Linear Algebra

Shan-Hung Wu

shwu@cs.nthu.edu.tw

Department of Computer Science,
National Tsing Hua University, Taiwan

Large-Scale ML, Fall 2016
Outline

1. Span & Linear Dependence
2. Norms
3. Eigendecomposition
4. Singular Value Decomposition
5. Traces and Determinant
Outline

1. Span & Linear Dependence
2. Norms
3. Eigendecomposition
4. Singular Value Decomposition
5. Traces and Determinant
A linear function (or map or transformation) $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ can be represented by a matrix $A$, $A \in \mathbb{R}^{m \times n}$, such that

$$f(x) = Ax = y, \forall x \in \mathbb{R}^n, y \in \mathbb{R}^m$$
A linear function (or map or transformation) $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ can be represented by a matrix $A$, $A \in \mathbb{R}^{m \times n}$, such that

$$f(x) = Ax = y, \forall x \in \mathbb{R}^n, y \in \mathbb{R}^m$$

$\text{span}(A_{:,1}, \cdots, A_{:,n})$ is called the \textit{column space} of $A$.

$\text{rank}(A) = \text{dim}(\text{span}(A_{:,1}, \cdots, A_{:,n}))$
System of Linear Equations

- Given $A$ and $y$, solve $x$ in $Ax = y$
System of Linear Equations

- Given $A$ and $y$, solve $x$ in $Ax = y$
- What kind of $A$ that makes $Ax = y$ always have a solution?
System of Linear Equations

- Given $A$ and $y$, solve $x$ in $Ax = y$
- What kind of $A$ that makes $Ax = y$ always have a solution?
  - Since $Ax = \sum_i x_i A_{:,i}$, the column space of $A$ must contain $\mathbb{R}^m$, i.e.,
    \[ \mathbb{R}^m \subseteq \text{span}(A_{:,1}, \cdots, A_{:,n}) \]
  - Implies $n \geq m$
System of Linear Equations

- Given $A$ and $y$, solve $x$ in $Ax = y$
- What kind of $A$ that makes $Ax = y$ always have a solution?
  - Since $Ax = \sum ix_iA_{:,i}$, the column space of $A$ must contain $\mathbb{R}^m$, i.e.,
    $$\mathbb{R}^m \subseteq span(A_{:,1}, \cdots, A_{:,n})$$
  - Implies $n \geq m$
- When does $Ax = y$ always have exactly one solution?
Given $A$ and $y$, solve $x$ in $Ax = y$

What kind of $A$ that makes $Ax = y$ always have a solution?

- Since $Ax = \sum_ix_iA_{i,:}$, the column space of $A$ must contain $\mathbb{R}^m$, i.e.,
  $\mathbb{R}^m \subseteq \text{span}(A_{1,:}, \cdots, A_{n,:})$
- Implies $n \geq m$

When does $Ax = y$ always have exactly one solution?

- $A$ has at most $m$ columns; otherwise there is more than one $x$ parametrizing each $y$
- Implies $n = m$ and the columns of $A$ are linear independent with each other
System of Linear Equations

- Given $A$ and $y$, solve $x$ in $Ax = y$
- What kind of $A$ that makes $Ax = y$ always have a solution?
  - Since $Ax = \Sigma_i x_i A_{:,i}$, the column space of $A$ must contain $\mathbb{R}^m$, i.e.,
    \[ \mathbb{R}^m \subseteq \text{span}(A_{:,1}, \cdots, A_{:,n}) \]
  - Implies $n \geq m$
- When does $Ax = y$ always have exactly one solution?
  - $A$ has at most $m$ columns; otherwise there is more than one $x$
    parametrizing each $y$
  - Implies $n = m$ and the columns of $A$ are \textit{linear independent} with each other
  - $A^{-1}$ exists at this time, and $x = A^{-1}y$
Outline

1. Span & Linear Dependence
2. Norms
3. Eigendecomposition
4. Singular Value Decomposition
5. Traces and Determinant
Vector Norms

- A **norm** of vectors is a function $\| \cdot \|$ that maps vectors to non-negative values satisfying
  - $\|x\| = 0 \Rightarrow x = 0$
  - $\|x + y\| \leq \|x\| + \|y\|$ (the triangle inequality)
  - $\|cx\| = |c| \cdot \|x\|, \forall c \in \mathbb{R}$
Vector Norms

- A norm of vectors is a function \( \| \cdot \| \) that maps vectors to non-negative values satisfying
  - \( \| \mathbf{x} \| = 0 \Rightarrow \mathbf{x} = \mathbf{0} \)
  - \( \| \mathbf{x} + \mathbf{y} \| \leq \| \mathbf{x} \| + \| \mathbf{y} \| \) (the triangle inequality)
  - \( \| c \mathbf{x} \| = |c| \cdot \| \mathbf{x} \|, \forall c \in \mathbb{R} \)

- E.g., the \( L^p \) norm

\[
\| \mathbf{x} \|_p = \left( \sum_i |x_i|^p \right)^{1/p}
\]

- \( L^2 \) (Euclidean) norm: \( \| \mathbf{x} \| = (\mathbf{x}^\top \mathbf{x})^{1/2} \)
- \( L^1 \) norm: \( \| \mathbf{x} \|_1 = \sum_i |x_i| \)
- Max norm: \( \| \mathbf{x} \|_\infty = \max_i |x_i| \)
Vector Norms

- A **norm** of vectors is a function \( \| \cdot \| \) that maps vectors to non-negative values satisfying
  - \( \| x \| = 0 \Rightarrow x = 0 \)
  - \( \| x + y \| \leq \| x \| + \| y \| \) (the triangle inequality)
  - \( \| cx \| = |c| \cdot \| x \|, \forall c \in \mathbb{R} \)

- E.g., the \( L^p \) norm

\[
\| x \|_p = \left( \sum_i |x_i|^p \right)^{1/p}
\]

- \( L^2 \) (Euclidean) norm: \( \| x \| = (x^\top x)^{1/2} \)
- \( L^1 \) norm: \( \| x \|_1 = \sum_i |x_i| \)
- Max norm: \( \| x \|_\infty = \max_i |x_i| \)

- \( x^\top y = \| x \| \| y \| \cos \theta \), where \( \theta \) is the angle between \( x \) and \( y \)
  - \( x \) and \( y \) are **orthonormal** iff \( x^\top y = 0 \) (orthogonal) and \( \| x \| = \| y \| = 1 \) (unit vectors)
Matrix Norms

- Frobenius norm
  \[ \|A\|_F = \sqrt{\sum_{i,j} A_{i,j}^2} \]

- Analogous to the \( L^2 \) norm of a vector

- An **orthogonal matrix** is a square matrix whose column (resp. rows) are mutually orthonormal, i.e.,
  \[ A^\top A = I = AA^\top \]

- Implies \( A^{-1} = A^\top \)
Outline

1. Span & Linear Dependence
2. Norms
3. Eigendecomposition
4. Singular Value Decomposition
5. Traces and Determinant
Decomposition

- Integers can be decomposed into prime factors
  - E.g., $12 = 2 \times 2 \times 3$
  - Helps identify useful properties, e.g., 12 is not divisible by 5
Decomposition

- Integers can be decomposed into prime factors
  - E.g., $12 = 2 \times 2 \times 3$
  - Helps identify useful properties, e.g., 12 is not divisible by 5
- Can we decompose matrices to identify information about their functional properties more easily?
An **eigenvector** of a square matrix $A$ is a non-zero vector $v$ such that multiplication by $A$ alters only the scale of $v$:

$$Av = \lambda v,$$

where $\lambda \in \mathbb{R}$ is called the **eigenvalue** corresponding to this eigenvector.
An eigenvector of a square matrix $A$ is a non-zero vector $v$ such that multiplication by $A$ alters only the scale of $v$:

$$Av = \lambda v,$$

where $\lambda \in \mathbb{R}$ is called the eigenvalue corresponding to this eigenvector.

- If $v$ is an eigenvector, so is any its scaling $cv, c \in \mathbb{R}, c \neq 0$
  - $cv$ has the same eigenvalue
  - Thus, we usually look for unit eigenvectors
Eigendecomposition I

• Every **real symmetric** matrix \( A \in \mathbb{R}^{n \times n} \) can be decomposed into

\[
A = Q \text{diag}(\lambda) Q^\top
\]

• \( \lambda \in \mathbb{R}^n \) consists of real-valued eigenvalues (usually sorted in descending order)

• \( Q = [v^{(1)}, \ldots, v^{(n)}] \) is an orthogonal matrix whose columns are corresponding eigenvectors

Eigendecomposition may not be unique
When any two or more eigenvectors share the same eigenvalue
Then any set of orthogonal vectors lying in their span are also eigenvectors with that eigenvalue

What can we tell after decomposition?
Eigendecomposition I

- Every *real symmetric* matrix \( A \in \mathbb{R}^{n \times n} \) can be decomposed into
  \[
  A = Q \text{diag}(\lambda) Q^\top
  \]

- \( \lambda \in \mathbb{R}^n \) consists of real-valued eigenvalues (usually sorted in descending order)
- \( Q = [v^{(1)}, \ldots, v^{(n)}] \) is an orthogonal matrix whose columns are corresponding eigenvectors

- Eigendecomposition may not be unique
  - When any two or more eigenvectors share the same eigenvalue
  - Then any set of orthogonal vectors lying in their span are also eigenvectors with that eigenvalue
Eigendecomposition I

- Every *real symmetric* matrix $A \in \mathbb{R}^{n \times n}$ can be decomposed into

  $$A = Q \text{diag}(\lambda) Q^\top$$

- $\lambda \in \mathbb{R}^n$ consists of real-valued eigenvalues (usually sorted in descending order)
- $Q = [v^{(1)}, \ldots, v^{(n)}]$ is an orthogonal matrix whose columns are corresponding eigenvectors

- Eigendecomposition may not be unique
  - When any two or more eigenvectors share the same eigenvalue
  - Then any set of orthogonal vectors lying in their span are also eigenvectors with that eigenvalue

- What can we tell after decomposition?
Eigendecomposition II

Because \( Q = [\mathbf{v}^{(1)}, \cdots, \mathbf{v}^{(n)}] \) is an orthogonal matrix, we can think of \( A \) as scaling space by \( \lambda_i \) in direction \( \mathbf{v}^{(i)} \)
Rayleigh’s Quotient

Theorem (Rayleigh’s Quotient)
Given a symmetric matrix $A \in \mathbb{R}^{n \times n}$, then $\forall x \in \mathbb{R}^n$,

$$\lambda_{\min} \leq \frac{x^\top Ax}{x^\top x} \leq \lambda_{\max},$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of $A$.

- $\frac{x^\top Px}{x^\top x} = \lambda_i$ when $x$ is the corresponding eigenvector of $\lambda_i$. 
Suppose $A = Q \text{diag}(\lambda) Q^\top$, then $A^{-1} = Q \text{diag}(\lambda)^{-1} Q^\top$
Singularity

- Suppose $A = Q \text{diag}(\lambda) Q^\top$, then $A^{-1} = Q \text{diag}(\lambda)^{-1} Q^\top$
- $A$ is non-singular (invertible) iff none of the eigenvalues is zero
Positive Definite Matrices I

- A is **positive semidefinite** (denoted as $A \succeq O$) iff its eigenvalues are all non-negative
  - $x^\top Ax \geq 0$ for any $x$
- A is **positive definite** (denoted as $A \succ O$) iff its eigenvalues are all positive
  - Further ensures that $x^\top Ax = 0 \Rightarrow x = 0$

- Why these matter?
Positive Definite Matrices II

A function $f$ is **quadratic** iff it can be written as

$$f(x) = \frac{1}{2}x^\top Ax - b^\top x + c,$$

where $A$ is symmetric.
Positive Definite Matrices II

- A function $f$ is **quadratic** iff it can be written as
  
  $$f(x) = \frac{1}{2}x^\top Ax - b^\top x + c,$$
  
  where $A$ is symmetric

- $x^\top Ax$ is called the **quadratic form**
A function $f$ is **quadratic** iff it can be written as

$$f(x) = \frac{1}{2}x^\top Ax - b^\top x + c,$$

where $A$ is symmetric.

$x^\top Ax$ is called the **quadratic form**

**Figure:** Graph of a quadratic form when $A$ is a) positive definite; b) negative definite; c) positive semidefinite (singular); d) indefinite matrix.
Outline

1. Span & Linear Dependence
2. Norms
3. Eigendecomposition
4. Singular Value Decomposition
5. Traces and Determinant
Singular Value Decomposition (SVD)

- Eigendecomposition requires square matrices. What if $A$ is not square?

\[
A = UDV^T,
\]
where $U \in \mathbb{R}^{m \times m}$, $D \in \mathbb{R}^{m \times n}$, and $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices, and their columns are called the left- and right-singular vectors respectively.

Elements along the diagonal of $D$ are called the singular values.

Left-singular vectors of $A$ are eigenvectors of $AA^T$.

Right-singular vectors of $A$ are eigenvectors of $A^TA$.

Non-zero singular values of $A$ are square roots of eigenvalues of $AA^T$ (or $A^TA$).
Singular Value Decomposition (SVD)

- Eigendecomposition requires square matrices. What if $A$ is not square?
- Every real matrix $A \in \mathbb{R}^{m \times n}$ has a \textit{singular value decomposition}:

$$A = U D V^\top,$$

where $U \in \mathbb{R}^{m \times m}$, $D \in \mathbb{R}^{m \times n}$, and $V \in \mathbb{R}^{n \times n}$

- $U$ and $V$ are orthogonal matrices, and their columns are called the \textit{left-} and \textit{right-singular vectors} respectively.
- Elements along the diagonal of $D$ are called the \textit{singular values}.
Singular Value Decomposition (SVD)

- Eigendecomposition requires square matrices. What if \( A \) is not square?
- Every real matrix \( A \in \mathbb{R}^{m \times n} \) has a **singular value decomposition**:
  \[
  A = UDV^\top,
  \]
  where \( U \in \mathbb{R}^{m \times m}, D \in \mathbb{R}^{m \times n}, \) and \( V \in \mathbb{R}^{n \times n} \)
  - \( U \) and \( V \) are orthogonal matrices, and their columns are called the **left-** and **right-singular vectors** respectively
  - Elements along the diagonal of \( D \) are called the **singular values**

- Left-singular vectors of \( A \) are eigenvectors of \( AA^\top \)
- Right-singular vectors of \( A \) are eigenvectors of \( A^\top A \)
- Non-zero singular values of \( A \) are square roots of eigenvalues of \( AA^\top \) (or \( A^\top A \))
Moore-Penrose Pseudoinverse I

- Matrix inversion is not defined for matrices that are not square
Matrix inversion is not defined for matrices that are not square.

Suppose we want to make a left-inverse $B \in \mathbb{R}^{n \times m}$ of a matrix $A \in \mathbb{R}^{m \times n}$ so that we can solve a linear equation $Ax = y$ by left-multiplying each side to obtain $x = By$.
Moore-Penrose Pseudoinverse I

- Matrix inversion is not defined for matrices that are not square.
- Suppose we want to make a left-inverse $B \in \mathbb{R}^{n \times m}$ of a matrix $A \in \mathbb{R}^{m \times n}$ so that we can solve a linear equation $Ax = y$ by left-multiplying each side to obtain $x = By$.
  - If $m > n$, then it is possible to have no such $B$.
  - If $m < n$, then there could be multiple $B$'s.

Shan-Hung Wu (CS, NTHU)  Linear Algebra  Large-Scale ML, Fall 2016 20 / 26
Matrix inversion is not defined for matrices that are not square.

Suppose we want to make a left-inverse $B \in \mathbb{R}^{n \times m}$ of a matrix $A \in \mathbb{R}^{m \times n}$ so that we can solve a linear equation $Ax = y$ by left-multiplying each side to obtain $x = By$

- If $m > n$, then it is possible to have no such $B$
- If $m < n$, then there could be multiple $B$'s

By letting $B = A^\dagger$ the Moore-Penrose pseudoinverse, we can make headway in these cases:

- When $m = n$ and $A^{-1}$ exists, $A^\dagger$ degenerates to $A^{-1}$
Moore–Penrose Pseudoinverse I

- Matrix inversion is not defined for matrices that are not square
- Suppose we want to make a left-inverse $B \in \mathbb{R}^{n \times m}$ of a matrix $A \in \mathbb{R}^{m \times n}$ so that we can solve a linear equation $Ax = y$ by left-multiplying each side to obtain $x = By$
  - If $m > n$, then it is possible to have no such $B$
  - If $m < n$, then there could be multiple $B$’s
- By letting $B = A^\dagger$ the Moore–Penrose pseudoinverse, we can make headway in these cases:
  - When $m = n$ and $A^{-1}$ exists, $A^\dagger$ degenerates to $A^{-1}$
  - When $m > n$, $A^\dagger$ returns the $x$ for which $Ax$ is closest to $y$ in terms of Euclidean norm $\|Ax - y\|$
Moore-Penrose Pseudoinverse I

- Matrix inversion is not defined for matrices that are not square.
- Suppose we want to make a left-inverse $B \in \mathbb{R}^{n \times m}$ of a matrix $A \in \mathbb{R}^{m \times n}$ so that we can solve a linear equation $Ax = y$ by left-multiplying each side to obtain $x = By$.
  - If $m > n$, then it is possible to have no such $B$.
  - If $m < n$, then there could be multiple $B$'s.
- By letting $B = A^\dagger$, the **Moore-Penrose pseudoinverse**, we can make headway in these cases:
  - When $m = n$ and $A^{-1}$ exists, $A^\dagger$ degenerates to $A^{-1}$.
  - When $m > n$, $A^\dagger$ returns the $x$ for which $Ax$ is closest to $y$ in terms of Euclidean norm $\|Ax - y\|$.
  - When $m < n$, $A^\dagger$ returns the solution $x = A^\dagger y$ with minimal Euclidean norm $\|x\|$ among all possible solutions.
The Moore-Penrose pseudoinverse is defined as:

\[ A^\dagger = \lim_{\alpha \to 0} (A^\top A + \alpha I_n)^{-1} A^\top \]

\[ A^\dagger A = I \]
The Moore-Penrose pseudoinverse is defined as:

$$A^\dagger = \lim_{\alpha \to 0} (A^\top A + \alpha I_n)^{-1} A^\top$$

- $A^\dagger A = I$

In practice, it is computed by $A^\dagger = V D^\dagger U^\top$, where $UDV^\top = A$

- $D^\dagger \in \mathbb{R}^{n \times m}$ is obtained by taking the inverses of its non-zero elements then taking the transpose
Outline

1. Span & Linear Dependence
2. Norms
3. Eigendecomposition
4. Singular Value Decomposition
5. Traces and Determinant
Traces

- $\text{tr}(A) = \sum_i A_{i,i}$
Traces

- $\text{tr}(A) = \sum_i A_{i,i}$
- $\text{tr}(A) = \text{tr}(A^\top)$
- $\text{tr}(aA + bB) = a\text{tr}(A) + b\text{tr}(B)$
- $\|A\|_F^2 = \text{tr}(AA^\top) = \text{tr}(A^\top A)$
- $\text{tr}(ABC) = \text{tr}(BCA) = \text{tr}(CAB)$
  - Holds even if the products have different shapes
Determinant I

- Determinant $\det(\cdot)$ is a function that maps a square matrix $A \in \mathbb{R}^{n \times n}$ to a real value:

$$\det(A) = \sum_i (-1)^{i+1} A_{1,i} \det(A_{-1,-i}),$$

where $A_{-1,-i}$ is the $(n-1) \times (n-1)$ matrix obtained by deleting the $i$-th row and $j$-th column.
Determinant I

- Determinant $\det(\cdot)$ is a function that maps a square matrix $A \in \mathbb{R}^{n \times n}$ to a real value:

$$\det(A) = \sum_i (-1)^{i+1} A_{1,i} \det(A_{-1,-i}),$$

where $A_{-1,-i}$ is the $(n-1) \times (n-1)$ matrix obtained by deleting the $i$-th row and $j$-th column.

- $\det(A^\top) = \det(A)$
- $\det(A^{-1}) = 1/\det(A)$
- $\det(AB) = \det(A) \det(B)$
Determinant I

- Determinant $\det(\cdot)$ is a function that maps a square matrix $A \in \mathbb{R}^{n \times n}$ to a real value:

$$\det(A) = \sum_i (-1)^{i+1} A_{1,i} \det(A_{-1,-i}),$$

where $A_{-1,-i}$ is the $(n-1) \times (n-1)$ matrix obtained by deleting the $i$-th row and $j$-th column.

- $\det(A^\top) = \det(A)$
- $\det(A^{-1}) = 1 / \det(A)$
- $\det(AB) = \det(A) \det(B)$
- $\det(A) = \prod_i \lambda_i$
- What does it mean?
Determinant I

- Determinant \( \det(\cdot) \) is a function that maps a square matrix \( A \in \mathbb{R}^{n \times n} \) to a real value:

\[
\det(A) = \sum_i (-1)^{i+1} A_{1,i} \det(A_{-1,-i}),
\]

where \( A_{-1,-i} \) is the \( (n-1) \times (n-1) \) matrix obtained by deleting the \( i \)-th row and \( j \)-th column.

- \( \det(A^\top) = \det(A) \)
- \( \det(A^{-1}) = 1/\det(A) \)
- \( \det(AB) = \det(A) \det(B) \)
- \( \det(A) = \prod_i \lambda_i \)

- What does it mean? \( \det(A) \) can be also regarded as the \textit{signed area of the image of the “unit square”}. 

Shan-Hung Wu (CS, NTHU)  Linear Algebra  Large-Scale ML, Fall 2016  24 / 26
Determinant II

Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, we have $[1, 0]A = [a, b]$, $[0, 1]A = [c, d]$, and $\det(A) = ad - bc$

**Figure:** The area of the parallelogram is the absolute value of the determinant of the matrix formed by the images of the standard basis vectors representing the parallelogram’s sides.
The absolute value of the determinant can be thought of as a measure of how much multiplication by the matrix expands or contracts space.
The absolute value of the determinant can be thought of as a measure of how much multiplication by the matrix expands or contracts space.

If \( \det(A) = 0 \), then space is contracted completely along at least one dimension.

- \( A \) is invertible iff \( \det(A) \neq 0 \).
Determinant III

- The absolute value of the determinant can be thought of as a measure of how much multiplication by the matrix expands or contracts space.
- If \( \det(A) = 0 \), then space is contracted completely along at least one dimension.
  - \( A \) is invertible iff \( \det(A) \neq 0 \).
- If \( \det(A) = 1 \), then the transformation is volume-preserving.