CSE 392: Matrix and Tensor Algorithms for Data

Instructor: Shashanka Ubaru

University of Texas, Austin

Spring 2024

Lecture 11: Randomized SVD

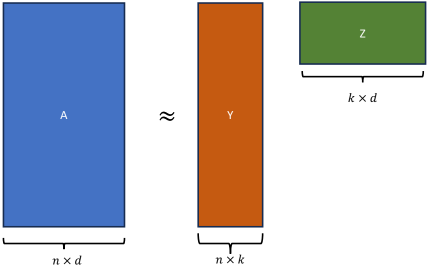
Outline

1. Low rank approximation
2. Randomized SVD - sampling
3. Randomized SVD - sketching

Low rank approximation

Given a large data matrix, we wish to compute its low rank approximation for:

- Compression
- De-noising
- Pattern finding - clustering
- Make hard problems tractable, e.g., matrix completion.



Types of low rank approximations

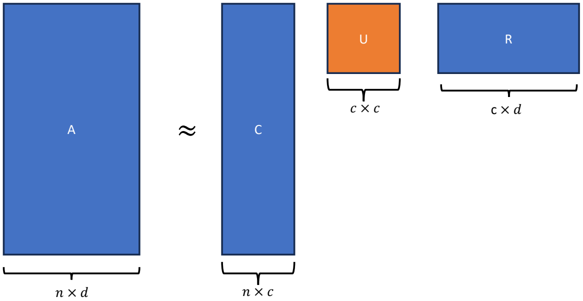
Depending on the applications, we can consider different types of low rank matrix approximations. Most common ones are:

- Truncated SVD (PCA)
- CUR decomposition
- Non-negative matrix factorization

CUR decomposition

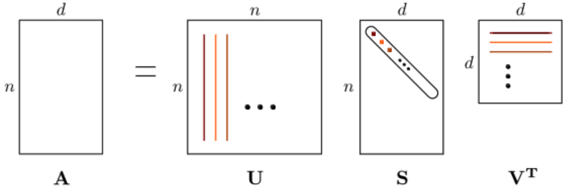
Given $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, a particular type of low rank approximation:

- A row sampling matrix $\boldsymbol{S}_1 \in \mathbb{R}^{c \times n}$, and $\boldsymbol{R} = \boldsymbol{S}_1 \boldsymbol{A} \in \mathbb{R}^{c \times d}$
- A column sampling matrix $\boldsymbol{S}_2 \in \mathbb{R}^{d \times c}$, and $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{S}_2 \in \mathbb{R}^{n \times c}$
- A matrix $\boldsymbol{U} \in \mathbb{R}^{c \times c}$, such that $\boldsymbol{A} \approx \boldsymbol{C}\boldsymbol{U}\boldsymbol{R}$ and $c \ll \{n, d\}$.



Low rank approximation

- Given a data matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ and integer k, find a rank-k approximation of $\boldsymbol{A}$.
- $\boldsymbol{A}_k = \boldsymbol{U}_k \Sigma_k \boldsymbol{V}_k^\top = \boldsymbol{U}_k \boldsymbol{U}_k^\top \boldsymbol{A} = \boldsymbol{A} \boldsymbol{V}_k \boldsymbol{V}_k^\top$.



$$\boldsymbol{U}_k = \arg \min_{\boldsymbol{U} \in \mathbb{R}^{n \times k}} \|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{A}\|_F^2 = \arg \max_{\boldsymbol{U} \in \mathbb{R}^{n \times k}} \|\boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{A}\|_F^2.$$

$$\|\boldsymbol{A} - \boldsymbol{A}_k\|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2.$$

Randomized SVD: Proto-algorithm

Input: Data matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ and target rank k .
Output: Approximate rank-k SVD: $\boldsymbol{H}_k \hat{\Sigma}_k \boldsymbol{W}_k^\top$.

- Draw a random matrix $\boldsymbol{S} \in \mathbb{R}^{d \times m}$.
- Form the sketch $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{S} \in \mathbb{R}^{n \times m}$.
- Compute an orthonormal matrix $\boldsymbol{Q}$ such that $\boldsymbol{Y} = \boldsymbol{Q}\boldsymbol{R}$.
- Form $m \times d$ matrix $\boldsymbol{B} = \boldsymbol{Q}^\top \boldsymbol{A}$.
- Compute SVD of the small matrix $\boldsymbol{B} = \hat{\boldsymbol{H}}_k \hat{\Sigma}_k \boldsymbol{W}_k^\top$.
- Form $\boldsymbol{H}_k = \boldsymbol{Q}\hat{\boldsymbol{H}}_k$.

$\boldsymbol{S}$ is a sampling/sketching matrix.

LinearTimeSVD

- **Input:** Data matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ and integers m, k such that $1 \leq k \leq m \leq d$, and $\{p_i\}_{i=1}^{d}$ with $p_i \geq 0$ and $\sum_i p_i = 1$.
- **Output:** $\boldsymbol{H}_k$ and $\hat{\Sigma}_k$.
- For $t = 1$ to m,
  Pick $i \in [d]$, with $\Pr[i = j] = p_j$.
  Set $\boldsymbol{C}_{*t} = \boldsymbol{A}_{*i} / \sqrt{m p_i}$
- Compute $\boldsymbol{C}^\top \boldsymbol{C}$ and its SVD: $\boldsymbol{C}^\top \boldsymbol{C} = \boldsymbol{W}_k \hat{\Sigma}_k \boldsymbol{W}_k^\top$.
- Compute $\boldsymbol{H}_k = \boldsymbol{C} \boldsymbol{W}_k \hat{\Sigma}_k^{-1}$.

Single pass over $\boldsymbol{A}$.

Sampling - Analysis

Given $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{H}_k$ is computed from the LinearTimeSVD algorithm, then

$$\|\boldsymbol{A} - \boldsymbol{H}_k \boldsymbol{H}_k^\top \boldsymbol{A}\|_F^2 \leq \|\boldsymbol{A} - \boldsymbol{A}_k\|_F^2 + 2\sqrt{k}\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{C}\boldsymbol{C}^\top\|_F$$
$$\|\boldsymbol{A} - \boldsymbol{H}_k \boldsymbol{H}_k^\top \boldsymbol{A}\|_2^2 \leq \|\boldsymbol{A} - \boldsymbol{A}_k\|_2^2 + 2\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{C}\boldsymbol{C}^\top\|_2$$

These results hold for any p_i's.

Sampling - Analysis

Given $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{H}_k$ is computed from the LinearTimeSVD algorithm, then

$$\|\boldsymbol{A} - \boldsymbol{H}_k \boldsymbol{H}_k^\top \boldsymbol{A}\|_F^2 \leq \|\boldsymbol{A} - \boldsymbol{A}_k\|_F^2 + 2\sqrt{k}\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{C}\boldsymbol{C}^\top\|_F$$
$$\|\boldsymbol{A} - \boldsymbol{H}_k \boldsymbol{H}_k^\top \boldsymbol{A}\|_2^2 \leq \|\boldsymbol{A} - \boldsymbol{A}_k\|_2^2 + 2\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{C}\boldsymbol{C}^\top\|_2$$

These results hold for any p_i's.

Proof: First, we note that

$$\|\boldsymbol{A} - \boldsymbol{H}_k \boldsymbol{H}_k^\top \boldsymbol{A}\|_F^2 = \|\boldsymbol{A}\|_F^2 - \|\boldsymbol{A}^\top \boldsymbol{H}_k\|_F^2$$

Next, we relate $\|\boldsymbol{A}^\top \boldsymbol{H}_k\|_F^2$ and $\sum_{t=1}^k \sigma_t^2(\boldsymbol{C})$ as:

$$\left| \|\boldsymbol{A}^\top \boldsymbol{H}_k\|_F^2 - \sum_{t=1}^k \sigma_t^2(\boldsymbol{C}) \right| \leq \sqrt{k} \|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{C}\boldsymbol{C}^\top\|_F.$$

We also have

$$\left| \sum_{t=1}^k \sigma_t^2(\boldsymbol{C}) - \sum_{t=1}^k \sigma_t^2(\boldsymbol{A}) \right| \leq \sqrt{k} \|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{C}\boldsymbol{C}^\top\|_F.$$

Combining,

$$\left| \|\boldsymbol{A}^\top \boldsymbol{H}_k\|_F^2 - \sum_{t=1}^k \sigma_t^2(\boldsymbol{A}) \right| \leq 2\sqrt{k} \|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{C}\boldsymbol{C}^\top\|_F.$$

Length squared sampling

Given $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, let $\boldsymbol{H}_k$ is computed from the LinearTimeSVD algorithm using *length squared sampling*, i.e., $p_i = \frac{\beta\|\boldsymbol{A}_{*i}\|^2}{\|\boldsymbol{A}\|_F^2}$ for some $\beta \leq 1$. If $m \geq ck/\beta\epsilon^2$, then

$$\mathbb{E}[\|\boldsymbol{A} - \boldsymbol{H}_k\boldsymbol{H}_k^\top \boldsymbol{A}\|_F^2] \leq \|\boldsymbol{A} - \boldsymbol{A}_k\|_F^2 + \epsilon\|\boldsymbol{A}\|_F^2,$$

and if $m \geq c_1 k \log(1/\delta)/\beta\epsilon^2$ with probability $1-\delta$:

$$\|\boldsymbol{A} - \boldsymbol{H}_k\boldsymbol{H}_k^\top \boldsymbol{A}\|_F^2 \leq \|\boldsymbol{A} - \boldsymbol{A}_k\|_F^2 + \epsilon\|\boldsymbol{A}\|_F^2$$

In addition, if $m \geq c_2 \log(1/\delta)/\beta\epsilon^2$ with probability $1-\delta$:

$$\|\boldsymbol{A} - \boldsymbol{H}_k\boldsymbol{H}_k^\top \boldsymbol{A}\|_2^2 \leq \|\boldsymbol{A} - \boldsymbol{A}_k\|_2^2 + \epsilon\|\boldsymbol{A}\|_F^2.$$

Recall AMM

For length squared sampling, i.e., compute $\boldsymbol{C} \in \mathbb{R}^{n \times m}$ with $p_i = \frac{\beta \|\boldsymbol{A}_{*i}\|^2}{\|\boldsymbol{A}\|_F^2}$, then we have

$$\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{C}\boldsymbol{C}^\top\|_F \leq \frac{1}{\sqrt{\beta m}} \|\boldsymbol{A}\|_F^2.$$

Combining with the previous results, we get the expectation bounds.
We then use the Markov's inequality to get the probabilistic bounds.

Recall, we obtained similar results for CUR decomposition using length squared sampling.

Randomized SVD using sketching

Suppose $\hat{\boldsymbol{A}}_k = \boldsymbol{H}_k \hat{\Sigma}_k \boldsymbol{W}_k^\top$ is the rank-k approximation we obtain from randomized SVD. We wish to obtain relative error guarantees of the form:

$$\|\boldsymbol{A} - \hat{\boldsymbol{A}}_k\|_F \leq (1+\epsilon)\|\boldsymbol{A} - \boldsymbol{A}_k\|_F$$

We use sketching and the subspace embedding property.

SVD by sketching

Given $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, let $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ be a sketching matrix such that if it is a Countsketch matrix with $m = O(k^2/\epsilon)$ or SRHT with $m = O(k \log k/\epsilon)$, or Gaussian sketch with $m = O(k/\epsilon)$, then

$$\|\boldsymbol{A} - \hat{\boldsymbol{A}}_k\|_F \leq (1+\epsilon)\|\boldsymbol{A} - \boldsymbol{A}_k\|_F,$$

where $\hat{\boldsymbol{A}}_k$ is a rank-k approximation in rowspace of $\boldsymbol{S}\boldsymbol{A}$.

SVD by sketching

Given $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, let $\boldsymbol{S} \in \mathbb{R}^{m\times n}$ be a sketching matrix such that if it is a Countsketch matrix with $m = O(k^2/\epsilon)$ or SRHT with $m = O(k\log k/\epsilon)$, or Gaussian sketch with $m = O(k/\epsilon)$, then

$$\|\boldsymbol{A} - \hat{\boldsymbol{A}}_k\|_F \leq (1+\epsilon)\|\boldsymbol{A} - \boldsymbol{A}_k\|_F,$$

where $\hat{\boldsymbol{A}}_k$ is a rank-k approximation in rowspace of $\boldsymbol{S}\boldsymbol{A}$.

Proof: Let $\boldsymbol{U}_k$ be the top k left singular vectors of $\boldsymbol{A}$. Consider:

$$\|\boldsymbol{U}_k(\boldsymbol{S}\boldsymbol{U}_k)^\dagger \boldsymbol{S}\boldsymbol{A} - \boldsymbol{A}\|_F^2.$$

We wish to show this is $(1+\epsilon)\|\boldsymbol{A} - \boldsymbol{A}_k\|_F^2$.

Since the columns of $\boldsymbol{A} - \boldsymbol{A}_k$ are orthogonal to the columns of $\boldsymbol{U}_k$, by the matrix Pythagorean theorem :

$$
\begin{aligned}
\|\boldsymbol{U}_k(\boldsymbol{S}\boldsymbol{U}_k)^\dagger \boldsymbol{S}\boldsymbol{A} - \boldsymbol{A}\|_F^2 &= \\
&=
\end{aligned}
$$

We have to show $\|(\boldsymbol{S}\boldsymbol{U}_k)^\dagger \boldsymbol{S}\boldsymbol{A} - \Sigma_k \boldsymbol{V}_k^\top\|_F^2 = O(\epsilon)\|\boldsymbol{A} - \boldsymbol{A}_k\|_F^2$.

Since the columns of $\boldsymbol{A}-\boldsymbol{A}_k$ are orthogonal to the columns of $\boldsymbol{U}_k$, by the matrix Pythagorean theorem :

$$\begin{aligned}\|\boldsymbol{U}_k(\boldsymbol{S}\boldsymbol{U}_k)^{\dagger}\boldsymbol{S}\boldsymbol{A}-\boldsymbol{A}\|_F^2 &= \\ &= \end{aligned}$$

We have to show $\|(\boldsymbol{S}\boldsymbol{U}_k)^{\dagger}\boldsymbol{S}\boldsymbol{A}-\Sigma_k\boldsymbol{V}_k^{\top}\|_F^2 = O(\epsilon)\|\boldsymbol{A}-\boldsymbol{A}_k\|_F^2$.

We have $\boldsymbol{A}=\boldsymbol{U}\Sigma\boldsymbol{V}^{\top}=\boldsymbol{U}_k\Sigma_k\boldsymbol{V}_k^{\top}+\boldsymbol{U}_{n-k}\Sigma_{r-k}\boldsymbol{V}_{d-k}^{\top}$.

Since the columns of $\boldsymbol{A} - \boldsymbol{A}_k$ are orthogonal to the columns of $\boldsymbol{U}_k$, by the matrix Pythagorean theorem :

$$\begin{aligned} \|\boldsymbol{U}_k(\boldsymbol{S}\boldsymbol{U}_k)^\dagger \boldsymbol{S}\boldsymbol{A} - \boldsymbol{A}\|_F^2 &= \\ &= \end{aligned}$$

We have to show $\|(\boldsymbol{S}\boldsymbol{U}_k)^\dagger \boldsymbol{S}\boldsymbol{A} - \Sigma_k \boldsymbol{V}_k^\top\|_F^2 = O(\epsilon)\|\boldsymbol{A} - \boldsymbol{A}_k\|_F^2$.

We have $\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^\top = \boldsymbol{U}_k\Sigma_k\boldsymbol{V}_k^\top + \boldsymbol{U}_{n-k}\Sigma_{r-k}\boldsymbol{V}_{d-k}^\top$.

...

We need $\|(\boldsymbol{S}\boldsymbol{U}_k)^\dagger \boldsymbol{S}\boldsymbol{U}_{n-k}\Sigma_{r-k}\boldsymbol{V}_{d-k}^\top\|_F^2 = O(\epsilon)\|\boldsymbol{A} - \boldsymbol{A}_k\|_F^2$.

Note that, $(\boldsymbol{S}\boldsymbol{U}_k)^\dagger$ and $(\boldsymbol{S}\boldsymbol{U}_k)^\top$ have the same row space. We can write $(\boldsymbol{S}\boldsymbol{U}_k)^\top = \boldsymbol{G}(\boldsymbol{S}\boldsymbol{U}_k)^\dagger$.

For a subspace embedding $\boldsymbol{S}$, we have

$$\|\boldsymbol{U}_k^\top \boldsymbol{S}^\top \boldsymbol{S}\boldsymbol{U}_k - \boldsymbol{I}\|_2 \leq \frac{1}{2}.$$

We can show $\|\boldsymbol{G}^{-1}\| \leq 4$.
Hence, we need to show $\|(\boldsymbol{S}\boldsymbol{U}_k)^\top \boldsymbol{S}\boldsymbol{U}_{n-k}\Sigma_{r-k}\boldsymbol{V}_{d-k}^\top\|_F^2 = O(\epsilon)\|\boldsymbol{A} - \boldsymbol{A}_k\|_F^2$.

For a subspace embedding $\boldsymbol{S}$, we have

$$\|\boldsymbol{U}_k^\top \boldsymbol{S}^\top \boldsymbol{S} \boldsymbol{U}_k - \boldsymbol{I}\|_2 \leq \frac{1}{2}.$$

We can show $\|\boldsymbol{G}^{-1}\| \leq 4$.
Hence, we need to show $\|(\boldsymbol{S}\boldsymbol{U}_k)^\top \boldsymbol{S}\boldsymbol{U}_{n-k} \Sigma_{r-k} \boldsymbol{V}_{d-k}^\top\|_F^2 = O(\epsilon)\|\boldsymbol{A} - \boldsymbol{A}_k\|_F^2$.

Using the AMM property, we have with high probability

$$\|(\boldsymbol{S}\boldsymbol{U}_k)^\top \boldsymbol{S}\boldsymbol{U}_{n-k} \Sigma_{r-k} \boldsymbol{V}_{d-k}^\top\|_F^2 \leq 9\frac{\epsilon}{k}\|\boldsymbol{U}_k\|_F^2\|\boldsymbol{A} - \boldsymbol{A}_k\|_F^2 \leq 9\epsilon\|\boldsymbol{A} - \boldsymbol{A}_k\|_F^2.$$

Two sided sketching

Let $\boldsymbol{S}_2$ be a $O(\epsilon)$- subspace embedding for the row space of $\boldsymbol{S}_1\boldsymbol{A}$, where $\boldsymbol{S}_1$ is as in the above result. Then,

$$\|\boldsymbol{A}\boldsymbol{S}_2(\boldsymbol{S}_1\boldsymbol{A}\boldsymbol{S}_2)^{\dagger}\boldsymbol{S}_1\boldsymbol{A} - \boldsymbol{A}\|_F^2 \leq (1+\epsilon)\|\boldsymbol{A} - \boldsymbol{A}_k\|_F^2.$$

We can compute $\boldsymbol{A}\boldsymbol{S}_2, (\boldsymbol{S}_1\boldsymbol{A}\boldsymbol{S}_2)^{\dagger}, \boldsymbol{S}_1\boldsymbol{A}$ in $O(nnz(\boldsymbol{A}) + (n+d)poly(k/\epsilon))$ time.

Further Reading

- Halko, Nathan, Per-Gunnar Martinsson, and Joel A. Tropp. “Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions.” SIAM review 53.2 (2011): 217-288.
- Clarkson, Kenneth L., and David P. Woodruff. “Low-rank approximation and regression in input sparsity time.” Journal of the ACM (JACM) 63.6 (2017): 1-45.

Matlab Demo