<u>Recall</u> The SVD of an $n \times d$ matrix, A, $(n \geq d)$
is $A = USV^T$, where $U \in \mathbb{R}^{n \times d}$ has orthonormal columns*,
S is $d \times d$ nonnegative diagonal with entries $\sigma_1 \geq \cdots \geq \sigma_d$
V is a $d \times d$ orthogonal matrix.

*: they are pairwise orthogonal unit vectors,
and look like the first d columns of an orthogonal matrix.

<u>Claim</u> Let $u_1, \ldots, u_d$ and $v_1, \ldots v_d$ be the columns of U and V. Then $A = \sum_i \sigma_i u_i v_i^T$

<u>proof</u>

Let e_i be elementary unit vector in dimension i.
So, $e_i e_i^T$ = all zeros, but with a 1 in the (i,i) entry.

$$S = \sum_i \sigma_i e_i e_i^T \Rightarrow USV^T = U\left(\sum_i \sigma_i e_i e_i^T\right)V^T$$

$$= \sum_i \sigma_i \left(U e_i e_i^T V^T\right)$$

$$= \sum_i \sigma_i (Ue_i)(Ve_i)^T$$

$$= \sum_i \sigma_i u_i v_i^T$$

Note: not unique: can replace u_i by $-u_i$, and might have $\sigma_i = \sigma_{i+1}$.

Recall $Av_i = \sigma_i u_i$,
because $\left(\sum_j \sigma_j u_j v_j^T\right) v_i = \sum_j \sigma_j u_j (v_j^T v_i) = \sigma_i u_i$

and similarly $A^T u_i = \sigma v_i$

Existence of SVD follows from

<u>The Spectral Theorem</u>

Every square symmetric matrix A can be written

$A = V\Lambda V^T$ where V is an orthogonal matrix of eigenvectors and Λ is a diagonal matrix of eigenvalues, $\lambda_1, \dots, \lambda_n$.

The ith column of V, v_i, satisfies $Av_i = \lambda_i v_i$.

We won't prove this, but we review part:

If A is symmetric, $Av = \lambda v$, $Au = \mu u$, and $\lambda \neq \mu$ then $v^T u = 0$.

<u>proof</u>: Consider $u^T A v = u^T \lambda v = \lambda u^T v$, and, as symmetric

$= u^T A^T v = \mu u^T v$. So either $\mu = \lambda$ or $u^T v = 0$.

proof SVD exists: (for simplicity when A is non-singular)

let $\lambda_1, \dots, \lambda_d$ and $v_1, \dots, v_d$ be eigvals / vecs of $A^T A$

$\lambda_i \geq 0$ because $v_i^T A^T A v_i = v_i^T \lambda_i v_i = \lambda_i v_i^T v_i = \lambda_i$,

and $= (Av_i)^T (Av_i) = \|Av_i\|_2^2$

And, $v_1 \dots v_d$ are orthonormal.

For i s.t. $\lambda_i > 0$, let $u_i = \frac{1}{\sqrt{\lambda_i}} A v_i$, and $\sigma_i = \sqrt{\lambda_i}$

We then have $A^T u_i = \frac{1}{\sqrt{\lambda_i}} A^T A v_i = \sqrt{\lambda_i}\, v_i$

We will show that these u_i are eigenvectors of AA^T:

$$AA^T u_i = \frac{1}{\sqrt{\lambda_i}} A(A^T A) v_i = \frac{1}{\sqrt{\lambda_i}} \lambda_i A v_i = \lambda_i u_i$$

Thus, the u_i are eigenvectors of AA^T of eigenvalue λ_i, and are thus orthogonal.

Set $S = \begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n} \end{pmatrix}$, so USV^T has the right form.

To finish, let $\tilde{A} = \sum_i \sqrt{\lambda_i} u_i v_i^T$

To see that $A = \tilde{A}$, recall that a matrix is determined by its action on a basis, and we have

$\tilde{A} v_i = A v_i = \sqrt{\lambda_i} u_i$ for all i

Geometric view of USV^T as an operator:

$V^T x$ rotates x.

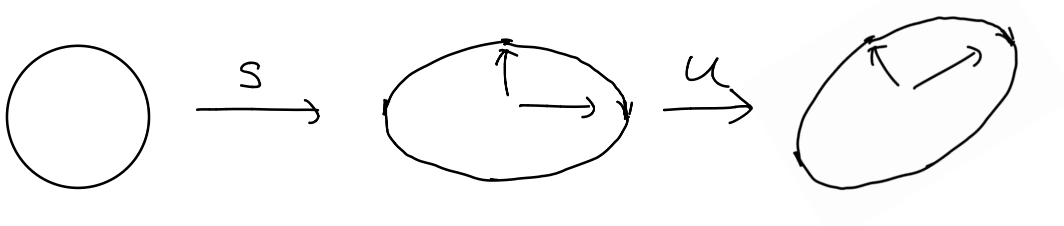
$SV^T x$ then scales its coordinates

$USV^T x$ rotates it again.

If apply to unit sphere, V^T does not change

S scales to ellipse

U rotates it.



Recall $v_1 = \arg\max_{\|x\|=1} \|Ax\|$

Generalization: $v_i = \underset{\substack{\|x\|=1 \\ x \perp v_1, \dots, v_{i-1}}}{\arg\max} \|Ax\|$

Proof. v_i is orth to $v_1 \dots v_{i-1}$, and $\|Av_i\| = \sigma_i$

and every x orth to $v_1 \dots v_{i-1}$ maps to a $U^T x$

whose coordinates are scaled by $\sigma_i \geq \dots \geq \sigma_n$

Two (related) meanings of the SVD:

1. A partial sum, $A_r \stackrel{\text{def}}{=} \sum_{i \leq r} \sigma_i u_i v_i^T$, is the rank-$r$ matrix that is closest to A.

   We care because a lot of data is low-rank + noise.

   Think of movie preferences, with n people and k movies.

   $A(p,m)$ = how much person p likes movie m.

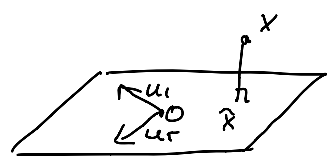
   Might measure r properties of a movie, $v_m(1), \dots, v_m(r)$ and person p's weighting of those properties $u_p(1), \dots, u_p(r)$, and hope $A(p,m) \approx \sum_{i \leq r} u_p(i) v_m(i)$

2. If view rows of A as vectors, $a_1^T, \dots, a_n^T$

   $u_1, \dots, u_r$ span the rank-r subspace that comes closest to these vectors.

Projection: let $S \subseteq \mathbb{R}^d$ be a subspace of dimension r and let $u_1, \ldots, u_r$ be an orthonormal basis of S. For any $x \in \mathbb{R}^d$, we want $\hat{x} = \arg\min_{y \in S} \operatorname{dist}(x, y)$.

By the Pythagorean theorem, we know that $\hat{x} - x$ is orthogonal to the vectors in S



This implies $\hat{x} = \sum_{i=1}^{r} u_i (u_i^T x)$, because is in S and

for all j, $u_j^T \hat{x} = \sum_{i=1}^{r} u_j^T u_i (u_i^T x) = u_j^T x$ as $u_j^T u_i = \begin{cases} 1 & i = j \\ 0 & \text{ow.} \end{cases}$

So, $u_j^T(\hat{x} - x) = 0 \Rightarrow u_j$ is $\perp$ to $\hat{x} - x$

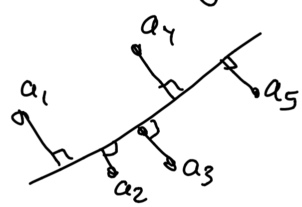
The approach we took last lecture amounts to computing an orthonormal basis $u_{r+1}, \ldots, u_d$ of the space orthogonal to S, and setting $\hat{x} = x - \sum_{i=r+1}^{n} u_i (u_i^T x)$

This is the same because

$$\sum_{i=1}^{n} u_i (u_i^T x) = \left(\sum_i u_i u_i^T \right) X = UU^T x = I,$$

where $U = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_d \\ | & & | \end{pmatrix}$

For $r=1$, want line in $\mathbb{R}^d$ through O minimizing sum of squares of distances to it.



Specify a line by span(u) where u is a unit vector.
The point on the line closest to a_i is $(a_i^T u) \cdot u$

Let $\delta_i = a_i - (a_i^T u) u$, so distance from a_i to line is $\|\delta_i\|$
We want to find the u that minimizes $\sum_i \|\delta_i\|^2$

As $\|\delta_i\|^2 = \|a_i\|^2 - (a_i^T u)^2 \|u\|^2 = \|a_i\|^2 - (a_i^T u)^2$,
this is equivalent to maximizing $\sum_i (a_i^T u)^2 = \|A^T u\|_2^2$.

We proved in Lect 4 that $\arg\max_u \|A^T u\|_2^2 = u_1$

Theorem For any r, a r-dimensional subspace S that minimizes $\sum_i \text{dist}(a_i, S)^2$ is $\text{span}(u_1, \dots, u_r)$

proof Induction on r. Did $r=1$.

Let S minimize this, and let w_r be a unit vec in S orthogonal to $u_1, \dots, u_{r-1}$. (exists by $\dim(S) = r$)
Let $w_1, \dots, w_r$ be an orthonormal basis of S.
The projection of a_i onto S is $\sum_{j=1}^{r} (a_i^T w_j) w_j$

By assumption, (and Pythagoras) S maximizes

$$\sum_i \|\text{Proj}_S(a_i)\|^2 = \sum_i \sum_j (a_i^T w_j)^2 = \sum_{j=1}^{r} \|A w_j\|_2^2$$

By induction, $\sum_{j=1}^{r-1} \|A u_j\|_2^2 \geq \sum_{j=1}^{r-1} \|A w_j\|_2^2$

So, we may assume $w_1, \dots, w_{r-1} = u_1 \dots u_{r-1}$.

As $u_r = \underset{\substack{\|x\|=1 \\ x \perp v_1 \dots v_{r-1}}}{\arg\max} \|Ax\|_2^2$, and $w_r \perp u_1, \dots, u_{r-1}$

$$\|A w_r\|_2^2 \leq \|A u_r\|_2^2 \Rightarrow \sum_{j=1}^{r} \|A w_j\|_2^2 \leq \sum_{j=1}^{r} \|A u_j\|_2^2$$

See BHK Thm 3.5 for a related proof that A_r minimizes $\|A-A_r\|_2$ and $\|A-A_r\|_F$ over rank r matrices.

Computing the first r terms of the SVD.

What does it mean to compute u_1 and v_1?
Are not unique if $\sigma_1 = \sigma_2$.
We really just want $\widetilde{A}_r = \sum_{i \leq r} \sigma_i u_i v_i^T$ s.t.
$\|A - \widetilde{A}_r\|_F \leq (1+\varepsilon)$ optimal.

Iteratively: first compute u_1 and v_1,
then recursively work on $A - \sigma_1 u_1 v_1^T$

Focus on finding $\|x\|=1$ s.t. $\|Ax\|_2 \geq (1-\varepsilon)\sigma_1$

Start with random unit u_1^0.
Compute the σ_1^0, v_1^0 that minimizes $\|A - \sigma_1^0 u_1^0 v_1^{0T}\|_F$
$= \sum_i \|a_i - \sigma_1 u_1 v_1(i)\|^2$
By previous, $\sigma_1 v_1 = A^T u_1$.
Now, iterate. Compute best u_1^1 for v_1^0, and best v_1^1 for u_1^1...

After t iterations get vector $u_1^t = \dfrac{(AA^T)^t u_1^0}{\|(AA^T)^t u_1^0\|}$

This is the Power Method.

To understand it, let x be the initial random vector, and write $x = \sum_i \alpha_i u_i$, where $\alpha_i = u_i^T x$, $\sum \alpha_i^2 = 1$

So, $(AA^T) u_1^0 = \sum \alpha_i \sigma_i^2 u_i$, and $(AA^T)^t u_1^0 = \sum \alpha_i \sigma_i^{2t} u_i$

$\Rightarrow$ big σ_i dominate.

Choose k so that $\sigma_k \geq (1-\varepsilon)\sigma_1 > \sigma_{k+1}$

Let $S = \text{span}(u_1, \ldots, u_k)$.

<u>Theorem</u> If $t \geq \frac{1}{2\varepsilon} \ln(1/\varepsilon\alpha_1)$, then

$$\text{dist}(S, u_1^t) \leq \varepsilon$$

<u>proof</u>: let initial random unit vec be x_0, and $x_t = (AA^T) x_0$, $u_1^t = x_t / \|x_t\|$.

1. $\|x_t\|_2^2 = \left\| \sum_i \alpha_i \sigma_i^{2t} u_i \right\|_2^2 \geq \alpha_1^2 \sigma_1^{4t}$

2. $\text{dist}(S, u_1^t)^2 = \left\| \sum_{i>k} u_i \cdot u_i^T x_t / \|x_t\| \right\|^2$

$$= \frac{1}{\|x_t\|^2} \sum_{i>k} \left(\alpha_i \sigma_i^{2t}\right)^2 \leq \frac{1}{\|x_t\|^2} (1-\varepsilon)^{4t} \sigma_1^{4t} \sum_{i>k} \alpha_i^2$$

$$\leq \frac{1}{\|x_t\|^2} (1-\varepsilon)^{4t} \sigma_1^{4t} \leq \frac{(1-\varepsilon)^{4t}}{\alpha_1^2}$$

So, $\text{dist}(S, u_1^t) \leq \frac{(1-\varepsilon)^{2t}}{\alpha_1} \leq \frac{\exp(-2\varepsilon t)}{\alpha_1} \leq \varepsilon$

Note: α_i is unlikely to be small:

the chance a random unit vector is close to a hyperplane is like the chance for an appropriately scaled Gaussian,

Lem (see Lem B.1 in Sankar-Spielman-Teng '06)

For $\beta \le 1$ and a random unit vector x,

$$\Pr\left[|a_i^T x| \le \beta/\sqrt{d} \right] \le 2\beta$$