Singular Value Decomposition. (SVD)

- SVD is a matrix factorization method that is useful for many applications, e.g. recommendation systems, least squares problems, etc.
- Using SVD, we can check:
  - rank of a matrix
  - a given matrix near a simpler one (e.g. a matrix of smaller rank).
  -

- There are many algorithms to compute SVD but most of them are expensive (slow to compute).

$\Rightarrow$ How to compute a good approximation to the SVD of a <u>big matrix fast</u> is an active research topic in numerical linear algebra!

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

" The image of the unit sphere under any $m \times n$ matrix is a hyperellipsoid".

For simplicity, assume $m \geq n$ and $\text{rank}(A) = n$. (Full rank, tall matrix)

- <u>Def:</u> The <u>singular values</u> of A are the lengths of the n principal semiaxes of the hyperellipsoid (A(sphere)). σ_1

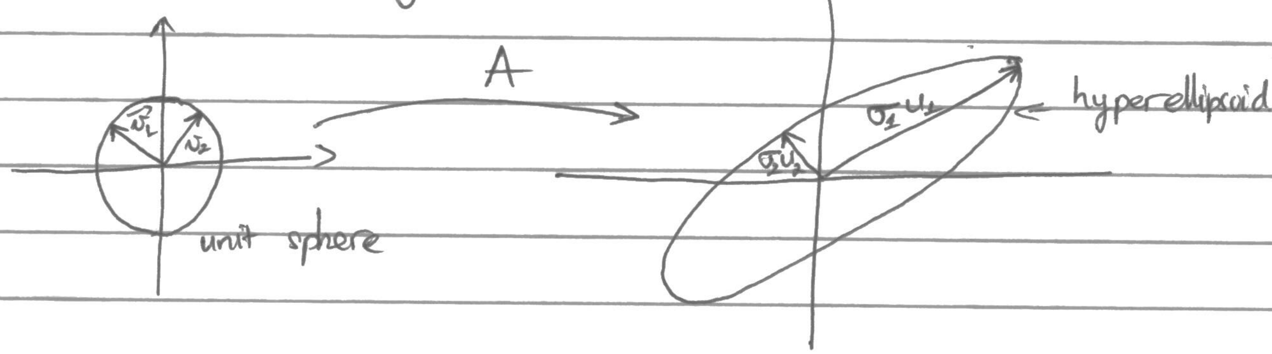
$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n > 0.$$

Def: the n left singular vectors of A are the (orthonormal) unit vectors in $\mathbb{R}^n$ along the principal semiaxes of A(sphere). We denote them by $\{u_1, \dots, u_n\}$.
$\Rightarrow$ $\sigma_i u_i$ is the ith largest pricipal semiaxes of A(sphere).

Def: The n right singular vector of A are the (orthonormal) unit vectors $\{v_1, \dots, v_n\}$ such that

$$A v_i = \sigma_i u_i, \qquad i = 1, \dots, n.$$

Geometric meaning:



· Reduced SVD:

$$A v_i = \sigma_i u_i \qquad i = 1, \dots, n.$$

$$[A v_1 \ A v_2 \ \dots \ A v_n] = [\sigma_1 u_1 \ \sigma_2 u_2 \ \dots \ \sigma_n u_n].$$

$$\underbrace{A [v_1 \ v_2 \ \dots \ v_n]}_{} = [u_1 \ u_2 \ \dots \ u_n] \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{bmatrix}$$

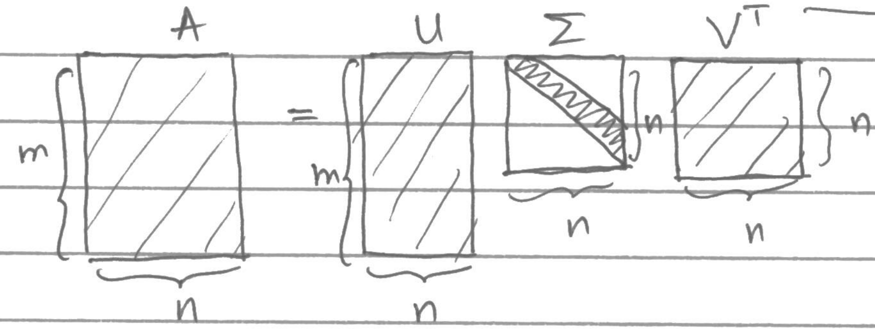
$$\underset{m \times n}{A} \ \underset{n \times n}{V} = \underset{m \times n}{U} \ \underset{n \times n}{\Sigma}.$$

columns of U are orthonormal.

$$\Rightarrow \quad AV = U\Sigma.$$

Since V is an orthogonal matrix,

$$A = U\Sigma V^T.$$ The <u>reduced SVD</u> of A.



* <u>Pseudoinverse via SVD</u>:

$$\boxed{A^+ = V\Sigma^+ U^T}$$

where $\Sigma^+ = \begin{bmatrix} 1/\sigma_1 & & \\ & \ddots & \\ & & 1/\sigma_n \end{bmatrix}$

[<u>Exer</u>: Show that $AA^+ = UU^T$
and $A^+A = VV^T$.]

<u>The Moore-Penrose conditions</u>: For a given matrix $A \in \mathbb{R}^{m\times n}$, if $B \in \mathbb{R}^{n\times m}$ satisfies the following:

i) $ABA = A$
ii) $BAB = B$
iii) $(AB)^T = AB$
iv) $(BA)^T = BA$.

Then B is called the pseudoinverse (or the Moore-Penrose inverse) of A and written as A^+.

Thm: (SVD) Any $m \geq n$ matrix A, with $m \geq n$, can be factorized

(Full version) $$A = U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^T.$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal

$\Sigma \in \mathbb{R}^{n \times n}$ is diagonal.

$$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$$

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0.$$

Remark: There are accurate algorithms for computing the SVD.

Proposition: The 2-norm of a matrix is given by

$$\|A\|_2 = \sigma_1 \leftarrow \text{the largest singular value of } A.$$

Pf: (Sketch).

1) The norm is invariant under orthogonal transformatio

$$\Rightarrow \quad \|A\|_2 = \|\Sigma\|_2.$$

2) 2-norm of a diagonal matrix is equal to the absolute value of the largest diagonal element.

* Matrix properties via SVD.

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

$p = \min\{m, n\}$.

$r = \#$ of nonzero singular values

$(\Rightarrow r \leq p)$.

. Thm: $\operatorname{rank}(A) = r$

$\operatorname{range}(A) = \operatorname{span}\{u_1, \dots, u_r\}$.

$\operatorname{Null}(A) = \operatorname{span}\{v_{r+1}, \dots, v_n\}$.

$\|A\|_2 = \sigma_1$ and $\|A\|_F = \sqrt{\sigma_1^2 + \dots + \sigma_r^2}$

. Thm: If $A^T = A$, $\sigma_i(A) = |\lambda_i(A)|$.

$\uparrow$ ith eigenvalue of A.

. Thm: For $A \in \mathbb{R}^{m \times m}$, (square matrix)

$$|\det(A)| = \prod_{i=1}^{m} \sigma_i$$

Pf: Recall.

1) $\det(AB) = \det(A)\det(B)$.
2) $\det(A^T) = \det(A)$.
3) $\det(\operatorname{diag}(a_1, \dots, a_n)) = \prod_{i=1}^{n} a_i$.
4) For any orthogonal Q, $|\det(Q)| = 1$.

$$\Rightarrow \det(A) = \det(U \Sigma V^T) = \det(U)\det(\Sigma)\det(V^T)$$
$$= \det(\Sigma) = \prod_{i=1}^{m} \sigma_i.$$

* Outer product:

Let $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$.

Then the outer product between u and v is:

$$uv^T = \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} \begin{bmatrix} v_1 & \dots & v_n \end{bmatrix} = \begin{bmatrix} u_1v_1 & u_1v_2 & \dots & u_1v_n \\ u_2v_1 & u_2v_2 & \dots & u_2v_n \\ \vdots & & & \vdots \\ u_mv_1 & u_mv_2 & \dots & u_mv_n \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

This matrix has rank 1 because

$$\vec{u}\vec{v}^T = \begin{bmatrix} v_1\vec{u} & v_2\vec{u} & \dots & v_n\vec{u} \end{bmatrix}.$$

each column is just a scalar multiple of the same vector $\vec{u}$.

⇒ we can view a matrix A as:

$$A = U\Sigma V^T = \underbrace{\sum_{i=1}^{r} \sigma_i u_i v_i^T}_{\text{sum of rank-1 matrices.}}, \quad r = \operatorname{rank}(A).$$

(Exer: show this).

* <u>Best Rank-k Approximations:</u>

Let $A \in \mathbb{R}^{n \times d}$. (That is, the rows of A are n points in d dimensional space).
Suppose that $\text{rank}(A) = r$. Then the SVD of A is

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T.$$

For $k \in \{1, 2, \ldots, r\}$, let

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T.$$

A_k ~~is called~~ the sum truncated after k terms.
<u>Fact:</u> A_k has rank k (why?).

<u>Lemma:</u> The rows of A_k are the projections of the rows of A onto the subspace V_k spanned by the first k singular vectors of A, (i.e., $V_k = \text{span}\{v_1, \ldots, v_k\}$)

<u>pf:</u> Let $\vec{x}$ be an arbitrary vector in $\mathbb{R}^d$.
Since $\vec{v}_i$ are orthonormal, the orthogonal projection of $\vec{x}$ onto V_k is given by

$$\sum_{i=1}^{k} \langle \vec{v}_i, \vec{x} \rangle \vec{v}_i$$

$$= \left(\sum_{i=1}^{k} v_i v_i^T \right) \vec{x}$$

$$\Rightarrow P = \sum_{i=1}^{k} v_i v_i^T \quad \text{is the projection matrix onto } V_k.$$

Let $\vec{a}_1, \dots, \vec{a}_n$ be the columns of A^T.

$\Rightarrow$ their transposes are the rows of A.

Then

$$PA^T = P[\vec{a}_1 \ \vec{a}_2 \ \dots \ \vec{a}_n]$$
$$= [P\vec{a}_1 \ P\vec{a}_2 \ \dots \ P\vec{a}_n].$$

Transpose both sides:

$$AP = \begin{bmatrix} (Pa_1)^T \\ \vdots \\ (Pa_n)^T \end{bmatrix}$$

$P^T = P$

$\rightarrow$ AP is the matrix whose rows are the projections of the rows of A onto V_k.

Since $P = \sum_{i=1}^{k} v_i v_i^T$

$$AP = \sum_{i=1}^{k} A v_i v_i^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T = A_k.$$

since $A v_i = \sigma_i u_i$.

Thm: For any matrix B of rank at most k.

$$\|A - A_k\|_F \le \|A - B\|_F.$$

That is, A_k is the best rank k approximation to A, where error is measured in the Frobenius norm.

(See textbook: Foundation of Data science Theorem 3.6, p. 48).

Pf. Let B minimize $\|A-B\|_F^2$ among all rank k or less matrices.

Let V be the space spanned by the rows of B.

$\Rightarrow \dim(V) \leq k$.

Let a_i^T be the ith row of A.

b_i^T be the ith row of B.

$$\Rightarrow \|A-B\|_F^2 = \sum_{i=1}^{n} \|a_i^T - b_i^T\|_2^2.$$

Since $b_i \in V$, $\|a_i^T - b_i^T\|_2^2$ is minimized if b_i is the projection of a_i onto V.

$\Rightarrow$ each row of B is the projection of the corresponding row of A onto V.

Then $\|A-B\|_F^2$ is the sum of squared distance of rows of A to V. Since A_k minimizes the sum of squared distance of rows of A to any k-dim subspace, by the previous lemma, $B = A_k$.

* Applications:

- Image Compression:

As each image is represented by a matrix, it is possible to compress an image by approximating it by a lower rank matrix.

* Power Method for SVD.

Computing the SVD is an important research topic in ~~numb~~ numerical linear algebra.

power method is a basic method to ~~establish~~ find the approximate SVD of a matrix A in polynomial time.

Let A be matrix whose SVD is $\sum_{i=1}^{r} \sigma_i u_i v_i^T$.

Let $B = A^T A$. Then

$$B = A^T A$$

$$= \left(\sum_i \sigma_i v_i u_i^T \right)\left(\sum_j \sigma_j u_j v_j^T \right).$$

$$= \sum_{i,j} \sigma_i \sigma_j v_i u_i^T u_j v_j^T.$$

$$\Rightarrow \quad B = \sum_i \sigma_i^2 v_i v_i^T. \quad \text{since } u_i \text{ are orthonormal.}$$

Facts: 1) The matrix B is square and symmetric, and has the same left and right-singular vectors.

2) v_j is an eigenvector of B with eigenvalue σ_j^2.

$$B v_j = \left(\sum_i \sigma_i^2 v_i v_i^T \right) v_j.$$

$$v_i^T v_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \qquad = \sum_i \sigma_i^2 v_i v_i^T v_j.$$

$$= \sigma_j^2 v_j v_j^T v_j.$$

$$\Rightarrow \quad B v_j = \sigma_j^2 v_j.$$

Now compute B^2,

$$B^2 = \left(\sum_i \sigma_i^2 v_i v_i^T\right)\left(\sum_j \sigma_j^2 v_j v_j^T\right)$$

$$= \sum_{i,j} \sigma_i^2 \sigma_j^2 v_i v_i^T v_j v_j^T.$$

$$= \sum_i \sigma_i^4 v_i v_i^T.$$

<u>Fact:</u> $\quad B^k = \sum_i \sigma_i^{2k} v_i v_i^T.$

If $\sigma_1 > \sigma_2$, when k is large ($k \to \infty$), the first term $\sigma_1^{2k} v_1 v_1^T$ in the summation dominates.

$$\Rightarrow \quad B^k \longrightarrow \sigma_1^{2k} v_1 v_1^T \quad \text{as} \quad k \to \infty.$$

→ ~~To find~~

⇒ To estimate v_1: one can compute B^k, then normalize the first column of B^k.

But in practice, A may be a very large, sparse matrix, E.g. Netflix Rating matrix

$$\begin{array}{r c c c c c}
 & \text{Matrix} & \text{Alien} & \text{Star Wars} & \multicolumn{2}{c}{\overbrace{\text{Casablanca} \quad \text{Titanic}}^{\text{drama}}} \\
\text{Joe} & 1 & 1 & 1 & 0 & 0 \\
\text{Jim} & 3 & 3 & 3 & 0 & 0 \\
\text{John} & 4 & 4 & 4 & 0 & 0 \\
\text{Jack} & 5 & 5 & 5 & 0 & 0 \\
\text{Jill} & 0 & 0 & 0 & 4 & 4 \\
\text{Jenny} & 0 & 0 & 0 & 5 & 5 \\
\text{Jane} & 0 & 0 & 0 & 2 & 2
\end{array}$$

in practice, A may be a $10^8 \times 10^8$ matrix with 10^9 nonzero entries.

Though A is sparse, B need not be and in the worse case may have all 10^{16} entries nonzero.

$\Rightarrow$ computing B^k is very costly.

Faster way:

(randomly) choose a vector x and compute $B^k x$

This is a little bit faster than computing B^k because you do matrix-vector multiplication:

compute Bx = this is a ~~vector~~ matrix-vector.

Then compute $B(Bx)$

and so on.

How to get v_1?

Note that $x = \sum_{i=1}^{d} c_i v_i$ (suppose that B is of full rank)

since $\{v_1, \dots, v_d\}$ forms a basis for $\mathbb{R}^d$.

Then for k large

$$B^k x \approx (\sigma_1^{2k} v_1 v_1^T)\left(\sum_{i=1}^{d} c_i v_i\right) = \sigma_1^{2k} c_1 v_1$$

$$\Rightarrow B^k x \approx \sigma_1^{2k} c_1 v_1$$

$\Rightarrow$ normalizing $B^k x$, we can obtain an approximate vector of v_1.

* SVD and condition number:

Recall the condition number for a square nonsingular matrix A is defined by

$$\kappa(A) := \|A\|_2 \, \|A^{-1}\|_2.$$

- $\kappa(A)$ small, then A is said well-conditioned.
- $\kappa(A)$ large, then A is said ill-conditioned

Finding The solution of $Ax = b$ by using computers may ~~be~~ give us a bad approximation

How is the condition number related to SVD?

Recall that

$$\|A\|_2 = \sigma_1.$$

and

$$\|A^{-1}\|_2 = \frac{1}{\sigma_m} \quad \text{~~where~~}$$

why? $A = U \Sigma V^T$

and $A^{-1} = V \Sigma^{-1} U^T$

$$= V \, \mathrm{diag}\left(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_m}\right) U^T.$$

since $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m > 0$,

$$\frac{1}{\sigma_m} \geq \frac{1}{\sigma_{m-1}} \geq \dots \geq \frac{1}{\sigma_1}.$$

largest singular value of A^{-1}.

$$\Rightarrow \|A^{-1}\|_2 = \frac{1}{\sigma_m}$$

We can generalize the definition of the condition ~~nu~~ number for a rectangular matrix $A \in \mathbb{R}^{m \times n}$ using the pseudo-inverse A^+ and SVDs.

$$\kappa(A) = \|A\|_2 \, \|A^+\|_2$$
$$= \frac{\sigma_1}{\sigma_r} \qquad \text{where } r = \operatorname{rank}(A).$$