4. Singular value decomposition

- singular value decomposition
- related eigendecompositions
- matrix properties from singular value decomposition
- min–max and max–min characterizations
- low-rank approximation
- sensitivity of linear equations
every $m \times n$ matrix $A$ can be factored as

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

- $U$ is $m \times m$ and orthogonal
- $V$ is $n \times n$ and orthogonal
- $\Sigma$ is $m \times n$ and “diagonal”: diagonal with diagonal elements $\sigma_1, \ldots, \sigma_n$ if $m = n$,

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix} \quad \text{if } m > n, \quad \Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_m & 0 & \cdots & 0 \end{bmatrix} \quad \text{if } m < n$$

- the diagonal entries of $\Sigma$ are nonnegative and sorted:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{m,n\}} \geq 0$$
Singular values and singular vectors

\[ A = U \Sigma V^T \]

- the numbers \( \sigma_1, \ldots, \sigma_{\min\{m,n\}} \) are the *singular values* of \( A \)
- the \( m \) columns \( u_i \) of \( U \) are the *left singular vectors*
- the \( n \) columns \( v_i \) of \( V \) are the *right singular vectors*

if we write the factorization \( A = U \Sigma V^T \) as

\[ AV = U \Sigma, \quad A^T U = V \Sigma^T \]

and compare the \( i \)th columns on the left- and right-hand sides, we see that

\[ A v_i = \sigma_i u_i \quad \text{and} \quad A^T u_i = \sigma_i v_i \quad \text{for } i = 1, \ldots, \min\{m,n\} \]

- if \( m > n \) the additional \( m - n \) vectors \( u_i \) satisfy \( A^T u_i = 0 \) for \( i = n + 1, \ldots, m \)
- if \( n > m \) the additional \( n - m \) vectors \( v_i \) satisfy \( A v_i = 0 \) for \( i = m + 1, \ldots, n \)
Reduced SVD

often \( m \gg n \) or \( n \gg m \), which makes one of the orthogonal matrices very large

**Tall matrix:** if \( m > n \), the last \( m - n \) columns of \( U \) can be omitted to define

\[
A = U \Sigma V^T
\]

- \( U \) is \( m \times n \) with orthonormal columns
- \( V \) is \( n \times n \) and orthogonal
- \( \Sigma \) is \( n \times n \) and diagonal with diagonal entries \( \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0 \)

**Wide matrix:** if \( m < n \), the last \( n - m \) columns of \( V \) can be omitted to define

\[
A = U \Sigma V^T
\]

- \( U \) is \( m \times m \) and orthogonal
- \( V \) is \( m \times n \) with orthonormal columns
- \( \Sigma \) is \( m \times m \) and diagonal with diagonal entries \( \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0 \)

we refer to these as *reduced* SVDs (and to the factorization on p. 4.2 as a *full* SVD)
Outline

- singular value decomposition
- related eigendecompositions
- matrix properties from singular value decomposition
- min–max and max–min characterizations
- low-rank approximation
- sensitivity of linear equations
Eigendecomposition of Gram matrix

suppose $A$ is an $m \times n$ matrix with full SVD

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

the SVD is related to the eigendecomposition of the Gram matrix $A^T A$:

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

- $V$ is an orthogonal $n \times n$ matrix
- $\Sigma^T \Sigma$ is a diagonal $n \times n$ matrix with (non-increasing) diagonal elements

$$\sigma_1^2, \sigma_2^2, \ldots, \sigma_{\min\{m,n\}}^2, 0, 0, \ldots, 0$$

$n - \min\{m,n\}$ times

- the $n$ diagonal elements of $\Sigma^T \Sigma$ are the eigenvalues of $A^T A$
- the right singular vectors (columns of $V$) are corresponding eigenvectors
the SVD also gives the eigendecomposition of $AA^T$:

$$AA^T = U \Sigma \Sigma^T U^T$$

- $U$ is an orthogonal $m \times m$ matrix
- $\Sigma \Sigma^T$ is a diagonal $m \times m$ matrix with (non-increasing) diagonal elements

$$\sigma_1^2, \sigma_2^2, \ldots, \sigma_{\min\{m,n\}}^2, 0, 0, \cdots, 0_{m - \min\{m,n\} \text{ times}}$$

- the $m$ diagonal elements of $\Sigma \Sigma^T$ are the eigenvalues of $AA^T$
- the left singular vectors (columns of $U$) are corresponding eigenvectors

in particular, the first $\min\{m,n\}$ eigenvalues of $A^T A$ and $AA^T$ are the same:

$$\sigma_1^2, \sigma_2^2, \ldots, \sigma_{\min\{m,n\}}^2$$
Example

scatter plot shows $m = 500$ points from the normal distribution on page 3.34

$$
\mu = \begin{bmatrix} 5 \\ 4 \end{bmatrix}, \quad \Sigma_{ex} = \frac{1}{4} \begin{bmatrix} 7 & \sqrt{3} \\ \sqrt{3} & 5 \end{bmatrix}
$$

- we define an $m \times 2$ data matrix $X$ with the $m$ vectors as its rows
- the centered data matrix is $X_c = X - (1/m)11^T X$

sample estimate of mean is

$$
\hat{\mu} = \frac{1}{m}X^T1 = \begin{bmatrix} 5.01 \\ 3.93 \end{bmatrix}
$$

sample estimate of covariance is

$$
\hat{\Sigma} = \frac{1}{m}X_c^TX_c = \begin{bmatrix} 1.67 & 0.48 \\ 0.48 & 1.35 \end{bmatrix}
$$
Example

\[ A = \frac{1}{\sqrt{m}} X_c \]

- Eigenvectors of \( \hat{\Sigma} \) are right singular vectors \( v_1, v_2 \) of \( A \) (and of \( X_c \))
- Eigenvalues of \( \hat{\Sigma} \) are squares of the singular values of \( A \)
Existence of singular value decomposition

the Gram matrix connection gives a proof that every matrix has an SVD

- assume $A$ is $m \times n$ with $m \geq n$ and rank $r$
- the $n \times n$ matrix $A^T A$ has rank $r$ (page 2.5) and an eigendecomposition
  \[ A^T A = V \Lambda V^T \]  
  \(\Lambda\) is diagonal with diagonal elements $\lambda_1 \geq \cdots \geq \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_n$
- define $\sigma_i = \sqrt{\lambda_i}$ for $i = 1, \ldots, n$, and an $n \times n$ matrix
  \[ U = \begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_1} A v_1 & \frac{1}{\sigma_2} A v_2 & \cdots & \frac{1}{\sigma_r} A v_r & u_{r+1} & \cdots & u_n \end{bmatrix} \]
  where $u_{r+1}, \ldots, u_n$ form any orthonormal basis for $\text{null}(A^T)$
- \(1\) and the choice of $u_{r+1}, \ldots, u_n$ imply that $U$ is orthogonal
- \(1\) also implies that $A v_i = 0$ for $i = r+1, \ldots, n$
- together with the definition of $u_1, \ldots, u_r$ this shows that $AV = U \Sigma$
Non-uniqueness of singular value decomposition

the derivation from the eigendecomposition

\[ A^T A = V \Lambda V^T \]

shows that the singular value decomposition of \( A \) is almost unique

**Singular values**

- the singular values of \( A \) are uniquely defined
- we have also shown that \( A \) and \( A^T \) have the same singular values

**Singular vectors** (assuming \( m \geq n \): see the discussion on page 3.14

- right singular vectors \( v_i \) with the same positive singular value span a subspace
- in this subspace, any other orthonormal basis can be chosen
- the first \( r = \text{rank}(A) \) left singular vectors then follow from \( \sigma_i u_i = A v_i \)
- the remaining vectors \( v_{r+1}, \ldots, v_n \) can be any orthonormal basis for \( \text{null}(A) \)
- the remaining vectors \( u_{r+1}, \ldots, u_m \) can be any orthonormal basis for \( \text{null}(A^T) \)
Exercise

suppose $A$ is an $m \times n$ matrix with $m \geq n$, and define

$$B = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$$

1. suppose $A = U \Sigma V^T$ is a full SVD of $A$; verify that

$$B = \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} 0 & \Sigma \\ \Sigma^T & 0 \end{bmatrix} \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}^T$$

2. derive from this an eigendecomposition of $B$

*Hint*: if $\Sigma_1$ is square, then

$$\begin{bmatrix} 0 & \Sigma_1 \\ \Sigma_1 & 0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} I & I \\ I & -I \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & -\Sigma_1 \end{bmatrix} \begin{bmatrix} I & I \\ I & -I \end{bmatrix}$$

3. what are the $m + n$ eigenvalues of $B$?
Outline

• singular value decomposition

• related eigendecompositions

• **matrix properties from singular value decomposition**

• min–max and max–min characterizations

• low-rank approximation

• sensitivity of linear equations
the number of positive singular values is the rank of a matrix

• suppose there are \( r \) positive singular values:

\[
\sigma_1 \geq \cdots \geq \sigma_r > 0 = \sigma_{r+1} = \cdots = \sigma_{\min\{m,n\}}
\]

• partition the matrices in a full SVD of \( A \) as

\[
A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^T
= U_1\Sigma_1V_1^T
\]

(2)

\( \Sigma_1 \) is \( r \times r \) with the positive singular values \( \sigma_1, \ldots, \sigma_r \) on the diagonal

• since \( U_1 \) and \( V_1 \) have orthonormal columns, the factorization (2) proves that

\[
\text{rank}(A) = r
\]

(see page 1.12)
Inverse and pseudo-inverse

we use the same notation as on the previous page

\[ A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^T = U_1\Sigma_1V_1^T \]

diagonal entries of \( \Sigma_1 \) are the positive singular values of \( A \)

- pseudo-inverse follows from page 1.36:

\[ A^\dagger = V_1\Sigma_1^{-1}U_1^T \]

\[ = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} \Sigma_1^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} \]

\[ = V\Sigma^\dagger U^T \]

- if \( A \) is square and nonsingular, this reduces to the inverse

\[ A^{-1} = (U\Sigma V^T)^{-1} = V\Sigma^{-1}U^T \]
we continue with the same notation for the SVD of an $m \times n$ matrix $A$ with rank $r$:

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^T$$

the diagonal entries of $\Sigma_1$ are the positive singular values of $A$

the SVD provides orthonormal bases for the four subspaces associated with $A$

- the columns of the $m \times r$ matrix $U_1$ are a basis of $\text{range}(A)$
- the columns of the $m \times (m - r)$ matrix $U_2$ are a basis of $\text{range}(A)^\perp = \text{null}(A^T)$
- the columns of the $n \times r$ matrix $V_1$ are a basis of $\text{range}(A^T)$
- the columns of the $n \times (n - r)$ matrix $V_2$ are a basis of $\text{null}(A)$
Image of unit ball

define $\mathcal{E}_y$ as the image of the unit ball under the linear mapping $y = Ax$:

$$\mathcal{E}_y = \{Ax \mid \|x\| \leq 1\} = \{U \Sigma V^T x \mid \|x\| \leq 1\}$$
Control interpretation

system $A$ maps input $x$ to output $y = Ax$

$\begin{array}{c}
x \\ \rightarrow \\ A \\ \rightarrow \\ y = Ax
\end{array}$

- if $\|x\|^2$ represents input energy, the set of outputs realizable with unit energy is

$$\mathcal{E}_y = \{Ax \mid \|x\| \leq 1\}$$

- assume $\text{rank}(A) = m$: every desired $y$ can be realized by at least one input

- the most energy-efficient input that generates output $y$ is

$$x_{\text{eff}} = A^\dagger y = A^T (AA^T)^{-1} y, \quad \|x_{\text{eff}}\|^2 = y^T (AA^T)^{-1} y = \sum_{i=1}^{m} \frac{(u_i^T y)^2}{\sigma_i^2}$$

- (if $\text{rank}(A) = m$) the set $\mathcal{E}_y$ is an ellipsoid $\mathcal{E}_y = \{y \mid y^T (AA^T)^{-1} y \leq 1\}$
Inverse image of unit ball

define $E_x$ as the inverse image of the unit ball under the linear mapping $y = Ax$:

$$E_x = \{ x \mid \|Ax\| \leq 1 \} = \{ x \mid \|U\Sigma V^T x\| \leq 1 \}$$
Estimation interpretation

measurement $A$ maps unknown quantity $x_{\text{true}}$ to observation

$$y_{\text{obs}} = A(x_{\text{true}} + v)$$

where $v$ is unknown but bounded by $\|Av\| \leq 1$

- if $\text{rank}(A) = n$, there is a unique estimate $\hat{x}$ that satisfies $A\hat{x} = y_{\text{obs}}$
- uncertainty in $y$ causes uncertainty in estimate: true value $x_{\text{true}}$ must satisfy

$$\|A(x_{\text{true}} - \hat{x})\| \leq 1$$

- the set $\mathcal{E}_x = \{x \mid \|A(x - \hat{x})\| \leq 1\}$ is the uncertainty region around estimate $\hat{x}$
Frobenius norm and 2-norm

for an $m \times n$ matrix $A$ with singular values $\sigma_i$:

$$\|A\|_F = \left( \sum_{i=1}^{\min\{m,n\}} \sigma_i^2 \right)^{1/2}, \quad \|A\|_2 = \sigma_1$$

this readily follows from the unitary invariance of the two norms:

$$\|A\|_F = \|U\Sigma V^T\|_F = \|\Sigma\|_F = \left( \sum_{i=1}^{\min\{m,n\}} \sigma_i^2 \right)^{1/2}$$

and

$$\|A\|_2 = \|U\Sigma V^T\|_2 = \|\Sigma\|_2 = \sigma_1$$
Exercise

Exercise 1: express $\|A^\dagger\|_2$ and $\|A^\dagger\|_F$ in terms of the singular values of $A$

Exercise 2: the condition number of a square nonsingular matrix $A$ is defined as

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

express $\kappa(A)$ in terms of the singular values of $A$

Exercise 3: give an SVD and the 2-norm of the matrix

$$A = ab^T$$

where $a$ is an $n$-vector and $b$ is an $m$-vector
Outline

• singular value decomposition

• related eigendecompositions

• matrix properties from singular value decomposition

• min–max and max–min characterizations

• low-rank approximation

• sensitivity of linear equations
First singular value

the first singular value is the maximal value of several functions:

\[
\sigma_1 = \max_{\|x\|=1} \|Ax\| = \max_{\|x\|=\|y\|=1} y^T Ax = \max_{\|y\|=1} \|A^T y\| \tag{3}
\]

- the first and last expressions follow from page 3.24 and

\[
\sigma_1^2 = \lambda_{\text{max}}(A^T A) = \max_{\|x\|=1} x^T A^T Ax, \quad \sigma_1^2 = \lambda_{\text{max}}(AA^T) = \max_{\|y\|=1} y^T AA^T y
\]

- second expression in (3) follows from the Cauchy–Schwarz inequality:

\[
\|Ax\| = \max_{\|y\|=1} y^T (Ax), \quad \|A^T y\| = \max_{\|x\|=1} x^T (A^T y)
\]
First singular value

alternatively, we can use an SVD of $A$ to solve the maximization problems in

$$\sigma_1 = \max_{\|x\|=1} \|Ax\| = \max_{\|x\|=\|y\|=1} y^T A x = \max_{\|y\|=1} \|A^T y\|$$

(4)

• suppose $A = USV^T$ is a full SVD of $A$

• if we define $\tilde{x} = V^T x$, $\tilde{y} = U^T y$, then (4) can be written as

$$\sigma_1 = \max_{\|\tilde{x}\|=1} \|\Sigma \tilde{x}\| = \max_{\|\tilde{x}\|=\|\tilde{y}\|=1} \tilde{y}^T \Sigma \tilde{x} = \max_{\|\tilde{y}\|=1} \|\Sigma^T \tilde{y}\|$$

• an optimal choice for $\tilde{x}$ and $\tilde{y}$ is $\tilde{x} = (1, 0, \ldots, 0)$ and $\tilde{y} = (1, 0, \ldots, 0)$

• therefore $x = v_1$ and $y = u_1$ are optimal for each of the maximizations in (4)
Last singular value

two of the three expressions in (3) have a counterpart for the last singular value

• for an $m \times n$ matrix $A$, the last singular $\sigma_{\min\{m,n\}}$ can be written as follows:

$$\begin{align*}
\text{if } m \geq n: \quad \sigma_n &= \min_{\|x\|=1} \|Ax\|, \\
\text{if } n \geq m: \quad \sigma_m &= \min_{\|y\|=1} \|A^T y\|
\end{align*}$$

(5)

• if $m \neq n$, we need to distinguish the two cases because

$$\begin{align*}
\min_{\|x\|=1} \|Ax\| &= 0 \quad \text{if } n > m, \\
\min_{\|y\|=1} \|A^T y\| &= 0 \quad \text{if } m > n
\end{align*}$$

to prove (5), we substitute full SVD $A = U \Sigma V^T$, and define $\tilde{x} = V^T x$, $\tilde{y} = U^T y$:

$$\begin{align*}
\text{if } m \geq n: \quad \min_{\|\tilde{x}\|=1} \|\Sigma \tilde{x}\| &= \min_{\|\tilde{x}\|=1} \left( \sigma_1^2 \tilde{x}_1^2 + \cdots + \sigma_n^2 \tilde{x}_n^2 \right)^{1/2} = \sigma_n \\
\text{if } n \geq m: \quad \min_{\|\tilde{y}\|=1} \|\Sigma^T \tilde{y}\| &= \min_{\|\tilde{y}\|=1} \left( \sigma_1^2 \tilde{y}_1^2 + \cdots + \sigma_m^2 \tilde{y}_m^2 \right)^{1/2} = \sigma_m
\end{align*}$$

optimal choices for $x$ and $y$ in (5) are $x = v_n$, $y = u_m$
Max–min characterization

we extend (3) to a max–min characterization of the other singular values:

\[
\sigma_k = \max_{X^TX=I_k} \sigma_{\min}(AX) \tag{6a}
\]

\[
= \max_{X^TX=Y^TY=I_k} \sigma_{\min}(Y^TA^XY) \tag{6b}
\]

\[
= \max_{Y^TY=I_k} \sigma_{\min}(A^TY) \tag{6c}
\]

- \(\sigma_k\) for \(k = 1, \ldots, \min\{m,n\}\) are the singular values of the \(m \times n\) matrix \(A\)
- \(X\) is \(n \times k\) with orthonormal columns, \(Y\) is \(m \times k\) with orthonormal columns
- \(\sigma_{\min}(B)\) denotes the smallest singular value of the matrix \(B\)
- in the three expressions in (6) \(\sigma_{\min}(\cdot)\) denotes the \(k\)th singular value
- for \(k = 1\), we obtain the three expressions for \(\sigma_1\) in (3)

- these can be derived from the min–max theorems for eigenvalues (p. 3.29)
- or we can find an optimal choice for \(X, Y\) from an SVD of \(A\) (we skip the details)
Min–max characterization

we extend (5) to a min–max characterization of all singular values

Tall or square matrix: if $A$ is $m \times n$ with $m \geq n$

$$\sigma_{n-k+1} = \min_{X^TX=I_k} \|AX\|_2, \quad k = 1, \ldots, n$$

• we minimize over $n \times k$ matrices $X$ with orthonormal columns
• $\|AX\|_2$ is the maximum singular value of an $m \times k$ matrix
• for $k = 1$, this is the first expression in (5)

Wide or square matrix ($A$ is $m \times n$ with $m \leq n$)

$$\sigma_{m-k+1} = \min_{Y^TY=I_k} \|A^TY\|_2, \quad k = 1, \ldots, m$$

• we minimize over $n \times k$ matrices $Y$ with orthonormal columns
• $\|A^TY\|_2$ is the maximum singular value of an $n \times k$ matrix
• for $k = 1$, this is the second expression in (5)
Proof of min–max characterization (for \( m \geq n \))

we use a full SVD \( A = U\Sigma V^T \) to solve the optimization problem

\[
\begin{align*}
\text{minimize} & \quad \|AX\|_2 \\
\text{subject to} & \quad X^TX = I
\end{align*}
\]

changing variables to \( \tilde{X} = V^TX \) gives the equivalent problem

\[
\begin{align*}
\text{minimize} & \quad \|\Sigma\tilde{X}\|_2 \\
\text{subject to} & \quad \tilde{X}^T\tilde{X} = I
\end{align*}
\]

• we show that the optimal value is \( \sigma_{n-k+1} \)

• an optimal solution for \( \tilde{X} \) is formed from the last \( k \) columns of the \( n \times n \) identity

\[
\tilde{X}_{\text{opt}} = \begin{bmatrix} e_{n-k+1} & \cdots & e_{n-1} & e_n \end{bmatrix}
\]

• an optimal solution for \( X \) is formed from the last \( k \) columns of \( V \):

\[
X_{\text{opt}} = \begin{bmatrix} v_{n-k+1} & \cdots & v_{n-1} & v_n \end{bmatrix}
\]
Proof of min–max characterization (for $m \geq n$)

- we first note that $\Sigma \tilde{X}_{\text{opt}}$ are the last $k$ columns of $\Sigma$, so $\|\Sigma \tilde{X}_{\text{opt}}\|_2 = \sigma_{n-k+1}$
- to show that this is optimal, consider any other $n \times k$ matrix $\tilde{X}$ with $\tilde{X}^T \tilde{X} = I$
- find a nonzero $k$-vector $u$ for which $y = \tilde{X}u$ has last $k - 1$ components zero:

$$
\tilde{X}u = (y_1, \ldots, y_{n-k+1}, 0, \ldots, 0)
$$

this is possible because a $(k - 1) \times k$ matrix has linearly dependent columns
- normalize $u$ so that $\|u\| = \|y\| = 1$, and use it to lower bound $\|\Sigma \tilde{X}\|_2$:

$$
\|\Sigma \tilde{X}\|_2 \geq \|\Sigma \tilde{X}u\|_2 = \|\Sigma y\|_2 = \left(\sigma_1^2 y_1^2 + \cdots + \sigma_{n-k+1}^2 y_{n-k+1}^2\right)^{1/2} \geq \sigma_{n-k+1} \left(y_1^2 + \cdots + y_{n-k+1}^2\right)^{1/2} = \sigma_{n-k+1}
$$

Singular value decomposition 4.27
Outline

• singular value decomposition

• related eigendecompositions

• matrix properties from singular value decomposition

• min–max and max–min characterizations

• low-rank approximation

• sensitivity of linear equations
Rank-$r$ approximation

let $A$ be an $m \times n$ matrix with rank($A$) > $r$ and full SVD

$$A = U\Sigma V^T = \sum_{i=1}^{\min\{m,n\}} \sigma_i u_i v_i^T, \quad \sigma_1 \geq \cdots \geq \sigma_{\min\{m,n\}} \geq 0, \quad \sigma_{r+1} > 0$$

the best rank-$r$ approximation of $A$ is the sum of the first $r$ terms in the SVD:

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

- $B$ is the best approximation for the Frobenius norm: for every $C$ with rank $r$,

$$\|A - C\|_F \geq \|A - B\|_F = \left( \sum_{i=r+1}^{\min\{m,n\}} \sigma_i^2 \right)^{1/2}$$

- $B$ is also the best approximation for the 2-norm: for every $C$ with rank $r$,

$$\|A - C\|_2 \geq \|A - B\|_2 = \sigma_{r+1}$$
Rank-\(r\) approximation in Frobenius norm

the approximation problem in \(|| \cdot ||_F\) is a nonlinear least squares problem:

\[
\text{minimize} \quad ||A - YX^T||_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \left( A_{ij} - \sum_{k=1}^{r} Y_{ik}X_{jk} \right)^2
\]  \hspace{1cm} (8)

• matrix \(B\) is written as \(B = YX^T\) with \(Y\) of size \(m \times r\) and \(X\) of size \(n \times r\)
• we optimize over \(X\) and \(Y\)

Outline of the solution

• the first order (necessary but not sufficient) optimality conditions are

\[
AX = Y(X^TX), \quad A^TY = X(Y^TY)
\]

• can assume \(X^TX = Y^TY = D\) with \(D\) diagonal (e.g., get \(X, Y\) from SVD of \(B\))
• then columns of \(X, Y\) are formed from \(r\) pairs of right/left singular vectors of \(A\)
• optimal choice for (8) is to take the first \(r\) singular vector pairs

Singular value decomposition
Rank-$r$ approximation in 2-norm

to show that $B$ is the best approximation in 2-norm, we prove that

$$
\|A - C\|_2 \geq \|A - B\|_2 \quad \text{for all } C \text{ with } \text{rank}(C) = r
$$
on the right-hand side

$$
A - B = \sum_{i=1}^{\min\{m,n\}} \sigma_i u_i v_i^T - \sum_{i=1}^r \sigma_i u_i v_i^T
= \sum_{i=r+1}^{\min\{m,n\}} \sigma_i u_i v_i^T
= \begin{bmatrix} u_{r+1} & \cdots & u_{\min\{m,n\}} \end{bmatrix} \begin{bmatrix} \sigma_{r+1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{\min\{m,n\}} \end{bmatrix} \begin{bmatrix} v_{r+1} \\ \vdots \\ v_{\min\{m,n\}} \end{bmatrix}^T
$$

and the 2-norm of the difference is $\|A - B\|_2 = \sigma_{r+1}$
on the next page we show that $\|A - C\|_2 \geq \sigma_{r+1}$ if $C$ has rank $r$
Proof: we prove that $\|A - C\|_2 \geq \sigma_{r+1}$ for all $m \times n$ matrices $C$ of rank $r$

- we assume that $m \geq n$ (otherwise, first take the transpose of $A$ and $C$)
- if $\text{rank}(C) = r$, the nullspace of $C$ has dimension $n - r$
- define an $n \times (n - r)$ matrix $\hat{X}$ with orthonormal columns that span null($C$)
- use the min–max characterization on page 4.25 to bound $\|A - C\|_2$:

\[
\|A - C\|_2 = \max_{\|x\| = 1} \|(A - C)x\| \\
\geq \max_{\|w\| = 1} \|(A - C)\hat{X}w\| \\
= \|(A - C)\hat{X}\|_2 \\
= \|A\hat{X}\|_2 \\
\geq \min_{X^TX = I_{n-r}} \|AX\|_2 \\
= \sigma_{r+1} \quad \text{(apply (7) with } k = n - r)\
\]
Outline

• singular value decomposition

• related eigendecompositions

• matrix properties from singular value decomposition

• min–max and max–min characterizations

• low-rank approximation

• sensitivity of linear equations
SVD of square matrix

for the rest of the lecture we assume that $A$ is $n \times n$ and nonsingular with SVD

$$A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T$$

- 2-norm of $A$ is $\|A\|_2 = \sigma_1$
- $A$ is nonsingular if and only if $\sigma_n > 0$
- inverse of $A$ and 2-norm of $A^{-1}$ are

$$A^{-1} = V \Sigma^{-1} U^T = \sum_{i=1}^{n} \frac{1}{\sigma_i} v_i u_i^T, \quad \|A^{-1}\|_2 = \frac{1}{\sigma_n}$$

- condition number of $A$ is

$$\kappa(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n} \geq 1$$

$A$ is called *ill-conditioned* if the condition number is very high.
Sensitivity to right-hand side perturbations

linear equation with right-hand side $b \neq 0$ and perturbed right-hand side $b + e$:

$$Ax = b, \quad Ay = b + e$$

- bound on distance between the solutions:

$$\|y - x\| = \|A^{-1}e\| \leq \|A^{-1}\|_2\|e\|$$

recall that $\|Bx\| \leq \|B\|_2\|x\|$ for matrix 2-norm and Euclidean vector norm

- bound on relative change in the solution, in terms of $\delta_b = \|e\|/\|b\|:

$$\frac{\|y - x\|}{\|x\|} \leq \|A\|_2\|A^{-1}\|_2 \frac{\|e\|}{\|b\|} = \kappa(A) \delta_b$$

in the first step we use $\|b\| = \|Ax\| \leq \|A\|_2\|x\|$

large $\kappa(A)$ indicates that the solution can be very sensitive to changes in $b$
Worst-case perturbation of right-hand side

\[ \frac{\|y - x\|}{\|x\|} \leq \kappa(A) \delta_b \quad \text{where} \quad \delta_b = \frac{\|e\|}{\|b\|} \]

- the upper bound is often very conservative
- however, for every $A$ one can find $b, e$ for which the bound holds with equality

- choose $b = u_1$ (first left singular vector of $A$): solution of $Ax = b$ is

\[ x = A^{-1}b = V \Sigma^{-1} U^T u_1 = \frac{1}{\sigma_1} v_1 \]

- choose $e = \delta_b u_n$ ($\delta_b$ times left singular vector $u_n$): solution of $Ay = b + e$ is

\[ y = A^{-1}(b + e) = x + \frac{\delta_b}{\sigma_n} v_n \]

- relative change is

\[ \frac{\|y - x\|}{\|x\|} = \frac{\sigma_1 \delta_b}{\sigma_n} = \kappa(A) \delta_b \]
Nearest singular matrix

the singular matrix closest to $A$ is

$$\sum_{i=1}^{n-1} \sigma_i u_i v_i^T = A + E \quad \text{where} \quad E = -\sigma_n u_n v_n^T$$

• this gives another interpretation of the condition number:

$$\|E\|_2 = \sigma_n = \frac{1}{\|A^{-1}\|_2}, \quad \frac{\|E\|_2}{\|A\|_2} = \frac{\sigma_n}{\sigma_1} = \frac{1}{\kappa(A)}$$

$1/\kappa(A)$ is the relative distance of $A$ to the nearest singular matrix

• this also implies that a perturbation $A + E$ of $A$ is certainly nonsingular if

$$\|E\|_2 < \frac{1}{\|A^{-1}\|_2} = \sigma_n$$
Bound on inverse

on the next page we prove the following inequality:

\[
\|(A + E)^{-1}\|_2 \leq \frac{\|A^{-1}\|_2}{1 - \|A^{-1}\|_2\|E\|_2} \quad \text{if } \|E\|_2 < \frac{1}{\|A^{-1}\|_2}
\] (9)

using \(\|A^{-1}\|_2 = 1/\sigma_n\):

\[
\|(A + E)^{-1}\|_2 \leq \frac{1}{\sigma_n - \|E\|_2}
\]

if \(\|E\|_2 < \sigma_n\)
Proof:

• the matrix $Y = (A + E)^{-1}$ satisfies

$$ (I + A^{-1}E)Y = A^{-1}(A + E)Y = A^{-1} $$

• therefore

$$ \|Y\|_2 = \|A^{-1} - A^{-1}EY\|_2 $$

$$ \leq \|A^{-1}\|_2 + \|A^{-1}EY\|_2 \quad \text{(triangle inequality)} $$

$$ \leq \|A^{-1}\|_2 + \|A^{-1}E\|_2 \|Y\|_2 $$

in the last step we use the property $\|CD\|_2 \leq \|C\|_2 \|D\|_2$ of the matrix 2-norm

• rearranging the last inequality for $\|Y\|_2$ gives

$$ \|Y\|_2 \leq \frac{\|A^{-1}\|_2}{1 - \|A^{-1}E\|_2} \leq \frac{\|A^{-1}\|_2}{1 - \|A^{-1}\|_2 \|E\|_2} $$

in the second step we again use the property $\|A^{-1}E\|_2 \leq \|A^{-1}\|_2 \|E\|_2$
Sensitivity to perturbations of coefficient matrix

linear equation with matrix $A$ and perturbed matrix $A + E$:

$$Ax = b, \quad (A + E)y = b$$

• we assume $\|E\|_2 < 1/\|A^{-1}\|_2$, which guarantees that $A + E$ is nonsingular

• bound on distance between the solutions:

$$\|y - x\| = \|(A + E)^{-1}(b - (A + E)x)\|$$

$$= \|(A + E)^{-1}Ex\|$$

$$\leq \|(A + E)^{-1}\|_2 \|E\|_2 \|x\|$$

$$\leq \frac{\|A^{-1}\|_2 \|E\|_2}{1 - \|A^{-1}\|_2 \|E\|_2} \|x\| \quad \text{(applying (9))}$$

• bound on relative change in solution in terms of $\delta_A = \|E\|_2/\|A\|_2$:

$$\frac{\|y - x\|}{\|x\|} \leq \frac{\kappa(A) \delta_A}{1 - \kappa(A)\delta_A} \quad (10)$$

Singular value decomposition 4.38
Worst-case perturbation of coefficient matrix

an example where the upper bound (10) is sharp (from SVD $A = \sum_{i=1}^{n} \sigma_i u_i v_i^T$)

- choose $b = u_n$: the solution of $Ax = b$ is
  \[
x = A^{-1} b = (1/\sigma_n)v_n
  \]

- choose $E = -\delta_A \sigma_1 u_n v_n^T$ with $\delta_A < \sigma_n/\sigma_1 = 1/\kappa(A)$:
  \[
  A + E = \sum_{i=1}^{n-1} \sigma_i u_i v_i^T + (\sigma_n - \delta_A \sigma_1) u_n v_n^T
  \]

- solution of $(A + E)y = b$ is
  \[
y = (A + E)^{-1} b = \frac{1}{\sigma_n - \delta_A \sigma_1} v_n
  \]

- relative change in solution is
  \[
  \frac{\|y - x\|}{\|x\|} = \sigma_n \left( \frac{1}{\sigma_n - \delta_A \sigma_1} - \frac{1}{\sigma_n} \right) = \frac{\delta_A \kappa(A)}{1 - \delta_A \kappa(A)}
  \]
Exercises

Exercise 1

to evaluate the sensitivity to changes in $A$, we can also look at the residual

$$||(A + E)x - b||$$

where $x = A^{-1}b$ is the solution of $Ax = b$

1. show that

$$\frac{||(A + E)x - b||}{||b||} \leq \kappa(A)\delta_A \quad \text{where } \delta_A = \frac{||E||}{||A||}$$

2. show that for every $A$ there exist $b, E$ for which the inequality is sharp
**Exercises**

**Exercise 2:** consider perturbations in \(A\) and \(b\)

\[
Ax = b, \quad (A + E)y = b + e
\]

assuming \(\|E\|_2 < 1/\|A^{-1}\|_2\), show that

\[
\frac{\|y - x\|}{\|x\|} \leq \frac{(\delta_A + \delta_b)\kappa(A)}{1 - \delta_A\kappa(A)}
\]

where \(\delta_b = \|e\|/\|b\|\) and \(\delta_A = \|E\|_2/\|A\|_2\)