two alternative strategies for subset selection based on the Nyström method were presented, along with accompanying algorithms and error bounds.
- Simulation studies demonstrated improvements in performance, relative to existing adaptive (and non-adaptive) methods.

Ongoing work is focused on applying these results to statistical signal processing and machine learning applications.