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

Outline

Matrix Representation of Linear Functions

- A linear function (or map or transformation) $f : \mathbb{R}^n \to \mathbb{R}^m$ can be represented by a matrix $\boldsymbol{A}$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, such that

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

- $span(\boldsymbol{A}_{:,1}, \cdots, \boldsymbol{A}_{:,n})$ is called the ***column space*** of $\boldsymbol{A}$
- $rank(\boldsymbol{A}) = dim(span(\boldsymbol{A}_{:,1}, \cdots, \boldsymbol{A}_{:,n}))$

System of Linear Equations

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

Outline

Vector Norms

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

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

  - L^2(Euclidean) norm: $\|\boldsymbol{x}\| = (\boldsymbol{x}^\top \boldsymbol{x})^{1/2}$
  - L^1norm: $\|\boldsymbol{x}\|_1 = \sum_i |x_i|$
  - Max norm: $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$
- $\boldsymbol{x}^\top \boldsymbol{y} = \|\boldsymbol{x}\|\|\boldsymbol{y}\|\cos\theta$, where θ is the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$
  - $\boldsymbol{x}$ and $\boldsymbol{y}$ are ***orthonormal*** iff $\boldsymbol{x}^\top \boldsymbol{y} = 0$ (orthogonal) and $\|\boldsymbol{x}\| = \|\boldsymbol{y}\| = 1$ (unit vectors)

Matrix Norms

- Frobenius norm
$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}$$
  - 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.,
$$\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I} = \boldsymbol{A}\boldsymbol{A}^\top$$
  - Implies $\boldsymbol{A}^{-1} = \boldsymbol{A}^\top$

Outline

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?

Eigenvectors and Eigenvalues

- An ***eigenvector*** of a square matrix $\boldsymbol{A}$ is a non-zero vector $\boldsymbol{v}$ such that multiplication by $\boldsymbol{A}$ alters only the scale of $\boldsymbol{v}$:

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v},$$

  where $\lambda \in \mathbb{R}$ is called the ***eigenvalue*** corresponding to this eigenvector
- If $\boldsymbol{v}$ is an eigenvector, so is any its scaling $c\boldsymbol{v}, c \in \mathbb{R}, c \neq 0$
  - $c\boldsymbol{v}$ has the same eigenvalue
  - Thus, we usually look for unit eigenvectors

Eigendecomposition I

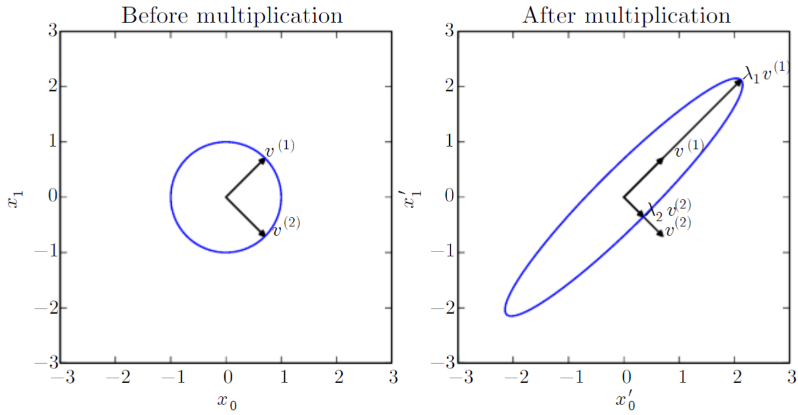
- Every ***real symmetric*** matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ can be decomposed into

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

  - $\boldsymbol{\lambda} \in \mathbb{R}^n$ consists of real-valued eigenvalues (usually sorted in descending order)
  - $\boldsymbol{Q} = [\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(n)}]$ 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 $\boldsymbol{Q} = [\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(n)}]$ is an orthogonal matrix, we can think of $\boldsymbol{A}$ as scaling space by λ_i in direction $\boldsymbol{v}^{(i)}$



Rayleigh's Quotient

Theorem (Rayleigh's Quotient)

Given a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$*, then* $\forall \boldsymbol{x} \in \mathbb{R}^n$*,*

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

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

- $\frac{\boldsymbol{x}^\top \boldsymbol{P} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \lambda_i$ when $\boldsymbol{x}$ is the corresponding eigenvector of λ_i

Singularity

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

Positive Definite Matrices I

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

Positive Definite Matrices II

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

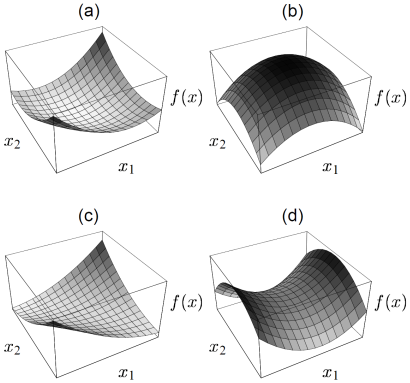


Figure: Graph of a quadratic form when $\boldsymbol{A}$ is a) positive definite; b) negative definite; c) positive semidefinite (singular); d) indefinite matrix.

Outline

Singular Value Decomposition (SVD)

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

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top,$$

  where $\boldsymbol{U} \in \mathbb{R}^{m \times m}$, $\boldsymbol{D} \in \mathbb{R}^{m \times n}$, and $\boldsymbol{V} \in \mathbb{R}^{n \times n}$
  - $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, and their columns are called the ***left-*** and ***right-singular vectors*** respectively
  - Elements along the diagonal of $\boldsymbol{D}$ are called the ***singular values***
- Left-singular vectors of $\boldsymbol{A}$ are eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$
- Right-singular vectors of $\boldsymbol{A}$ are eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$
- Non-zero singular values of $\boldsymbol{A}$ are square roots of eigenvalues of $\boldsymbol{A}\boldsymbol{A}^\top$ (or $\boldsymbol{A}^\top\boldsymbol{A}$)

Moore-Penrose Pseudoinverse I

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

Moore-Penrose Pseudoinverse II

- The Moore-Penrose pseudoinverse is defined as:

$$\boldsymbol{A}^{\dagger} = \lim_{\alpha \searrow 0} (\boldsymbol{A}^{\top}\boldsymbol{A} + \alpha \boldsymbol{I}_n)^{-1}\boldsymbol{A}^{\top}$$

  - $\boldsymbol{A}^{\dagger}\boldsymbol{A} = \boldsymbol{I}$
- In practice, it is computed by $\boldsymbol{A}^{\dagger} = \boldsymbol{V}\boldsymbol{D}^{\dagger}\boldsymbol{U}^{\top}$, where $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\top} = \boldsymbol{A}$
  - $\boldsymbol{D}^{\dagger} \in \mathbb{R}^{n \times m}$ is obtained by taking the inverses of its non-zero elements then taking the transpose

Outline

Traces

- $\mathrm{tr}(\boldsymbol{A}) = \sum_i \boldsymbol{A}_{i,i}$
- $\mathrm{tr}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}^\top)$
- $\mathrm{tr}(a\boldsymbol{A} + b\boldsymbol{B}) = a\mathrm{tr}(\boldsymbol{A}) + b\mathrm{tr}(\boldsymbol{B})$
- $\|\boldsymbol{A}\|_F^2 = \mathrm{tr}(\boldsymbol{A}\boldsymbol{A}^\top) = \mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{A})$
- $\mathrm{tr}(\boldsymbol{ABC}) = \mathrm{tr}(\boldsymbol{BCA}) = \mathrm{tr}(\boldsymbol{CAB})$
  - Holds even if the products have different shapes

Determinant I

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

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

  where $\boldsymbol{A}_{-1,-i}$ is the $(n-1) \times (n-1)$ matrix obtained by deleting the i-th row and j-th column
- $\det(\boldsymbol{A}^\top) = \det(\boldsymbol{A})$
- $\det(\boldsymbol{A}^{-1}) = 1/\det(\boldsymbol{A})$
- $\det(\boldsymbol{A}\boldsymbol{B}) = \det(\boldsymbol{A}) \det(\boldsymbol{B})$
- $\det(\boldsymbol{A}) = \prod_i \lambda_i$
- What does it mean? $\det(\boldsymbol{A})$ can be also regarded as the ***signed area of the image of the "unit square"***

Determinant II

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

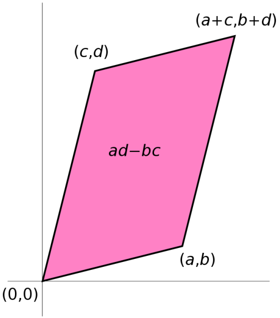


Figure: The area of the parallelogram is the absolute value of the determinant of the matrix formed by the images of the standard basis vectors representing the parallelogram's sides.

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 $\det(\boldsymbol{A}) = 0$, then space is contracted completely along at least one dimension
  - $\boldsymbol{A}$ is invertible iff $\det(\boldsymbol{A}) \neq 0$
- If $\det(\boldsymbol{A}) = 1$, then the transformation is volume-preserving