Low Rank Approximation using Error Correcting Coding Matrices

Shashanka Ubaru, Arya Mazumdar and Yousef Saad

University of Minnesota-Twin Cities, USA

July 8, 2015

International Conference on Machine Learning 2015, Lille, France

Outline

Low Rank Approximation

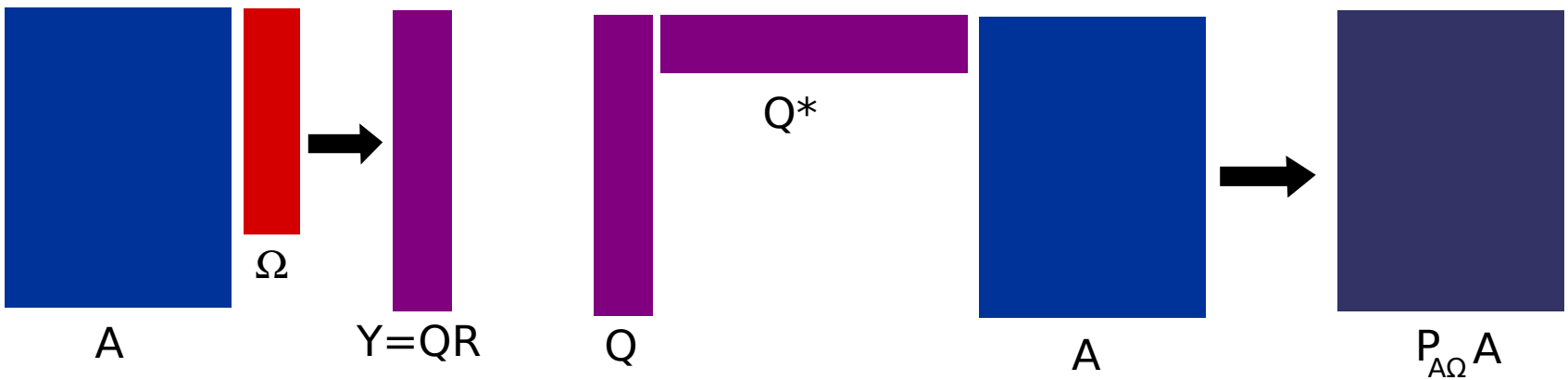
Goal:

Given a large input matrix $A \in \mathbb{R}^{m \times n}$ and $k \ll \min\{m, n\}$, find a low-rank approximation to A with rank about k.

- Low-rank matrix approximation is an integral component of tools such as principal component analysis (PCA).
- Many data processing applications: machine learning, computer vision, signal processing, recommender systems, information retrieval, web search modeling, DNA microarray data.
- Recently, research focussed on developing techniques that use randomization for computing low rank approximations and matrix decompositions of such matrices [Halko et al., 2011].

Randomized Techniques

- Compute a basis that approximately spans the range of A, using a sampling matrix Ω of size $n \times \ell$.
- Form a matrix $Y = A\Omega$ and compute the orthonormal basis $Y = QR$.
- Next, project A onto this space Q to reduce rank. Then we wish $P_{A\Omega}A = QQ^*A$ is close to A.
- Can also estimate approximate rank-k SVD of A from Q^*A.



Choice of sampling matrix

- **Fully random matrices:** ± 1 entries with equal probabilities, i.i.d. Gaussian matrices [Halko et al., 2011].
  1. Computation of $A\Omega$ in $O(mn\ell)$ time for general A.
  2. Need $O(n\ell)$ random numbers. Highly impractical for large matrices.
  3. The columns of A are well-mixed.
  4. $\ell = O(k)$.
- **Structured random matrices:** Subsampled randomized Fourier / Hadamard transform (SRFT/SRHT) [Woolfe et al., 2008].
  1. Computation of $A\Omega$ will take $O(mn \log_2 \ell)$.
  2. Need $O(n)$ random numbers.
  3. Mixing might be not as uniform, potential accuracy loss.
  4. $\ell = O(k \ln k)$.

Here, we show how matrices from error correcting codes can be used.

Error Correcting Code Matrices

- In digital communication, information is encoded by adding redundancy into codewords to facilitates detection and correction of errors due to noisy channels.
- Coding schemes generate codewords that maintain a fixed minimum distance between each other, called the **distance** of the code.
- A code is a subspace of a vector space, and the null-space of the code is another well-defined subspace, called the *dual* of the code.
- Minimum distance of the dual code is called the **dual distance** of the code.
- An $[\ell, r]$-code will have 2^r unique codewords of length ℓ. Form a code matrix Φ of size $2^r \times \ell$ by stacking up all codewords and setting $1 \to \frac{-1}{\sqrt{2^r}}$ and $0 \to \frac{1}{\sqrt{2^r}}$ (BPSK mapping).

Properties of Code Matrices

- Depending on the coding schemes used, the code matrices Φ have a variety of favorable properties. For e.g., low coherence used in Compressed Sensing.
- A code matrix Φ with the dual distance ≥ 3 has orthonormal columns.
- Codewords need to be widespread and distinct. Act like random numbers. Can define probability measures.
- A code with a dual distance $> k$ supports a k-**wise independent** probability measure.
- k-wise independent measure: Any k or fewer of n binary random variables are independent, equally likely to be $0/1$.
- In a k-wise independent code, for any k entries of each codeword $\mathbf{c} = \{c_{i_1}, c_{i_2}, \ldots, c_{i_k}\}$ and for any k bit binary string α, we have

$$\mathbf{Pr}[\mathbf{c} = \alpha] = 2^{-k}$$

Construction of Subsampled Code Matrix

$$\Omega = \sqrt{\frac{2^r}{\ell}} DS\Phi, \tag{1}$$

where

- D is a random $n \times n$ diagonal matrix whose entries are independent random signs, i.e., random variables uniformly distributed on $\{\pm 1\}$.
- S is the uniformly random downsampler, an $n \times 2^r$matrix whose n rows are randomly selected from a $2^r \times 2^r$ identity matrix.
- Φ is the $2^r \times \ell$ code matrix, generated using an $[\ell, r]$- coding scheme, with BPSK mapping and scaled by $2^{-r/2}$ such that all columns have unit norm.

Purpose of D is to flatten out input vectors, and of scaling is to make rows unit vectors.

Proto-Algorithm

Input: An $m \times n$ matrix A, and a target rank k.

Output: Rank-k factors U, Σ and V in an approximate SVD $A \approx U\Sigma V^*$.

1. Form an $n \times \ell$ subsampled code matrix Ω, as described in (1), using an $[\ell, r]-$ coding scheme, where $\ell > k$ and $r \geq \lceil \log_2 n \rceil$.

2. Form the $m \times \ell$ sample matrix $Y = A\Omega$.

3. Form an $m \times \ell$ orthonormal matrix Q such that $Y = QR$.

4. Form the $\ell \times n$ matrix $B = Q^*A$.

5. Compute the SVD of the small matrix $B = \hat{U}\Sigma V^*$.

6. Form the matrix $U = Q\hat{U}$.

- Algorithm requires two passes over A. Can also be modified to one pass. Better performance if $Y = (AA^*)^q A\Omega$, but requires $2(q+1)$ passes over A.

Cost Analysis

- Most structured codes can be decoded using Fast Fourier Transform. In $\mathbb{F}_2$, Φ from such codes have all columns as some columns of a $2^r \times 2^r$ Hadamard matrix (after BPSK mapping).
- Matrix product $Y = A\Omega$ with such code matrices can be achieved in $O(mn \log_2 \ell)$ time for a general dense A using the technique described in [Ailon and Liberty, 2009].
- If cyclic code scheme used, Ω is made of blocks of circulant matrices. Matrix product with circulant matrices can be achieved in $O(m\ell \log_2 \ell)$ time for each block (n/ℓ of them).

Cost

Total cost of the algorithm when **structured code** or **cyclic code** matrices are used will be $O(mn \log_2 \ell + k^2(m+n))$ for a general dense A.

Deterministic Error Bound

- A is $m \times n$ with SVD partition as

$$A = U \begin{array}{c} \begin{matrix} k & n-k \end{matrix} \\ \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \end{array} \begin{array}{c} \begin{matrix} n \end{matrix} \\ \begin{bmatrix} V_1^* \\ V_2^* \end{bmatrix} \begin{matrix} k \\ n-k \end{matrix} \end{array}.$$

- Test (sampling) matrix Ω, is decomposed as

$$\Omega_1 = V_1^* \Omega \quad \text{and} \quad \Omega_2 = V_2^* \Omega.$$

- For $Y = A\Omega$ and basis Q for $range(Y)$.

Theorem 9.1 in [Halko et al., 2011] *When* Ω_1 *is full rank,* $\xi \in \{2, F\}$,

$$\|A - QQ^*A\|_\xi \leq \|\Sigma_2\|_\xi^2 + \|\Sigma_2 \Omega_2 \Omega_1^\dagger\|_\xi^2.$$

Challenge is to show that Ω_1 will be **full rank**.

Connection to Johnson-Lindenstrauss Transform (JLT)

- A matrix $\Omega \in \mathbb{R}^{n \times \ell}$ is JLT(ϵ, δ) if for a vector $v \in \mathbb{R}^n$, it holds

$$(1-\epsilon)\|v\|^2 \leq \|v^*\Omega\|^2 \leq (1+\epsilon)\|v\|^2$$

  with probability $1-\delta$, under certain conditions on ℓ.
- [Sarlos, 2006, Corollary 11] *If Ω is a JLT from $\mathbb{R}^n$ to $O(k\ln(k/\epsilon))$, then for an orthonormal matrix $V \in \mathbb{R}^{n \times k}$,*

$$\mathbf{Pr}(\forall \in [1..k] : |1-\sigma_i(V^*\Omega)| \leq \epsilon) \geq 1-\delta$$

- [Ailon and Liberty, 2009] showed a matrix Ω which is 4-wise independent will be a JLT. Give dual BCH codes of distance 5 as examples. But have conditions on v.
- [Clarkson and Woodruff, 2009] showed if Ω is $4\lceil\log(\sqrt{(}2)/\delta)\rceil$-wise independent, then Ω will be a JLT.
- Thus, any error correcting code matrix with dual distance > 4 will preserve geometry of V (i.e., Ω_1 is full rank). But, need $\ell = O(k\ln(k/\epsilon))$.

Code matrices as sign matrices

- Can treat code matrices as random sign matrices with certain probabilistic measures.
- Indeed a code with dual distance above k supports a k-wise independent probability measure.
- [Clarkson and Woodruff, 2009, Lemma 3.4] *If* Ω *is* $\rho(k + \ln(1/\delta))$*-wise independent with a constant* $\rho > 1$ *then, with* $\ell = O(k \ln(1/\delta)/\epsilon)$ *and an orthonormal matrix* $V \in \mathbb{R}^{n \times k}$,

$$\|V^* \Omega \Omega^* V - I\|_2 \leq \epsilon$$

- Thus, a code matrix with dual distance $> k$ will preserve the geometry of V with $\ell = O(k)$.
- Significant theoretical result that shows $O(k)$ can be achieved in the number of samples required with structured random matrices.
- Example: for $k = 100$, need $\ell \approx 500$ for SRHT. Can use a 55 error correcting dual BCH code of length $\ell = 255$. Especially significant when input is sparse.

Error Analysis

- [Tropp, 2011] gives bounds on the singular values when orthonormal matrices are subsampled. The code matrix Φ will be orthonormal, if the dual distance of the code is ≥ 3. The top singular value of subsampled code matrix Ω will be $\sigma_1(\Omega) \leq \sqrt{\frac{(1+\eta)n}{\ell}}$ for a small $\eta > 0$.
- For any code matrices Ω with **dual distance** > 4 **and length** $\ell = O(k \ln(k/\epsilon))$, the approximation error satisfies for $\xi \in \{2, F\}$,

$$\|A - QQ^*A\|_\xi \leq \|A - A_k\|_\xi \sqrt{1 + \frac{(1+\eta)n}{(1-\epsilon)^2\ell}}$$

with failure probability δ.

- For any code matrices Ω with **dual distance** $> k + \ln(1/\delta)$ **and length** $\ell = O(k/\epsilon)$, the approximation error satisfies

$$\|A - QQ^*A\|_F \leq \|A - A_k\|_F(1+\epsilon)$$

with failure probability δ, using [Clarkson and Woodruff, 2009, Theorem 4.2].

Numerical Experiments

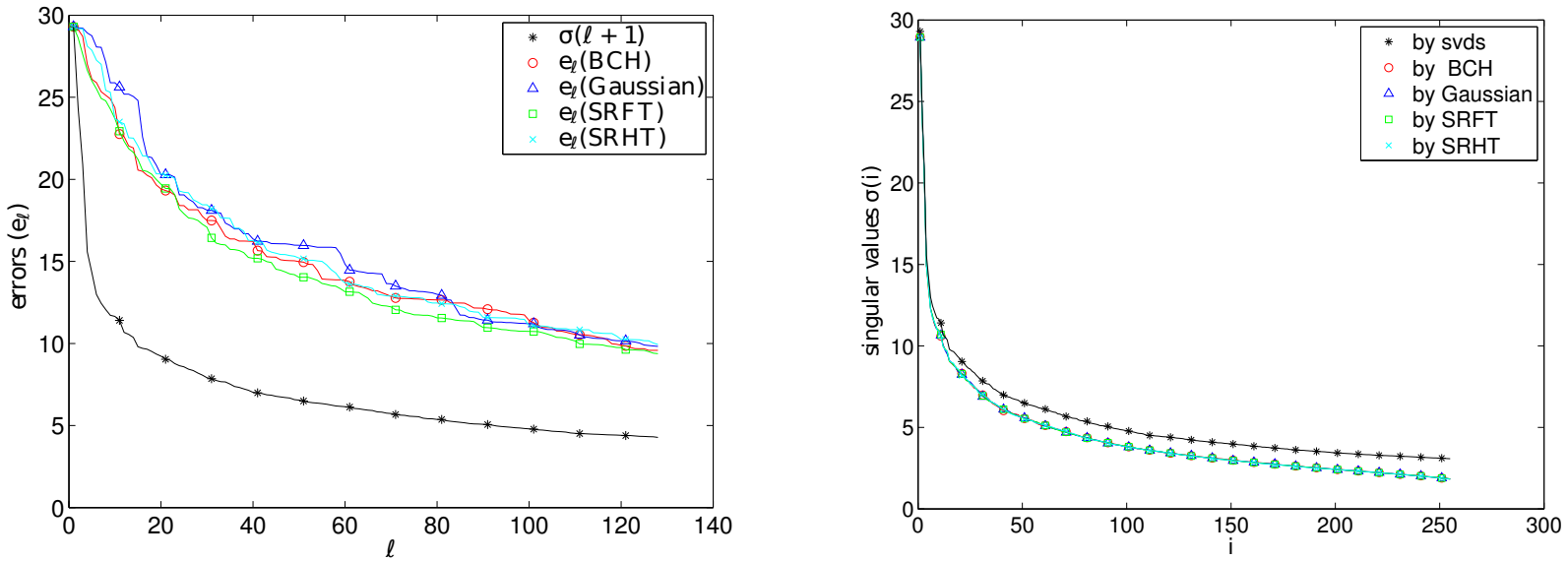


Figure: The approximate errors and the singular values obtained using dual BCH code, Gaussian, SRFT and SRHT matrices. Kohonen matrix 4772×4772 from University of Florida Database.

Numerical Experiments

Table: Comparison of errors

Matrix	ℓ	dual BCH	Gaussian	SRFT
lpiceria3d	31	21.8779	23.7234	23.3688
S80PI	63	3.8148	3.8492	3.7975
Delaunay	63	6.3864	6.3988	6.3829
lpiceria3d	63	15.4865	18.3882	16.3619
deter3	127	9.2602	9.2658	9.2984
EPA	255	5.5518	5.5872	5.4096
Kohonen	511	4.2977	4.2934	4.2610

Matrices from University of Florida Database. Matrices lpi_ceria3d (4400×3576) and deter3 (21777×7647) linear programming problems. S80PI_n1 (4028×4028) an eigenvalue/model reduction problem. Delaunay (4096×4096), EPA (4772×4772) and Kohonen are graph Laplacian matrices. $k = \ell - 5$ and average over 5 trials.

Conclusion

Properties	Gaussian	SRFT/SRHT	Code Matrices
Low Randomness $O(n)$		✓	✓
Fast Multiplication $O(mn \log_2 \ell)$		✓	✓
$\ell = O(k)$	✓		✓

- An introduction of coding theory and techniques to the applications of machine learning and data analysis.
- Code matrices perform equally well as fully random Gaussian matrices or complex Fourier matrices.
- Possible to improve the error bounds.
- Other applications where such code matrices may be used: matrix products, least squares regression, compressed sensing.

References

Ailon, N. and Liberty, E. (2009).

Fast dimension reduction using rademacher series on dual bch codes.

Discrete & Computational Geometry, 42(4):615–630.

Clarkson, K. L. and Woodruff, D. P. (2009).

Numerical linear algebra in the streaming model.

In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 205–214. ACM.

Halko, N., Martinsson, P., and Tropp, J. (2011).

Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions.

SIAM Review, 53(2):217–288.

Sarlos, T. (2006).

Improved approximation algorithms for large matrices via random projections.

In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 143–152. IEEE.

Tropp, J. A. (2011).

Improved analysis of the subsampled randomized hadamard transform.

Advances in Adaptive Data Analysis, 3(01n02):115–126.

Woolfe, F., Liberty, E., Rokhlin, V., and Tygert, M. (2008).

A fast randomized algorithm for the approximation of matrices.

Applied and Computational Harmonic Analysis, 25(3):335–366.

Questions?