

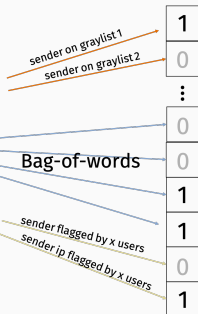
CS-GY 9223 D: Lecture 9

Low-rank approximation and singular value decomposition

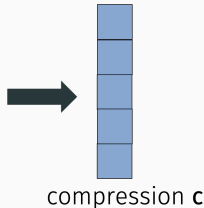
NYU Tandon School of Engineering, Prof. Christopher Musco

Return to data compression:

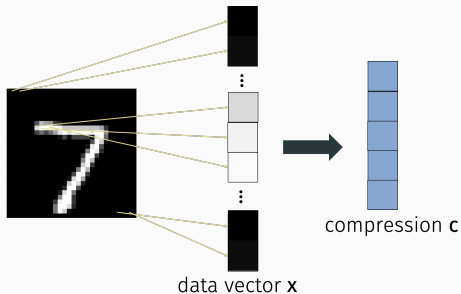
MIME-Version: 1.0 Date: Mon, 7 Oct 2019
 14:51:30 -0400 Message-ID: <CAUVp1stQpge-Q-
 39MLANh0ry29_jaas5QwumWb1ccrFmYD0A8@mail.gsa
 il.com> Subject: 92231 Reading Group, Meeting
 2, tomorrow at 10am From: Christopher Musco
 <cmusco@nyu.edu> To: algnldsfnyu.edu Content-
 Type: multipart/alternative;
 boundary="00000000000078ec240594568a53" --
 00000000000078ec240594568a53 Content-Type:
 text/plain charset="UTF-8" I hope everyone
 had a good weekend! Tomorrow at 10am in 370
 Jay St. #1114* we will meet for the second
 installment of the CS-0Y 92231 reading
 group. Nick Peng will be leading a discussion
 about the paper Simple Analysis of the Sparse
 Johnson-Lindenstrauss Transform
 <http://drops.dagstuhl.de/opus/Volltexte/2018/
 /8305/pdf/OAS2ca-2018-15.pdf>. Please
 read the abstract and introduction before the
 meeting. Best, - CM *Christopher Musco,
 Assistant Professor* *New York University,
 Tandon School of Engineering* *4011 578
 2541* --00000000000078ec240594568a53 Content-
 Type: text/html charset="UTF-8" Content-
 Transfer-Encoding: quoted-printable



data vector x

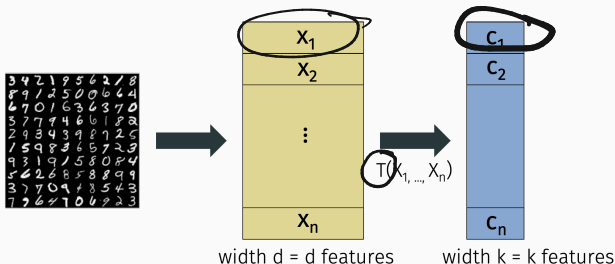


Return to data compression:



Main difference from randomized methods:

$$C_1 = \prod x_1$$



In this section, we will discuss data dependent transformations. Johnson-Lindenstrauss, MinHash, SimHash were all data oblivious.

Advantages of data **independent** methods:

- less computational cost
- easy to analyze, clean + general bounds
- more easily distributed + streaming

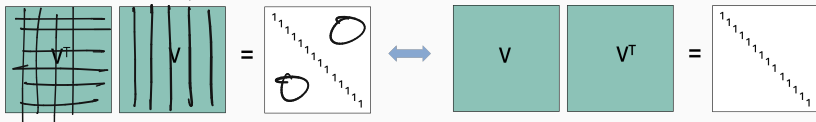
Advantages of data **dependent** methods:

- better compression, in practice
- don't use random oracles
- more interpretable.

LINEAR ALGEBRA REMINDER

If a square matrix has orthonormal rows, it also have orthonormal columns:

→ orthogonal matrix



$$V^T V = I = V V^T$$

→ I

$$(Vx)^T Vx = x^T V^T V x = x^T x = \|x\|_2^2$$

Implies that for any vector x , $\|Vx\|_2^2 = \|x\|_2^2$ and $\|V^T x\|_2^2 = \|x\|_2^2$.

Equivalently, any vector x , $\|x^T V^T\|_2^2 = \|x\|_2^2$ and $\|x^T V\|_2^2 = \|x\|_2^2$.

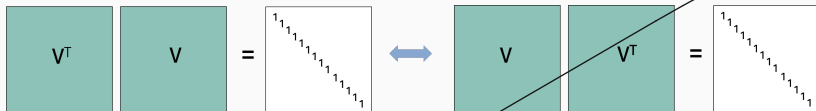
Same thing goes for Frobenius norm: for any matrix X ,

$$\|VX\|_F^2 = \|X\|_F^2 \text{ and } \|V^T X\|_F^2 = \|X\|_F^2.$$

$$\begin{aligned} & \hookrightarrow \|x_1\|_2^2 + \|x_2\|_2^2 + \dots + \|x_n\|_2^2 = \|X\|_F^2 \\ & = \|Vx_1\|_2^2 + \|Vx_2\|_2^2 + \dots + \|Vx_n\|_2^2 \quad \text{where } x_i \text{ is } i^{\text{th}} \text{ column of } X. \end{aligned}$$

LINEAR ALGEBRA REMINDER

If a square matrix has orthonormal rows, it also have orthonormal columns:



$$V^T V = I = V V^T$$

Implies that for any vector \mathbf{x} , $\|\mathbf{V}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2$ and $\|\mathbf{V}^T \mathbf{x}\|_2^2$.

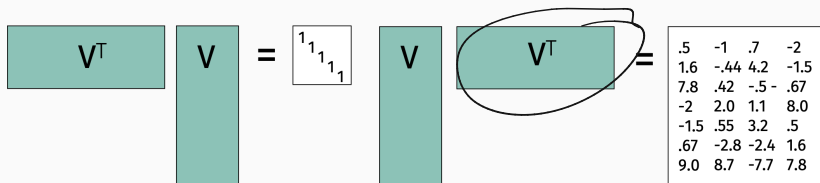
Equivalently, any vector \mathbf{x} , $\|\mathbf{x}^T \mathbf{V}^T\|_2^2 = \|\mathbf{x}\|_2^2$ and $\|\mathbf{x}^T \mathbf{V}\|_2^2 = \|\mathbf{x}\|_2^2$.

Same thing goes for Frobenius norm: for any matrix \mathbf{X} ,

$$\|\mathbf{V}\mathbf{X}\|_F^2 = \|\mathbf{X}\|_F^2 \text{ and } \|\mathbf{V}^T \mathbf{X}\|_F^2 = \|\mathbf{X}\|_F^2.$$

LINEAR ALGEBRA REMINDER

The same is not true for rectangular matrices:



$$V^T V = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$V V^T = \begin{bmatrix} .5 & -1 & .7 & -2 \\ 1.6 & -.44 & 4.2 & -1.5 \\ 7.8 & .42 & -.5 & .67 \\ -2 & 2.0 & 1.1 & 8.0 \\ -1.5 & .55 & 3.2 & .5 \\ .67 & -2.8 & -2.4 & 1.6 \\ 9.0 & 8.7 & -7.7 & 7.8 \end{bmatrix}$$

$$V^T V = I$$

but

$$V V^T \neq I$$

$$\|Vx\|_2^2 = x^T V^T V x = x^T x$$

For any x , $\|Vx\|_2^2 = \|x\|_2^2$ but $\|V^T x\|_2^2 \neq \|x\|_2^2$ in general.

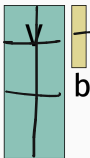
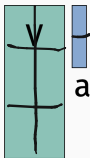
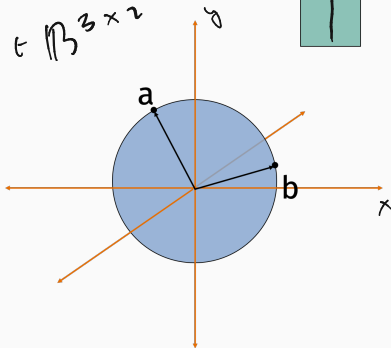
Equivalently, $x, \|x^T V^T\|_2^2 = \|x\|_2^2$ but $\|x^T V\|_2^2 \neq \|x\|_2^2$ in general.

$$(V^T x)^T (V^T x) = x^T \underline{V V^T} x \neq x^T x$$

LINEAR ALGEBRA REMINDER

Multiplying a vector by V with orthonormal columns rotates and/or reflects the vector.

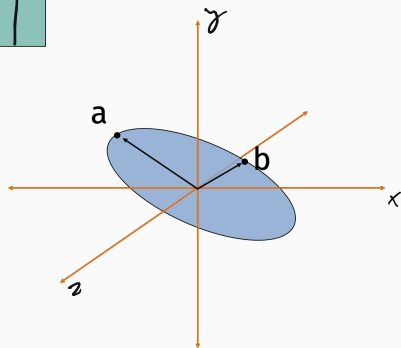
$$\begin{aligned} a &\in \mathbb{R}^2 \\ V a &\in \mathbb{R}^3 \\ V &\in \mathbb{R}^{3 \times 2} \end{aligned}$$



$$\langle V a, V b \rangle$$

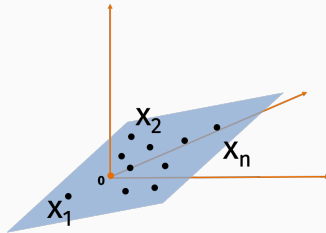
$$= a^T V^T V b$$

$$= a^T b = \langle a, b \rangle$$

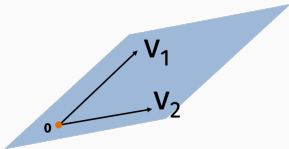


LOW-RANK DATA

Suppose $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ lie on a low-dimensional subspace S through the origin. I.e. our data set is **rank k** for $k < d$.



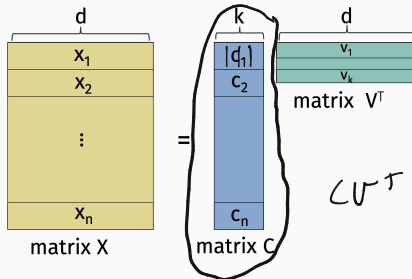
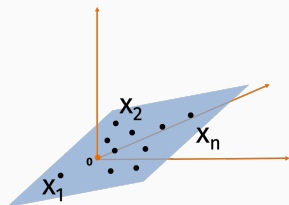
Let $\mathbf{v}_1, \dots, \mathbf{v}_k$ be orthogonal unit vectors spanning S .



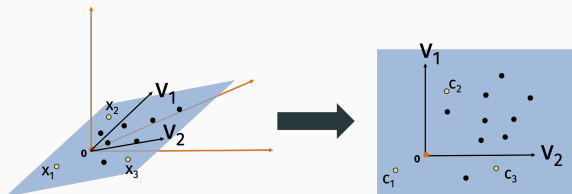
For all i , we can write:

$$\underline{\mathbf{x}_i} = \underline{C_{i,1}} \mathbf{v}_1 + \dots + \underline{C_{i,k}} \mathbf{v}_k.$$

LOW-RANK DATA



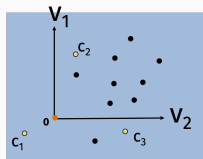
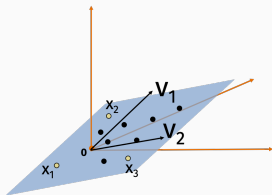
What are c_1, \dots, c_n ?



LOW-RANK DATA

$$X = CV^T \quad XV = CV^TV \quad X(V) \subset C$$

$$V^TV = I$$



$$y = Vz$$

Lots of information preserved:

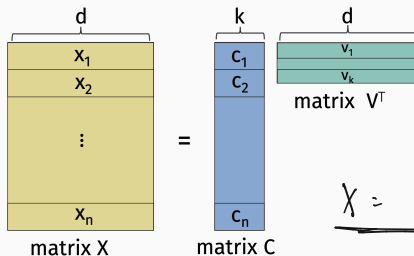
- $\|x_i - x_j\|_2 = \|c_i - c_j\|_2$ for all i, j .
- $x_i^T x_j = c_i^T c_j$ for all i, j .
- Norms preserved, linear separability preserved,

$$\min \|Xy - b\| = \min \|Cz - b\|, \text{ etc., etc.}$$

$$\begin{aligned} \|Xy - b\|_2^2 &= \|(V^T Vz - b)\|_2^2 \\ &= \|Cz - b\|_2^2 \\ &= \|(c_i - c_j) V^T\|_2^2 \\ &= \|V^T (c_i - c_j)\|_2^2 \\ &= \|c_i - c_j\|_2^2 \end{aligned}$$

$$X = CV^T \quad \int \int \frac{V^T}{\gamma} = V^T \gamma$$

LOW-RANK DATA



$$\underline{X = CV^T}$$

Formally, $\underline{C = XV^T}$.

$$XV = C \underbrace{V^T V}$$

$$XV = C$$

$$X = CV^T \Rightarrow XV = CV^T V$$

Since V 's columns are an orthonormal basis, $V^T V = I$.

$$X = CV^T$$

$$\downarrow$$

$$XV = XVV^T$$

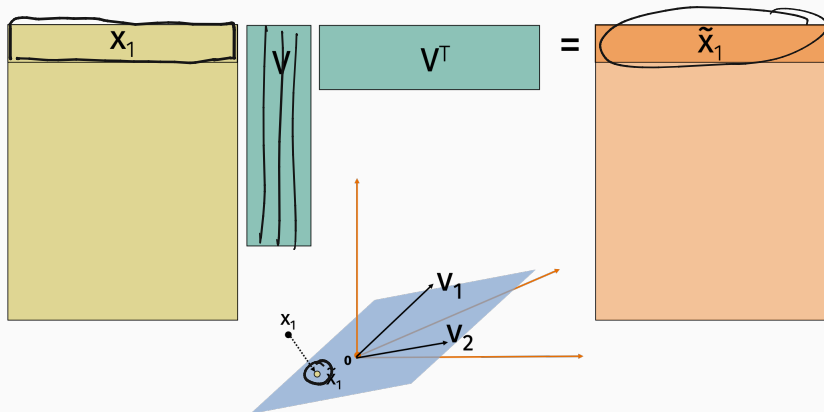
So $\underline{X = XVV^T}$.

$$\underline{A = AUVV^T}$$

PROJECTION MATRICES

VV^T is a symmetric projection matrix.

$$x_i^T (VV^T)$$



When all data points already lie in the subspace spanned by V 's columns, projection doesn't do anything. So $X = XVV^T$.

LOW-RANK APPROXIMATION

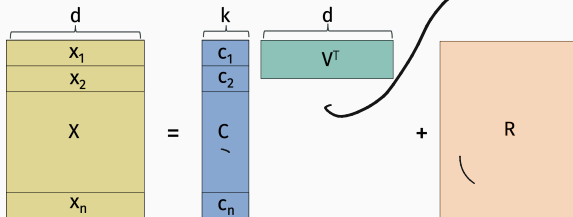
When X 's rows lie close to a k dimensional subspace, we can still approximate

$$X \approx XVV^T.$$

XVV^T is a low-rank approximation for X .

For a given subspace \mathcal{V} spanned by the columns in V ,

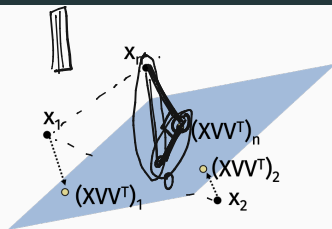
$$XV = \arg \min_C \|X - CV^T\|_F^2 = \sum_{i,j} (x_{i,j} - (CV^T)_{i,j})^2.$$



LOW-RANK APPROXIMATION

$$\|V a\|_2^2 = \|a\|_2^2$$

$$\|a^T V^T\|_2 = \|a\|_2$$



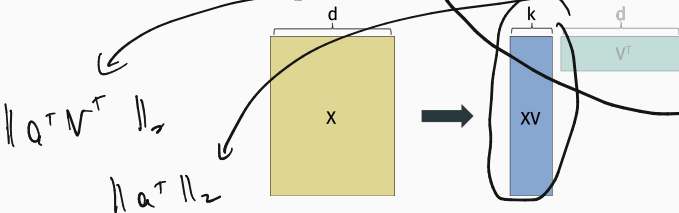
$$(XV)_i$$

$$= X_i^T V$$

$$\|X_i^T V - X_j^T V\|_2$$

$$= \|(X_i - X_j)^T V\|_2$$

$$\|x_i - x_j\|_2 \approx \|(XVV^T)_i - (XVV^T)_j\|_2 = \|(XV^T)_i - (XV^T)_j\|_2$$



$$\|(x_i - x_j)^T V\|_2$$

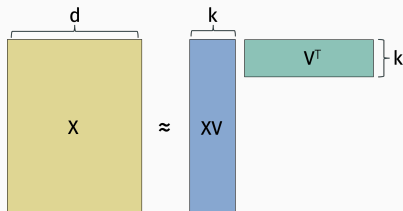
$$a^T = (x_i - x_j)^T V$$

XV can be used as a compressed version of data matrix X .

WHY IS DATA APPROXIMATELY LOW-RANK?

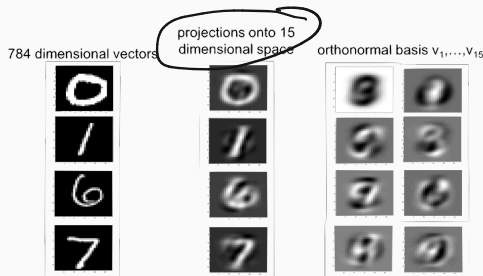
DUAL VIEW

Rows of \mathbf{X} (data points) are approximately spanned by k vectors. Columns of \mathbf{X} (data features) are approximately spanned by k vectors.



ROW REDUNDANCY

If a data set only had k unique data points, it would be exactly rank k . If it has k “clusters” of data points (e.g. the 10 digits) it’s often very close to rank k .



COLUMN REDUNDANCY

Colinearity/correlation of data features leads to a low-rank data matrix.

| | bedrooms | bathrooms | sq.ft. | floors | list price | sale price |
|--------|----------|-----------|--------|--------|------------|------------|
| home 1 | 2 | 2 | 1800 | 2 | 200,000 | 195,000 |
| home 2 | 4 | 2.5 | 2700 | 1 | 300,000 | 310,000 |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| home n | 5 | 3.5 | 3600 | 3 | 450,000 | 450,000 |

OTHER REASONS FOR LOW-RANK STRUCTURE

$$\| (c_i - c_j) V^T \|_2^2$$

transpose

$$\| V (c_i - c_j)^T \|_2^2 = \| V \alpha \|_2^2$$

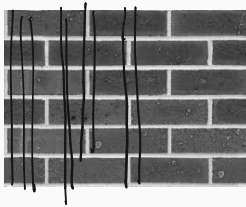
$$= \alpha^T V^T V \alpha$$

$$= \| \alpha \|_2^2$$

When encoded as a matrix, which image has lower approximate rank?



1



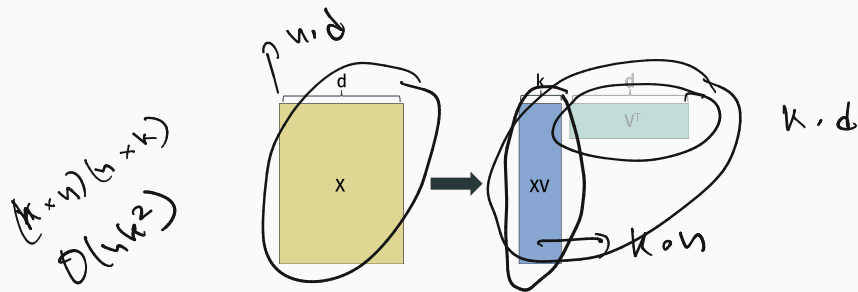
2



3

2, 1, 3

APPLICATIONS OF LOW-RANK APPROXIMATION



- $XV \cdot V^T$ takes $O(k(n+d))$ space to store instead of $O(nd)$.
- Regression problems involving $XV \cdot V^T$ can be solved in $\underline{O(nk^2)}$ instead of $\underline{O(nd^2)}$ time.
- XV can be used for visualization when $k=2,3$.

$$\min \|XV V^T - b\|_2^2$$

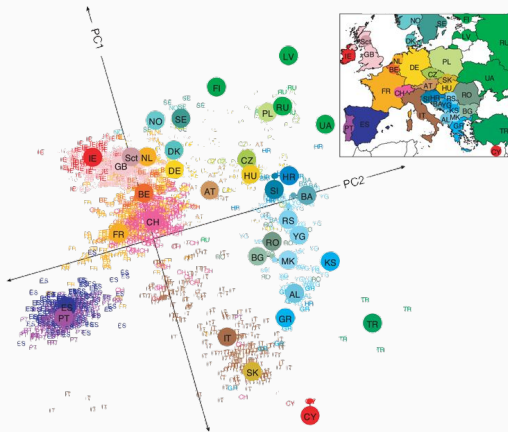
$$\min_z \|XV z - b\|_2^2$$

$$y = \underline{Vz} \quad A \quad (\underline{A^+A})^{-1} A^+ b$$

$$\rightarrow O(nk)$$

APPLICATIONS OF LOW-RANK APPROXIMATION

“Genes Mirror Geography Within Europe” – Nature, 2008.



Each data vector \mathbf{x}_i contains genetic information for one person in Europe. Set $k = 2$ and plot $(XV)_i$ for each i on a 2-d plane. Color points by what country they are from.

COMPUTATIONAL QUESTION

Given a subspace \mathcal{V} spanned by the k columns in \mathbf{V} ,

$$\|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2 = \min_{\mathbf{C}} \|\mathbf{X} - \mathbf{C}\mathbf{V}^T\|_F^2 \quad \checkmark$$

We want to find the best $\mathbf{V} \in \mathbb{R}^{d \times k}$:

$$\min_{\text{orthonormal } \mathbf{V} \in \mathbb{R}^{d \times k}} \|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2 \quad (1)$$

$\|\mathbf{X}\|_F^2 - \|\mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2$

Note that $\|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2 = \|\mathbf{X}\|_F^2 - \|\mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2$ for all orthonormal \mathbf{V} (since $\mathbf{V}\mathbf{V}^T$ is a projection). Equivalent form:

$$\max_{\text{orthonormal } \mathbf{V} \in \mathbb{R}^{d \times k}} \|\mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2 = \|\mathbf{X}\mathbf{V}\|_F^2 \quad (2)$$

$$\text{tr}((\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T)^T (\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T))$$

If $k = 1$, want to find a single vector \mathbf{v}_1 which maximizes:

$$\| \underbrace{\mathbf{X}\mathbf{v}_1\mathbf{v}_1^T}_{\text{rank 1 matrix}} \|_F^2 = \| \underbrace{\mathbf{X}\mathbf{v}_1}_{\text{vector}} \|_F^2 = \| \underbrace{\mathbf{X}\mathbf{v}_1}_{\text{vector}} \|_2^2 = \underbrace{\mathbf{v}_1^T \mathbf{X}^T \mathbf{X} \mathbf{v}_1}_{\text{scalar}}.$$

Choose \mathbf{v}_1 to be the top eigenvector of $\mathbf{X}^T \mathbf{X}$.

What about higher k ?

SINGULAR VALUE DECOMPOSITION

One-stop shop for computing optimal low-rank approximations.

Any matrix X can be written:

The diagram illustrates the Singular Value Decomposition (SVD) of a matrix X . Matrix X is shown as a rectangle with height n and width d . It is equal to the product of three matrices: U (left singular vectors), Σ (singular values), and V^T (right singular vectors). Matrix U is a rectangle with height n and width d , filled with vertical green stripes. Matrix Σ is a square with side length d , filled with a yellow diagonal band containing the singular values $\sigma_1, \sigma_2, \dots, \sigma_{d-1}, \sigma_d$, with zeros elsewhere. Matrix V^T is a rectangle with height d and width d , filled with horizontal blue stripes. Below the matrices, the following equation is written:

$$\|U \Sigma V^T\|_F^2 = \|\Sigma\|_F^2$$

Where $U^T U = I$, $V^T V = I$, and $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_d \geq 0$.

Note that $\sum_{i=1}^d \sigma_i^2 = \|X\|_F^2$.

CONNECTION TO EIGENDECOMPOSITION

- \mathbf{V}_k 's columns are called the “top right singular vectors of \mathbf{X} ”
- \mathbf{U}_k 's columns are called the “top left singular vectors of \mathbf{X} ”
- $\sigma_1, \dots, \sigma_k$ are the “top singular values”. $\sigma_1, \dots, \sigma_d$ are sometimes called the “spectrum of \mathbf{X} ” (although this is more typically used to refer to eigenvalues).
- \mathbf{U} contains the orthonormal eigenvectors of $\mathbf{X}\mathbf{X}^T$
- \mathbf{V} contains the orthonormal eigenvectors of $\mathbf{X}^T\mathbf{X}$.
- $\sigma_i^2 = \lambda_i(\mathbf{X}\mathbf{X}^T) = \lambda_i(\mathbf{X}^T\mathbf{X})$

Exercise: Check this can be checked directly.

SINGULAR VALUE DECOMPOSITION

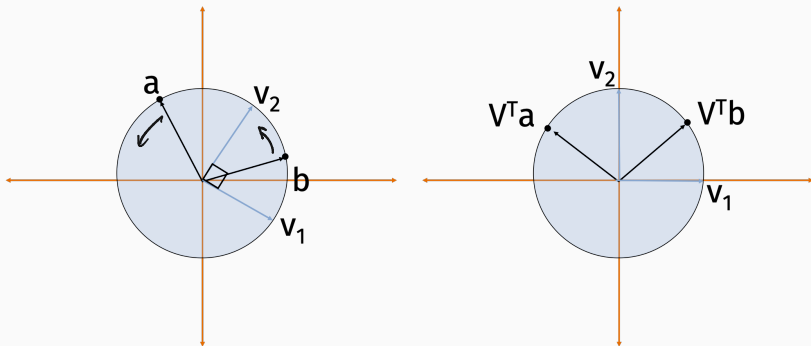
Important take away from singular value decomposition.

$$(U \cdot (\Sigma \cdot (V^T a)))$$

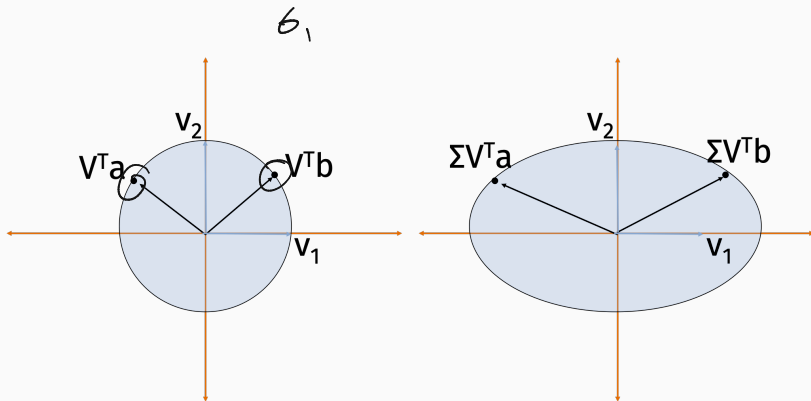
Multiplying any vector \mathbf{a} by a matrix \mathbf{X} to form \mathbf{Xa} can be viewed as a composition of 3 operations:

1. Rotate/reflect the vector (multiplication by \mathbf{V}^T).
2. Scale the coordinates (multiplication by $\mathbf{\Sigma}$).
3. Rotate/reflect the vector again (multiplication by \mathbf{U}).

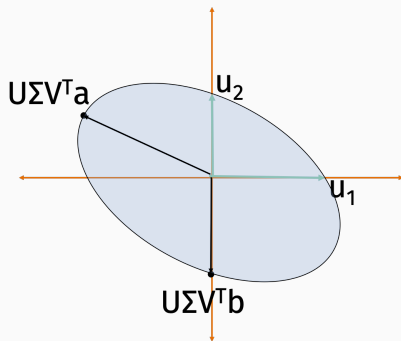
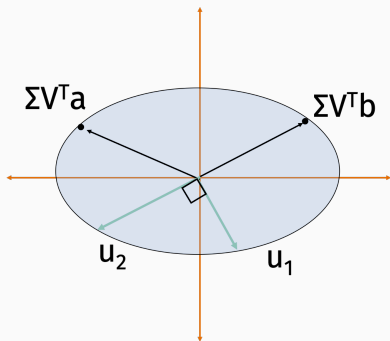
SINGULAR VALUE DECOMPOSITION: ROTATE/REFLECT



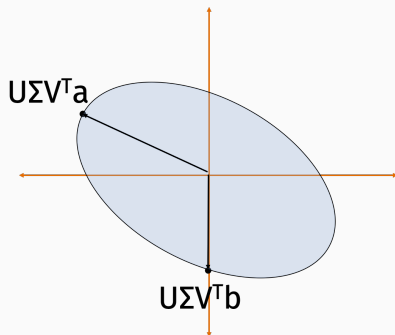
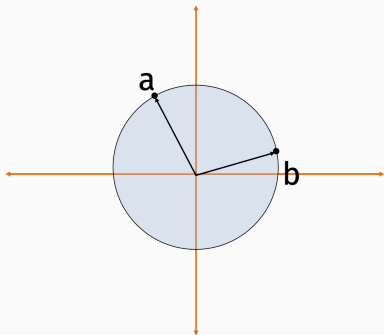
SINGULAR VALUE DECOMPOSITION: STRETCH



SINGULAR VALUE DECOMPOSITION: ROTATE/REFLECT



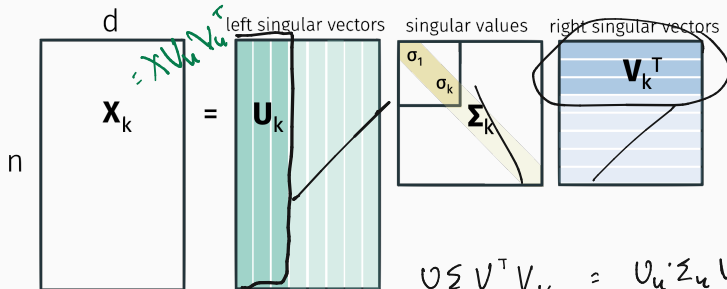
SINGULAR VALUE DECOMPOSITION



SINGULAR VALUE DECOMPOSITION

$$XU = U \Sigma W^T = U \begin{bmatrix} \sigma_1 & & \\ & \sigma_k & \\ & & \dots \end{bmatrix} = \begin{bmatrix} U_1 \sigma_1 & U_2 \sigma_2 & \dots & U_k \sigma_k \end{bmatrix}$$

Can read off optimal low-rank approximations from the SVD:



$$U \Sigma V^T V_k = U_k \Sigma_k V_k^T$$

$$X_k = U_k \Sigma_k V_k^T = U_k U_k^T X = \underbrace{X V_k^T}_{\text{circled}} V_k$$

$$V_k = \arg \min_{\text{orthonormal } V \in \mathbb{R}^{d \times k}} \|X - X V V^T\|_F^2 = \arg \max_{\text{orthonormal } V \in \mathbb{R}^{d \times k}} \|X V V^T\|_F^2$$

$$V_k = \begin{bmatrix} v_{k1} & \dots & v_{kn} \end{bmatrix}$$

$$X V_k V_k^T$$

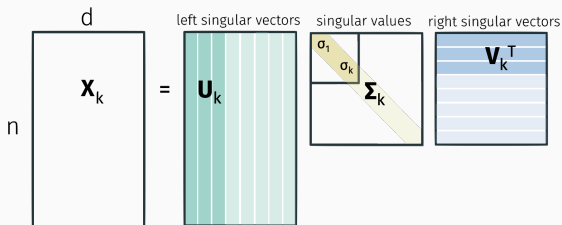
$$U \Sigma V^T V_k$$

$$\begin{bmatrix} V_k^T \\ \vdots \end{bmatrix} \begin{bmatrix} V_k \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots \\ 0 & 1 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} \rightarrow W$$

Connection to **Principal Component Analysis**:

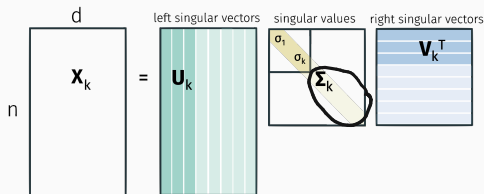
- Let $\bar{\mathbf{X}} = \mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T$ where $\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$. I.e. $\bar{\mathbf{X}}$ is obtained by mean centering \mathbf{X} 's rows.
- Let $\bar{\mathbf{U}}\bar{\boldsymbol{\Sigma}}\bar{\mathbf{V}}^T$ be the SVD of $\bar{\mathbf{X}}$. $\bar{\mathbf{U}}$'s first columns are the “top principal components” of \mathbf{X} . $\bar{\mathbf{V}}$'s first columns are the “weight vectors” for these principal components.

USEFUL OBSERVATIONS



Observation 1: The optimal compression \mathbf{XV}_k has orthogonal columns.

USEFUL OBSERVATIONS



Observation 2: The optimal low-rank approximation error

$E_k = \|\mathbf{X} - \mathbf{X}\mathbf{V}_k\mathbf{V}_k^T\|_F^2 = \|\mathbf{X}\|_F^2 - \|\mathbf{X}\mathbf{V}_k\mathbf{V}_k^T\|_F^2$ can be written:

$$E_k = \sum_{i=k+1}^d \sigma_i^2.$$

$$\sum_{i=1}^d \sigma_i^2 - \sum_{i=1}^k \sigma_i^2 = \sum_{i=k+1}^d \sigma_i^2$$

SPECTRAL PLOTS

Observation 2: The optimal low-rank approximation error

$E_k = \|\mathbf{X} - \mathbf{X}\mathbf{V}_k\mathbf{V}_k^T\|_F^2 = \|\mathbf{X}\|_F^2 - \|\mathbf{X}\mathbf{V}_k\mathbf{V}_k^T\|_F^2$ can be written:

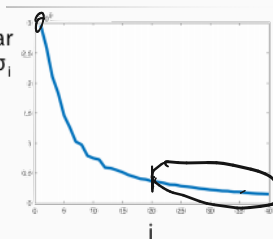
$$E_k = \sum_{i=k+1}^d \sigma_i^2.$$

Can immediately get a sense of “how low-rank” a matrix is from it’s spectrum:

784 dimensional vectors



singular
value σ_i



SPECTRAL PLOTS

Observation 2: The optimal low-rank approximation error

$E_k = \|\mathbf{X} - \mathbf{X}\mathbf{V}_k\mathbf{V}_k^T\|_F^2 = \|\mathbf{X}\|_F^2 - \|\mathbf{X}\mathbf{V}_k\mathbf{V}_k^T\|_F^2$ can be written:

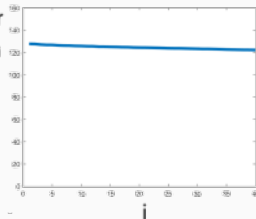
$$E_k = \sum_{i=k+1}^d \sigma_i^2.$$

Can immediately get a sense of “how low-rank” a matrix is from it’s spectrum:

784 dimensional vectors



singular
value σ_i



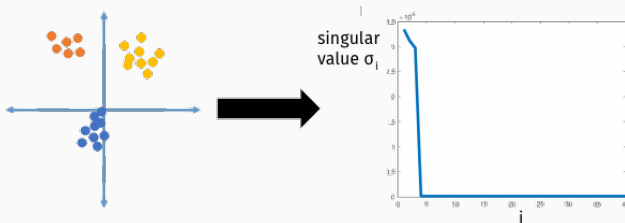
SPECTRAL PLOTS

Observation 2: The optimal low-rank approximation error

$E_k = \|\mathbf{X} - \mathbf{X}\mathbf{V}_k\mathbf{V}_k^T\|_F^2 = \|\mathbf{X}\|_F^2 - \|\mathbf{X}\mathbf{V}_k\mathbf{V}_k^T\|_F^2$ can be written:

$$E_k = \sum_{i=k+1}^d \sigma_i^2.$$

Can immediately get a sense of “how low-rank” a matrix is from it’s spectrum:



COMPUTING THE SVD

Suffices to compute right singular vectors V :

- Compute $X^T X$.

- Find eigendecomposition $\underline{V \Lambda V^T} = \underline{X^T X}$.

- Compute $L = XV$. Set $\sigma_i = \|L_i\|_2$ and $U_i = L_i / \|L_i\|_2$.

$$\begin{aligned}(U \Sigma U^T)^T U \Sigma V^T &= V^T \cancel{U^T U} \Sigma U^T \\ &= V^T \Sigma^2 V^T\end{aligned}$$

$$XV = U \Sigma U^T V = U \Sigma \cdot \underline{\Sigma^{-1}} = U$$

$$\begin{aligned}\text{Total runtime} &\approx O(nd^2) \\ &\quad + O(d^3)\end{aligned}$$

$$O(\underline{nd} \cdot \underline{\text{(something small)}})$$

COMPUTING THE SVD (FASTER)

- Compute approximate solution.
- Only compute top k singular vectors/values. Runtime will depend on k . When $k = d$ we can't do any better than classical algorithms based on eigendecomposition.
- Iterative algorithms achieve runtime $\approx \underline{O(ndk)}$ vs. $\underline{O(nd^2)}$ time.
 - Krylov subspace methods like the Lanczos method are most commonly used in practice.
 - Power method is the simplest Krylov subspace method, and still works very well.

$K=1$

What we won't discuss today: sketching methods and stochastic methods (which are faster in some settings).

POWER METHOD

Today: What about when $k = 1$?

Goal: Find some $\underline{\underline{z}} \approx \mathbf{v}_1$. $\|z - \mathbf{v}_1\|_2 \leq \epsilon$

Input: $X \in \mathbb{R}^{n \times d}$ with SVD $U \Sigma V^T$.

Power method:

• Choose $\underline{\underline{z}}^{(0)}$ randomly. E.g. $z_0 \sim \mathcal{N}(0, 1)$.

• $\underline{\underline{z}}^{(0)} = \underline{\underline{z}}^{(0)} / \|\underline{\underline{z}}^{(0)}\|_2$

• For $i = 1, \dots, T$

• $\underline{\underline{z}}^{(i)} = X^T \cdot (X \underline{\underline{z}}^{(i-1)})$

• $\underline{\underline{n}}_i = \|\underline{\underline{z}}^{(i)}\|_2$

• $\underline{\underline{z}}^{(i)} = \underline{\underline{z}}^{(i)} / \underline{\underline{n}}_i$

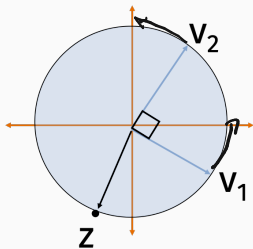
Return $\underline{\underline{z}}^{(T)}$

$$\underline{\underline{V}} \underline{\underline{\Sigma^2}} \underline{\underline{V}}^T$$

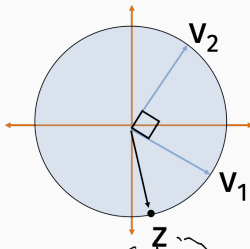
POWER METHOD INTUITION

$$X^T X z \approx V \Sigma^2 V^T z$$

0 iterations

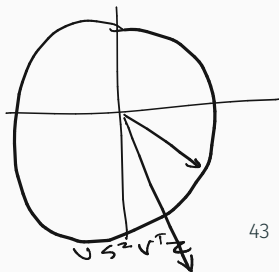
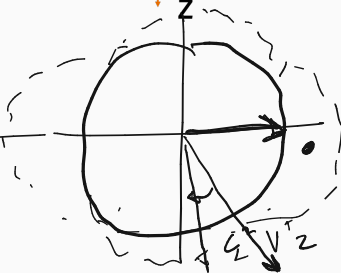
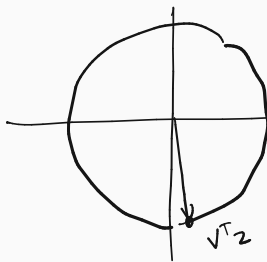
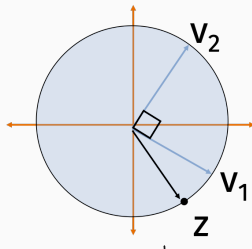


1 iterations



$$b^2 = .001$$

2 iterations



POWER METHOD FORMAL CONVERGENCE

Theorem (Basic Power Method Convergence)

Let $\gamma = \frac{\sigma_1 - \sigma_2}{\sigma_1}$ be parameter capturing the “gap” between the first and second largest singular values. If Power Method is initialized with a random Gaussian vector then, with high probability, after $T = O\left(\frac{\log d/\epsilon}{\gamma}\right)$ steps, we have either:

$$\|\underline{\mathbf{v}}_1 - \underline{\mathbf{z}}^{(T)}\|_2 \leq \epsilon \quad \text{or} \quad \|\underline{\mathbf{v}}_1 - (-\mathbf{z}^{(T)})\|_2 \leq \epsilon.$$

Total runtime: $O\left(\underline{nd} \cdot \frac{\log d/\epsilon}{\gamma}\right)$

$$1 - \frac{\sigma_2}{\sigma_1}$$

ONE STEP ANALYSIS OF POWER METHOD

Write $\underline{\mathbf{z}}^{(i)}$ in the right singular vector basis:

$$\underline{\mathbf{z}}^{(0)} = c_1^{(0)} \underline{\mathbf{v}}_1 + c_2^{(0)} \underline{\mathbf{v}}_2 + \dots + c_d^{(0)} \underline{\mathbf{v}}_d$$

$$\underline{\mathbf{z}}^{(1)} = c_1^{(1)} \underline{\mathbf{v}}_1 + c_2^{(1)} \underline{\mathbf{v}}_2 + \dots + c_d^{(1)} \underline{\mathbf{v}}_d$$

$$\vdots$$

$$\underline{\mathbf{z}}^{(i)} = c_1^{(i)} \underline{\mathbf{v}}_1 + c_2^{(i)} \underline{\mathbf{v}}_2 + \dots + c_d^{(i)} \underline{\mathbf{v}}_d$$

Note: $[c_1^{(i)}, \dots, c_d^{(i)}] = \mathbf{c}^{(i)} = \mathbf{V}^T \mathbf{z}^{(i)}$.

Also: $\sum_{j=1}^d \left(c_j^{(i)}\right)^2 = 1$.

ONE STEP ANALYSIS OF POWER METHOD

Claim: After update $\underline{z}^{(i)} = \underline{X}^T \underline{X} \underline{z}^{(i-1)}$, $(\underline{U} \underline{\Sigma} \underline{V}^T)^T \underline{U} \underline{\Sigma} \underline{V}^T$

$\underline{c}_j^{(i)} = \sigma_j^2 \underline{c}_j^{(i-1)}$

$= \underline{V} \underline{\Sigma} \underline{U}^T \underline{U} \underline{\Sigma} \underline{V}^T$
 $= \underline{V} \underline{\Sigma}^2 \underline{V}^T$

$$\underline{z}^{(i)} = \frac{1}{n_1} \left[\underline{c}_1^{(i-1)} \sigma_1^2 \cdot \underline{v}_1 + \underline{c}_2^{(i-1)} \sigma_2^2 \cdot \underline{v}_2 + \dots + \underline{c}_d^{(i-1)} \sigma_d^2 \cdot \underline{v}_d \right]$$

$$\underline{V} \underline{\Sigma}^2 \underline{V}^T \underline{z}^{(i-1)}$$

$$[\underline{c}_1^{(i-1)} \dots \underline{c}_d^{(i-1)}]$$

$$\underline{V} [\sigma_1^2 \underline{c}_1^{(i-1)} \dots \sigma_d^2 \underline{c}_d^{(i-1)}]$$

MULTI-STEP ANALYSIS OF POWER METHOD

Claim: After T updates:

$$\mathbf{z}^{(T)} = \frac{1}{\prod_{i=1}^T n_i} \left[\underline{c_1^{(0)}} \underline{\sigma_1^{2T}} \cdot \mathbf{v}_1 + c_2^{(0)} \sigma_2^{2T} \cdot \mathbf{v}_2 + \dots + c_d^{(0)} \sigma_d^{2T} \cdot \mathbf{v}_d \right]$$

$$\frac{\|\mathbf{z}^{(T)}\|_2}{\|\mathbf{v}_1\|_2} \downarrow \approx 1$$

$$2^{(1)}$$

$$\mathbf{v}_j$$

Let $\underline{\alpha_j} = \frac{1}{\prod_{i=1}^T n_i} \underline{c_j^{(0)} \sigma_j^{2T}}$. Goal: Show that $\underline{\alpha_j} \ll \alpha_1$ for all $j \neq 1$.

POWER METHOD FORMAL CONVERGENCE

Since $\mathbf{z}^{(T)}$ is a unit vector, $\sum_{i=1}^d \alpha_i^2 = 1$. So $\alpha_1 \leq 1$.

If we can prove that $\frac{\alpha_j}{\alpha_1} \leq \sqrt{\frac{\epsilon}{d}}$ then: $\alpha_j \leq \sqrt{\frac{\epsilon}{d}}$

$$\alpha_1^2 \geq 1 - d \cdot \left(\sqrt{\frac{\epsilon}{d}} \right)^2 \Rightarrow \underline{|\alpha_1| \geq 1 - \epsilon}$$

$$\alpha_1^2 = 1 - \sum_{j=2}^d \alpha_j^2 \geq 1 - d \cdot \left(\sqrt{\frac{\epsilon}{d}} \right)^2$$

$$\underline{\|\mathbf{v}_1 - \mathbf{z}^{(T)}\|_2} = \underline{2 - 2 \langle \mathbf{v}_1, \mathbf{z}^{(T)} \rangle} \leq 2\epsilon$$

$$\alpha_1 \approx 1 - \epsilon$$

$$2 - 2(1 - \epsilon) \leq 2\epsilon$$

POWER METHOD FORMAL CONVERGENCE

Lets prove that $\frac{\alpha_j}{\alpha_1} \leq \sqrt{\frac{\epsilon}{d}}$ where $\alpha_j = \frac{1}{\prod_{i=1}^T n_i} c_j^{(0)} \sigma_j^{2T}$

First observation: Starting coefficients are all roughly equal.

For all j $\underline{O(1/d^3)} \leq \underline{c_j^{(0)}} \leq \underline{1}$

with probability $1 - \frac{1}{d}$. This is a very loose bound, but it's all that we will need. **Prove at home.**

$$(1-x)^{1/x} \approx \frac{1}{e}$$

$$\frac{\alpha_j}{\alpha_1} = \frac{\sigma_j^{2T}}{\sigma_1^{2T}} \frac{c_j^{(0)}}{c_1^{(0)}} \leq$$

$$\frac{G_j^{2T}}{G_1^{2T}} \cdot \underline{O(d^3)}$$

with $1 - \frac{1}{d}$ probability

$$\leq \sqrt{\frac{\epsilon}{d}}$$

$$\left(\frac{G_j}{G_1}\right)^{2T} \leq \left(\frac{\sqrt{\epsilon}}{d^{3.5}}\right)^{2T}$$

Need $T =$

$$O\left(\frac{1}{1 - \frac{\epsilon_j}{\epsilon_1}} \cdot \log(d/\epsilon)\right)$$

$$\left(\frac{1}{1 - \frac{\epsilon_j}{\epsilon_1}}\right) \cdot \log(d^{3.5}/\sqrt{\epsilon})$$

Theorem (Gapless Power Method Convergence)

If Power Method is initialized with a random Gaussian vector then, with high probability, after $T = O\left(\frac{\log d/\epsilon}{\epsilon}\right)$ steps, we obtain a \mathbf{z} satisfying:

$$\|\mathbf{X} - \mathbf{X}\mathbf{z}\mathbf{z}^T\|_F^2 \leq (1 + \epsilon) \|\mathbf{X} - \mathbf{X}\mathbf{v}_1\mathbf{v}_1^T\|_F^2$$

GENERALIZATIONS TO LARGER k

- Block Power Method aka Simultaneous Iteration aka Subspace Iteration aka Orthogonal Iteration

Power method:

- Choose $\mathbf{G} \in \mathbb{R}^{d \times k}$ be a random Gaussian matrix.
- $\mathbf{Z}_0 = \text{orth}(\mathbf{G})$.
- For $i = 1, \dots, T$
 - $\mathbf{Z}^{(i)} = \mathbf{X}^T \cdot (\mathbf{X} \mathbf{Z}^{(i-1)})$
 - $\mathbf{Z}^{(i)} = \text{orth}(\mathbf{Z}^{(i)})$

Return $\mathbf{Z}^{(T)}$

Runtime: $O\left(\frac{\log d/\epsilon}{\epsilon}\right)$ iterations to obtain a nearly optimal low-rank approximation:

$$\|\mathbf{X} - \mathbf{X} \mathbf{Z} \mathbf{Z}^T\|_F^2 \leq (1 + \epsilon) \|\mathbf{X} - \mathbf{X} \mathbf{v}_k \mathbf{v}_k^T\|_F^2.$$

