

CSE 392/CS 395T/M 397C: Matrix and Tensor Algorithms for Data

Instructor: Shashanka Ubaru

University of Texas, Austin
Spring 2025

Lecture 12: Subspace iteration (power) method

- 1 Iterative methods
- 2 Subspace iteration methods
 - Power method
 - Block power method

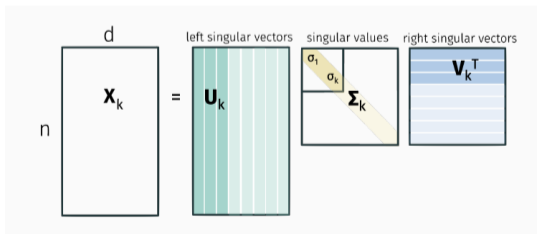
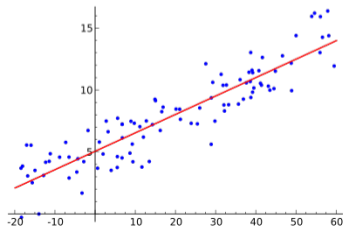
Covered so far:

- Linear least squares regression and Low rank approximation.
- *Linear Regression*: Given a data matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ and a column vector $\mathbf{b} \in \mathbb{R}^n$, *least-squares* regression solves:

$$\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2. \quad (1)$$

- *Low rank approximation*: Given a data matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ and integer k , find a rank- k approximation of \mathbf{A} , such that.

$$\mathbf{A}_k = \arg \min_{\mathbf{W}: \text{rank}(\mathbf{W})=k} \|\mathbf{A} - \mathbf{W}\|_F. \quad (2)$$

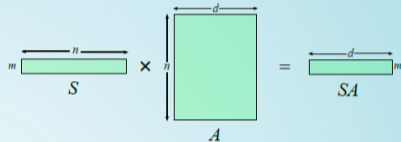


Covered so far: Sketching

SKETCH AND SOLVE

Generic scheme using sketching:

```
generate sketching matrix  $\mathbf{S} \in \mathbb{R}^{m \times n}$ ,  
compute  $\mathbf{SA}$  and  $\mathbf{Sb}$   
return  $\tilde{\mathbf{x}} := \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{S}(\mathbf{Ax} - \mathbf{b})\|$ 
```



- Oblivious sketching - subspace embedding property.
- $\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\| \leq (1 + \epsilon)\|\mathbf{Ax}^* - \mathbf{b}\|$.
- Similarly for low rank approximation: Suppose $\tilde{\mathbf{A}}_k$ is rank k approximation obtained using sketching \mathbf{AS} , then

$$\|\mathbf{A} - \tilde{\mathbf{A}}_k\|_F \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F.$$

- *Skylark project*: open source library for distributed randomized numerical linear algebra, funded through XDATA program by **DARPA** and **Air Force Research Laboratory**.

Iterative methods

- *Sketching methods* : Single pass over data. Advantageous when data is too large to fit in memory. Streaming settings.
- *Sketch size*: For rank- k approximation, for dense input matrices - Gaussian - $O\left(\frac{k}{\epsilon}\right)$ or SRFT/SRHT - $O\left(\frac{k \log(k/\epsilon)}{\epsilon}\right)$.
Sparse matrices - Countsketch - $O\left(\frac{k^2}{\epsilon}\right)$.

Iterative methods

- *Sketching methods* : Single pass over data. Advantageous when data is too large to fit in memory. Streaming settings.
- *Sketch size*: For rank- k approximation, for dense input matrices - Gaussian - $O\left(\frac{k}{\epsilon}\right)$ or SRFT/SRHT - $O\left(\frac{k \log(k/\epsilon)}{\epsilon}\right)$.
Sparse matrices - Countsketch - $O\left(\frac{k^2}{\epsilon}\right)$.
- *Iterative methods* - Multiple passes over data. Improved numerical results. Predate sketching methods.
- In *numerous fields* (system solvers, optimization, control systems, PDE solvers, scientific computing, NLP, etc.) and *many industry* (oil refineries, auto modeling, electronics, Google and Twitter (X?) and many more.)
- Partial SVD - compute top k singular vectors/values.
 - 1 Subspace iteration or block power method.
 - 2 Krylov subspace method.

Recall : PageRank

- PageRank value of a page is given as:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)},$$

p_1, p_2, \dots, p_N are the pages, $M(p_i)$ = set of pages that link to p_i , $L(p_j)$ = number of outbound links on page p_j , N = total number of pages, and d = damping factor.

- The values are the entries of the dominant right eigenvector of the modified adjacency matrix rescaled so that each column adds up to one.

$$\mathbf{r} = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{bmatrix}$$

- \mathbf{r} is the solution of the equation

$$\mathbf{r} = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} \ell(p_1, p_1) & \ell(p_1, p_2) & \cdots & \ell(p_1, p_N) \\ \ell(p_2, p_1) & \ddots & & \vdots \\ \vdots & & \ell(p_i, p_j) & \\ \ell(p_N, p_1) & \cdots & & \ell(p_N, p_N) \end{bmatrix} \mathbf{r}$$

the adjacency function $\ell(p_i, p_j)$ is the ratio between number of links outbound from page j to page i to the total number of outbound links of page j .

- $$\sum_{i=1}^N \ell(p_i, p_j) = 1,$$

The matrix is a stochastic matrix. Closely related to the problem of finding the stationary points of Markov processes. It is also a variant of the eigenvector centrality measure used commonly in network analysis.

Subspace iteration methods

Questions

- Given a symmetric matrix \mathbf{A} with eigen-decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, then
 - ① What are the eigenvalues/eigenvectors of \mathbf{A}^q for a given integer power q ?
 - ② If \mathbf{A} is nonsingular what are the eigenvalues/eigenvectors of \mathbf{A}^{-1} ?
 - ③ What are the eigenvalues/eigenvectors of $p(\mathbf{A})$ for a polynomial $p(\cdot)$?
- If the matrix \mathbf{A} has a certain spectral gap $|\lambda_1 - \lambda_2|$, what can we say about the spectral gap of \mathbf{A}^2 ? Does it increase, decrease or remain the same in general?
- Similarly, for a general matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, with SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, what are the singular/eigen-values of $\mathbf{A}^\top \mathbf{A}$?

Power Method

- Let us start with $k = 1$ (finding the top singular vector/value).
- Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, with SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, find a vector $\mathbf{z} \approx \mathbf{v}_1$.

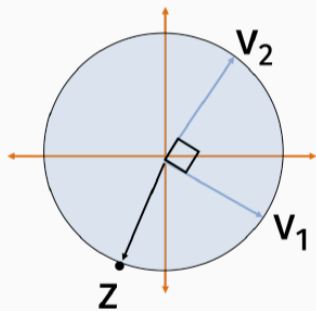
Power Method

- Choose a random vector \mathbf{z}_0 , E.g., $\mathbf{z}_0 \sim \mathcal{N}(0, 1)$.
- $\mathbf{z}_0 = \mathbf{z}_0 / \|\mathbf{z}_0\|_2$
- For $l = 1, \dots, q$
 - ▶ $\mathbf{z}_l = \mathbf{A}^\top (\mathbf{A}\mathbf{z}_{l-1})$
 - ▶ $\mathbf{z}_l = \mathbf{z}_l / \|\mathbf{z}_l\|_2$
- Return \mathbf{z}_q

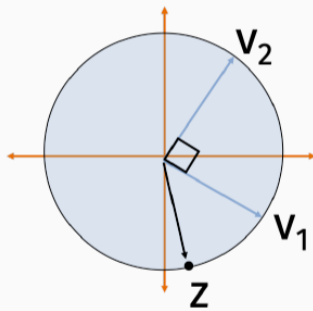
Runtime = ?

Power method intuition

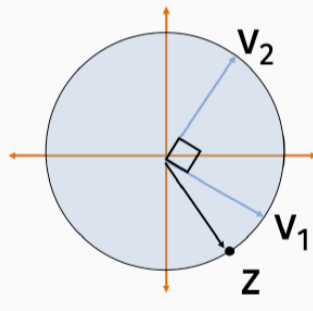
0 iterations



1 iterations



2 iterations



Convergence

Theorem (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 with $\mathbf{A} \in \mathbb{R}^{n \times d}$ then, with high probability, after $q = O\left(\frac{\log d/\epsilon}{\gamma}\right)$ steps, we have:

$$\|\mathbf{v}_1 - \mathbf{z}_q\|_2 \leq \epsilon.$$

Total runtime: $O(\text{nnz}(\mathbf{A})q) = O\left(\text{nnz}(\mathbf{A}) \cdot \frac{\log d/\epsilon}{\gamma}\right)$.

Above also implies, $\|\mathbf{A}\mathbf{z}_q\mathbf{z}_q^\top\|_F^2 \geq (1 - \epsilon)^2 \|\mathbf{A}\mathbf{v}_1\mathbf{v}_1^\top\|_F^2$.

Proof

- Let us write $\mathbf{z}_0 = \sum_{i=1}^d \mu_i \mathbf{v}_i$ in terms of the right singular vector basis.
- If $\boldsymbol{\mu} = [\mu_1, \dots, \mu_d]$, we have $\boldsymbol{\mu} = \mathbf{V}^\top \mathbf{g} / \|\mathbf{g}\|_2$ for random Gaussian \mathbf{g} .
- Since \mathbf{V} is orthogonal, we have $\|\boldsymbol{\mu}\|^2 = 1$.
- With high probability,

$$1/\text{poly}(d) \leq |\mu_i| \leq 1 \quad i = 1, \dots, d.$$

Note that $\boldsymbol{\mu}$ is Gaussian. We can show that $\text{poly}(d) \approx d^3$ with high probability.

- After q steps, we have $\mathbf{z}_q = c(\mathbf{A}^\top \mathbf{A})^q \mathbf{z}_0$ for some scaling c .
- If we write $\mathbf{z}_q = \sum_{i=1}^d \rho_i \mathbf{v}_i$, we have

$$\rho_i = c\sigma_i^{2q} \mu_i.$$

Since $\mathbf{A}^\top \mathbf{A} = \mathbf{V}\Sigma^2\mathbf{V}^\top$.

- After q steps, we have $\mathbf{z}_q = c(\mathbf{A}^\top \mathbf{A})^q \mathbf{z}_0$ for some scaling c .
- If we write $\mathbf{z}_q = \sum_{i=1}^d \rho_i \mathbf{v}_i$, we have

$$\rho_i = c\sigma_i^{2q} \mu_i.$$

Since $\mathbf{A}^\top \mathbf{A} = \mathbf{V}\Sigma^2\mathbf{V}^\top$.

- If the gap parameter is $\gamma = \frac{\sigma_1 - \sigma_2}{\sigma_1}$, we can show that, for all $j \geq 2$:

$$\frac{\sigma_j}{\sigma_1} \leq (1 - \gamma).$$

- For all $j \geq 2$,

$$\frac{|\rho_j|}{|\rho_1|} \leq (1 - \gamma)^{2q} \frac{|\mu_j|}{|\mu_1|} \leq (1 - \gamma)^{2q} \text{poly}(d).$$

- For any $0 < x \leq 1$, we can show that $(1 - x)^{\frac{q}{x}} \leq e^{-q}$.
(Hint: use Taylor series for $\log(1 - x)$).
- If we set $q = \frac{\log(\text{poly}(d)\sqrt{d/\epsilon})}{\gamma} = O\left(\frac{\log d/\epsilon}{\gamma}\right)$, then we get $\frac{|\rho_j|}{|\rho_1|} \leq \sqrt{\epsilon/d}$.
- Since \mathbf{z}_q is a unit vector, we have $\sum_i \rho_i^2 = 1$, and $|\rho_1| \leq 1$, hence

$$\rho_1^2 \geq 1 - d(\sqrt{\epsilon/d})^2 \implies |\rho_1| \geq 1 - \epsilon.$$

Therefore,

$$\|\mathbf{v}_1 - \mathbf{z}_q\|_2^2 = 2 - 2\langle \mathbf{v}_1, \mathbf{z}_q \rangle \leq 2\epsilon.$$

Analysis without gap

Theorem (Gapless Power Method Convergence)

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

$$\|\mathbf{A} - \mathbf{A}\mathbf{z}_q\mathbf{z}_q^\top\|_F^2 \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}\mathbf{v}_1\mathbf{v}_1^\top\|_F^2.$$

Gap γ might be too small. Then, we do not care to find \mathbf{v}_1 . Say, $\sigma_1 = \sigma_2$, then \mathbf{v}_2 is as good as \mathbf{v}_1 .

Proof:

We know that $\|\mathbf{A} - \mathbf{A}\mathbf{z}_q\mathbf{z}_q^T\|_F^2 = \|\mathbf{A}\|_F^2 - \|\mathbf{A}\mathbf{z}_q\mathbf{z}_q^T\|_F^2$.

So, to prove the above, we need to show $\|\mathbf{A}\mathbf{z}_q\|_2^2 \geq (1 - \epsilon)^2 \sigma_1^2$.

We have,

$$\|\mathbf{A}\mathbf{z}_q\|_2^2 = \mathbf{z}_q^T \mathbf{A}^T \mathbf{A} \mathbf{z}_q = \sum_{i=1}^d \rho_i^2 \sigma_i^2,$$

where $\rho_i = \mathbf{v}_i^T \mathbf{z}_q$.

Proof:

We know that $\|\mathbf{A} - \mathbf{A}\mathbf{z}_q\mathbf{z}_q^T\|_F^2 = \|\mathbf{A}\|_F^2 - \|\mathbf{A}\mathbf{z}_q\mathbf{z}_q^T\|_F^2$.

So, to prove the above, we need to show $\|\mathbf{A}\mathbf{z}_q\|_2^2 \geq (1 - \epsilon)^2 \sigma_1^2$.

We have,

$$\|\mathbf{A}\mathbf{z}_q\|_2^2 = \mathbf{z}_q^T \mathbf{A}^T \mathbf{A} \mathbf{z}_q = \sum_{i=1}^d \rho_i^2 \sigma_i^2,$$

where $\rho_i = \mathbf{v}_i^T \mathbf{z}_q$.

For $q = O\left(\frac{\log d/\epsilon}{\epsilon}\right)$, from our previous analysis we have $\rho_1 \geq (1 - \epsilon)$. Hence,

$$\|\mathbf{A}\mathbf{z}_q\|_2^2 = \sum_{i=1}^d \rho_i^2 \sigma_i^2 \geq \rho_1^2 \sigma_1^2 \geq (1 - \epsilon)^2 \sigma_1^2.$$

Subspace iteration

- For larger $k \geq 1$ (finding the top- k singular vectors/values).
- Block Power Method aka Simultaneous Iteration aka Subspace Iteration aka Orthogonal Iteration.

Block Power Method

- Choose $\mathbf{S} \in \mathbb{R}^{d \times k}$ a random Gaussian matrix .
- $\mathbf{Z}_0 = \text{orth}(\mathbf{S})$
- For $l = 1, \dots, q$
 - ▶ $\mathbf{Z}_l = \mathbf{A}^\top (\mathbf{A} \mathbf{Z}_{l-1})$
 - ▶ $\mathbf{Z}_l = \text{orth}(\mathbf{Z}_l)$.
- Return \mathbf{Z}_q

Total runtime: $O(\text{nnz}(\mathbf{A})kq)$.

Subspace iteration

- Equivalent to sketching with input $(\mathbf{A}^\top \mathbf{A})^q$.
- With $q = O\left(\frac{\log d/\epsilon}{\epsilon}\right)$, we obtain a nearly optimal low-rank approximation:

$$\|\mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^\top\|_F^2 \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}\mathbf{V}_k\mathbf{V}_k^\top\|_F^2.$$

- For $q = O\left(\frac{\log(nd)}{\epsilon}\right)$, we have

$$\|\mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^\top\|_2 \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_2.$$

Further Reading:

- *Sketching as a Tool for Numerical Linear Algebra* by David Woodruff.
- *Subspace iteration randomization and singular value problems* by Ming Gu.
- *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions* by N Halko, P. Martinsson and J. Tropp.

Matlab Demo