

# Randomized Matrix Computations Lecture 6

Daniel Kressner

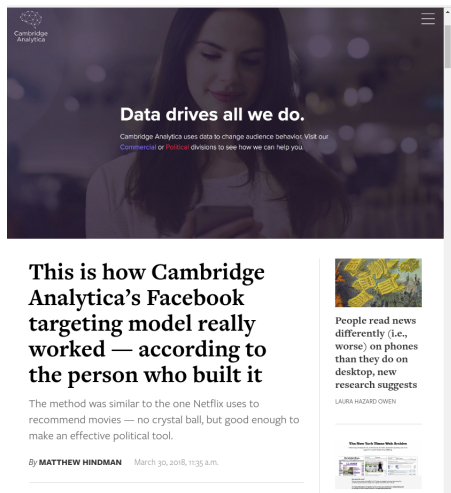
Chair for Numerical Algorithms and HPC

Institute of Mathematics, EPFL

`daniel.kressner@epfl.ch`



From <http://www.niemanlab.org>



**Cambridge Analytica**

**Data drives all we do.**

Cambridge Analytica uses data to change audience behavior. Visit our [Commercial](#) or [Political](#) divisions to see how we can help you.

## This is how Cambridge Analytica's Facebook targeting model really worked — according to the person who built it

The method was similar to the one Netflix uses to recommend movies — no crystal ball, but good enough to make an effective political tool.

By **MATTHEW HINDMAN** March 30, 2018, 11:35 a.m.

People read news differently (i.e., worse) on phones than they do on desktop, new research suggests

LAURA HAZARD OWEN

**New York Times Web Site**

... his [Aleksandr Kogan's] message went on to confirm that his approach was indeed similar to **SVD or other matrix factorization** methods, like in the Netflix Prize competition, and the Kosinski-Stillwell-Graepel Facebook model. **Dimensionality reduction** of Facebook data was the core of his model.

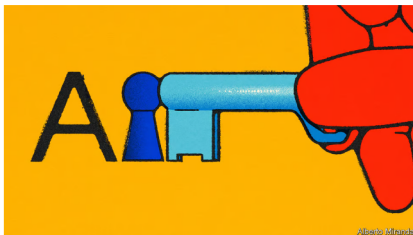
# Leaked Internal Google Document, May 2023



Leaders | A stochastic parrot in every pot

## What does a leaked Google memo reveal about the future of AI?

Open-source AI is booming. That makes it less likely that a handful of firms will control the technology



But the uncomfortable truth is, we aren't positioned to win this arms race and neither is OpenAI. While we've been squabbling, a third faction has been quietly eating our lunch... Open-source models are faster, more customizable, more private, and pound-for-pound more capable. They are doing things with \$100 and 13B params that we struggle with at \$10M and 540B. And they are doing so in weeks, not months.

...

In both cases, low-cost public involvement was enabled by a vastly cheaper mechanism for fine tuning called [low rank adaptation, or LoRA](#) [arXiv:2106.09685] ...

# Randomized Low-Rank Approximation

1. Foundations
2. Randomized SVD
3. Beyond the randomized SVD
4. Sampling-based techniques

# 1. Foundations

- ▶ Matrix rank
- ▶ SVD
- ▶ Best low-rank approximation
- ▶ Low-rank and subspace approximation
- ▶ When (not) to expect good low-rank approximations

References: [Golub/Van Loan'2013]<sup>1</sup>, [Horn/Johnson'2013]<sup>2</sup>

---

<sup>1</sup>Golub2013.

<sup>2</sup>Horn2013.

# Rank and basic properties

Let  $A \in \mathbb{R}^{m \times n}$ . Then

$$\text{rank}(A) := \dim(\text{range}(A)).$$

# Rank and basic properties

Let  $A \in \mathbb{R}^{m \times n}$ . Then

$$\text{rank}(A) := \dim(\text{range}(A)).$$

## Quiz

1. What is the rank of this matrix?



# Rank and basic properties

Let  $A \in \mathbb{R}^{m \times n}$ . Then

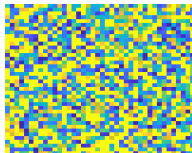
$$\text{rank}(A) := \dim(\text{range}(A)).$$

## Quiz

1. What is the rank of this matrix?



2. What is the rank of `randn(40)`?





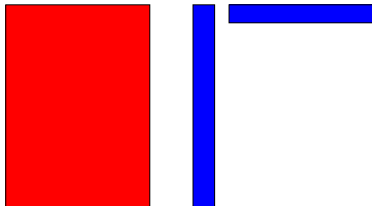
# Rank and matrix factorizations

**Lemma.** A matrix  $A \in \mathbb{R}^{m \times n}$  of rank  $r$  admits a factorization of the form

$$A = BC^T, \quad B \in \mathbb{R}^{m \times r}, \quad C \in \mathbb{R}^{n \times r}.$$

We say that  $A$  has **low rank** if  $\text{rank}(A) \ll m, n$ .

Illustration of low-rank factorization:



	$A$	$BC^T$
#entries	$mn$	$mr + nr$

- ▶ Generically (and in most applications),  $A$  has **full rank**, that is,  $\text{rank}(A) = \min\{m, n\}$ .
- ▶ Aim instead at **approximating**  $A$  by a low-rank matrix.

# The singular value decomposition

**Theorem (SVD).** Let  $A \in \mathbb{R}^{m \times n}$  with  $m \geq n$ . Then there are orthogonal matrices  $U \in \mathbb{R}^{m \times m}$  and  $V \in \mathbb{R}^{n \times n}$  such that

$$A = U \Sigma V^T, \quad \text{with} \quad \Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & 0 & \end{bmatrix} \in \mathbb{R}^{m \times n}$$

and  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$ .

- ▶  $\sigma_1, \dots, \sigma_n$  are called singular values
- ▶  $u_1, \dots, u_n$  are called *left* singular vectors
- ▶  $v_1, \dots, v_n$  are called *right* singular vectors
- ▶  $Av_i = \sigma_i u_i$ ,  $A^T u_i = \sigma_i v_i$  for  $i = 1, \dots, n$ .
- ▶ Singular values are always uniquely defined by  $A$ .
- ▶ Singular values are *never* unique. If  $\sigma_1 > \sigma_2 > \dots > \sigma_n > 0$  then unique up to  $u_i \leftarrow \pm u_i$ ,  $v_i \leftarrow \pm v_i$ .

# The singular value decomposition

**Theorem (SVD).** Let  $A \in \mathbb{R}^{m \times n}$  with  $m \geq n$ . Then there are orthogonal matrices  $U \in \mathbb{R}^{m \times m}$  and  $V \in \mathbb{R}^{n \times n}$  such that

$$A = U \Sigma V^T, \quad \text{with} \quad \Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & 0 & \end{bmatrix} \in \mathbb{R}^{m \times n}$$

and  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$ .

**Quiz:** Which properties of  $A$  can be extracted from the SVD?

# The singular value decomposition

**Theorem (SVD).** Let  $A \in \mathbb{R}^{m \times n}$  with  $m \geq n$ . Then there are orthogonal matrices  $U \in \mathbb{R}^{m \times m}$  and  $V \in \mathbb{R}^{n \times n}$  such that

$$A = U \Sigma V^T, \quad \text{with} \quad \Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & 0 & \end{bmatrix} \in \mathbb{R}^{m \times n}$$

and  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$ .

**Quiz:** Which properties of  $A$  can be extracted from the SVD?

$r = \text{rank}(A)$  = number of nonzero singular values of  $A$ ,

$\text{kernel}(A) = \text{span}\{v_{r+1}, \dots, v_n\}$ ,  $\text{range}(A) = \text{span}\{u_1, \dots, u_r\}$

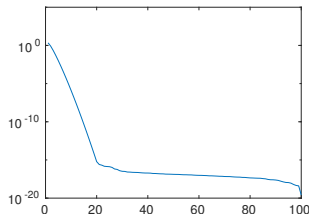
$\|A\|_2 = \sigma_1$ ,  $\|A^\dagger\|_2 = 1/\sigma_r$ ,  $\|A\|_F^2 = \sigma_1^2 + \dots + \sigma_n^2$

$\sigma_1^2, \dots, \sigma_n^2$  eigenvalues of  $AA^T$  and  $A^T A$ .

# SVD: Computational aspects

- ▶ Standard implementations (LAPACK, Matlab's `svd`, `scipy.linalg.svd`, ...) require  $\mathcal{O}(mn^2)$  operations to compute (economy size) SVD of  $m \times n$  matrix  $A$ .
- ▶ Beware of roundoff error when interpreting singular value plots.

Example: `semilogy(svd(hilb(100)))`



- ▶ Kink is caused by roundoff error and does not reflect true behavior of singular values.
- ▶ Exact singular values are known to decay exponentially.<sup>3</sup>
- ▶ *Sometimes more accuracy possible.*<sup>4</sup>

---

<sup>3</sup>Beckermann, B. The condition number of real Vandermonde, Krylov and positive definite Hankel matrices. Numer. Math. 85 (2000), no. 4, 553–577.

<sup>4</sup>Drmač, Z.; Veselić, K. New fast and accurate Jacobi SVD algorithm. I. SIAM J. Matrix Anal. Appl. 29 (2007), no. 4, 1322–1342

# Best low-rank approximation

For  $k < n$ , partition SVD as

$$U\Sigma V^T = \begin{bmatrix} U_k & * \end{bmatrix} \begin{bmatrix} \Sigma_k & 0 \\ 0 & * \end{bmatrix} \begin{bmatrix} V_k & * \end{bmatrix}^T, \quad \Sigma_k = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}$$

Rank- $k$  truncation:

$$A \approx \mathcal{T}_k(A) := U_k \Sigma_k V_k^T.$$

has rank at most  $k$ . By unitary invariance of  $\|\cdot\| \in \{\|\cdot\|_2, \|\cdot\|_F\}$ :

$$\|\mathcal{T}_k(A) - A\| = \|\text{diag}(0, \dots, 0, \sigma_{k+1}, \dots, \sigma_n)\|.$$

In particular:

$$\|A - \mathcal{T}_k(A)\|_2 = \sigma_{k+1}, \quad \|A - \mathcal{T}_k(A)\|_F = \sqrt{\sigma_{k+1}^2 + \dots + \sigma_n^2}.$$

Nearly equal iff singular values decay quickly.

# Best low-rank approximation

**Theorem (Schmidt-Mirsky).** Let  $A \in \mathbb{R}^{m \times n}$ . Then

$$\|A - \mathcal{T}_k(A)\| = \min \{ \|A - B\| : B \in \mathbb{R}^{m \times n} \text{ has rank at most } k \}$$

holds for any unitarily invariant norm  $\|\cdot\|$ .

*Proof:* See Section 7.4.9 in [Horn/Johnson'2013] for general case.

*Proof for  $\|\cdot\|_F$ :* Let  $\sigma(A), \sigma(B)$  denote the vectors of singular values of  $A$  and  $B$  and use the matrix inner product  $\langle A, B \rangle = \text{trace}(B^T A)$ . Then von Neumann's trace inequality states that

$$|\langle A, B \rangle| \leq \langle \sigma(A), \sigma(B) \rangle$$

Hence,

$$\begin{aligned} \|A - B\|_F^2 &= \langle A - B, A - B \rangle = \|A\|_F^2 - 2\langle A, B \rangle + \|B\|_F^2 \\ &\geq \|\sigma(A)\|_2^2 - 2\langle \sigma(A), \sigma(B) \rangle + \|\sigma(B)\|_2^2 \\ &= \sum_{i=1}^n (\sigma_i(A) - \sigma_i(B))^2 \geq \|A - \mathcal{T}_k(A)\|_F^2. \end{aligned}$$

# Best low-rank approximation

Theorem (Schmidt-Mirsky). Let  $A \in \mathbb{R}^{m \times n}$ . Then

$$\|A - \mathcal{T}_k(A)\| = \min \{ \|A - B\| : B \in \mathbb{R}^{m \times n} \text{ has rank at most } k \}$$

holds for any unitarily invariant norm  $\|\cdot\|$ .

Quiz. Is the best rank- $k$  approximation unique if  $\sigma_k > \sigma_{k+1}$ ?



# Best low-rank approximation

**Theorem (Schmidt-Mirsky).** Let  $A \in \mathbb{R}^{m \times n}$ . Then

$$\|A - \mathcal{T}_k(A)\| = \min \{ \|A - B\| : B \in \mathbb{R}^{m \times n} \text{ has rank at most } k \}$$

holds for any unitarily invariant norm  $\|\cdot\|$ .

**Quiz.** Is the best rank- $k$  approximation unique if  $\sigma_k > \sigma_{k+1}$ ?

- ▶ If  $\sigma_k > \sigma_{k+1}$  best rank- $k$  approximation unique wrt  $\|\cdot\|_F$ .
- ▶ Wrt  $\|\cdot\|_2$  only unique if  $\sigma_{k+1} = 0$ . For example,  $\text{diag}(2, 1, \epsilon)$  with  $0 < \epsilon < 1$  has infinitely many best rank-two approximations:

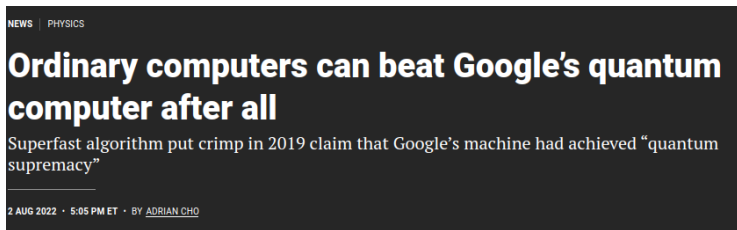
$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 2 - \epsilon/2 & 0 & 0 \\ 0 & 1 - \epsilon/2 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 2 - \epsilon/3 & 0 & 0 \\ 0 & 1 - \epsilon/3 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \dots$$

- ▶ If  $\sigma_k = \sigma_{k+1}$  best rank- $k$  approximation never unique.  
 $I_3$  has several best rank-two approximations:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

# Some uses of low-rank approximation

- ▶ Data compression.
- ▶ Fast solvers for linear systems: Kernel matrices, integral operators, under the hood of sparse direct solvers (MUMPS, PaStiX), ...
- ▶ Fast solvers for dynamical systems: Dynamical low-rank method.
- ▶ Low-rank compression / training of neural nets.
- ▶ Defeating quantum supremacy claims by Google/IBM. Science'2022:



# Approximating the range of a matrix

Aim at finding a matrix  $Q \in \mathbb{R}^{m \times k}$  with orthonormal columns such that

$$\text{range}(Q) \approx \text{range}(A).$$

$QQ^T$  is orthogonal projector onto  $\text{range}(Q) \leadsto$  Aim at solving

$$\min \{ \|A - QQ^T A\| : Q^T Q = I_k \}$$

for  $\|\cdot\| \in \{\|\cdot\|_2, \|\cdot\|_F\}$ . Because  $\text{rank}(QQ^T A) \leq k$ ,

$$\|A - QQ^T A\| \geq \|A - \mathcal{T}_k(A)\|.$$

Setting  $Q = U_k$  one obtains

$$U_k U_k^T A = U_k U_k^T U \Sigma V^T = U_k \Sigma_k V_k^T = \mathcal{T}_k(A).$$

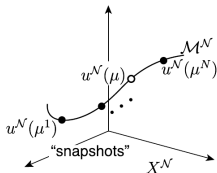
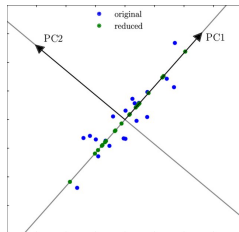
$\leadsto Q = U_k$  is optimal.

Low-rank approximation and range approximation  
are essentially the same tasks!

# Two popular uses of range approximation

## Principal component analysis (PCA):

Dominant left singular vectors of data matrix  $X = [x_1, \dots, x_n]$  (with mean subtracted) provide directions of maximum variance, 2nd maximum variance, etc.



Proper orthogonal decomposition (POD), reduced basis methods: Collect snapshots of time-dependent and/or parameter-dependent equations and perform model reduction by projection to dominant left singular vectors  $U_k$  of snapshot matrix.

# When to expect good low-rank approximations

## Smoothness.

Example 1: **Snapshot matrix** with snapshots depending smoothly on time/parameter

$$A = \begin{bmatrix} u(t_1) & u(t_2) & \cdots & u(t_n) \end{bmatrix}$$
$$\approx \underbrace{\begin{bmatrix} p_1 & p_2 & \cdots & p_k \end{bmatrix}}_{\text{low-dim. polynomial basis}} \times \underbrace{\begin{bmatrix} \ell_1(t_1) & \ell_1(t_2) & \cdots & \ell_1(t_n) \\ \ell_2(t_1) & \ell_2(t_2) & \cdots & \ell_2(t_n) \\ \vdots & \vdots & & \vdots \\ \ell_k(t_1) & \ell_k(t_2) & \cdots & \ell_k(t_n) \end{bmatrix}}_{\text{Vandermonde-like matrix}}$$

where  $u(t) \approx p(t) = p_1 \ell_1(t) + \cdots + p_k \ell_k(t)$  (polynomial approximation of degree  $k$  in basis of Lagrange polynomials).

If  $u : [-1, 1] \rightarrow \mathbb{R}^n$  admits analytic extension to Bernstein ellipse  $\mathcal{E}_\rho$  (foci  $\pm 1$  and sum of half axes equal to  $\rho > 1$ ) then polynomial approximation implies

$$\sigma_k(A) \lesssim \max_{z \in \mathcal{E}_\rho} \|u(z)\|_2 \cdot \rho^{-k}.$$

Exponential decay of singular values!

# When to expect good low-rank approximations

## Smoothness.

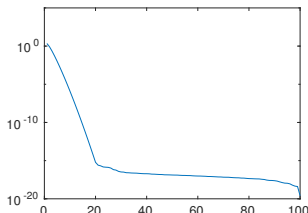
Example 2: **Kernel matrix** for smooth (low-dimensional) kernel:

$$K = \begin{bmatrix} \kappa(x_1, x_1) & \cdots & \kappa(x_1, x_n) \\ \vdots & & \vdots \\ \kappa(x_n, x_1) & \cdots & \kappa(x_n, x_n) \end{bmatrix}, \quad \kappa : \Omega \times \Omega \rightarrow \mathbb{R}.$$

Hilbert matrix:

$$K = \left[ \frac{1}{i+j-1} \right]_{i,j=1}^n$$

Kernel  $\kappa(x, y) = 1/(x + y - 1)$ .



Exponential singular value decay established through Taylor expansion [Börm'2010] or exponential sum approximation [Braess/Hackbusch'2005]:

$$\frac{1}{x+y} \approx \sum_{i=1}^k \gamma_i \exp(\beta_i(x+y)) = \sum_{i=1}^k \gamma_i \exp(\beta_i x) \cdot \exp(\beta_i y).$$

# When to expect good low-rank approximations

## Algebraic structure.

If  $X$  satisfies low-rank Sylvester matrix equation:

$$AX + XB = \text{low rank}$$

and spectra of  $A, B$  are disjoint then singular values of  $X$  (usually) decay exponentially<sup>5</sup>.

- ▶ Basis of fast solvers for matrix equations.
- ▶ Captures many structured matrices: Vandermonde, Cauchy, Pick, ... matrices, canonical Krylov bases, ...

---

<sup>5</sup>Beckermann2017.

# When *not* to expect good low-rank approximations

In most over situations:

- ▶ Kernel matrices with singular/non-smooth kernels
- ▶ Snapshot matrices for time-dependent / parametrized solutions featuring a slowly decaying Kolmogoroff  $N$ -width.
- ▶ Images
- ▶ White noise
- ▶ ...

⊃ Exceptions to these rules:



Also: Low-rank methods are often used even when there is no notable singular value decay in, e.g., statistical inference.



# When *not* to expect good low-rank approximations

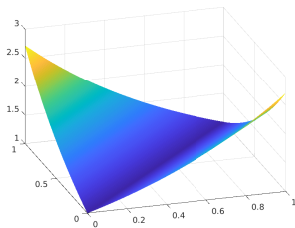
Consider **kernel matrix**

$$K = \begin{bmatrix} \kappa(x_1, x_1) & \cdots & \kappa(x_1, x_n) \\ \vdots & & \vdots \\ \kappa(x_n, x_1) & \cdots & \kappa(x_n, x_n) \end{bmatrix}, \quad \kappa : D \times D \rightarrow \mathbb{R}.$$

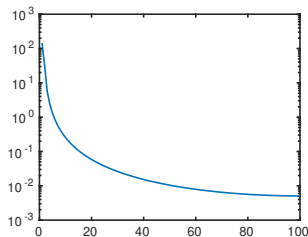
for 1D-kernel  $\kappa$  with diagonal singularity/non-smoothness. Example:

$$\kappa(x, y) = \exp(-|x - y|), \quad x, y \in [0, 1]$$

Function



Singular values



# But not everything is lost..

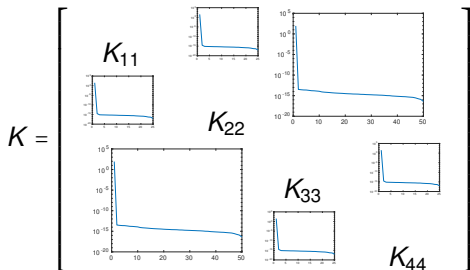
Block partition  $K$ . Level 1:

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} = \left[ \begin{array}{c} \begin{array}{c} K_{11} \\ \begin{array}{c} \text{Plot of } K_{11} \end{array} \end{array} \quad \begin{array}{c} \begin{array}{c} \text{Plot of } K_{12} \end{array} \\ K_{22} \end{array} \right]$$

The figure shows the block partitioning of matrix  $K$  at Level 1. The matrix is represented as a 2x2 block matrix. The top-left block is  $K_{11}$ , which is associated with a plot showing a sharp initial drop followed by a gradual decay on a logarithmic scale. The bottom-right block is  $K_{22}$ , also associated with a similar plot. The off-diagonal blocks  $K_{12}$  and  $K_{21}$  are represented by empty plots, indicating they are zero or negligible. The plots for  $K_{11}$  and  $K_{22}$  have a y-axis ranging from  $10^{-20}$  to  $10^5$  and an x-axis ranging from 0 to 50.

# But not everything is lost..

Block partition  $K$ . Level 2:



etc.  $\leadsto$  HODLR. More general constructions [Hackbusch'2015]:

- ▶  $\mathcal{H}$ -matrices = general recursive block partition.
- ▶ HSS/ $\mathcal{H}^2$ -matrices impose additional nestedness conditions on the low-rank factors on different levels of the recursion.

Exciting news: Recovery of such matrices from mat-vec products<sup>6</sup>.

---

<sup>6</sup>Halikias, Levitt.

# Low-rank matrix approximation algorithms

Landscape

# Landscape of algorithms

Choice of algorithm for performing low-rank approximation of  $A$  depends critically on how  $A$  is accessed:

1. **Small matrices:** If  $m, n = O(10^2)$ , don't think twice, apply  $\text{svd}$ .
2. **Mat-vecs:**  $A$  is accessed through matrix-vector products  $v \mapsto Av$ . massive dense matrices, sparse matrices, implicit representation (e.g., through matrix functions, Schur complements, ...).

## Randomized SVD and friends

3. **Entry-by-entry:** Individual entries  $A(i, j)$  can be directly computed but it is too expensive to compute/hold the whole matrix. kernel matrices, distances matrices, discretizations of nonlocal equations (integral eqns, fractional diff eqns), ...

## Sampling-based techniques.

4. **Semi-analytical techniques:** Polynomial approximation, exponential sum approximation, Random Fourier features.
5. **Implicit:**  $A$  satisfies linear system/eigenvalue problem/opt problem/...

## Alternating optimization, Riemannian optimization, ...

## 2. Randomized SVD

# Basic randomized algorithm for low-rank approx

**Must read:** Halko/Martinsson/Tropp'2011: Finding Structure with Randomness...

Randomized Algorithm:

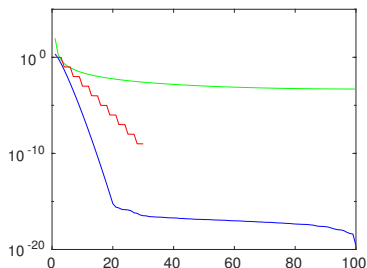
1. Draw Gaussian random matrix  $\Omega \in \mathbb{R}^{n \times r}$ .
2. Perform block mat-vec  $Y = A\Omega$ .
3. Compute (economic) QR decomposition  $Y = QR$ .
4. Form  $B = Q^T A$ .
5. Return  $\hat{A} = QB$  (in factorized form)

**Exact recovery:** If  $A$  has rank  $r$ , we recover  $\hat{A} = A$  with probability 1.

# Three test matrices

- (a) The  $100 \times 100$  Hilbert matrix  $A$  defined by  $A(i, j) = 1/(i + j - 1)$ .
- (b) The matrix  $A$  defined by  $A(i, j) = \exp(-\gamma|i - j|/n)$  with  $\gamma = 0.1$ .
- (c)  $30 \times 30$  diagonal matrix with diagonal entries

$$1, 0.99, 0.98, \frac{1}{10}, \frac{0.99}{10}, \frac{0.98}{10}, \frac{1}{100}, \frac{0.99}{100}, \frac{0.98}{100}, \dots$$



Singular values of test matrices



# Randomized algorithm applied to test matrices

errors measured in spectral norm:

(a) Hilbert matrix,  $r = 5$ :

Exact	mean	std
0.0019	0.0092	0.0099

(b) Matrix with slower decay,  $r = 25$ :

Exact	mean	std
0.0034	0.012	0.002

(c) Matrix with staircase sv,  $r = 7$ :

Exact	mean	std
0.010	0.038	0.025

# Randomized algorithm applied to test matrices

errors measured in Frobenius norm:

(a) Hilbert matrix,  $r = 5$ :

Exact	mean	std
0.0019	0.0093	0.0099

(b) Matrix with slower decay,  $r = 25$ :

Exact	mean	std
0.011	0.024	0.001

(c) Matrix with staircase sv,  $r = 7$ :

Exact	mean	std
0.014	0.041	0.024

# Basic randomized algorithms for low-rank approx

Add oversampling. (usually small) integer  $p$

Randomized Algorithm:

1. Draw standard Gaussian random matrix  $\Omega \in \mathbb{R}^{n \times (r+p)}$ .
2. Perform block mat-vec  $Y = A\Omega$ .
3. Compute (economic) QR decomposition  $Y = QR$ .
4. Form  $B = Q^T A$ .
5. Return  $\hat{A} = QB$  (in factorized form)

**Problem:**  $\hat{A}$  has rank  $r + p > r$ .

**Solution:** Compress  $B \approx \mathcal{T}_r(B) \rightsquigarrow Q\mathcal{T}_r(B)$  has rank  $r$ .

Error:

$$\begin{aligned}\|Q\mathcal{T}_r(B) - A\| &= \|Q\mathcal{T}_r(B) - QB + QB - A\| \\ &\leq \|\mathcal{T}_r(B) - B\| + \|(I - QQ^T)A\|\end{aligned}$$

# Basic randomized algorithms for low-rank approx

Add oversampling. (usually small) integer  $p$

Randomized Algorithm:

1. Draw standard Gaussian random matrix  $\Omega \in \mathbb{R}^{n \times (r+p)}$ .
2. Perform block mat-vec  $Y = A\Omega$ .
3. Compute (economic) QR decomposition  $Y = QR$ .
4. Form  $B = Q^T A$ .
5. Return  $\hat{A} = Q\mathcal{T}_r(B)$  (in factorized form)

Gold standard best rank- $r$  approximation error:

- ▶ spectral norm:  $\sigma_{r+1}$
- ▶ Frobenius norm:  $\sqrt{\sigma_{r+1}^2 + \dots + \sigma_n^2}$ .

# Randomized algorithm applied to test matrices

errors measured in spectral norm:

(a) Hilbert matrix,  $r = 5$ :

Exact	mean	std	
0.0019	0.0092	0.0099	$p = 0$
0.0019	0.0026	0.0019	$p = 1$
0.0019	0.0019	0.0001	$p = 2$

(b) Matrix with slower decay,  $r = 25$ :

Exact	mean	std	
0.0034	0.012	0.002	$p = 0$
0.0034	0.011	0.0017	$p = 1$
0.0034	0.010	0.0015	$p = 2$
0.0034	0.0064	0.0008	$p = 10$
0.0034	0.0037	0.0002	$p = 25$

(c) Matrix with staircase sv,  $r = 7$ :

Exact	mean	std	
0.010	0.038	0.025	$p = 0$
0.010	0.021	0.012	$p = 1$
0.010	0.012	0.005	$p = 2$

# Analysis: general considerations

**Goal:** Say something sensible about  $\|(I - QQ^T)A\|$ . Expected value, tail bounds, ... wrt random matrix  $\Omega$ .

Significantly more complicated than what we have seen so far!

Analysis of randomized low-rank approximation can be separated into two phases:

1. **Structural bound:** Derive bound that holds for (almost) *every*  $\Omega$ .

This bound usually depends on  $\Omega$  and dependence needs to be simple enough to facilitate 2nd phase.

2. **Stochastic analysis:** Derive expected value, tail bounds for structural bound using random matrix theory, concentration results, ...

# Recap on pseudo-inverses

For a matrix  $A \in \mathbb{R}^{n \times r}$  of *full column rank*, consider least-squares problem

$$\min \|Ax - b\|_2.$$

Equivalent to solving normal equations

$$(A^T A)x = A^T b \quad \Rightarrow \quad x = (A^T A)^{-1} A^T b.$$

Pseudo-inverse of  $A$  defined as

$$A^\dagger := (A^T A)^{-1} A^T.$$

- ▶ If  $A$  is square and invertible,  $A^\dagger = A^{-1}$ .
- ▶ If  $A$  is ONB,  $A^\dagger = A^T$ .

If  $A$  has full row rank:  $A^\dagger := A^T (AA^T)^{-1}$ .

# Recap on projectors

Consider  $r$ -dimensional subspace  $\mathcal{V} \subset \mathbb{R}^n$ . Then

- ▶  $\Pi \in \mathbb{R}^{n \times n}$  is a **projector** onto  $\mathcal{V}$  if  $\Pi x \in \mathcal{V}$  for all  $x \in \mathbb{R}^n$  and  $\Pi^2 = \Pi$ .
- ▶  $\Pi$  is an **orthogonal projector** if  $\Pi$  is projector and  $\Pi = \Pi^T$ .

Let  $V \in \mathbb{R}^{n \times r}$  be basis of  $\mathcal{V}$ . Then  $\Pi = VX$  for some  $r \times n$  matrix  $X$ .  
Condition  $\Pi^2 = \Pi$  becomes

$$X = XVX.$$

Some choices for  $X$ :

- ▶  $X = V^\dagger = (V^T V)^{-1} V^T \rightsquigarrow$  orthogonal projector
- ▶  $X = (W^T V)^{-1} W^T$  for some  $n \times r$  matrix  $W$  such that  $W^T V$  is invertible.  $\rightsquigarrow$  oblique projector  
(Projection is along  $\text{span}(W)^\perp$  instead of  $\mathcal{V}^\perp$ .)
- ▶  $X = (W^T V)^\dagger W^T$  for some  $n \times s$  matrix  $W$  such that  $W^T V \in \mathbb{R}^{s \times r}$  has full *column* rank.  $\rightsquigarrow$  oblique projector



# Basic properties of projectors

Fun properties of a projector  $\Pi$ :

- ▶ If  $\mathcal{V} = \text{span}\{e_1, \dots, e_r\}$  then  $\Pi$  takes the form

$$\Pi = \begin{bmatrix} I_r & Y \\ 0 & 0 \end{bmatrix}$$

and  $\Pi$  is an orthogonal projector if and only if  $Y = 0$ .

- ▶  $\|\Pi\|_2 \leq 1$  and  $\|\Pi\|_2 = 1$  if and only if  $\Pi$  is an orthogonal projector.
- ▶  $\|\Pi\|_2 = \|I - \Pi\|_2$  unless  $\Pi$  is trivial.

# Orthogonal vs. oblique projectors

An orthogonal projector  $\Pi$  is optimal in the sense that

$$\|x - \Pi x\|_2 = \min\{\|x - v\|_2 : v \in \mathcal{V}\}.$$

Note also that  $\|x\|_2^2 = \|\Pi x\|_2^2 + \|x - \Pi x\|_2^2$ .

The (quasi)-optimality of an oblique projector  $\tilde{\Pi}$  is determined by its norm:

$$\begin{aligned}\|(I - \tilde{\Pi})x\|_2 &= \|(I - \tilde{\Pi})(x - \Pi x)\|_2 \leq \|I - \tilde{\Pi}\|_2 \|x - \Pi x\|_2 \\ &= \|\tilde{\Pi}\|_2 \|x - \Pi x\|_2 = \|\tilde{\Pi}\|_2 \cdot \min\{\|x - v\|_2 : v \in \mathcal{V}\}.\end{aligned}$$

If  $\tilde{\Pi} = V(W^T V)^{-1} W^T$  for ONB  $V, W$  then

$$\|\tilde{\Pi}\|_2 = \|(W^T V)^{-1}\|_2 = 1/\sigma_{\min}(W^T V)$$

# Analysis: structural bound

**Goal:** Bound  $\|(I - \Pi_{A\Omega})A\|_F$ , where  $\Pi_{A\Omega} = QQ^T$  is orthogonal projector onto range of  $A\Omega$ .

Important observation: Because of

$$(I - \Pi_{A\Omega})A\Omega = 0,$$

we have for the oblique projector  $\tilde{\Pi} = V(\Omega^T V)^\dagger \Omega^T$  (for some  $V \in \mathbb{R}^{n \times r}$  that

$$\|(I - \Pi_{A\Omega})A\|_F = \|(I - \Pi_{A\Omega})A(I - \tilde{\Pi}^T)\|_F \leq \|A(I - \tilde{\Pi}^T)\|_F.$$

Because  $\tilde{\Pi}$  is a projector onto  $\text{span}(V)$ , it follows that  $\tilde{\Pi}V = V$  and, hence,

$$(I - VV^T)(I - \tilde{\Pi}^T) = (I - \tilde{\Pi}^T).$$

## Analysis: structural bound

$$\|(I - \Pi_{A\Omega})A\|_F \leq \|A(I - VV^T)(I - \tilde{\Pi}^T)\|_F.$$

We now consider SVD of  $A$

$$A = [u_1, \dots, u_r, u_{r+1}, \dots, u_m] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} [v_1, \dots, v_r, v_{r+1}, \dots, v_n].$$

We will choose  $V = [v_1, \dots, v_r]$ , the  $r$  principal right singular vectors of  $A$ . We will also use  $V_\perp = [v_{r+1}, \dots, v_n]$ .

Quick but suboptimal argument:

$$\|A(I - VV^T)(I - \tilde{\Pi}^T)\|_F \leq \|A(I - VV^T)\|_F \|(I - \tilde{\Pi}^T)\|_2 = \|\Sigma_2\|_F \|\tilde{\Pi}\|_2$$

Deviation from gold standard  $\|\Sigma_2\|_F$  determined by

$$\|\tilde{\Pi}\|_2 \leq \|(\Omega^T V)^\dagger\|_2 \|\Omega\|_2.$$

Drawback: Involves big matrix  $\Omega$ , which will lead to suboptimal constants.

# Analysis: structural bound

More refined argument:

$$\begin{aligned}\|A(I - VV^T)(I - \tilde{\Pi}^T)\|_F^2 &= \|A(I - VV^T)\|_F^2 + \|A(I - VV^T)\Pi^T\|_F^2 \\ &= \|\Sigma_2\|_F^2 + \|\Sigma_2(V_\perp^T\Omega)(V^T\Omega)^\dagger\|_F^2\end{aligned}$$

Final structural bound:

$$\|(I - QQ^T)A\|_F^2 \leq \|\Sigma_2\|_F^2 + \|\Sigma_2(V_\perp^T\Omega)(V^T\Omega)^\dagger\|_F^2.$$

By orthogonal invariance of Gaussian random matrices:

## Lemma

*Let  $[V, V_\perp] \in \mathbb{R}^{n \times n}$  be orthogonal and let  $\Omega$  be an  $n \times m$  Gaussian random matrix. Then  $V^T\Omega$  and  $V_\perp^T\Omega$  are independent Gaussian random matrices.*

# Bounding expectation

**Goal:** Bound expected value of

$$\|(I - QQ^T)A\|_F^2 \leq \|\Sigma_2\|_F^2 + \|\Sigma_2\Omega_2\Omega_1^\dagger\|_F^2 \quad (1)$$

for independent Gaussian random matrices  $\Omega_1, \Omega_2$ .

To analyze red term, we use

$$\mathbb{E}\|\Sigma_2\Omega_2\Omega_1^\dagger\|_F^2 = \mathbb{E}(\mathbb{E}(\|\Sigma_2\Omega_2\Omega_1^\dagger\|_F^2 \mid \Omega_1)) = \|\Sigma_2\|_F^2 \cdot \mathbb{E}\|\Omega_1^\dagger\|_F^2.$$

(EFY:  $\mathbb{E}\|A\Omega B\|_F^2 = \|A\|_F^2\|B\|_F^2$  for Gaussian matrix  $\Omega$  and constant matrices  $A, B$ .)

## Analysis: $r = 1, p = 0$

For  $r = 1, p = 0$ , we have

$$(V^T \Omega)^\dagger = \omega_1^{-1}, \quad \omega_1 \sim \mathcal{N}(0, 1).$$

**Problem:**  $\omega_1^{-1}$  (reciprocal of standard normal random variable) is Cauchy distribution with undefined mean and variance.

Need to consider  $p \geq 2$ .

## Analysis: $r = 1, p \geq 2$

For  $r = 1$  we have  $\|\Omega_1^\dagger\|_F^2 = 1/\|\Omega_1\|_F^2$ , where  $\|\Omega_1\|_F^2$  is a sum of  $p + 1$  squared independent standard normal random variables.

$\|\Omega_1\|_F^2 \sim \chi_{p+1}^2$  chi-squared distribution with  $p + 1$  d.o.f.; pdf

$$f_{\Omega_1}(x) = \frac{2^{-(p+1)/2}}{\Gamma((p+1)/2)} x^{(p+1)/2-1} \exp(-x/2), \quad x > 0.$$

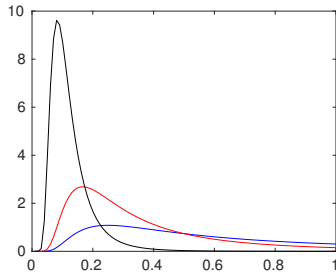


## Analysis: $r = 1, p \geq 2$

$$\|\Omega_1^\dagger\|_F^2 = \frac{1}{\|\Omega_1\|_F^2} = \left( \sum_{i=1}^p \Omega_{1,i}^2 \right)^{-1} \sim \text{Inv} - \chi^2(p+1),$$

the inverse-chi-squared distribution with  $p+1$  degrees of freedom.  
Pdf given by

$$\frac{2^{-(p+1)/2}}{\Gamma((p+1)/2)} x^{-(p+1)/2-1} \exp(-1/(2x)).$$



pdf for  $p = 1$ ,  $p = 3$ ,  $p = 9$

## Analysis: $r = 1, p \geq 2$

Textbook results:

$$\blacktriangleright \mathbb{E} \|\Omega_1\|_F^2 = p + 1, \quad \mathbb{E} \|\Omega_1^\dagger\|_F^2 = (p - 1)^{-1}$$

Tail bound by [Laurent/Massart'2000]:

$$\blacktriangleright \mathbb{P} \left[ \|\Omega_1\|_F^2 \leq p + 1 - t \right] \leq \exp \left( -\frac{t^2}{4(p+1)} \right)$$

### Theorem

For  $r = 1, p \geq 2$ , we have

$$\mathbb{E} \|(I - QQ^T)A\|_F \leq \sqrt{1 + \frac{1}{p-1}} \|\Sigma_2\|_F.$$

Probability of deviating from this upper bound decays exponentially, as *indicated* by tail bound for  $\chi_{p+1}^2$ .

## Analysis: general $r, p \geq 2$

Again use

$$\mathbb{E} \|\Sigma_2 \Omega_2 \Omega_1^\dagger\|_F^2 = \|\Sigma_2\|_F^2 \cdot \mathbb{E} \|\Omega_1^\dagger\|_F^2.$$

By standard results in multivariate statistics, we have

$$\mathbb{E} \|\Omega_1^\dagger\|_F^2 = \frac{r}{p-1}.$$

Sketch of argument:

- ▶  $\Omega_1 \Omega_1^T \sim W_r(r+p)$  (Wishart distribution with  $r+p$  degrees of freedom)
- ▶  $(\Omega_1 \Omega_1^T)^{-1} \sim \mathcal{W}_r^{-1}(r+p)$  (inverse Wishart distribution with  $r+p$  degrees of freedom)
- ▶  $\mathbb{E}(\Omega_1 \Omega_1^T)^{-1} = \frac{1}{p-1} I_r$ ; see Page 96 in [Muirhead'1982]<sup>7</sup>
- ▶ Result follows from  $\|\Omega_1^\dagger\|_F^2 = \|\Omega_1^T (\Omega_1 \Omega_1^T)^{-1}\|_F^2 = \text{trace}((\Omega_1 \Omega_1^T)^{-1})$

---

<sup>7</sup>R. J. Muirhead, Aspects of Multivariate Statistical Theory, Wiley, New York, NY, 1982.

## Analysis: general $r, p \geq 2$

Together with  $\mathbb{E}\|(I - QQ^T)A\|_F \leq \sqrt{\mathbb{E}\|(I - QQ^T)A\|_F^2}$ , we obtain:

### Theorem

*For  $p \geq 2$ , we have*

$$\mathbb{E}\|(I - QQ^T)A\|_F \leq \sqrt{1 + \frac{r}{p-1}} \|\Sigma_2\|_F,$$

$$\mathbb{E}\|(I - QQ^T)A\|_2 \leq \left(1 + \sqrt{\frac{r}{p-1}}\right) \|\Sigma_2\|_2 + \frac{e\sqrt{r+p}}{p} \|\Sigma_2\|_F.$$

For proof of spectral norm and tail bounds, see [Halko/Martinsson/Tropp'2011].

# Randomized subspace iteration

1. Draw standard Gaussian random matrix  $\Omega \in \mathbb{R}^{n \times (k+p)}$ .
2. Perform block mat-vec  $Y_0 = A\Omega$ .
3. Perform  $q$  steps of subspace iteration:  $Y = (AA^T)^q Y_0$ .
4. Compute (economic) QR decomposition  $Y = QR$ .
5. Form  $B = Q^T A$ .
6. Return  $\hat{A} = Q\mathcal{T}_r(B)$  (in factorized form)

Algorithm is essentially equivalent with randomized algorithm applied to  $(AA^T)^q A$ . Using Hölder's inequality and result from the exercises, we have

$$\begin{aligned}\mathbb{E}\|(I - QQ^T)A\|_2 &\leq (\mathbb{E}\|(I - QQ^T)A\|_2^{2q+1})^{1/(2q+1)} \\ &\leq (\mathbb{E}\|(I - QQ^T)(AA^T)^{2q}A\|_2)^{1/(2q+1)}.\end{aligned}$$

The singular values of  $(AA^T)^{2q}A$  are  $\sigma_1^{2q+1}, \dots, \sigma_n^{2q+1}$ .

# Randomized subspace iteration

Combination with previous result for randomized algorithm gives:

## Theorem

For  $p \geq 2$ ,  $\mathbb{E}\|(I - QQ^T)A\|_2$  is bounded by

$$\begin{aligned} & \left[ \left(1 + \sqrt{\frac{r}{p-1}}\right) \sigma_{r+1}^{2q+1} + \frac{e\sqrt{r+p}}{p} (\sigma_{k+1}^{2(2q+1)} + \dots)^{1/2} \right]^{1/(2q+1)} \\ & \leq \sigma_{r+1} \left[ 1 + \sqrt{\frac{r}{p-1}} + \frac{e\sqrt{(r+p)(n-r)}}{p} \right]^{1/(2q+1)} \end{aligned}$$

Numerical experiments reveal that, in most situations of practical interest, subspace iteration is *not* a wise way to spend matrix-vector products. Usually preferable to spend them on oversampling in the randomized SVD.

# A posteriori error estimate and adaptive choice of $k$

## Lemma

Let  $\omega^{(i)}$ ,  $i = 1, \dots, s$  be  $n$ -dimensional random Gaussian vectors. Then for any  $m \times n$  matrix  $C$  the inequality

$$\|C\|_2 \leq 10\sqrt{2/\pi} \max_{i=1, \dots, s} \|C\omega^{(i)}\|_2.$$

holds with probability  $1 - 10^{-s}$ .

Proof. A misleading proof idea can be found in [Halko/Martinsson/Tropp'2011]. See exercises for the full proof.

Given ONB  $Q$  returned by randomized algorithm, apply result of lemma to  $C = (I - QQ^T)A$ . Can be combined into adaptive algorithm for choosing  $k$ .

**Error estimate obtained from the lemma is not great.** See Epperly/Tropp SISC'2024 for more sophisticated error estimators (e.g., leave-one-out error estimator).

Assuming that one knows  $\|A\|_F$  the Frobenius norm error can be trivially estimated from

$$\|A\|_F^2 = \|QQ^T A\|_F^2 + \|(I - QQ^T)A\|_F^2.$$

# 3. Beyond the randomized SVD

- ▶ Other random sketching matrices
- ▶ Symmetric matrices, Nyström
- ▶ Generalized Nyström



# Sketching beyond Gaussians

Can replace  $\Omega^T$  in randomized SVD by any of the OSEs discussed in L5. Subsampled trigonometric transforms perform very well in practice but bounds are weaker (due to loss of orthogonal invariance). Analysis for general OSEs starts from a variant of (1):

$$\|(I - \Pi_{A\Omega})A\|_2 \leq \|\Sigma_2\|_2 + \|\Sigma_2\Omega_2\Omega_1^\dagger\|_2^2 \leq \sigma_{r+1}(1 + \|\Omega_2\|_2\|\Omega_1^\dagger\|_2)$$

with  $\Omega_1 = \Omega^T V$ ,  $\Omega_2 = \Omega^T V_\perp$ , where  $V \in \mathbb{R}^{n \times r}$  contains leading  $r$  right singular vectors.

Let  $\Omega^T$  be subsampled trigonometric transform (e.g., SRHT) satisfying  $(r, 1/2, 0.95)$ -OSE property. By L5, this is ensured if  $\Omega$  has  $O(r \log(r + \log n))$  columns<sup>8</sup>. For this choice,  $\|\Omega_2\|_2 \leq 1$ ,  $\|\Omega_1^\dagger\|_2 \leq \sqrt{2}$ , and

$$\|(I - \Pi_{A\Omega})A\|_2 \leq (1 + \sqrt{2})\sigma_{r+1}$$

holds with probability 95%.

---

<sup>8</sup>This estimate is very pessimistic.

# General symmetric matrices

If  $A$  is symmetric, the approximation  $QQ^T A$  of the randomized SVD does *not* preserve symmetry.

**Idea:** Use  $Q$  from randomized SVD and approximate range+co-range of  $A$  at the same time:

$$QQ^T AQQ^T.$$

This produces an error on the level of the randomized SVD error  $\varepsilon := \|A - QQ^T A\|_F$ :

$$\begin{aligned}\|A - QQ^T AQQ^T\|_F^2 &= \|A - QQ^T A + QQ^T A - QQ^T AQQ^T\|_F^2 \\ &= \|(I - QQ^T)A\|_F^2 + \|QQ^T A(I - QQ^T)\|_F^2 \\ &\leq 2\|(I - QQ^T)A\|_F^2,\end{aligned}$$

and hence  $\|A - QQ^T AQQ^T\|_F \leq \sqrt{2}\varepsilon$ .

# SPSD matrices: Nyström

For a symmetric positive semi-definite (SPSD) matrix  $A$ , one can improve upon  $QQ^T A QQ^T$  on two aspects: (1) halving #matvecs, and (2) ensuring streaming property.

**Basic idea:** If SPSPD  $A$  has rank  $r$  and  $\Omega$  is  $n \times (r + p)$  Gaussian with  $p \geq 0$  then  $A\Omega(\Omega^T A\Omega)^\dagger \Omega^T$  is almost surely an oblique projector onto range of  $A$ . Hence,

$$\hat{A} := A\Omega(\Omega^T A\Omega)^\dagger (A\Omega)^T$$

is (almost surely) equal to  $A$ .

For general SPSPD  $A$ , approximation  $\hat{A}$  is called **Nyström approximation**. Choose  $p \geq 2$ .

Highly recommended reading on Nyström: Tropp/Webber arXiv'2023.

# SPSD matrices: Error analysis of Nystrom

Error of Nystrom satisfies

$$A - \hat{A} = A^{1/2}(I - \Pi_{A^{1/2}\Omega})A^{1/2}, \quad \Pi_{A^{1/2}\Omega} := A^{1/2}\Omega(\Omega^T A \Omega)^\dagger (A^{1/2}\Omega)^T$$

Note:  $\Pi_{A^{1/2}\Omega}$  is orth. projector (onto span of  $\Pi_{A^{1/2}\Omega}$ ). This implies:

$I - \Pi_{A^{1/2}\Omega}$  is orth. projector  $\Rightarrow I - \Pi_{A^{1/2}\Omega}$  is PSD  $\Rightarrow A - \hat{A}$  is PSD.

Measure error in nuclear norm:

$$\|A - \hat{A}\|_* = \sigma_1(A - \hat{A}) + \dots + \sigma_n(A - \hat{A}) = \lambda_1(A - \hat{A}) + \dots + \lambda_n(A - \hat{A}) = \text{trace}(A - \hat{A}).$$

Hence,

$$\|A - \hat{A}\|_* = \|A^{1/2}(I - \Pi_{A^{1/2}\Omega})(I - \Pi_{A^{1/2}\Omega})A^{1/2}\|_* = \|(I - \Pi_{A^{1/2}\Omega})A^{1/2}\|_F^2.$$

But this is the error of the randomized SVD applied to  $A^{1/2}$ ! Hence,

$$\begin{aligned} \mathbb{E}\|A - \hat{A}\|_* &= \mathbb{E}\|(I - \Pi_{A^{1/2}\Omega})A^{1/2}\|_F^2 \leq \left(1 + \frac{r}{p-1}\right) \|A^{1/2} - \mathcal{T}_r(A^{1/2})\|_F^2 \\ &= \left(1 + \frac{r}{p-1}\right) (\lambda_{r+1}(A) + \dots + \lambda_n(A)) \end{aligned}$$

Second term is best rank- $r$  approx error in nuclear norm.

# SPSD matrices: Implementation of Nyström

The pseudo-inverse  $(\Omega^T A \Omega)^\dagger$  is never computed explicitly when forming  $\hat{A}$ ! Consider Cholesky decomposition

$$\Omega^T A \Omega = C^T C \Rightarrow \hat{A} = (A \Omega C^\dagger)(A \Omega C^\dagger)^T.$$

In most situations,  $\Omega^T A \Omega$  is invertible  $C^\dagger = C^{-1}$ . In extreme situations, it is recommended to regularize  $A$  with shift  $\varepsilon \approx 10^{-16} \cdot \text{trace}(A)$ .

**Nyström:**

1. Draw standard Gaussian random matrix  $\Omega \in \mathbb{R}^{n \times (r+p)}$ .
2. Perform block mat-vec  $Y = A\Omega + \varepsilon\Omega$ .
3. Compute Cholesky decomposition  $\Omega^T Y = C^T C$ .
4. Compute  $Z = YC^{-1}$  by triangular solves.
5. Compute economy-sized SVD  $Z = U\Sigma V^T$ .
6. Return  $A \approx U\Lambda U^T$  with  $\Lambda = \max\{0, \Sigma^2 - \varepsilon I\}$ .

**Streaming property:** If  $A$  gets updated  $A \leftarrow A + \Delta$  with sparse  $\Delta$ , cheap to update most expensive part of Nyström:  $Y \leftarrow Y + \Delta\Omega$ . Randomized SVD does not have this property! (because of QR decomposition)

# Generalized Nyström

The streaming property of Nyström can be extended to nonsymmetric matrix  $A \in \mathbb{R}^{m \times n}$ .

**Idea:** Instead of orthogonal projector  $\Pi_{A\Omega}$  onto span of  $A\Omega$  use *oblique* projector

$$\hat{A} := A\Omega(\Psi^T A\Omega)^\dagger (A^T \Psi)^T = Y(\Psi^T Y)^\dagger Z^T, \quad Y = A\Omega, \quad Z = A^T \Psi.$$

for another Gaussian random matrix  $\Psi \in \mathbb{R}^{m \times (r+p+\ell)}$  with  $p, \ell \geq 2$ .

$\hat{A}$  is called **Generalized Nyström approximation**.

Generalized Nyström does *not* reduce #matvec products but avoids QR decomposition and has **streaming property**: If  $A$  gets updated  $A \leftarrow A + \Delta$  with sparse  $\Delta$ , cheap to update most expensive parts  $Y \leftarrow Y + \Delta\Omega$ ,  $Z \leftarrow Z + \Delta^T \Psi$ .

EFY: Analysis of generalized Nyström.

Highly recommended reading: Nakatsukasa arXiv'2020.

## 4. Sampling-based techniques

- ▶ Column Subset Selection / QR with pivoting
- ▶ CUR
- ▶ Nyström revisited

# Column Subset Selection



**Goal:** Find family<sup>9</sup> of column indices  $J \in \{1, \dots, n\}^k$  such that corresponding columns  $A(:, J)$  from good approximation of the whole column range of  $A$ :

$$\Pi_J A = A(:, J) A(:, J)^\dagger \approx A,$$

where  $\Pi_J$  is orthogonal projection onto range of  $A(:, J)$ .

$\Pi_J A$  is rank- $k$  approx, built from subspace spanned by orig data.

$\leadsto$  Accuracy limited by best rank- $k$  approx error (=“gold standard”)

---

<sup>9</sup>For all practical purposes, it is sufficient to work with a subset, that is, to not allow for duplicate column indices.



# Column Subset Selection

Can column selection (nearly) attain gold standard?

For Frobenius norm error, probabilistic argument by [Deshpande/Rademacher/Vempala/Wang'2006]:

Let

$$\mathbb{P}(X = J) = \frac{\text{Vol}^2(A(:, J))}{\sum_{K \in \{1, \dots, n\}^k} \text{Vol}^2(A(:, K))},$$

where  $\text{Vol}^2(B) = \sigma_1(B)^2 \cdots \sigma_k(B)^2$ .

Theorem [DRVW'2006].

$$\mathbb{E}_X[\|A - \Pi_X A\|_F^2] \leq (k+1)(\sigma_{k+1}(A)^2 + \cdots + \sigma_n(A)^2).$$

Corollary:  $\exists J \in \{1, \dots, n\}^k$  such that

$$\|A - \Pi_J A\|_F \leq \sqrt{k+1} \|A - \mathcal{T}_k(A)\|_F.$$

Gold standard missed by at most  $\sqrt{k+1}$ .

Expensive to sample. Still expensive: Derandomized alg by [Deshpande/Rademacher'2010], [Cortinovis/DK'2020].

# Random column sampling

To simplify notation, use  $C$  instead of  $A(:, J)$ :

$$A = [a_1 \quad \cdots \quad a_n], \quad C = [c_1 \quad \cdots \quad c_k]$$

*General column sampling method (for product distribution):*

**Input:**  $A \in \mathbb{R}^{m \times n}$ , probabilities  $p_1, \dots, p_n \neq 0$ , integer  $k$ .

**Output:**  $C \in \mathbb{R}^{m \times k}$  containing selected columns of  $A$ .

- 1: **for**  $t = 1, \dots, k$  **do**
- 2:   Pick  $s_t \in \{1, \dots, n\}$  with  $\mathbb{P}[s_t = \ell] = p_\ell$ ,  $\ell = 1, \dots, n$ ,  
independently and with replacement.
- 3:   Set  $c_t = a_{s_t} / \sqrt{k p_{j_t}}$ .
- 4: **end for**
- 5: Compute SVD  $C = U \Sigma V^T$  and set  $Q = U_r \in \mathbb{R}^{m \times r}$ .
- 6: Return low-rank approximation  $QQ^T A$ .

Step 5 accounts for the fact that, in general,  $k$  needs to be chosen significantly larger than  $r$ . If this step is omitted, one can simply compute  $Q$  from a QR decomposition of  $C$  (and the scaling of the columns of  $C$  is not needed).

# Random column sampling and matrix Monte Carlo

## Lemma

For any matrix  $C \in \mathbb{R}^{m \times r}$ , let  $Q$  be the matrix computed above. Then

$$\|A - QQ^T A\|_2^2 \leq \sigma_{r+1}(A)^2 + 2\|AA^T - CC^T\|_2.$$

*Proof.* We have

$$\begin{aligned} & (A - QQ^T A)(A - QQ^T A)^T \\ = & (I - QQ^T)CC^T(I - QQ^T) + (I - QQ^T)(AA^T - CC^T)(I - QQ^T) \end{aligned}$$

Hence,

$$\begin{aligned} \|A - QQ^T A\|_2^2 &= \lambda_{\max}((A - QQ^T A)(A - QQ^T A)^T) \\ &\leq \lambda_{\max}((I - QQ^T)CC^T(I - QQ^T)) + \|AA^T - CC^T\|_2 \\ &= \sigma_{r+1}(C)^2 + \|AA^T - CC^T\|_2. \end{aligned}$$

The proof is completed by applying Weyl's inequality:

$$\sigma_{r+1}(C)^2 = \lambda_{r+1}(CC^T) \leq \lambda_{r+1}(AA^T) + \|AA^T - CC^T\|_2.$$

# Random column sampling and matrix Monte Carlo

Combined with the results from L4 S19 and S21, we obtain the following estimates.

$$\mathbb{E} \|A - QQ^T A\|_2^2 \leq \sigma_{r+1}(A)^2 + 4\varepsilon \|A\|_2^2$$

holds for:

uniform sampling  $p_\ell \propto 1$  if

$$k \geq \max \{ \varepsilon^{-2}, (3\varepsilon)^{-1} \} \cdot \mu(A) \cdot \log(2m).$$

$$\text{with } \mu(A) = n \cdot \max_j \|a_j\|_2^2 / \|A\|_2^2;$$

importance sampling  $p_\ell \propto \|a_\ell\|_2$  if

$$k \geq 2 \max \{ 2\varepsilon^{-2}, (6\varepsilon)^{-1} \} \cdot \text{srank}(A) \cdot \log(2m).$$

$$\text{with } \text{srank}(A) = \|A\|_F^2 / \|A\|_2^2.$$

( $\log(2m)$  factor can be removed [Drineas/Kannan/Mahoney'2006])

# Random column sampling and matrix Monte Carlo

$$\mathbb{E}\|A - QQ^T A\|_2^2 \leq \sigma_{r+1}(A)^2 + 4\varepsilon\|A\|_2^2$$

$k \sim \varepsilon^{-2}$  and **dependence of absolute error on  $\varepsilon$**  make these results useless when aiming at small error.

EFY: Study behavior of importance sampling for  $r = 2$  for “bad” matrix

$$A = \begin{pmatrix} \frac{1}{\sqrt{n}} \mathbf{e}_1 & \frac{1}{\sqrt{n}} \mathbf{e}_1 & \cdots & \frac{1}{\sqrt{n}} \mathbf{e}_1 & \frac{1}{\sqrt{n}} \mathbf{e}_2 \end{pmatrix} \in \mathbb{R}^{n \times (n+1)}.$$

Expensive fix [Drineas/Mahoney/Muthukrishnan'2007]: Let  $V_r$  contain  $r$  dominant right singular vectors of  $A$ . Setting

$$p_\ell = \|V_r(\ell, :)\|_2^2 / r, \quad \ell = 1, \dots, n$$

and sampling  $\mathcal{O}(r^2(\log 1/\delta)/\varepsilon^2)$  columns<sup>10</sup> yields

$$\|A - QQ^T A\|_F \leq (1 + \varepsilon) \|A - \mathcal{T}_r(A)\|_F$$

with probability  $1 - \delta$ . **Relative error bound!**

---

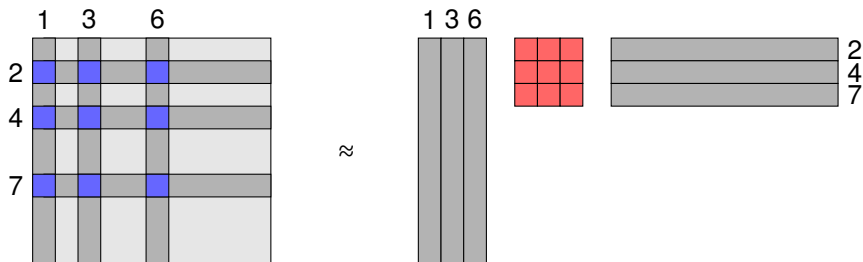
<sup>10</sup>There are variants that improve this to  $\mathcal{O}(k \log k \log(1/\delta)/\varepsilon^2)$ .

# CUR

When approximating matrix  $A$  (and not only its range) Column Subset Selection has two disadvantages:

1. co-range of  $QQ^T A$  not spanned by rows of original data (interpretability affected);
2. building approximation  $QQ^T A$  requires access to full matrix  $A$ .

**CUR decomposition:**  $A \approx A(:, J) U A(I, :)^T$



## CUR: Choice of $U$ – first attempt

After applying column subset selection to  $A$  (gives  $J$ ) and  $A^T$  (gives  $I$ )

$\leadsto$

$$A \approx A(:, J) A(:, J)^\dagger A A(I, :)^{\dagger} A(I, :) =: A(:, J) U A(I, :).$$

**Good.** Error bound follows directly from columns subset selection error bounds:

$$\begin{aligned} & \|A - A(:, J) U A(I, :)\|_2 \\ = & \|A - A(:, J) A(:, J)^\dagger A + A(:, J) A(:, J)^\dagger A - A(:, J) U A(I, :)\|_2 \\ \leq & \|A - A(:, J) A(:, J)^\dagger A\|_2 + \|A(:, J) A(:, J)^\dagger A - A(:, J) U A(I, :)\|_2 \\ \leq & \|A - A(:, J) A(:, J)^\dagger A\|_2 + \|A - A A(I, :)^{\dagger} A(I, :)\|_2 \end{aligned}$$

**Bad.** Fixes first issue of column subset selection **but not second** because building  $U$  requires access to full  $A$ .

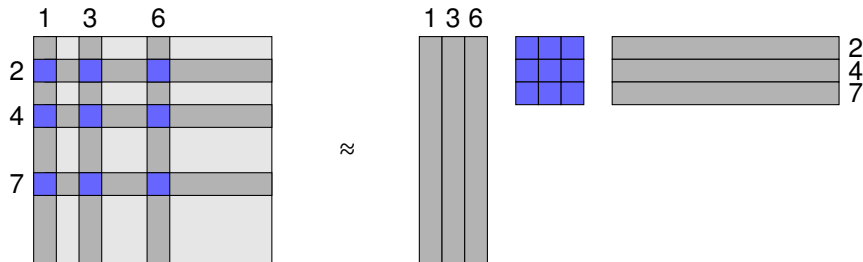
# CUR: Choice of $U$ – second attempt

Choice of  $S = (A(I, J))^{-1}$  in CUR<sup>11</sup>  $\rightsquigarrow$  Remainder term

$$R := A - A(:, J)(A(I, J))^{-1}A(I, :)$$

has zero rows at  $I$  and zero columns at  $J$ .

Cross approximation:



---

<sup>11</sup> $A(I, J)$  might not be invertible but this is definitely not a great choice and will be ignored



# CUR: Existence result for cross approximation

Extension of probabilistic argument [Deshpande et al.'2006]:

Let

$$\mathbb{P}(X = I, Y = J) \propto \text{Vol}^2(A(I, J)) = \det(A(I, J))^2$$

Theorem [Zamarashkin/Osinsky'2018].

$$\mathbb{E}_{X,Y} [\|A - A(:, Y)(A(X, Y))^{-1}A(X, :)\|_F^2] \leq (k+1)^2 (\sigma_{k+1}(A)^2 + \dots).$$

Corollary:  $\exists I, J \in \{1, \dots, n\}^k$  such that

$$\|A - A(:, J)(A(I, J))^{-1}A(I, :)\|_F \leq (k+1) \|A - \mathcal{T}_k(A)\|_F.$$

Gold standard missed by at most  $k+1$ .

Expensive to sample. Closely related to Determinantal Point Processes (Recommended reading: Poulson's talk on DPP<sup>12</sup>)

In practice: Greedy approach (= Gaussian elimination with complete pivoting) or heuristic searches for pivot elements (like Adaptive Cross Approximation)

---

<sup>12</sup>See <https://nhigham.com/wp-content/uploads/2019/05/talk06-poulson.pdf>.

# Nyström approximation revisited

Now consider SPSPD  $A \in \mathbb{R}^{n \times n}$ . Most reasonable choice:  $I = J$ :

$$A \approx A(:, I)(A(I, I))^{-1}A(I, :)$$

This will be called column Nyström (= Nyström with coordinate sampling). Closely related to the Cholesky factorization  $A = LL^T$  with  $L$  lower triangular.

Cholesky algorithm:

Initialize  $L = 0$

**for**  $i = 1, \dots, n$  **do**

$$L(i+1:n, i) := \frac{1}{\sqrt{a_{ii}}} A(i+1:n, i)$$

$$A(i+1:n, i+1:n) \leftarrow A(i+1:n, i+1:n) - L(i+1:n, i)L(i+1:n, i)^T$$

**end for**

## Cholesky revisited

After one step of Cholesky, the updated matrix is

$$A^{(1)} = A - L(:, 1)L(:, 1)^T = A - A(:, 1)a_{11}^{-1}A(1, :)$$

and  $A^{(1)}(1, :) = 0$ ,  $A^{(1)}(:, 1) = 0$ . After  $k$  steps of Cholesky, the updated matrix is  $A^{(k)} = A - L(:, l)L(:, l)^T$  with  $l = \{1, \dots, k\}$ . Can be rewritten using block partition:

$$A^{(k)} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} - \begin{bmatrix} L_{11} \\ L_{21} \end{bmatrix} \begin{bmatrix} L_{11} \\ L_{21} \end{bmatrix}^T = \begin{bmatrix} 0 & 0 \\ 0 & * \end{bmatrix}$$

with  $A_{11} = A(l, l)$ , etc. This implies

$$\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} = \begin{bmatrix} L_{11} \\ L_{21} \end{bmatrix} L_{11}^T \Rightarrow \begin{bmatrix} L_{11} \\ L_{21} \end{bmatrix} \begin{bmatrix} L_{11} \\ L_{21} \end{bmatrix}^T = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} A_{11}^{-1} \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix}^T.$$

Rewritten back in index notation  $\leadsto$

$$A^{(k)} = A - A(l, :)A(l, l)^{-1}A(:, l)$$

This is column Nyström! (for a particular  $l$ )

## Cholesky revisited

While (full) Cholesky for solving linear systems works just fine numerically, (partial) Cholesky for low-rank approximation can fail miserably, as can be seen from  $A = \begin{bmatrix} 0 & 0 \\ 0 & I_k \end{bmatrix}$ .

Pivoting produces a different index set  $I$ .

### Cholesky algorithm with pivoting:

Initialize  $L = 0 \in \mathbb{R}^{n \times k}$

**for**  $i = 1, \dots, k$  **do**

    Select a pivot  $s_i \in \{1, \dots, n\}$ .

    Fetch corresponding column  $c = A(:, s_i)$

    Subtract existing approx  $c \leftarrow c - L(:, 1:i-1)L(s_i, 1:i-1)^T$

    Normalize and store  $L(:, i) = c / \sqrt{c_{s_i}}$

**end for**

- ▶ This is equivalent to first permuting the columns/rows  $s_1, \dots, s_k$  to  $1, \dots, k$  and then applying  $k$  steps of Cholesky. In turn, we obtain the approximation

$$A \approx LL^T = A(I, :)A(I, I)^{-1}A(:, I), \quad I = \{s_1, \dots, s_k\}.$$

- ▶ Already know that  $A^{(k)} = A - A(I, :)A(I, I)^{-1}A(:, I)$  is SPSD (see S56).

# Choosing the pivots

- ▶ **Uniform sampling.**  $\mathbb{P}[s_i = j] \propto 1$  (sampling without replacement). Common strategy in ML, popularized by [Williams/Seeger'2000].

Fails if data is not evenly spread out, such as  $A = \begin{bmatrix} 0 & 0 \\ 0 & I_k \end{bmatrix}$ .

$A^{(i-1)}$  is SPSPD  $\leadsto$  (nuclear) norm is controlled through diagonal  
 $\leadsto$  favor pivots with large diagonal elements

- ▶ **Greedy sampling.** Choose

$$s_i = \operatorname{argmax} \{A^{(i-1)}(\ell, \ell) : \ell = 1, \dots, n\}.$$

“Greedy” because this choice greedily maximizes the volume of the submatrix  $A(I, I)$  in each step.

Usually works well in practice, but can yield approximation error up to a factor  $4^k$  larger than “gold standard” [Harbrecht/Peters/Schneider'2012].

# Choosing the pivots

- **Adaptive sampling.** Sample  $s_i$  from distribution

$$\mathbb{P}\{s_i = \ell\} = \frac{A^{(i-1)}(\ell, \ell)}{\text{trace}(A^{(i-1)})}, \quad \ell = 1, \dots, n.$$

Called **RPCholesky**, proposed and analyzed by [Chen, Epperly, Tropp, Webber. arXiv:2207.06503, 2022].

Illustration on next slide taken from paper.

**Cholesky algorithm with pivoting:**

Initialize  $L = 0 \in \mathbb{R}^{n \times k}$ ,  $d = \text{diag}(A)$

**for**  $i = 1, \dots, k$  **do**

    Sample pivot  $s_i \sim d / (d_1 + \dots + d_n)$ .

    Fetch corresponding column  $c = A(:, i)$

    Subtract existing approx  $c \leftarrow c - L(:, 1 : i - 1)L(s_i, 1 : i - 1)^T$

    Normalize and store  $L(:, i) = c / \sqrt{c_{s_i}}$

    Update diagonal  $d \leftarrow d - |L(:, i)|^2$    % Elementwise square

**end for**

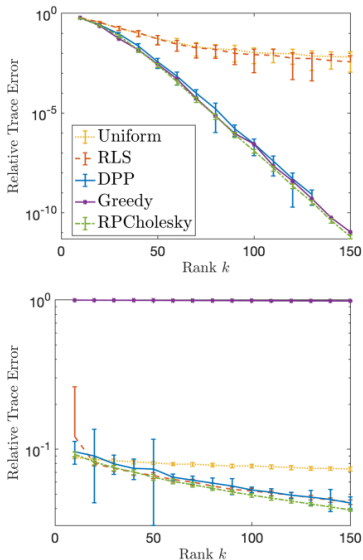


Figure 1: **Rank- $k$  approximation of Gaussian kernel matrices.** *Top: Smile data. Bottom: Spiral data.* *Left:* Mean relative trace-norm error  $\text{tr}(\mathbf{A} - \hat{\mathbf{A}}^{(k)}) / \text{tr} \mathbf{A}$  and one standard deviation error bars for several Nyström-based column approximation methods. *Right:* Selected pivots (colored stars) and data points (gray circles).

# Analysis of RPCholesky

Note that random pivots  $s_i$  are *not* independent for different  $i$ .

**Theorem.** Let  $\hat{A}^{(k)} = A(:, l)A(l, l)^{-1}A(l, :)$  with  $l = \{s_1, \dots, s_k\}$  denote output of RPCholesky. If

$$k \geq \frac{r}{\varepsilon} + r \log\left(\frac{1}{\varepsilon \eta}\right), \quad \text{where} \quad \eta := \text{trace}(A - \mathcal{T}_r(A))/\text{trace}(A),$$

then

$$\mathbb{E}[\text{trace}(A - \hat{A}^{(k)})] \leq (1 + \varepsilon) \cdot \text{trace}(A - \mathcal{T}_r(A)).$$

Proof based on “expected residual function” to track error  $A^{(1)} = A - \hat{A}^{(1)}$  of one step by RPCholesky for (random) matrix  $A$ .

EFY: Prove

$$\Phi(A) := \mathbb{E}[A^{(1)} | A] = A - \frac{A^2}{\text{trace } A}.$$



# Analysis of RPCholesky

**Lemma.** The function  $\Phi(A) = A - \frac{A^2}{\text{trace } A}$  is (1) nonnegative, (2) monotone, and (3) concave on spsd matrices.

*Proof of lemma.* (1) If  $\lambda$  is eigenvalue of  $A$  then  $\Phi(A)$  has eigenvalue  $\Phi(\lambda) = \lambda(1 - \lambda/\text{trace } A) \geq 0$  because  $\lambda \leq \text{trace } A$ .

(2) EFY: For SPSD  $A, H$  verify the identity

$$\Phi(A + H) = \Phi(A) + \Phi(H) + \frac{1}{\text{trace}(A + H)} \left[ \sqrt{\frac{\text{trace } H}{\text{trace } A}} A - \sqrt{\frac{\text{trace } A}{\text{trace } H}} H \right]^2.$$

This implies monotonicity:  $\Phi(A + H) \geq \Phi(A) + \Phi(H) \geq \Phi(A)$ .

(3) Follows from the fact that every *matrix* monotone function is matrix concave. EFY: Find a proof of this fact.<sup>13</sup>  $\diamond$

---

<sup>13</sup>ChatGPT claims that this fact is wrong, as of Nov 2024.

# Analysis of RPCholesky

**Lemma.** Let  $\Phi^{\circ k} := \Phi \circ \Phi \circ \dots \circ \Phi$  ( $k$  times). For  $0 < \epsilon < 1$ ,

$$\text{trace } \Phi^{\circ k}(A) \leq \text{trace}(A - \mathcal{T}_r(A)) + \epsilon \cdot \text{trace } A,$$

holds if

$$k \geq \frac{r \cdot \text{trace}(A - \mathcal{T}_r(A))}{\epsilon \cdot \text{trace } A} + r \log(1/\epsilon).$$

*Proof sketch of lemma.* W.l.o.g. one may assume that  $A$  is diagonal. By matrix concavity of  $\Phi$ , one may assume that

$$A = \begin{bmatrix} \alpha/r \cdot I_r & 0 \\ 0 & \beta/(n-r) \cdot I_{n-r} \end{bmatrix}, \quad \alpha = \text{trace}(\mathcal{T}_r(A)), \quad \beta = \text{trace}(A - \mathcal{T}_r(A)).$$

# Analysis of RPCholesky

By the definition of  $\Phi$ , we have that

$$\Phi^{\circ k}(A) = \begin{bmatrix} \alpha_k/r \cdot I_r & 0 \\ 0 & \beta_k/(n-r) \cdot I_{n-r} \end{bmatrix}$$

with

$$\alpha_k = \alpha_{k-1} - \frac{\alpha_{k-1}^2}{r(\alpha_{k-1} + \beta_{k-1})}, \quad \beta_k = \beta_{k-1} - \frac{\beta_{k-1}^2}{(n-r)(\alpha_{k-1} + \beta_{k-1})}$$

Both sequences are monotonically decreasing  $\leadsto \beta_k \leq \beta$  and, hence,  $\alpha_k \leq \bar{\alpha}_k$  with

$$\bar{\alpha}_k - \bar{\alpha}_{k-1} = \frac{\bar{\alpha}_{k-1}^2}{r(\bar{\alpha}_{k-1} + \beta)}, \quad \bar{\alpha}_0 = \alpha.$$

$\bar{\alpha}_k$  is upper bounded by  $\alpha(k)$ , where  $\alpha(t)$  solves ODE

$$\dot{\alpha}(t) = \frac{\alpha(t)^2}{r(\alpha(t) + \beta)}, \quad \bar{\alpha}(0) = \alpha.$$

This ODE admits explicit solution, which can be used to establish result of lemma.

◇

# Analysis of RPCholesky

*Proof of theorem.* By concavity of  $\Phi$ , we have

$$\mathbb{E}A^{(i)} = \mathbb{E}\Phi(A^{(i-1)}) \leq \Phi(\mathbb{E}A^{(i-1)}).$$

By monotonicity of  $\Phi$ , this implies

$$\mathbb{E}A^{(k)} \leq \Phi(\mathbb{E}A^{(k-1)}) \leq \Phi \circ \Phi(\mathbb{E}A^{(k-2)}) \leq \dots \leq \Phi^{\circ k}(A).$$

Finally, by properties of trace,

$$\mathbb{E} \text{trace } A^{(k)} \leq \text{trace } \Phi^{\circ k}(A).$$

Result follows from applying the lemma with  $\epsilon = \varepsilon\eta$ .

## Back to column subset selection

For  $B \in \mathbb{R}^{m \times n}$ , we can apply RPCholesky to  $A = B^T B$  to get a good column selection. Verifying this, is an EFY:

1. Let  $I$  be a column selection and set  $\hat{B} = B(:, I)B(:, I)^\dagger B$ . Then

$$\hat{B}^T \hat{B} = A(:, I)A(I, I)^{-1}A(I, :).$$

Hint: Assume w.l.o.g. that  $J = \{1, \dots, k\}$  and consider a QR factorization of  $B(:, J)$ .

2. Using Part 1, formulate a variant of the theorem from S76 that establishes an upper bound on  $\mathbb{E}\|B - \hat{B}\|_F^2$  in terms of properties of  $B$  only.