Lecture 16
SVD Applications

- general pseudo-inverse
- full SVD
- image of unit ball under linear transformation
- SVD in estimation/inversion
- sensitivity of linear equations to data error
- low rank approximation via SVD

General pseudo-inverse

if $A \neq 0$ has SVD $A = U\Sigma V^T$,

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

is the *pseudo-inverse* or *Moore-Penrose inverse* of A

if A is skinny and full rank,

$$A^\dagger = (A^TA)^{-1}A^T$$

gives the least-squares approximate solution $x_{\rm ls} = A^\dagger y$

if A is fat and full rank,

$$A^\dagger = A^T(AA^T)^{-1}$$

gives the least-norm solution $x_{\rm ln} = A^\dagger y$

in general case:

$$X_{\mathrm{ls}} = \{ \ z \mid \|Az - y\| = \min_{w} \ \|Aw - y\| \ \}$$

is set of least-squares approximate solutions

$x_{\mathrm{pinv}} = A^{\dagger}y \in X_{\mathrm{ls}}$ has minimum norm on X_{ls}, $i.e.$, x_{pinv} is the minimum-norm, least-squares approximate solution

Pseudo-inverse via regularization

for $\mu > 0$, let x_μ be (unique) minimizer of

$$\|Ax - y\|^2 + \mu\|x\|^2$$

i.e.,

$$x_\mu = \left(A^T A + \mu I\right)^{-1} A^T y$$

here, $A^T A + \mu I > 0$ and so is invertible

then we have $\lim_{\mu \to 0} x_\mu = A^\dagger y$

in fact, we have $\lim_{\mu \to 0} \left(A^T A + \mu I\right)^{-1} A^T = A^\dagger$

(check this!)

Full SVD

SVD of $A \in \mathbf{R}^{m\times n}$ with $\mathbf{Rank}(A) = r$:

$$A = U_1\Sigma_1V_1^T = \left[\begin{array}{ccc} u_1 & \cdots & u_r \end{array}\right] \left[\begin{array}{ccc} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{array}\right] \left[\begin{array}{c} v_1^T \\ \vdots \\ v_r^T \end{array}\right]$$

- find $U_2 \in \mathbf{R}^{m\times(m-r)}$, $V_2 \in \mathbf{R}^{n\times(n-r)}$ s.t. $U = [U_1\ U_2] \in \mathbf{R}^{m\times m}$ and $V = [V_1\ V_2] \in \mathbf{R}^{n\times n}$ are orthogonal

- add zero rows/cols to Σ_1 to form $\Sigma \in \mathbf{R}^{m\times n}$:

$$\Sigma = \left[\begin{array}{c|c} \Sigma_1 & 0_{r\times(n-r)} \\ \hline 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{array}\right]$$

then we have

$$A = U_1 \Sigma_1 V_1^T = \left[\begin{array}{c|c} U_1 & U_2 \end{array}\right] \left[\begin{array}{c|c} \Sigma_1 & 0_{r \times (n-r)} \\ \hline 0_{(m-r) \times r} & 0_{(m-r) \times (n-r)} \end{array}\right] \left[\begin{array}{c} V_1^T \\ \hline V_2^T \end{array}\right]$$

$i.e.$:

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

called *full SVD* of A

(SVD with positive singular values only called *compact SVD*)

Image of unit ball under linear transformation

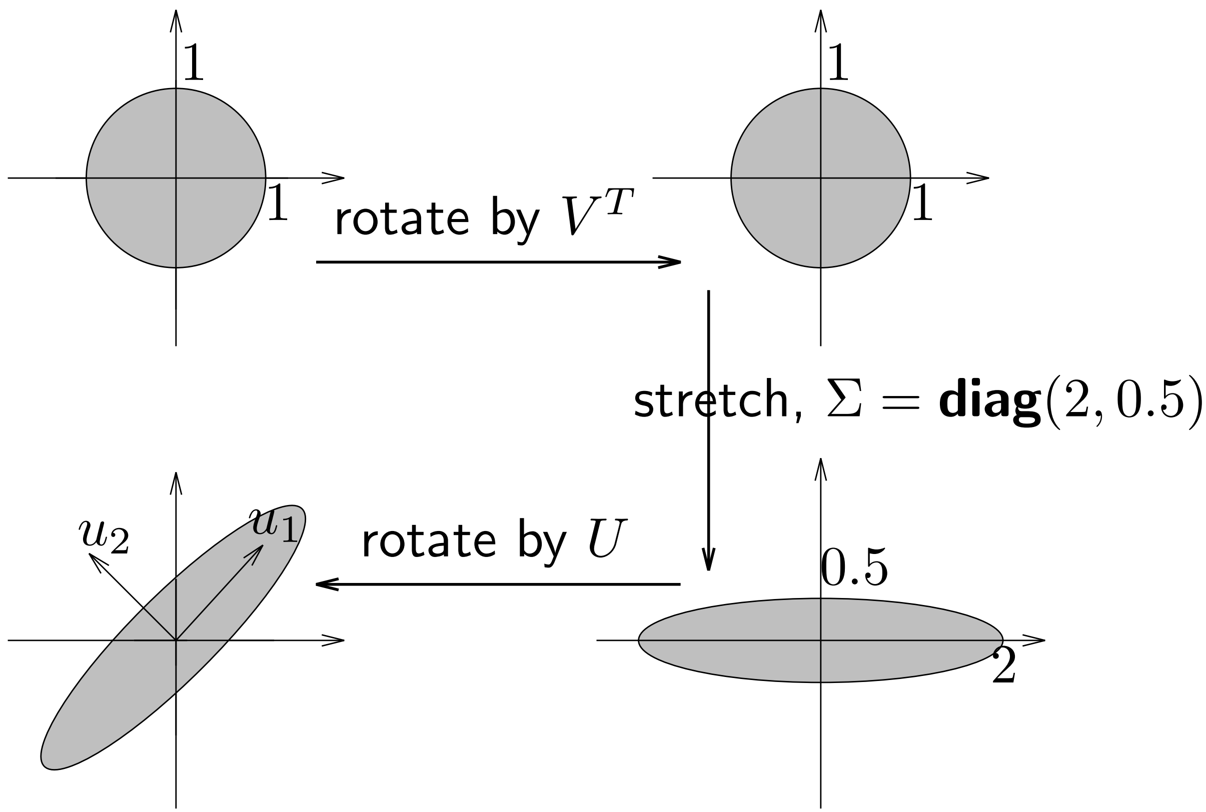
full SVD:

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

gives intepretation of $y = Ax$:

- rotate (by V^T)
- stretch along axes by σ_i ($\sigma_i = 0$ for $i > r$)
- zero-pad (if $m > n$) or truncate (if $m < n$) to get m-vector
- rotate (by U)

Image of unit ball under A



$\{Ax \mid \|x\| \leq 1\}$ is *ellipsoid* with principal axes $\sigma_i u_i$.

SVD in estimation/inversion

suppose $y = Ax + v$, where

- $y \in \mathbf{R}^m$ is measurement
- $x \in \mathbf{R}^n$ is vector to be estimated
- v is a measurement noise or error

'norm-bound' model of noise: we assume $\|v\| \leq \alpha$ but otherwise know nothing about v (α gives max norm of noise)

- consider estimator $\hat{x} = By$, with $BA = I$ (*i.e.*, unbiased)
- estimation or inversion error is $\tilde{x} = \hat{x} - x = Bv$
- set of possible estimation errors is ellipsoid

$$\tilde{x} \in \mathcal{E}_{\rm unc} = \{ \; Bv \mid \|v\| \leq \alpha \; \}$$

- $x = \hat{x} - \tilde{x} \in \hat{x} - \mathcal{E}_{\rm unc} = \hat{x} + \mathcal{E}_{\rm unc}$, *i.e.*:

  true x lies in *uncertainty ellipsoid* $\mathcal{E}_{\rm unc}$, centered at estimate $\hat{x}$
- 'good' estimator has 'small' $\mathcal{E}_{\rm unc}$ (with $BA = I$, of course)

semiaxes of $\mathcal{E}_{\mathrm{unc}}$ are $\alpha\sigma_i u_i$ (singular values & vectors of B)

e.g., maximum norm of error is $\alpha\|B\|$, *i.e.*, $\|\hat{x}-x\| \leq \alpha\|B\|$

optimality of least-squares: suppose $BA = I$ is any estimator, and $B_{\mathrm{ls}} = A^\dagger$ is the least-squares estimator

then:

- $B_{\mathrm{ls}}B_{\mathrm{ls}}^T \leq BB^T$
- $\mathcal{E}_{\mathrm{ls}} \subseteq \mathcal{E}$
- in particular $\|B_{\mathrm{ls}}\| \leq \|B\|$

i.e., the least-squares estimator gives the *smallest* uncertainty ellipsoid

Example: navigation using range measurements (lect. 4)

we have

$$y = -\begin{bmatrix} k_1^T \\ k_2^T \\ k_3^T \\ k_4^T \end{bmatrix} x + v$$

where $k_i \in \mathbf{R}^2$

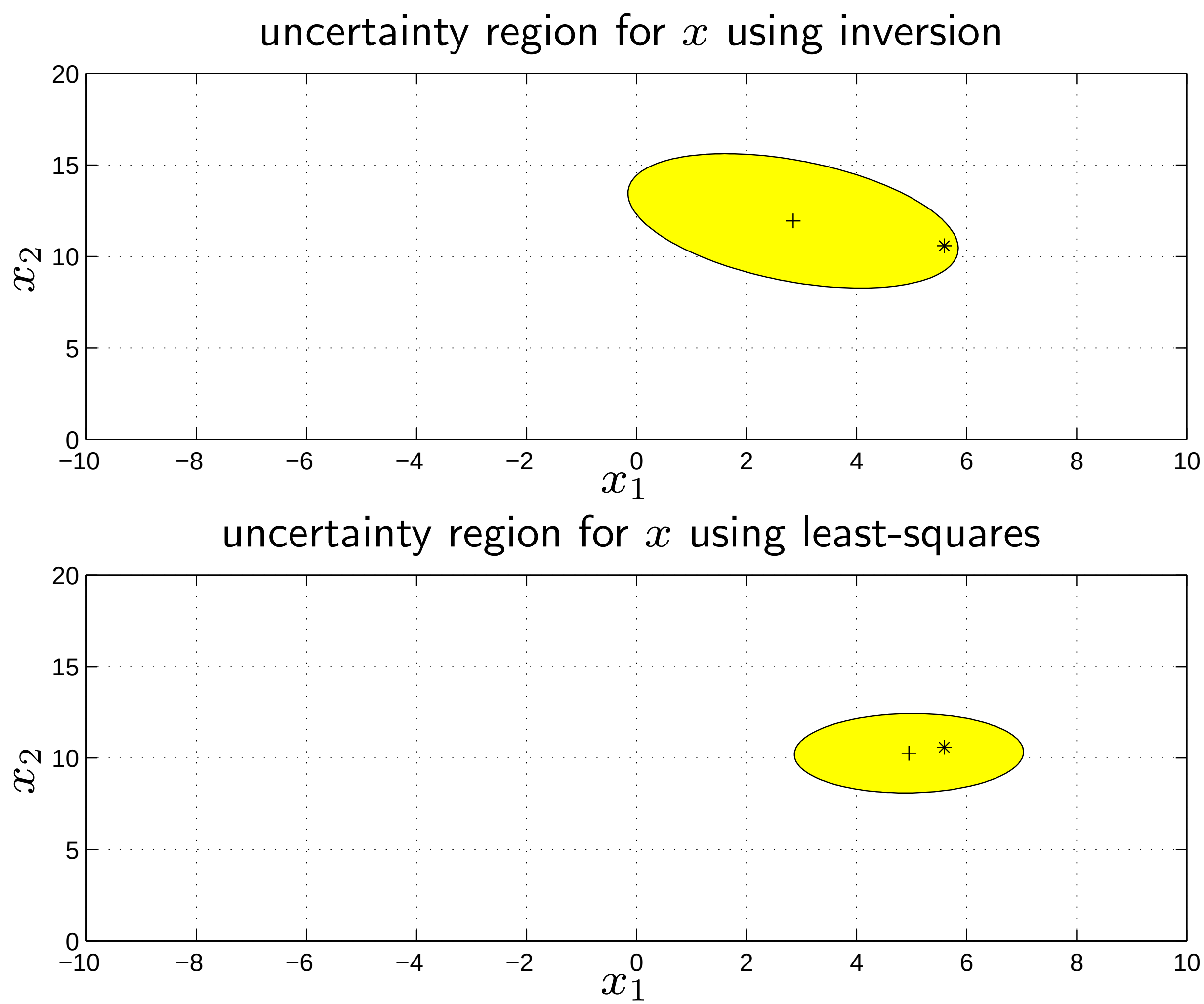
using first two measurements and inverting:

$$\hat{x} = -\begin{bmatrix} \begin{bmatrix} k_1^T \\ k_2^T \end{bmatrix}^{-1} & 0_{2\times 2} \end{bmatrix} y$$

using all four measurements and least-squares:

$$\hat{x} = A^\dagger y$$

uncertainty regions (with $\alpha = 1$):



Proof of optimality property

suppose $A \in \mathbf{R}^{m \times n}$, $m > n$, is full rank

SVD: $A = U\Sigma V^T$, with V orthogonal

$B_{\text{ls}} = A^\dagger = V\Sigma^{-1}U^T$, and B satisfies $BA = I$

define $Z = B - B_{\text{ls}}$, so $B = B_{\text{ls}} + Z$

then $ZA = ZU\Sigma V^T = 0$, so $ZU = 0$ (multiply by $V\Sigma^{-1}$ on right)

therefore

$$\begin{aligned} BB^T &= (B_{\text{ls}} + Z)(B_{\text{ls}} + Z)^T \\ &= B_{\text{ls}}B_{\text{ls}}^T + B_{\text{ls}}Z^T + ZB_{\text{ls}}^T + ZZ^T \\ &= B_{\text{ls}}B_{\text{ls}}^T + ZZ^T \\ &\geq B_{\text{ls}}B_{\text{ls}}^T \end{aligned}$$

using $ZB_{\text{ls}}^T = (ZU)\Sigma^{-1}V^T = 0$

Sensitivity of linear equations to data error

consider $y = Ax$, $A \in \mathbf{R}^{n \times n}$ invertible; of course $x = A^{-1}y$

suppose we have an error or noise in y, *i.e.*, y becomes $y + \delta y$

then x becomes $x + \delta x$ with $\delta x = A^{-1}\delta y$

hence we have $\|\delta x\| = \|A^{-1}\delta y\| \leq \|A^{-1}\|\|\delta y\|$

if $\|A^{-1}\|$ is large,

- small errors in y can lead to large errors in x
- can't solve for x given y (with small errors)
- hence, A can be considered singular in practice

a more refined analysis uses *relative* instead of *absolute* errors in x and y

since $y = Ax$, we also have $\|y\| \leq \|A\|\|x\|$, hence

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta y\|}{\|y\|}$$

$$\kappa(A) = \|A\|\|A^{-1}\| = \sigma_{\max}(A)/\sigma_{\min}(A)$$

is called the *condition number* of A

we have:

relative error in solution $x \leq$ condition number $\cdot$ relative error in data y

or, in terms of $\#$ bits of guaranteed accuracy:

$\#$ bits accuracy in solution $\approx$ $\#$ bits accuracy in data $- \log_2 \kappa$

we say

- A is well conditioned if κ is small
- A is poorly conditioned if κ is large

(definition of 'small' and 'large' depend on application)

same analysis holds for least-squares approximate solutions with A nonsquare, $\kappa = \sigma_{\max}(A)/\sigma_{\min}(A)$

Low rank approximations

suppose $A \in \mathbf{R}^{m \times n}$, $\mathbf{Rank}(A) = r$, with SVD $A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$

we seek matrix $\hat{A}$, $\mathbf{Rank}(\hat{A}) \leq p < r$, s.t. $\hat{A} \approx A$ in the sense that $\|A - \hat{A}\|$ is minimized

solution: optimal rank p approximator is

$$\hat{A} = \sum_{i=1}^{p} \sigma_i u_i v_i^T$$

- hence $\|A - \hat{A}\| = \left\| \sum_{i=p+1}^{r} \sigma_i u_i v_i^T \right\| = \sigma_{p+1}$
- interpretation: SVD dyads $u_i v_i^T$ are ranked in order of 'importance'; take p to get rank p approximant

proof: suppose $\mathbf{Rank}(B) \leq p$

then $\dim \mathcal{N}(B) \geq n - p$

also, $\dim \operatorname{span}\{v_1, \ldots, v_{p+1}\} = p + 1$

hence, the two subspaces intersect, $i.e.$, there is a unit vector $z \in \mathbf{R}^n$ s.t.

$$Bz = 0, \qquad z \in \operatorname{span}\{v_1, \ldots, v_{p+1}\}$$

$$(A - B)z = Az = \sum_{i=1}^{p+1} \sigma_i u_i v_i^T z$$

$$\|(A - B)z\|^2 = \sum_{i=1}^{p+1} \sigma_i^2 (v_i^T z)^2 \geq \sigma_{p+1}^2 \|z\|^2$$

hence $\|A - B\| \geq \sigma_{p+1} = \|A - \hat{A}\|$

Distance to singularity

another interpretation of σ_i:

$$\sigma_i = \min\{ \ \|A - B\| \mid \mathbf{Rank}(B) \leq i - 1 \ \}$$

i.e., the distance (measured by matrix norm) to the nearest rank $i - 1$ matrix

for example, if $A \in \mathbf{R}^{n \times n}$, $\sigma_n = \sigma_{\min}$ is distance to nearest singular matrix

hence, small $\sigma_{\min}$ means A is near to a singular matrix

application: model simplification

suppose $y = Ax + v$, where

- $A \in \mathbf{R}^{100\times 30}$ has SVs

$$10,\ 7,\ 2,\ 0.5,\ 0.01,\ \ldots, 0.0001$$

- $\|x\|$ is on the order of 1

- unknown error or noise v has norm on the order of 0.1

then the terms $\sigma_i u_i v_i^T x$, for $i = 5, \ldots, 30$, are substantially smaller than the noise term v

simplified model:

$$y = \sum_{i=1}^{4} \sigma_i u_i v_i^T x + v$$