Linear Algebra

Author: Wenxiao Yang

Institute: Department of Mathematics, University of Illinois at Urbana-Champaign

Date: 2022

All models are wrong, but some are useful.

Contents

Chapter 1 Vector Space

1.1 Vector Space $(V, +, \times)$ (over a field $\mathbb{F}$)

A vector space over a field $\mathbb{F}$ is a set V w/ an operation addition $+ : V \times V \to V$ and an operation scalar multiplication $\mathbb{F} \times V \to V$

(1) Addition is associative $\&$ commutative

(2) $\exists 0 \in V$, additive identity: $0 + v = v \forall v \in V$

(3) $1v = v \forall v \in V$(where $1 \in \mathbb{F}$ is multi. id. in $\mathbb{F}$)

(4) $\forall \alpha, \beta \in \mathbb{F},\ v \in V,\ \alpha(\beta v) = (\alpha\beta)v$

(5) $\forall v \in V,\ (-1)v = -v$ we have $v + (-v) = 0$

(6) $\forall \alpha \in \mathbb{F},\ v, u \in V,\ \alpha(v + u) = \alpha v + \alpha u$

(7) $\forall \alpha, \beta \in \mathbb{F},\ v \in V,\ (\alpha + \beta)v = \alpha v + \beta v$

1.2 A field is a vector space over its subfield

Example 1.1 $\mathbb{K} \subset \mathbb{F}$ is a subfield of a field $\mathbb{F}$. Then $\mathbb{F}$ is a vector space over $\mathbb{K}$. (Since $\mathbb{F} \subset \mathbb{F}[x]$, then $\mathbb{F}[x]$ is a vector space over $\mathbb{F}$.)

1.3 Vector subspace

Suppose that V is a vector space over $\mathbb{F}$. A vector subspace or just subspace is a nonempty subset $W \subset V$ closed under addition and scalar multiplication. i.e. $v + w \in W,\ av \in W,\ \forall v, w \in W,\ a \in \mathbb{F}$.

Example 1.2 $\mathbb{K} \subset \mathbb{L} \subset \mathbb{F}$, then $\mathbb{L}$ is a subspace of $\mathbb{F}$ over $\mathbb{K}$.

1.4 Linear independent, Linear combination

1.5 span V, basis, dimension

A set of elements $v_1, ..., v_n \in V$ is said to **span** V if every vector $v \in V$ can be expressed as a linear combination of $v_1, ..., v_n$. If $v_1, ..., v_n$ spans and is linearly independent, then we call the set a ***basis*** for V.

Proposition 1.1 (Proposition 2.4.10.)

Suppose V is a vector space over a field $\mathbb{F}$ having a basis $\{v_1, ..., v_n\}$ with $n \geq 1$. ♠

(i) For all $v \in V$, $v = a_1 v_1 + ... + a_n v_n$ for exactly one $(a_1, ..., a_n) \in \mathbb{F}^n$.

(ii) If $w_1, ..., w_n$ span V , then they are linearly independent.

(iii)If $w_1, ..., w_n$ are linearly independent, then they span V.

If a vector space V over $\mathbb{F}$ has a basis with n vectors, then V is said to be n-dimensional (over $\mathbb{F}$) or is said to have **dimension** n.

1.6 Standard basis vectors

$$e_1 = (1, 0, ..., 0), e_2 = (0, 1, 0, ..., 0), ..., e_n = (0, 0, ..., 0, 1) \in \mathbb{F}^n$$

are a basis for $\mathbb{F}^n$ called the **standard basis vectors**.

1.7 Linear transformation

Given two vector spaces V and W over $\mathbb{F}$ a **linear transformation** is a function $T : V \to W$ such that for all $a \in \mathbb{F}$ and $v, w \in V$, we have

$$T(av) = aT(v) \; and \; T(v + w) = T(v) + T(w)$$

Proposition 1.2 (Proposition 2.4.15.)

If V and W are vector spaces and $v_1, ..., v_n$ is a basis for V then any function from $\{v_1, ..., v_n\} \to W$ extends uniquely to a linear transformation $V \to W$. ♠

Any $v \in V$, $\exists (a_1, ..., a_n)$ s.t. $v = a_1 v_1 + ... + a_n v_n$. Then $T(v) = T(a_1 v_1 + ... + a_n v_n) = a_1 T(v_1) + ... + a_n T(v_n)$

1.8 A Linear Transformation Correponds to a Matrix

Corollary 1.1 (Corollary 2.4.16.)

If $v_1, ..., v_n$ is a basis for a vector space V and $w_1, ..., w_n$ is a basis for a vector space W (both over $\mathbb{F}$), then any linear transformation $T : V \to W$ determines (and is determined by) the $m \times n$ matrix:

$$A = A(T) = \begin{bmatrix} A_{11} & A_{12} & \ldots & A_{1n} \\ A_{21} & A_{22} & \ldots & A_{2n} \\ \vdots & \vdots & \ldots & \vdots \\ A_{m1} & A_{m2} & \ldots & A_{mn} \end{bmatrix}$$

$$\begin{bmatrix} w_1 & \cdots & w_m \end{bmatrix}^T = A \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}^T$$

$\mathcal{L}(V, M)$ denotes the set of all linear transformations from V to W; $M_{m\times n}(\mathbb{F})$ the set of $m\times n$ matrix with entries in $\mathbb{F}$. $T \to A(T)$ defines a *bijection* $\mathcal{L}(V, M) \to M_{m\times n}(\mathbb{F})$. $A(T)$ **represents the linear transformation** T.

Proposition 1.3 (Proposition 2.4.19)

Suppose that V , W, and U are vector spaces over $\mathbb{F}$, with fixed chosen bases. If $T : V \to W$ and $S : W \to U$ are linear transformations represented by matrices $A = A(T)$ and $B = B(S)$, then $ST = S \circ T : V \to U$ is a linear transformation represented by the matrix $BA = B(S)A(T)$.

1.9 GL(V): invertible linear transformations $V \to V$

Given a vector space V over F, we let $GL(V) \subset \mathcal{L}(V, V)$ denote the subset of **invertible linear transformations**.

$$GL(V) = \{T \in \mathcal{L}(V, V) | T \text{ is a bijection}\} = \mathcal{L}(V, V) \cap Sym(V)$$

Chapter 2 Basic Definition

2.1 Square Matrix $A_{n\times n}$: $det(A)$, singular

1. A is singular if $det(A) = 0$, else non-singular.
2. If $det(A) \neq 0$, A^{-1} exists and $A^{-1} = \frac{adj(A)}{det(A)}$
3. $det(AB) = det(A)det(B)$

2.2 Orthogonal Vectors

Two vectors a and b are orthogonal, if their dot product is equal to zero (they are perpendicular).

$$a \cdot b = 0$$

2.3 Orthonormal Vectors

Two vectors a and b are orthonormal, if they are orthogonal **unit vectors**.

Chapter 3 Eigenvalues Related

3.1 Eigenvalues, Eigenvectors Definition

A vector x is an **eigenvector** of a matrix A if Ax is parallel to x, that is if $Ax = \lambda x$ for some number $\lambda \in \mathbb{R}$.

The number λ is called an **eigenvalue** of A.

i.e. the root of $(A - \lambda I_n)x = 0 \Leftrightarrow det(A - \lambda I_n) = 0$

3.2 Diagonalizable Matrix

A $n \times n$ matrix A with n linearly independent eigenvalues u is said to be *diagonalizable*.

$$\begin{aligned}
AU &= A\begin{bmatrix} | & | & \cdots & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & \cdots & | \end{bmatrix} \\
&= \begin{bmatrix} | & | & \cdots & | \\ \lambda_1 u_1 & \lambda_2 u_2 & \cdots & \lambda_n u_n \\ | & | & \cdots & | \end{bmatrix} \\
&= \begin{bmatrix} | & | & \cdots & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \\
&= UD \\
\Rightarrow A &= UDU^{-1}
\end{aligned}$$

Theorem 3.1

If an $n \times n$ matrix A has n linearly independent eigenvectors $u_1, ..., u_n$ corresponding to eigenvalues $\lambda_1, ..., \lambda_n$, then $A = UDU^{-1}$ where D is diagonal with entries $\lambda_1, ..., \lambda_n$, and U has columns $u_1, ..., u_n$. ♡

A is **similar** to D ($\exists P$ s.t. $A = PDP^{-1}$).

Not *diagonalizable* is also called *defective*.

Theorem 3.2

An $n \times n$ matrix with n distinct eigenvalues is diagonalizable. ♡

(Because the n associated eigenvectors are always linearly independent.)

3.3 Eigen Decomposition of Symmetric Matrices Results

Let A be a symmetric $n \times n$ matrix, i.e. $A^T = A$

Proposition 3.1

All eigenvalues of A are real. ♠

Proposition 3.2

Eigenvectors corresponding to distinct eigenvalues are orthogonal. ♠

Proof 3.1

Let λ_1, λ_2 be eigenvalues s.t. $\lambda_1 \neq \lambda_2$.

$$\begin{aligned} Au_1 = \lambda_1 u_1; \ Au_2 &= \lambda_2 u_2 \\ \lambda_1 u_1^T u_2 = (Au_1)^T u_2 &= u_1^T A^T u_2 \\ &= u_1^T A u_2 = u_1^T (Au_2) = \lambda_2 u_1^T u_2 \\ &\Rightarrow u_1^T u_2 = 0 \text{ Since } \lambda_1 \neq \lambda_2 \end{aligned}$$

Proposition 3.3

If λ is an eigenvalue with multiplicity k, we can find k orthogonal eigenvectors for λ. ♠

Multiplicity: the number of times an element is repeated in a multiset.

3.4 Diagonalization of Real Symmetric Matrices

A real symmetric matrix $A_{n\times n}$ can be written as

$$\begin{aligned} A &= \sum_{i=1}^{n} \lambda_i u_i u_i^T \\ &= U \Omega U^T \end{aligned}$$

u_i are orthonormal eigenvectors. λ_i are eigenvalues.

Where $U = [u_1, u_2, ..., u_n]$, $\Omega = diag(\lambda_1, ..., \lambda_n)$

Since u_i are orthonormal eigenvectors, $U^T U = I \Rightarrow U^T = U^{-1}$. U is an orthogonal matrix.

3.4.1 Proposition: $\lambda_{\min}\|x\|^2 \le x^T Ax \le \lambda_{\max}\|x\|^2$

Proposition 3.4

For any $x \in \mathbb{R}^n$,

$$\lambda_{\min}\|x\|^2 \le x^T Ax \le \lambda_{\max}\|x\|^2$$

Proof 3.2

Since u_i are orthonormal and linearly independent. $x = \sum_{i=1}^n \alpha_i u_i$ *for some* $\alpha_i \in \mathbb{R}, i = 1, ..., n$

$$\begin{aligned} x^T Ax &= (\sum_{i=1}^n \alpha_i u_i)^T A(\sum_{j=1}^n \alpha_j u_j) \\ &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j u_i^T A u_j \\ &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j u_i^T (A u_j) \\ &= \sum_{i=1}^n \alpha_i^2 \lambda_i \end{aligned}$$

$$\Rightarrow \lambda_{\min}\|x\|^2 \le x^T Ax \le \lambda_{\max}\|x\|^2$$

The first equation holds if x is the eigenvector for $\lambda_{\min}$. The second equation holds if x is the eigenvector for $\lambda_{\max}$.

3.4.2 Proposition: λ^2 is the eigenvalue of A^2 and $A^T A$

Proposition 3.5

If λ is an eigenvalue of A, then λ^2 is the eigenvalue of A^2 and $A^T A$, the corresponding eigenvector doesn't change.

Proof 3.3

$$\begin{aligned} Ax_1 &= \lambda x_1 \\ A^2 x_1 = A(Ax_1) = \lambda A x_1 &= \lambda^2 x_1 \\ A^T A x_1 = A^T (Ax_1) = \lambda A^T x_1 &= \lambda^2 x_1 \end{aligned}$$

3.5 Trace

$A_{n\times n}$, $Tr(A) = \sum_{i=1}^n A_{kk}$

$$det(A) = \prod_{i=1}^n \lambda_i,\ Tr(A) = \sum_{i=1}^n \lambda_i$$

Proposition 3.6 (Invariance Property)

$A_{m\times n}$, $B_{n\times k}$, $C_{k\times m}$, $Tr(ABC) = Tr(CAB) = Tr(BCA)$. ♠

3.6 Jacobian matrix

Suppose $\mathbf{f} : \mathbf{R}^n \to \mathbf{R}^m$ is a function such that each of its first-order partial derivatives exist on $\mathbf{R}^n$. This function takes a point $\mathbf{x} \in \mathbf{R}^n$ as input and produces the vector $\mathbf{f}(\mathbf{x}) \in \mathbf{R}^m$ as output. Then the Jacobian matrix of $\mathbf{f}$ is defined to be an $m \times n$ matrix, denoted by $\mathbf{J}$, whose (i, j) th entry is $\mathbf{J}_{ij} = \frac{\partial f_i}{\partial x_j}$, or explicitly

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^{\mathrm{T}} f_1 \\ \vdots \\ \nabla^{\mathrm{T}} f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

where $\nabla^{\mathrm{T}} f_i$ is the transpose (row vector) of the gradient of the i component.

3.7 Hessian matrix

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a function taking as input a vector $\mathbf{x} \in \mathbb{R}^n$ and outputting a scalar $f(\mathbf{x}) \in \mathbb{R}$. If all second partial derivatives of f exist and are continuous over the domain of the function, then the Hessian matrix $\mathbf{H}$ of f is a square $n \times n$ matrix, usually defined and arranged as follows:

$$\mathbf{H}_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix},$$

or, by stating an equation for the coefficients using indices i and j,

$(\mathbf{H}_f)_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$.

The Hessian matrix is a symmetric matrix, since the hypothesis of continuity of the second derivatives implies that the order of differentiation does not matter (Schwarz's theorem).

The determinant of the Hessian matrix is called the Hessian determinant.

3.8 Positive Definite Matrices

3.8.1 Definition

We say that a symmetric $n \times n$ matrix A is:

(1). **positive semidefinite (PSD)** (written $A \succeq 0$) if $x^T Ax \geq 0$ for all x.

(2). **positive definite (PD)** (written $A \succ 0$) if $x^T Ax > 0$ for all $x \neq 0$.

(3). **negative semidefinite (NSD)** (written $A \preceq 0$) if $x^T Ax \leq 0$ for all x.

(4). **negative definite (ND)** (written $A \prec 0$) if $x^T Ax < 0$ for all $x \neq 0$.

(5). **indefinite** (not written in any particular way) if none of the above apply.

$x^T Ax$ is a function of x called the quadratic form associated to A.

A is ND(NSD) $\Leftrightarrow$ $-A$ is PD(PSD)

Note: $A^T A$ is **positive semidefinite**, since $x^T A^T Ax = \|Ax\|^2 \geq 0$.

Note: We can extend definition to non-symmetric $n \times n$

$$x^T Ax = x^T A^T x \Rightarrow x^T Ax = x^T (\frac{A + A^T}{2})x$$

3.8.2 Condition number (for PD matrix)

Condition number (for PD matrix):

$$\kappa(A) = \frac{\lambda_{max}}{\lambda_{min}} > 0$$

3.8.3 Diagonal matrix situation

$$D = \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_n \end{bmatrix}$$

Lemma 3.1

If $d_1, \dots d_n$ *are all nonnegative, then* $D \succeq 0$;

If $d_1, \dots d_n$ *are all positive, then* $D \succ 0$;

If $d_1, \dots d_n$ *are all nonpositive, then* $D \preceq 0$;

If $d_1, \dots d_n$ *are all negative, then* $D \prec 0$; ♡

3.8.4 Using eigenvalues

If A is an $n \times n$ symmetric matrix, then it can be factored as

$$A = Q^T \Lambda Q = Q^T \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \lambda_n \end{bmatrix} Q$$

where $\lambda_1, ..., \lambda_n$ are the eigenvalues of A and the columns of Q are the corresponding eigenvectors.

We can get $x^T Ax = x^T Q^T \Lambda Q x = (Qx)^T \Lambda (Qx)$

If we substitute $y = Qx$:

$x^T Ax = y^T \Lambda y = \lambda_1 y_1^2 + \lambda_2 y_2^2 + ... + \lambda_n y_n^2$

Theorem 3.3

If $\lambda_1, ...\lambda_n$ are all non-negative, then symmetric matrix $A \succeq 0$;

If $\lambda_1, ...\lambda_n$ are all positive, then $A \succ 0$;

If $\lambda_1, ...\lambda_n$ are all non-positive, then $A \preceq 0$;

If $\lambda_1, ...\lambda_n$ are all negative, then $A \prec 0$;

if it has both positive and negative eigenvalues, then A is indefinite ♡

3.8.5 Sylvester's Criterion

Consider a $n \times n$ matrix A:

$$A = \begin{bmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{bmatrix}$$

Denote its $k \times k$ submatrix $A^{(k)}$:

$$A^{(k)} = \begin{bmatrix} a_{11} & a_{12} & ... & a_{1k} \\ a_{21} & a_{22} & ... & a_{2k} \\ ... & ... & ... & ... \\ a_{k1} & a_{k2} & ... & a_{kk} \end{bmatrix}$$

Let $\Delta_k = det(A^{(k)})$

$$det(A - xI) = (\lambda_1 - x)(\lambda_2 - x)...(\lambda_n - x)$$

by setting $x = 0$ we get $det(A) = \lambda_1\lambda_2...\lambda_n$.

When $A \succ 0$, all the eigenvalues are positive, so $det(A) > 0$ as well.

$A \succ 0 \Rightarrow \vec{x}^T A\vec{x} > 0$ for all $\vec{x} \neq \vec{0}$. Consider $\vec{x} \in \mathbb{R}^n$ with $x_{k+1} = \cdots = x_n = 0$. $\vec{x} = [x_1, x_2, ..., x_k, 0, ...0]^T$.

Then,

$$\vec{x}^T A\vec{x} = [x_1, x_2, ..., x_k, 0, ...0] \begin{bmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \vec{x}^T A^{(k)}\vec{x}$$

Then we know $A \succ 0 \Rightarrow A^{(k)} \succ 0$

We expect $A^{(k)} \succ 0 \Rightarrow \Delta_k > 0$ for all k.

Theorem 3.4 (Sylvester's criterion)

Let A be $n \times n$ symmetric matrix

1. $A \succ 0$ *iff* $\Delta_i > 0\ \forall i = 1, ..., n$
2. $A \prec 0$ *iff* $(-1)^i\Delta_i > 0\ \forall i = 1, ..., n$
3. *A is indefinite if the first $\Delta_k$that breaks each pattern respectively is the wrong sign (rather than 0).* ♡

Proposition 3.7

1. *Symmetric matrix A is PD*

   $\Leftrightarrow$ *All eigenvalues of A are* > 0

   $\Leftrightarrow \Delta_i > 0\ \forall i = 1, ..., n$
2. *Symmetric matrix A is PSD*

   $\Leftrightarrow$ *All eigenvalues of A are* ≥ 0

   $\Leftrightarrow \Delta_i \geq 0\ \forall i = 1, ..., n$
3. *For ND and NSD, test $-A$ instead of A* ♠

3.9 Matrix Norm (Induced Norm) and Spectral Radius

$\|A\| = \max_{\|x\|=1} \|Ax\|$. i.e., find the column with the highest sum of absolute values.

Spectral Radius: for $n \times n$ matrix A,

$$S(A) = \max_{i=1,...,n} |\lambda_i|$$

Proposition 3.8

$S(A) \leq \|A\|$

Proof 3.4

$$\|A\| = \max_{\|x\|=1} \|Ax\| \geq \|Au\| = |\lambda|\|u\| = |\lambda|$$

Proposition 3.9

For symmetric $A_{n\times n}$, $S(A) = \|A\|$

Proof 3.5

$\forall x \in \mathbb{R}^n$, *decompose it by* u_i. *Since* u_i *are orthonormal and linearly independent.* $x = \sum_{i=1}^n \alpha_i u_i$ *for some* $\alpha_i \in \mathbb{R}, i = 1, ..., n$. $\|x\|^2 = \sum_{i=1}^n |\alpha_i|^2$

$$\begin{aligned} \|Ax\|^2 = \|\sum_{i=1}^n \alpha_i A u_i\|^2 = \|\sum_{i=1}^n \alpha_i \lambda_i u_i\|^2 = \sum_{i=1}^n |\alpha_i|^2 |\lambda_i|^2 \\ \leq \sum_{i=1}^n |\alpha_i|^2 S(A)^2 = S(A)^2 \Rightarrow \|A\| \leq S(A) \end{aligned}$$

Since we proved $S(A) \leq \|A\|$ *before,* $S(A) = \|A\|$.

Chapter 4 Euclidean geometry basics

4.1 Norm

4.1.1 Vector's Norm

Vector $x \in \mathbb{R}^n$-n-dim Euclidean space

$$x = (x_1, \ldots, x_n) \equiv \begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix}^\top = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Norm of x, $\|x\|$ satisfies properties:

(a) $\|x\| \geqslant 0$

(b) $\|x\| = 0 \Leftrightarrow x = 0$

(c) $\|cx\| = |c|\|x\|$, for $c \in \mathbb{R}$

(d) $\|x + y\| \leqslant \|x\| + \|y\| \longleftarrow$ Triangle Ineq.

Enclidean Norm (default $\rho = 2$): $\|x\| = \sqrt{x^\top x} = \sqrt{\sum_{i=1}^n x_i{}^2}$

Other norms:

1. l_1-norm : $\|x\|_1 = \sum_{i=1}^n |x_i|$
2. l_ρ-norm : $\|x\|_\rho = \sqrt[\rho]{\sum_{i=1}^n |x_i|^\rho}$
3. Supremum norm or l_∞-norm : $\|x\|_\infty = \max_i |x_i|$

4.1.2 Matrix's Norm

$A \in \mathbb{R}^{n \times m}$ is a matrix

$\|Ax\| \leqslant \|A\|\|x\|, \|AB\| \leqslant \|A\|\|B\|$

Default is $\rho = 1$: $\|A\| = \max_{\|x\|=1} \|Ax\|$.

$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ (Frobenius norm); Frobenius norm property: $\|A\|_F^2 =< A, A >= trace(A^T A)$

$\|A\|_1 = \max_j \sum_{i=1}^n |A_{ij}|$ i.e., find the column with the highest sum of absolute values.

$\|A\|_\infty = \max_j \sum_{j=1}^n |A_{ij}|$ i.e., find the row with the highest sum of absolute values.

$\|A\|_2 = \sqrt{\lambda_{\max}\left(A^T A\right)} = \sigma_{\max}(A)$. $\|A\|_2 = \max_k \sigma_k . \sigma_k$ is the singular value (square root of $A^T A$) of A (spectral norm, Euclidean norm)

$\|A\| = \max\left(\frac{\|Ax\|}{\|x\|}\right) \Rightarrow \|A\| \geqslant \frac{\|Ax\|}{\|x\|}$, $\|Ax\| \leqslant \|A\|\|x\|$

4.1.3 Difference between Spectral Radius and Spectral Norm

Spectral Radius: $S(A) = \max_{i=1,\ldots,n} |\lambda_i|$; Spectral Norm: $\|A\|_2 = \max_k \sigma_k$

For real symmetric matrices, $\|A\|_2 = S(A)$.

For general matrices, $\|A\|_2 \geq S(A)$.

4.2 Euclidean distance, inner product

Euclidean distance on $\mathbb{R}^n$:

$$|x-y| = \sqrt{(x_1-y_1)^2 + \ldots + (x_n-y_n)^2}$$

Euclidean inner product:

$$x \cdot y = x_1 y_1 + \cdots + x_n y_n = x^T y$$

Also written as $< x, y >$

Useful fact:

$$< x, y >= \cos(\theta)\|x\|_2\|y\|_2$$

θ is the angle between x and y.

Two important results for Euchidean norm:

1) Pythagorean Theorem: If $x^\top y = 0$,

$$\|x+y\|^2 = \|x\|^2 + \|y\|^2$$

2) Cauchy - Schwarz Inequality:

$$< x, y >= \left|x^\top y\right| \leqslant \|x\|_2\|y\|_2$$

$$" = " \text{ iff } x = \alpha y \text{ for some } \alpha \in \mathbb{R}$$

4.3 General Inner Products

4.3.1 Inner Product

Definition 4.1

An **inner product** $*$ is a function that maps two vectors $\vec{x}, \vec{y} \in \mathbb{R}^n$ to a single value $\vec{x} * \vec{y} \in \mathbb{R}$, satisfying the following axioms:

1. **Bilinear** (linearity in both arguments): for all $\vec{x}, \vec{y}, \vec{z} \in \mathbb{R}^n$ and $a, b \in \mathbb{R}$, we have

$$(a\vec{x} + b\vec{y}) * \vec{z} = a(\vec{x} * \vec{z}) + b(\vec{x}\vec{y} * \vec{z})$$

$$\vec{x} * (a\vec{y} + b\vec{z}) = a(\vec{x} * \vec{y}) + b(\vec{x} * \vec{z})$$

2. **Symmetric** i.e. for all $\vec{x}, \vec{y} \in \mathbb{R}^n$,

$$\vec{x} * \vec{y} = \vec{y} * \vec{x}$$

3. **Positivity** i.e. for all $\vec{x} \in \mathbb{R}^n$,

$$\vec{x} * \vec{x} \geq 0$$

   with equality if and only if $\vec{x} = \vec{0}$

Definition 4.2

Every inner product $*$ defines a corresponding norm $\| \cdot \|_*$ as $\|\vec{x}\|_* = \sqrt{\vec{x} * \vec{x}}$

4.3.2 Theorem: $*$ is inner product iff $\vec{x} * \vec{y} = \vec{x}^T H \vec{y}$ for some symmetric H

Theorem 4.1

An operation $* : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is an inner product if and only if it can be written as $\vec{x} * \vec{y} = \vec{x}^T H \vec{y}$ for some symmetric positive definite $n \times n$ matrix H.

Proof 4.1

Define $H := [H_{ij}] = [\vec{e}_i * \vec{e}_j]$

1. *Bilinear:*

$$\begin{aligned}
\vec{x} * \vec{y} &= \left(\sum_{i=1}^{n} x_i \vec{e}_i\right) * \left(\sum_{j=1}^{n} y_j \vec{e}_j\right) \\
&= \sum_{i=1}^{n} x_i \left(\vec{e}_i * \sum_{j=1}^{n} y_j \vec{e}_j\right) \quad \textit{(by linearity)} \\
&= \sum_{i=1}^{n} x_i \left(\sum_{j=1}^{n} y_j (\vec{e}_i * \vec{e}_j)\right) \quad \textit{(by linearity)} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} x_i \left(\vec{e}_i * \vec{e}_j\right) y_j \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} x_i H_{ij} y_j = \vec{x}^T H \vec{y}
\end{aligned}$$

2. *Symmetric* $\Leftrightarrow H = H^T$*:*

$$\vec{x} * \vec{y} = \vec{x}^T H \vec{y} = \left(\vec{x}^T H \vec{y}\right)^T = \vec{y}^T H^T \vec{x} = \vec{y}^T H \vec{x} = \vec{y} * \vec{x}$$

3. *Positivity* $\Leftrightarrow H \succ 0$*:* $\vec{x}^T H \vec{x} \geq 0$ *with equality only if* $\vec{x} = 0$

As we know that a symmetric matrix H is positive definite if and only if we can write $H = B^T B$ for some invertible matrix B.

$$\vec{x} * \vec{y} = \vec{x}^T H \vec{y} = \vec{x}^T B^T B \vec{y} = (B\vec{x})^T B \vec{y} = (B\vec{x}) \cdot (B\vec{y})$$

Definition 4.3

Given a positive definite matrix H*, let the associated inner product be* $\vec{x} \cdot_H \vec{y} = \vec{x}^T H \vec{y}$ *and the associated norm be* $\|\vec{x}\|_H = \sqrt{\vec{x}^T H \vec{x}}$ ♣

4.4 Isometry

An **isometry** of $\mathbb{R}^n$ is a bijection $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ that preserves distance, which means,

$$|\Phi(x) - \Phi(y)| = |x - y|, \; \forall x, y \in \mathbb{R}^n$$

We use $Isom(\mathbb{R}^n)$ denotes the set of all isometries of $\mathbb{R}^n$,

$$Isom(\mathbb{R}^n) = \{\Phi : \mathbb{R}^n \to \mathbb{R}^n | |\Phi(x) - \Phi(y)| = |x - y|, \; \forall x, y \in \mathbb{R}^n\}$$

Proposition 4.1

$\Phi, \Psi \in Isom(\mathbb{R}^n)$*, then* $\Phi \circ \Psi, \Phi^{-1} \in Isom(\mathbb{R}^n)$ ♠

Proof 4.2

Since Φ, Ψ are bijections, so is $\Phi \circ \Psi$. Moreover,

$$|\Phi \circ \Psi(x) - \Phi \circ \Psi(y)| = |\Phi(\Psi(x)) - \Phi(\Psi(y))| = |\Psi(x) - \Psi(y)| = |x - y|$$

Since $id \in Isom(\mathbb{R}^n)$,

$$|x - y| = |id(x) - id(y)| = |\Phi \circ \Phi^{-1}(x) - \Phi \circ \Phi^{-1}(y)| = |\Phi^{-1}(x) - \Phi^{-1}(y)|$$

4.5 Linear isometries i.e. orthogonal group

There is a matrix $A \in GL(n, \mathbb{R})$ i.e. a *invertible linear transofrmations* $T_A : \mathbb{R}^n \to \mathbb{R}^n$ is given by $T_A(v) = Av$.

$$T_A(v) \cdot T_A(w) = (Av) \cdot (Aw) = (Av)^t(Aw) = v^tA^tAw$$

$$A^tA = I \Leftrightarrow T_A(v) \cdot T_A(w) = v \cdot \Leftrightarrow T_A \in Isom(\mathbb{R}^n)$$

We define the all isometries in *invertible linear transofrmations* $\mathbb{R}^n \to \mathbb{R}^n$ as **orthogonal group**

$$O(n) = \{A \in GL(n, \mathbb{R}) | A^tA = I\} \subset GL(n, \mathbb{R})$$

4.6 Special orthogonal group

$O(n)$ are the matrices representing linear isometries of $\mathbb{R}^n$. $1 = det(I) = det(A^tA) = det(A^t)det(A) = det(A)^2 \Rightarrow det(A) = 1$ or $det(A) = -1$. We use **special orthogonal group** represents A with $det(A) = 1$,

$$SO(n) = \{A \in O(n) | det(A) = 1\}$$

4.7 translation

Define a *translation* by $v \in \mathbb{R}^n$,

$$\tau_v : \mathbb{R}^n \to \mathbb{R}^n, \; \tau_v(x) = x + v$$

Note *[Exercise 2.5.3] $\forall v \in \mathbb{R}^n, \tau_v$ is an isometry.*

Proof 4.3

$|\tau_v(x) - \tau_v(y)| = |(x + v) - (y + v)| = |x - y|$

4.8 All isometries can be represented by a composition of *a translation* and *an orthogonal transformation*

Since *the composition of isometries is an isometry,* $\forall A \in O(n)$ and $v \in \mathbb{R}^n$, the composition

$$\Phi_{A,v}(x) = \tau_v(T_A(x)) = Ax + v$$

is an isometry. **which could account for all isometries**.

Theorem 4.2

$Isom(\mathbb{R}^n) = \{\Phi_{A,v} | A \in O(n), v \in \mathbb{R}^n\}$ ♡

Chapter 5 Algebra Computation

5.1 Hessian Matrix

Definition 5.1

The Hessian of f at point x is an $n \times n$ symmetric matrix denoted by $\nabla^2 f(x)$ with $[\nabla^2 f(x)]_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ ♣

5.2 Taylor's Expansion

Definition 5.2 (Taylor's Expansion of Vector)

$$f(y) - f(x) = \nabla f(x)^T(y-x) + \frac{1}{2}(x-y)^T \nabla^2 f(x)(x-y) + o(\|x-y\|^2)$$

♣

5.3 Random Vectors and Random Matrices

5.3.1 Mean

Definition 5.3 (Mean of a random vector)

The mean of a $d-$dimensional random vector $\vec{x}$ is

$$\mathbb{E}(\vec{x}) = \begin{pmatrix} \mathbb{E}(x_1) \\ \mathbb{E}(x_2) \\ \cdots \\ \mathbb{E}(x_d) \end{pmatrix}$$

♣

Definition 5.4 (Mean of a random matrix)

The mean of a $d_1 \times d_2$ matrix with random entries $\vec{X}$ is

$$\mathbb{E}(\vec{X}) = \begin{pmatrix} \mathbb{E}(\vec{X}[1,1]) & \mathbb{E}(\vec{X}[1,2]) & \cdots & \mathbb{E}(\vec{X}[1,d_2]) \\ \mathbb{E}(\vec{X}[2,1]) & \mathbb{E}(\vec{X}[2,2]) & \cdots & \mathbb{E}(\vec{X}[2,d_2]) \\ \cdots & \cdots & \cdots & \cdots \\ \mathbb{E}(\vec{X}[d_1,1]) & \mathbb{E}(\vec{X}[d_1,2]) & \cdots & \mathbb{E}(\vec{X}[d_1,d_2]) \end{pmatrix}$$

♣

Lemma 5.1 (Linearity of expectation for random vectors and matrices)

Let $\vec{x}$ be a $d-$dimensional random vector and $b \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m\times d}$, then

$$\mathbb{E}(A\vec{x}+b) = A\mathbb{E}(\vec{x})+b$$

Similarly let, $\vec{X}$ be a $d_1 \times d_2$ random matrix and $B \in \mathbb{R}^{m\times d_2}$ and $A \in \mathbb{R}^{m\times d_1}$, then

$$\mathbb{E}(A\vec{X}+B) = A\mathbb{E}(\vec{X})+B$$

♡

Definition 5.5 (Sample mean of multivariate data)

Let $X := \{x_1, x_2, ..., x_n\}$ denote a set of $d-$dimensional vectors of real-valued data. The sample mean is the entry-wise average

$$\mu_X := \frac{\sum_{i=1}^n x_i}{n}$$

♣

5.3.2 Variance, Covariance

Definition 5.6 (Covariance matrix of a vector)

The covariance matrix of a $d-$dimensional random vector $\vec{x}$ is the $d \times d$ matrix

$$\Sigma_{\vec{x}} = \mathbb{E}[(\vec{x}-\mathbb{E}(\vec{x}))^T(\vec{x}-\mathbb{E}(\vec{x}))] = \begin{bmatrix} \text{Var}(\vec{x}[1]) & \cdots & \text{Cov}(\vec{x}[1],\vec{x}[d]) \\ \cdots & \cdots & \cdots \\ \text{Cov}(\vec{x}[d],\vec{x}[1]) & \cdots & \text{Var}(\vec{x}[d]) \end{bmatrix}$$

♣

Lemma 5.2

For any random vector $\vec{x}$ with covariance matrix $\Sigma_{\vec{x}}$, and any vector v:

$$\text{Var}(v^T\vec{x}) = v^T\Sigma_{\vec{x}}v$$

♡

Definition 5.7 (Sample covariance matrix)

Let $X := \{x_1, x_2, ..., x_n\}$ denote a set of $d-$dimensional vectors of real-valued data. The sample covariance matrix equals

♣

Variance-Covariance matrix Σ:

$$\Sigma_{m\times m} = Cov(\mathbf{Z}) = \mathbb{E}((\mathbf{Z}-\mu)(\mathbf{Z}-\mu)^T) = \begin{bmatrix} Var(Z_1) & \cdots & Cov(Z_1,Z_m) \\ \cdots & \cdots & \cdots \\ Cov(Z_m,Z_1) & \cdots & Var(Z_m) \end{bmatrix}$$

Affine Transformation

(1)

$$\mathbf{W} = \mathbf{a}_{n\times 1} + \mathbf{B}_{n\times m}\mathbf{Z}_{m\times 1}$$

$$\mathbb{E}(\mathbf{W}) = \mathbf{a} + \mathbf{B}\mu,\ Cov(\mathbf{W}) = \mathbf{B}\Sigma\mathbf{B}^T$$

(2)

$$\mathbf{W} = \mathbf{v}^T\mathbf{Z} = v_1 Z_1 + ... + v_m Z_m$$

$$\mathbb{E}(\mathbf{W}) = \mathbf{v}^T\mu = \sum_{i=1}^{m} v_i\mu_i$$

$$Var(\mathbf{W}) = \mathbf{v}^T\Sigma\mathbf{v} = \sum_{i=1}^{m} v_i^2 Var(Z_i) + 2\sum_{i<j} v_i v_j Cov(Z_i, Z_j)$$

$$\text{i.e. } \mathbb{E}(\mathbf{AZ}) = \mathbf{A}\mathbb{E}(Z);\ Var(\mathbf{AZ}) = \mathbf{A}Var(\mathbf{Z})\mathbf{A}^T$$

(3)

$$Cov(\mathbf{AX}, \mathbf{BY}) = \mathbb{E}[(\mathbf{AX}-\mathbf{A}\mathbb{E}(X))(\mathbf{BY}-\mathbf{B}\mathbb{E}(Y))^T] = \mathbf{A}\mathbb{E}[(\mathbf{X}-\mathbb{E}(X))(\mathbf{Y}-\mathbb{E}(Y))^T]\mathbf{B}^T = \mathbf{A}Cov(\mathbf{X}, \mathbf{Y})\mathbf{B}^T$$

5.4 Matrix Multiplication

(1). $A(BC) = (AB)C$.

(2). $A(B + C) = AB + AC$.

(2). $(B + C)A = BA + CA$.

(3). No commutative: $AB \neq BA$.

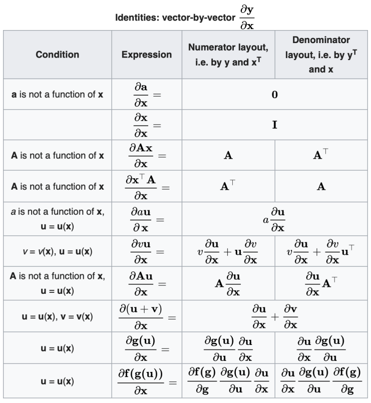
5.5 Matrix Derivation

$$\frac{\partial x^T Q x}{\partial x} = 2Qx$$

https://zhuanlan.zhihu.com/p/24709748

https://blog.csdn.net/daaikuaichuan/article/details/80620518

Vector by vector:



Identities: vector-by-vector $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$

Condition	Expression	Numerator layout, i.e. by y and x^{T}	Denominator layout, i.e. by y^{T} and x
a is not a function of **x**	$\frac{\partial \mathbf{a}}{\partial \mathbf{x}} =$	$\mathbf{0}$	
	$\frac{\partial \mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{I}$	
A is not a function of **x**	$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{A}$	$\mathbf{A}^{\top}$
A is not a function of **x**	$\frac{\partial \mathbf{x}^{\top}\mathbf{A}}{\partial \mathbf{x}} =$	$\mathbf{A}^{\top}$	$\mathbf{A}$
a is not a function of **x**, **u** = **u**(**x**)	$\frac{\partial a\mathbf{u}}{\partial \mathbf{x}} =$	$a\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	
v = *v*(**x**), **u** = **u**(**x**)	$\frac{\partial v\mathbf{u}}{\partial \mathbf{x}} =$	$v\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \mathbf{u}\frac{\partial v}{\partial \mathbf{x}}$	$v\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial v}{\partial \mathbf{x}}\mathbf{u}^{\top}$
A is not a function of **x**, **u** = **u**(**x**)	$\frac{\partial \mathbf{A}\mathbf{u}}{\partial \mathbf{x}} =$	$\mathbf{A}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\mathbf{A}^{\top}$
u = **u**(**x**), **v** = **v**(**x**)	$\frac{\partial (\mathbf{u}+\mathbf{v})}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$	
u = **u**(**x**)	$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}$
u = **u**(**x**)	$\frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{u}))}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{f}(\mathbf{g})}{\partial \mathbf{g}}$

Figure 5.1: Denominator layout means $x \in \mathbb{R}^{n\times 1}$

$$\begin{aligned}
\frac{\partial u}{\partial x^T} &= (\frac{\partial u^T}{\partial x})^T \\
\frac{\partial u^T v}{\partial x} &= \frac{\partial u^T}{\partial x} v + \frac{\partial v^T}{\partial x} u^T \\
\frac{\partial u v^T}{\partial x} &= \frac{\partial u}{\partial x} v^T + u \frac{\partial v^T}{\partial x} \\
\frac{\partial x^T x}{\partial x} &= 2x \\
\frac{\partial x^T A x}{\partial x} &= (A + A^T)x
\end{aligned}$$

where $x, u, v \in \mathbb{R}^{n\times 1}$

Note:

$$\frac{d\|Aw-b\|^2}{dw} = \frac{d(Aw-b)^T(Aw-b)}{dw} = \frac{d(Aw-b)^T}{dw}(Aw-b) + \frac{d(Aw-b)^T}{dw}(Aw-b) = 2A^T(Aw-b)$$

Matrix by vector:

$$\frac{\partial AB}{\partial x} = \frac{\partial A}{\partial x} B + A \frac{\partial B}{\partial x}$$

Matrix by matrix:

$$\begin{aligned}
\frac{\partial u^T X v}{\partial X} &= uv^T \\
\frac{\partial u^T X^T X u}{\partial X} &= 2Xuu^T \\
\frac{\partial [(Xu-v)^T(Xu-v)]}{\partial X} &= 2(Xu-v)u^T
\end{aligned}$$

Trace:

$$tr(a) = a$$

$$tr(AB) = tr(BA)$$

$$tr(ABC) = tr(CAB) = tr(BCA)$$

$$\frac{\partial tr(AB)}{\partial A} = B^T$$

$$tr(A) = tr(A^T)$$

$$\frac{\partial tr(ABA^TC)}{\partial A} = CAB + C^TAB^T$$

5.6 Matrix Inversion

5.6.1 Woodbury matrix identity

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1}VA^{-1},$$

where A, U, C and V are conformable matrices: A is $n \times n$, C is $k \times k$, U is $n \times k$, and V is $k \times n$. This can be derived using blockwise matrix inversion.

Blockwise inversion:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \\ -\left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} & \left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \end{bmatrix}$$

5.7 Linear Regression: Least Square

Minimize$_w \mathcal{R}(w) = \|Xw - y\|^2$

5.7.1 Normal Equations

$$\nabla_w \|Xw - y\|^2 = 2X^T(Xw - y) = 0$$

$$\Rightarrow X^TXw = X^Ty$$

These are called the **normal equations**.

Proposition 5.1

$\hat{w}$ satisfies $\mathcal{R}(\hat{w}) = \min_w \mathcal{R}(w)$ if and only if $\hat{w}$ satisfies the normal equations. (i.e. prove its is the global minimum) ♠

Proof 5.1

Consider $\boldsymbol{w}$ with $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{y}$, and any $\boldsymbol{w}'$; then

$$\begin{aligned}\left\|\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{y}\right\|^2 &= \left\|\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{X}\boldsymbol{w} + \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\right\|^2 \\ &= \left\|\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{X}\boldsymbol{w}\right\|^2 + 2\left(\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{X}\boldsymbol{w}\right)^\top (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) + \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2\end{aligned}$$

Since

$$\left(\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{X}\boldsymbol{w}\right)^\top (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) = \left(\boldsymbol{w}' - \boldsymbol{w}\right)^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^\top \boldsymbol{y}\right) = 0$$

then

$$\left\|\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{y}\right\|^2 = \left\|\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{X}\boldsymbol{w}\right\|^2 + \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 \geq \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$$

5.8 LU Decomposition (Restricted to Square)

Triangular matrix saves time when computing $Ax = b$.

Let A be a square matrix. An LU factorization refers to the factorization of A, with proper row and/or column orderings or permutations, into two factors - a lower triangular matrix L and an upper triangular matrix U: $A = LU$. In the lower triangular matrix all elements above the diagonal are zero, in the upper triangular matrix, all the elements below the diagonal are zero. For example, for a 3×3 matrix A, its LU decomposition looks like this:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}.$$

$$A = PLU$$

P is a permutation matrix (used to swap row, only one 1 in every row). P is orthogonal, so $P^{-1} = P^T$.

Solve $Ax = b$:

$$Ax = b$$

$$PLUx = b$$

$$\text{Let } y = Ux\text{, then solve PLy=b}$$

$$Ly = P^T b$$

Complexity: $O(n^3)$

5.9 SVD: Singular Value Decomposition

For a $n \times m$ matrix A with rank r,

$$\begin{aligned} A_{n\times m} =& U_{n\times n}\Sigma_{n\times m}V^T_{m\times m} \\ =& \sum_{i=1}^{r} s_i u_i v_i^T \\ =& \begin{bmatrix} | & | & \cdots & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} s_1 & & & & & & \\ & s_2 & & & & & \\ & & \ddots & & & & \\ & & & s_r & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix} \begin{bmatrix} | & | & \cdots & | \\ v_1 & v_2 & \cdots & v_m \\ | & | & \cdots & | \end{bmatrix}^T \end{aligned}$$

U, V are orthogonal matrices. $u_i \in \mathbb{R}^{n\times 1}$ are left singular vectors, $v_i \in \mathbb{R}^{m\times 1}$ are right singular vectors. $s_i, i = 1, ..., r$ are singular values (absolute values of eigenvalues of a normal matrix).

Complexity:$O(mn^2 + n^3)$

5.9.1 Pseudo-inverse

We can't compute the inverse matrix of a singular matrix. We can use pseudo-inverse matrix.

$$A^+_{m\times n} = \sum_{i=1}^{r} \frac{1}{s_i} v_i u_i^T = V\Sigma^+ U^T$$

Where

$$\Sigma^+ = \begin{bmatrix} \frac{1}{s_1} & & & & & & \\ & \frac{1}{s_2} & & & & & \\ & & \ddots & & & & \\ & & & \frac{1}{s_r} & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}$$

The SVD may not be unique, but the pseudo-inverse of A, A^+ is unique.

$$AA^+ = \sum_{i=1}^{r} u_i u_i^T = \begin{bmatrix} I_{r\times r} & O_{r\times n-r} \\ O_{n-r\times r} & O_{n-r\times n-r} \end{bmatrix}_{n\times n}$$

$$A^+A = \sum_{i=1}^{r} v_i v_i^T = \begin{bmatrix} I_{r\times r} & O_{r\times m-r} \\ O_{m-r\times r} & O_{m-r\times m-r} \end{bmatrix}_{m\times m}$$

If A^{-1} exists, $A^{-1} = A^+$.

5.9.2 Analysis of A^TA and AA^T

$$\begin{aligned} A^TA &= (U\Sigma V^T)^T(U\Sigma V^T) \\ &= V\Sigma^T U^T U \Sigma V^T \\ &= V\Sigma^T\Sigma V^T \\ &= V\Sigma^2 V^T \\ \Rightarrow V &= A^T U \Sigma^+ \end{aligned}$$

Columns of V are the eigenvectors of A^TA.

The diagonal entries of Σ^2, $s_1^2, s_2^2, ..., s_r^2$ are the eigenvalues of A^TA.

Similarly:

$$AA^T = U\Sigma^2U^T$$

$$\Rightarrow U = AV\Sigma^+$$

Columns of U are the eigenvectors of AA^T.

Fact: A^TA is positive semi-definite.

5.9.3 Solve Normal Equations

Solve $X^T X w = X^T y$,

$$\hat{w}_{ols} = X^+ y$$

$$X^T X \hat{w}_{ols} = X^T X X^+ y = (X^T (X X^+)) y = X^T y$$

5.9.4 Low-Rank Approximation

For a $n \times m$ matrix A with rank r, $A = \sum_{i=1}^{r} s_i u_i v_u^T$.

Rank-k approximation for A is

$$A_k = \sum_{i=1}^{k} s_i u_i v_u^T$$

Where $s_1 \geq s_2 \geq \cdots \geq 0$

Bibliography

[1] MATH 417: Christopher J Leininger Introduction to Abstract Algebra (Draft) 2017.

[2] MATH 484

[3] ECE 490

[4] STAT 425

[5] CS/MATH 357