Lecture 1
Basics of Linear Algebra

- Matrix Operations
- Matrix Derivative and Expectations
- Applications and Wrap-Up

Contents

Motivation

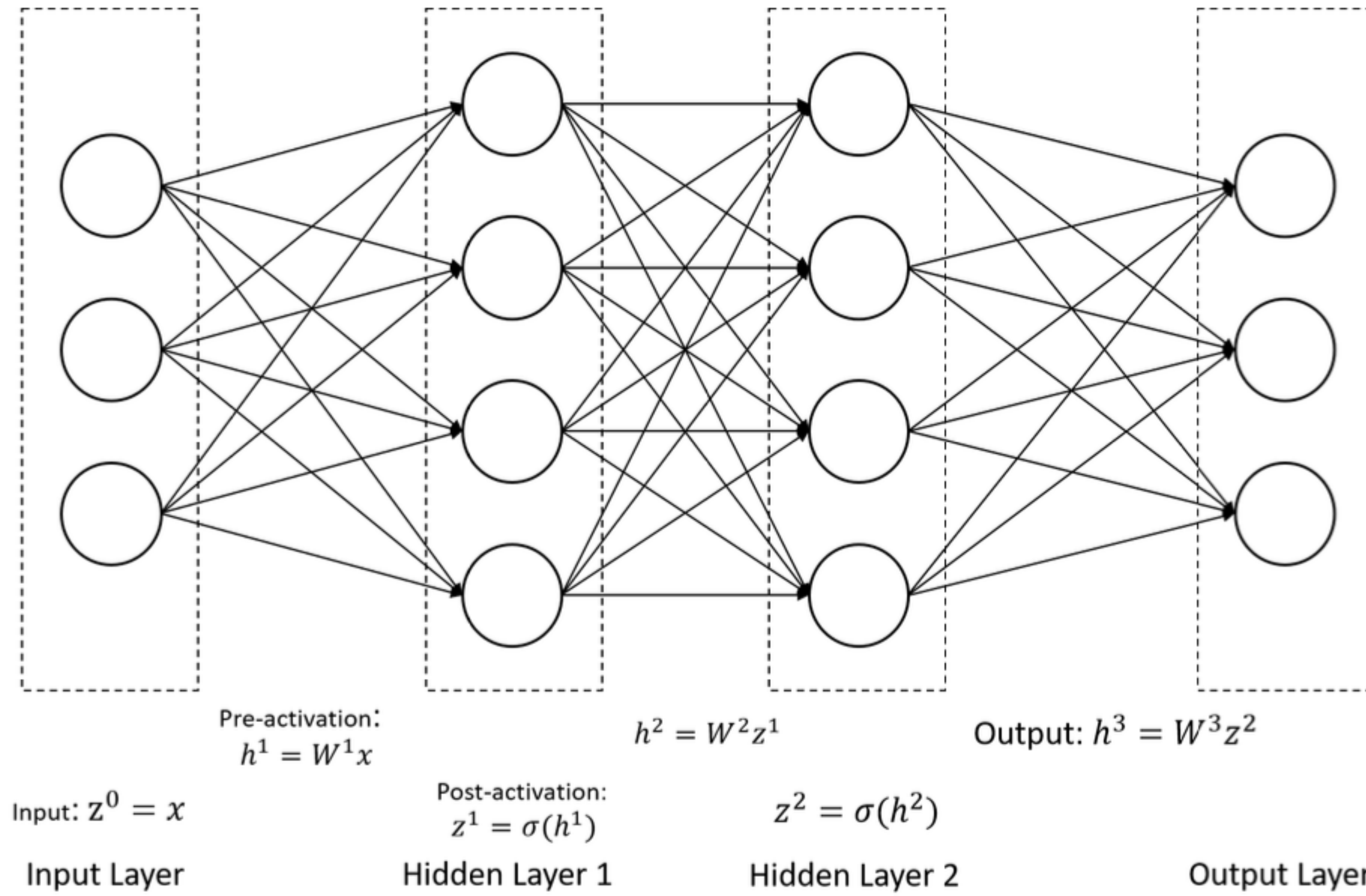


Figure: Example of a 3-layer fully-connected neural network. You should be able to understand its matrix representation.

What is a Matrix?

Let $A = (a_{ij})$ be an $m \times n$ matrix.

A= np. array([[1, 2], [3,4]])

- The jth column of A is denoted by a column vector $\mathbf{a}_j$, i.e.,

A= $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$

$$\mathbf{a}_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix}$$

- The ith row of A is denoted by a row vector $\vec{\mathbf{a}}_i$, i.e.,

$$\vec{\mathbf{a}}_i = (a_{i1}, a_{i2}, \ldots, a_{in})$$

- Matrix A can be represented in terms of either its columns and rows:

$$A = [\mathbf{a}_1, \cdots, \mathbf{a}_n] = \begin{bmatrix} \vec{\mathbf{a}}_1 \\ \vec{\mathbf{a}}_2 \\ \vdots \\ \vec{\mathbf{a}}_m \end{bmatrix}$$

Matrix-Vector Multiplication

For an $m \times n$ matrix A with the ith column $\mathbf{a}_i$, and a vector $\mathbf{u} = (u_1, u_2, \ldots, u_n)^\top$, the multiplication of A and $\mathbf{u}$ is defined as

$$A\mathbf{u} = u_1\mathbf{a}_1 + u_2\mathbf{a}_2 + \cdots + u_n\mathbf{a}_n$$

Example

$$\underbrace{\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \end{bmatrix}}_{2\times 4} \underbrace{\begin{bmatrix} 6 \\ -7 \\ 8 \\ -9 \end{bmatrix}}_{4\times 1} = 6\begin{bmatrix} 1 \\ 2 \end{bmatrix} - 7\begin{bmatrix} 2 \\ 3 \end{bmatrix} + 8\begin{bmatrix} 3 \\ 4 \end{bmatrix} - 9\begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

Inner Product

- Given a vector $\mathbf{a} = (a_1, \ldots, a_n)^\top$ and a vector $\mathbf{b} = (b_1, \ldots, b_n)^\top$, following the rule of matrix-vector product, we have

$$\mathbf{a}^\top \mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots a_n b_n$$

- We call this special vector-vector multiplication the **inner product** (scalar product) of $\mathbf{a}$ and $\mathbf{b}$ (denoted by $\mathbf{a}^\top \mathbf{b}$ or $\langle \mathbf{a}, \mathbf{b} \rangle$)
- Properties: Commutative, bilinear
- Application: Cosine similarity, $\cos\theta = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$

Inner Product

- Given a vector $\mathbf{a} = (a_1, \ldots, a_n)^\top$ and a vector $\mathbf{b} = (b_1, \ldots, b_n)^\top$, following the rule of matrix-vector product, we have

$$\mathbf{a}^\top \mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots a_n b_n$$

- We call this special vector-vector multiplication the **inner product** (scalar product) of $\mathbf{a}$ and $\mathbf{b}$ (denoted by $\mathbf{a}^\top \mathbf{b}$ or $\langle \mathbf{a}, \mathbf{b} \rangle$)
- Properties: Commutative, bilinear
- Application: Cosine similarity, $\cos\theta = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$

Row Perspective of Multiplication

The matrix-vector multiplication $A\mathbf{u}$ has a row formula as

$$A\mathbf{u} = \begin{bmatrix} \vec{\mathbf{a}}_1\mathbf{u} \\ \vec{\mathbf{a}}_2\mathbf{u} \\ \vdots \\ \vec{\mathbf{a}}_m\mathbf{u} \end{bmatrix}$$

- Consider $A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} 6 & -7 & 8 & -9 \end{bmatrix}^\top$.
- We calculate

$$\vec{\mathbf{a}}_1\mathbf{u} = 6 \cdot 1 - 7 \cdot 2 + 8 \cdot 3 - 9 \cdot 4 = -20$$

$$\vec{\mathbf{a}}_2\mathbf{u} = 6 \cdot 2 - 7 \cdot 3 + 8 \cdot 4 - 9 \cdot 5 = -22$$

- We see that $A\mathbf{u} = \begin{bmatrix} -20 & -22 \end{bmatrix}^\top$

Row Perspective of Multiplication

The matrix-vector multiplication $A\mathbf{u}$ has a row formula as

$$A\mathbf{u} = \begin{bmatrix} \vec{\mathbf{a}}_1\mathbf{u} \\ \vec{\mathbf{a}}_2\mathbf{u} \\ \vdots \\ \vec{\mathbf{a}}_m\mathbf{u} \end{bmatrix}$$

- Consider $A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} 6 & -7 & 8 & -9 \end{bmatrix}^\top$.
- We calculate

$$\vec{\mathbf{a}}_1\mathbf{u} = 6 \cdot 1 - 7 \cdot 2 + 8 \cdot 3 - 9 \cdot 4 = -20$$

$$\vec{\mathbf{a}}_2\mathbf{u} = 6 \cdot 2 - 7 \cdot 3 + 8 \cdot 4 - 9 \cdot 5 = -22$$

- We see that $A\mathbf{u} = \begin{bmatrix} -20 & -22 \end{bmatrix}^\top$

Row Perspective of Multiplication

The matrix-vector multiplication $A\mathbf{u}$ has a row formula as

$$A\mathbf{u} = \begin{bmatrix} \vec{\mathbf{a}}_1\mathbf{u} \\ \vec{\mathbf{a}}_2\mathbf{u} \\ \vdots \\ \vec{\mathbf{a}}_m\mathbf{u} \end{bmatrix}$$

- Consider $A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} 6 & -7 & 8 & -9 \end{bmatrix}^\top$.
- We calculate

$$\vec{\mathbf{a}}_1\mathbf{u} = 6 \cdot 1 - 7 \cdot 2 + 8 \cdot 3 - 9 \cdot 4 = -20$$

$$\vec{\mathbf{a}}_2\mathbf{u} = 6 \cdot 2 - 7 \cdot 3 + 8 \cdot 4 - 9 \cdot 5 = -22$$

- We see that $A\mathbf{u} = \begin{bmatrix} -20 & -22 \end{bmatrix}^\top$

Linear Systems as Matrix Equations

Write the following linear systems into compact matrix form:

$$\begin{cases} 2x_1 + x_2 + x_3 = 5 \\ 4x_1 - 6x_2 = -2 \\ -2x_1 + 7x_2 + 2x_3 = 9 \end{cases} \Rightarrow A\mathbf{x} = \mathbf{b}$$

where

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 5 \\ -2 \\ 9 \end{bmatrix}$$

Rank of a Matrix

- The rank of a matrix A is the number of linearly independent columns
- Equivalently, it is the number of linearly independent rows
- Example: $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ has rank 1
- Full rank: rank$(A) = \min(m, n)$ for $A \in \mathbb{R}^{m \times n}$
- Application: Determines solvability of linear systems $A\mathbf{x} = \mathbf{b}$

Rank of a Matrix

- The rank of a matrix A is the number of linearly independent columns
- Equivalently, it is the number of linearly independent rows
- Example: $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ has rank 1
- Full rank: $\text{rank}(A) = \min(m, n)$ for $A \in \mathbb{R}^{m \times n}$
- Application: Determines solvability of linear systems $A\mathbf{x} = \mathbf{b}$

Identity Matrix

- The identity matrix of order k, denoted by I or I_k, is a $k \times k$ square matrix whose diagonal elements are 1's and whose nondiagonal elements are 0's

$$I = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

- Properties: $AI = A$ for any compatible matrix A.

Identity Matrix

- The identity matrix of order k, denoted by I or I_k, is a $k \times k$ square matrix whose diagonal elements are 1's and whose nondiagonal elements are 0's

$$I = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

- Properties: $AI = A$ for any compatible matrix A.

Inverse of a Matrix

- Let A be a $k \times k$ matrix. The inverse of A, denoted by A^{-1}, is another $k \times k$ matrix such that

$$AA^{-1} = A^{-1}A = I$$

- If the inverse exists, it is unique
- Existence: A^{-1} exists if and only if $\det(A) \neq 0$ (or equivalently, rank$(A) = k$)
- For 2×2 matrix:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Inverse of a Matrix

- Let A be a $k \times k$ matrix. The inverse of A, denoted by A^{-1}, is another $k \times k$ matrix such that

$$AA^{-1} = A^{-1}A = I$$

$BA = AB = I$

Assume B, C $CA = AC = I$

$BAC = (BA)C = C$
$= B(AC) = B$

- If the inverse exists, it is unique
- Existence: A^{-1} exists if and only if $\det(A) \neq 0$ (or equivalently, $\text{rank}(A) = k$)
- For 2×2 matrix:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Inverse of a Matrix

- Let A be a $k \times k$ matrix. The inverse of A, denoted by A^{-1}, is another $k \times k$ matrix such that

$$AA^{-1} = A^{-1}A = I$$

- If the inverse exists, it is unique
- Existence: A^{-1} exists if and only if $\det(A) \neq 0$ (or equivalently, rank$(A) = k$)
- For 2×2 matrix:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Inverse of a Matrix

- Let A be a $k \times k$ matrix. The inverse of A, denoted by A^{-1}, is another $k \times k$ matrix such that

$$AA^{-1} = A^{-1}A = I$$

- If the inverse exists, it is unique
- Existence: A^{-1} exists if and only if $\det(A) \neq 0$ (or equivalently, rank$(A) = k$)
- For 2×2 matrix:

$\det(A) = ad - bc$

$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Transpose of a Matrix

- Let A be an $n \times k$ matrix. The transpose of A, denoted by $A^\top$, is a $k \times n$ matrix whose columns are the rows of A

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nk} \end{bmatrix} \Rightarrow A^\top = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{1k} & a_{2k} & \cdots & a_{nk} \end{bmatrix}$$

- Properties: $(A^\top)^\top = A$, $(AB)^\top = B^\top A^\top$

Transpose of a Matrix

- Let A be an $n \times k$ matrix. The transpose of A, denoted by $A^\top$, is a $k \times n$ matrix whose columns are the rows of A

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nk} \end{bmatrix} \Rightarrow A^\top = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{1k} & a_{2k} & \cdots & a_{nk} \end{bmatrix}$$

- Properties: $(A^\top)^\top = A$, $(AB)^\top = B^\top A^\top$

$A: n \times k \quad B: k \times m$

$$((AB)^T)_{i,j} = (AB)_{j,i} = \sum_e a_{j,e} b_{e,i}$$
$$= \sum_e b_{e,i} a_{j,e} = (B^T A^T)_{j,i}$$

Symmetric Matrices

- Let A be a $k \times k$ matrix. A is said to be symmetric if

$$A = A^\top$$

- Examples: Covariance matrices, Hessian matrices
- Properties: Real eigenvalues, orthogonal eigenvectors
- Spectral theorem: $A = Q\Lambda Q^\top$ where Q is orthogonal and Λ is diagonal

Symmetric Matrices

- Let A be a $k \times k$ matrix. A is said to be symmetric if

$$A = A^\top$$

$a = n \times 1$

$A = aa^T$

$A^T = (aa^T)^T$
$= (a^T)^T a^T$
$= aa^T = A$

- Examples: Covariance matrices, Hessian matrices
- Properties: Real eigenvalues, orthogonal eigenvectors
- Spectral theorem: $A = Q\Lambda Q^\top$ where Q is orthogonal and Λ is diagonal

Symmetric Matrices

- Let A be a $k \times k$ matrix. A is said to be symmetric if

$$A = A^\top$$

- Examples: Covariance matrices, Hessian matrices
- Properties: Real eigenvalues, orthogonal eigenvectors
- Spectral theorem: $A = Q\Lambda Q^\top$ where Q is orthogonal and Λ is diagonal

Symmetric Matrices

- Let A be a $k \times k$ matrix. A is said to be symmetric if

$$A = A^\top$$

- Examples: Covariance matrices, Hessian matrices
- Properties: Real eigenvalues, orthogonal eigenvectors
- Spectral theorem: $A = Q\Lambda Q^\top$ where Q is orthogonal and Λ is diagonal

Idempotent Matrices

- Let A be a $k \times k$ matrix. A is called idempotent if

$$A = AA$$

- If A is also symmetric, then A is called symmetric idempotent
- If A is symmetric idempotent, then $I - A$ is also symmetric idempotent
- Example: Projection matrices $P = X(X^\top X)^{-1} X^\top$

Idempotent Matrices

- Let A be a $k \times k$ matrix. A is called idempotent if

$$A = AA$$

- If A is also symmetric, then A is called symmetric idempotent
- If A is symmetric idempotent, then $I - A$ is also symmetric idempotent
- Example: Projection matrices $P = X(X^\top X)^{-1}X^\top$

Idempotent Matrices

$(I-A)^T = I^T - A^T$

$= I - A$

- Let A be a $k \times k$ matrix. A is called idempotent if

$$A = AA$$

- If A is also symmetric, then A is called symmetric idempotent
- If A is symmetric idempotent, then $I - A$ is also symmetric idempotent

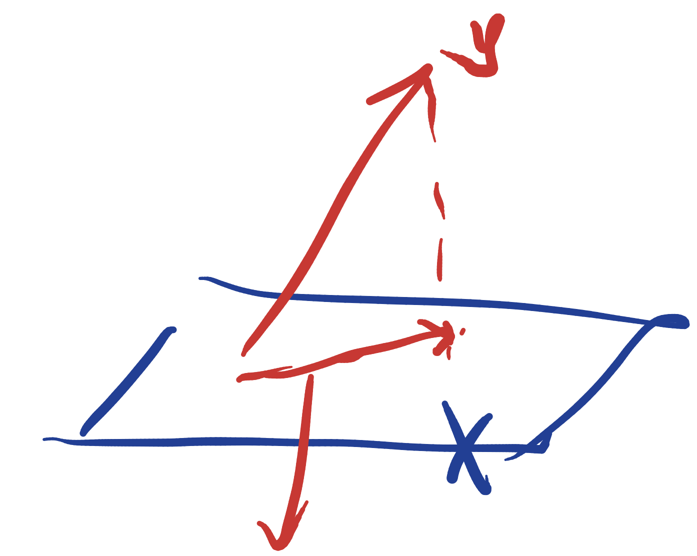
$(I-A)(I-A) = I(I-A) - A(I-A) = (I-A) + (-A + AA)$

$= (I-A) + (-A+A)$

$= I - A$

- Example: Projection matrices $P = X(X^\top X)^{-1}X^\top$

Idempotent Matrices



Py

- Let A be a $k \times k$ matrix. A is called idempotent if

$$A = AA$$

$PPy = Py, \ \forall y$

$PP = P.$

- If A is also symmetric, then A is called symmetric idempotent
- If A is symmetric idempotent, then $I - A$ is also symmetric idempotent
- Example: Projection matrices $P = X(X^\top X)^{-1}X^\top$

$P^T = P$

$$PP = (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T)$$

$$= X(X^TX)^{-1}\cancel{(X^TX)(X^TX)^{-1}}X^T = X(X^TX)^{-1}X^T = P$$

- $Ax = b$ what if this system have no solution ?

$$\min_x \|Ax-b\|_2^2 = \sum_{i=1}^{m} (a_i^T x - b_i)^2 \rightarrow F(x)$$

$$\frac{\partial F(x)}{\partial x} = 2A^T(Ax-b) = 0$$

$\Rightarrow$ $A^TAx = A^Tb$ (normal equation)

$x^* = (A^TA)^{-1}A^Tb$ (Assume A^TA inv.)

$$Ax^* \approx b$$

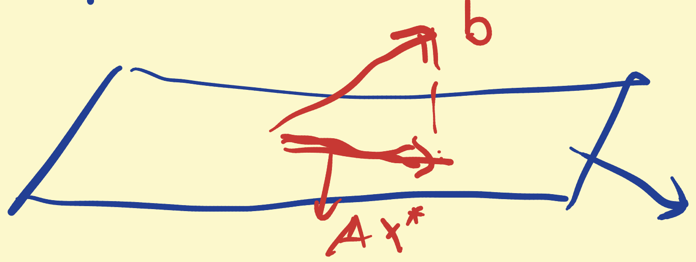
$C(A) = \text{span}\{a_1, \cdots, a_n\}$

$A = [a_1, \cdots, a_n]$

$$Ax^* = \underbrace{A(A^TA)^{-1}A^T}_{\text{projection matrix}} b$$

① $Ax^* = b$: $b \in$ column space of A $\Rightarrow$ $\|Ax^* - b\|_2^2 = 0$

② otherwise,



C(A)

$$\arg\min_{z \in C(A)} \|z - b\|_2^2$$

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

① $b \in C(A) \iff \exists\, x$ s.t. $Ax = b$

$$Pb = A(A^TA)^{-1}A^Tb$$

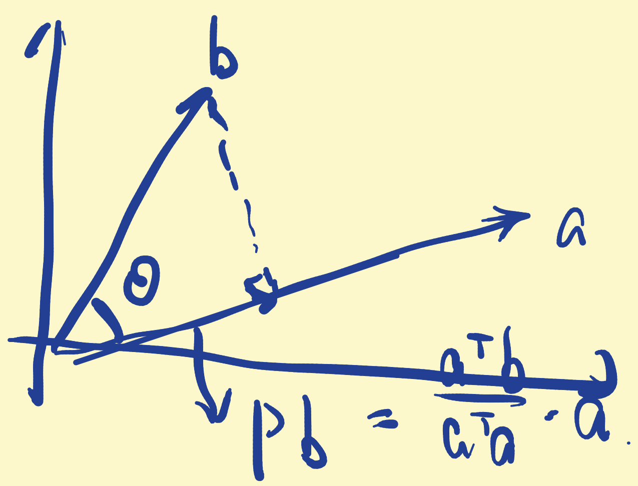
$$= A(A^TA)^{-1}A^TAx$$

$$= Ax = b$$

② $A = [a] \in R^{m\times 1}$

$$P = A(A^TA)^{-1}A^T = a(a^Ta)^{-1}a^T = \frac{aa^T}{a^Ta}$$

$$Pb = \frac{aa^Tb}{a^Ta} = \frac{\langle a, b\rangle}{\|a\|_2^2}\cdot a = \frac{\|a\|\|b\|\cos\theta}{\|a\|_2^2}\cdot a = \frac{a}{\|a\|}\cdot \|b\|\cos\theta$$



Orthonormal Matrices

- Let A be a $k \times k$ matrix. If A is an orthonormal matrix, then

$$A^\top A = I$$

- As a consequence, if A is an orthonormal matrix, then

$$A^{-1} = A^\top$$

- Properties: Preserves norms and angles ($\|A\mathbf{x}\| = \|\mathbf{x}\|$)
- Examples: Rotation matrices, permutation matrices

Orthonormal Matrices

- Let A be a $k \times k$ matrix. If A is an orthonormal matrix, then

$$A^\top A = I$$

- As a consequence, if A is an orthonormal matrix, then

$$A^{-1} = A^\top$$

- Properties: Preserves norms and angles ($\|A\mathbf{x}\| = \|\mathbf{x}\|$)
- Examples: Rotation matrices, permutation matrices

Orthonormal Matrices

- Let A be a $k \times k$ matrix. If A is an orthonormal matrix, then

$$A^\top A = I$$

- As a consequence, if A is an orthonormal matrix, then

$$A^{-1} = A^\top$$

$$\|Ax\|_2^2 = \langle Ax, Ax \rangle$$
$$= x^T A^T A x$$
$$= x^T x$$
$$= \|x\|_2^2$$

- Properties: Preserves norms and angles ($\|A\mathbf{x}\| = \|\mathbf{x}\|$)
- Examples: Rotation matrices, permutation matrices

Orthonormal Matrices

- Let A be a $k \times k$ matrix. If A is an orthonormal matrix, then

$$A^\top A = I$$

- As a consequence, if A is an orthonormal matrix, then

$$A^{-1} = A^\top$$

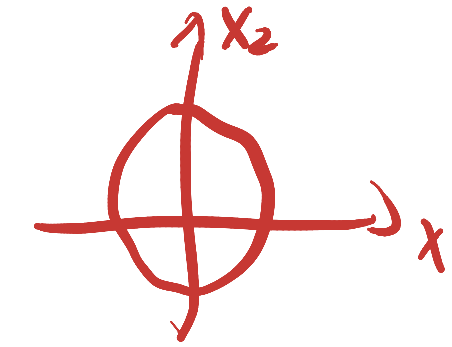
- Properties: Preserves norms and angles ($\|A\mathbf{x}\| = \|\mathbf{x}\|$)
- Examples: Rotation matrices, permutation matrices

Quadratic Forms

$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad y = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$y^T A y = 1 = x_1^2 + x_2^2$

- Let $\mathbf{y}$ be a $k \times 1$ vector, and let A be a $k \times k$ matrix. The function



$$\mathbf{y}^\top A\mathbf{y} = \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij} y_i y_j$$

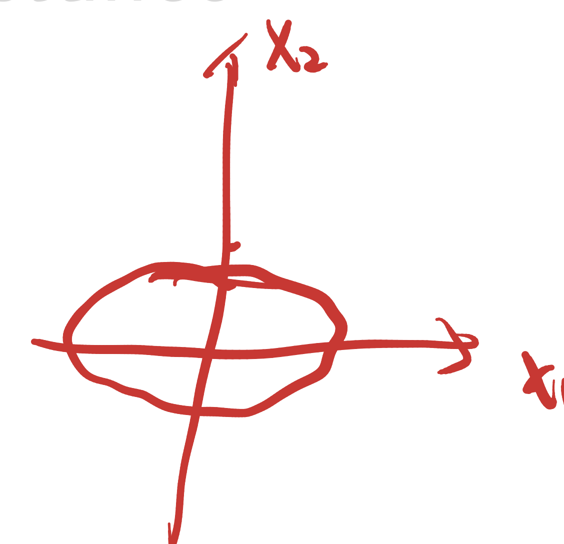
  is called a quadratic form

$A = \begin{pmatrix} 3 & 0 \\ 0 & 5 \end{pmatrix} \quad y = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

- Geometric interpretation: Ellipsoids in k-dimensional space

$y^T A y = 3x_1^2 + 5x_2^2 = 1$

- Example: Energy in physical systems, Mahalanobis distance



Quadratic Forms

- Let $\mathbf{y}$ be a $k \times 1$ vector, and let A be a $k \times k$ matrix. The function

$$\mathbf{y}^\top A\mathbf{y} = \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij} y_i y_j$$

  is called a quadratic form

- Geometric interpretation: Ellipsoids in k-dimensional space
- Example: Energy in physical systems, Mahalanobis distance

$$\|x-y\|_2^2 = (x+y)^T I (x-y)$$

$$\|x-y\|_A = \sqrt[3/2]{(x-y)^T A (x-y)}$$

Positive Definite and Positive Semidefinite Matrices

Let A be a $k \times k$ matrix.

- A is said to be *positive definite* if

(a) $A = A^\top$ (A is symmetric)

(b) $\mathbf{y}^\top A\mathbf{y} > 0 \quad \forall \mathbf{y} \in \mathbb{R}^k, \mathbf{y} \neq 0$

- A is said to be *positive semidefinite* if:

(a) $A = A^\top$ (A is symmetric)

(c) $\mathbf{y}^\top A\mathbf{y} \geq 0 \quad \forall \mathbf{y} \in \mathbb{R}^k$

- Tests: Eigenvalues > 0 (positive definite), eigenvalues ≥ 0 (positive semidefinite)

- Application: Convex optimization, kernel methods

Positive Definite and Positive Semidefinite Matrices

Let A be a $k \times k$ matrix.

- A is said to be *positive definite* if

(a) $A = A^\top$ (A is symmetric)

(b) $\mathbf{y}^\top A\mathbf{y} > 0 \quad \forall \mathbf{y} \in \mathbb{R}^k, \mathbf{y} \neq 0$

- A is said to be *positive semidefinite* if:

(a) $A = A^\top$ (A is symmetric)

(c) $\mathbf{y}^\top A\mathbf{y} \geq 0 \quad \forall \mathbf{y} \in \mathbb{R}^k$

- Tests: Eigenvalues > 0 (positive definite), eigenvalues ≥ 0 (positive semidefinite)

- Application: Convex optimization, kernel methods

Positive Definite and Positive Semidefinite Matrices

Let A be a $k \times k$ matrix.

- A is said to be *positive definite* if

  (a) $A = A^\top$ (A is symmetric)

  (b) $\mathbf{y}^\top A \mathbf{y} > 0 \quad \forall \mathbf{y} \in \mathbb{R}^k, \mathbf{y} \neq 0$

- A is said to be *positive semidefinite* if:

  (a) $A = A^\top$ (A is symmetric)

  (c) $\mathbf{y}^\top A \mathbf{y} \geq 0 \quad \forall \mathbf{y} \in \mathbb{R}^k$

- Tests: Eigenvalues > 0 (positive definite), eigenvalues ≥ 0 (positive semidefinite)

- Application: Convex optimization, kernel methods

A

$(\lambda, x) \Rightarrow \quad Ax = \lambda x$

eigenvalues

eigenvector

$(A - \lambda I)x = 0$

$\det(A - \lambda I) = 0$

Positive Definite and Positive Semidefinite Matrices

Let A be a $k \times k$ matrix.

- A is said to be *positive definite* if

(a) $A = A^\top$ (A is symmetric)

(b) $\mathbf{y}^\top A\mathbf{y} > 0 \quad \forall \mathbf{y} \in \mathbb{R}^k, \mathbf{y} \neq 0$

- A is said to be *positive semidefinite* if:

(a) $A = A^\top$ (A is symmetric)

(c) $\mathbf{y}^\top A\mathbf{y} \geq 0 \quad \forall \mathbf{y} \in \mathbb{R}^k$

- Tests: Eigenvalues > 0 (positive definite), eigenvalues ≥ 0 (positive semidefinite)

- Application: Convex optimization, kernel methods

① $A = B^T B$

$y^T A y = y^T B^T B y = \|By\|_2^2 \geq 0$

② $A = c_1 b_1 b_1^T + c_2 b_2 b_2^T + \dots + c_m b_m b_m^T$

$y^T A y = \sum_{i=1}^m c_i \, y^T b_i b_i^T y$

$c_1, \dots, c_m \geq 0$

$= \sum_{i=1}^m c_i (b_i^T y)^2 \geq 0$

Trace of a Matrix

Let A be a $k \times k$ matrix. The *trace* of A, denoted by trace(A) or tr(A), is the sum of the diagonal elements of A; thus,

$$\text{trace}(A) = \sum_{i=1}^{k} a_{ii}$$

Properties:

1. If A is an $m \times n$ matrix and B is an $n \times m$ matrix, then

$$\sum_{i=1}^{k} (AB)_{i,i} = \sum_{i=1}^{k} \sum_{\ell} a_{i,\ell} B_{\ell,i}$$

$$\text{trace}(AB) = \text{trace}(BA)$$

$$\sum_{i=1}^{k} (BA)_{i,i} = \sum_{i=1}^{k} \sum_{\ell} B_{i,\ell} A_{\ell,i} = \sum_{\ell} \sum_{i} A_{\ell,i} B_{i,\ell}$$

2. If the matrices are appropriately conformable, then

$$\text{trace}(ABC) = \text{trace}(CAB)$$

3. If A and B are $k \times k$ matrices and a and b are scalars, then

$$\text{trace}(aA + bB) = a\text{trace}(A) + b\text{trace}(B)$$

Trace of a Matrix

Let A be a $k \times k$ matrix. The *trace* of A, denoted by trace(A) or tr(A), is the sum of the diagonal elements of A; thus,

$$\text{trace}(A) = \sum_{i=1}^{k} a_{ii}$$

$A = X_1 (X_1^T X_1)^{-1} X_1^T$

$X_1 \in R^{m \times n}$

$\text{Trace}(A) = \text{Trace}(X_1 (X_1^T X_1)^{-1} X_1^T)$

$= \text{Trace}((X_1^T X_1)^{-1} X_1^T X_1)$

$= \text{Trace}(I_n)$

$= n$

Properties:

1. If A is an $m \times n$ matrix and B is an $n \times m$ matrix, then

$$\text{trace}(AB) = \text{trace}(BA)$$

2. If the matrices are appropriately conformable, then

$$\text{trace}(ABC) = \text{trace}(CAB) = \text{trace}(BCA)$$

3. If A and B are $k \times k$ matrices and a and b are scalars, then

$$\text{trace}(aA + bB) = a\text{trace}(A) + b\text{trace}(B)$$

Trace of a Matrix

Let A be a $k \times k$ matrix. The *trace* of A, denoted by trace(A) or tr(A), is the sum of the diagonal elements of A; thus,

$$\text{trace}(A) = \sum_{i=1}^{k} a_{ii}$$

Properties:

1. If A is an $m \times n$ matrix and B is an $n \times m$ matrix, then

$$\text{trace}(AB) = \text{trace}(BA)$$

2. If the matrices are appropriately conformable, then

$$\text{trace}(ABC) = \text{trace}(CAB)$$

3. If A and B are $k \times k$ matrices and a and b are scalars, then

$$\text{trace}(aA + bB) = a\text{trace}(A) + b\text{trace}(B)$$

Rank of an Idempotent Matrix

Assume (λ, x) is eigen-pair of A.

$Ax = \lambda x$

$AA = A$ $\quad$ $AAx = A(Ax) = A(\lambda x) = \lambda(Ax) = \lambda^2 x$

- Let A be an idempotent matrix. The rank of A is equal to its trace

$$\text{rank}(A) = \text{trace}(A)$$

$\Rightarrow \lambda x = \lambda^2 x$

$(\lambda - \lambda^2) x = 0$

- Proof sketch: Use the fact that idempotent matrices are diagonalizable with eigenvalues 0 or 1

$\Rightarrow \lambda = 0$ or $\lambda = 1$.

① $\text{Trace}(A) = \sum_{i=1}^{n} \lambda_i$

- Application: In regression, $\text{rank}(X) = \text{trace}(H)$ where $H = X(X^\top X)^{-1}X^\top$ is the hat matrix

$=$ # of 1s of eigenvalues

② $\text{rank}(A) =$ # of 1s of eigenvalues

Rank of an Idempotent Matrix

- Let A be an idempotent matrix. The rank of A is equal to its trace

$$\text{rank}(A) = \text{trace}(A)$$

- Proof sketch: Use the fact that idempotent matrices are diagonalizable with eigenvalues 0 or 1
- Application: In regression, $\text{rank}(X) = \text{trace}(H)$ where $H = X(X^\top X)^{-1}X^\top$ is the hat matrix

An Important Identity for a Partitioned Matrix

Let $\mathbf{X}$ be an $n \times p$ matrix partitioned such that

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$$

We note that

$$\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X} = \mathbf{X}$$

$$\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top[\mathbf{X}_1 \;\; \mathbf{X}_2] = \mathbf{X}$$

$$\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top[\mathbf{X}_1 \;\; \mathbf{X}_2] = [\mathbf{X}_1 \;\; \mathbf{X}_2]$$

Consequently,

$$\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}_1 = \mathbf{X}_1 \quad \text{and} \quad \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}_2 = \mathbf{X}_2$$

Similarly,

$$\mathbf{X}_1^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{X}_1^\top \quad \text{and} \quad \mathbf{X}_2^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{X}_2^\top$$

Inverse of a Partitioned Matrix

Consider a matrix of the form

$$\mathbf{X}^\top\mathbf{X} = \begin{bmatrix} \mathbf{X}_1^\top\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{X}_2 \\ \mathbf{X}_2^\top\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{X}_2 \end{bmatrix}$$

It can be shown that the inverse of this matrix is $(\mathbf{X}^\top\mathbf{X})^{-1}$ that equals

$$\begin{bmatrix} (\mathbf{X}_1^\top\mathbf{X}_1)^{-1} + (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{X}_2 G \mathbf{X}_2^\top\mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1} & -(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{X}_2 G \\ -G\mathbf{X}_2^\top\mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1} & G \end{bmatrix}$$

where

$$\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top \quad \text{and} \quad G = \left[\mathbf{X}_2^\top(\mathbf{I} - \mathbf{H}_1)\mathbf{X}_2\right]^{-1}$$

Application: Regression analysis with multiple groups of predictors

We will show that

$$\begin{bmatrix} (X_1^TX_1)^{-1}+(X_1^TX_1)^{-1}X_1^TX_2\,G\,X_2^TX_1(X_1^TX_1)^{-1} & -(X_1^TX_1)^{-1}X_1^TX_2\,G \\ -G\,X_2^TX_1\,(X_1^TX_1)^{-1} & G \end{bmatrix}\begin{bmatrix} X_1^TX_1 & X_1^TX_2 \\ X_2^TX_1 & X_2^TX_2 \end{bmatrix} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} = I.$$

① We can verify

$$M_{11} = \left[(X_1^TX_1)^{-1} + (X_1^TX_1)^{-1}X_1^TX_2\,G\,X_2^TX_1(X_1^TX_1)^{-1}\right]X_1^TX_1 + \left[-(X_1^TX_1)^{-1}X_1^TX_2G\right]X_2^TX_1$$

$$= I + (X_1^TX_1)^{-1}X_1^TX_2\,G\,X_2^TX_1 - (X_1^TX_1)^{-1}X_1^TX_2\,G\,X_2^TX = I$$

②
$$M_{12} = \left[(X_1^TX_1)^{-1} + (X_1^TX_1)^{-1}X_1^TX_2\,G\,X_2^TX_1(X_1^TX_1)^{-1}\right]X_1^TX_2 + \left[-(X_1^TX_1)^{-1}X_1^TX_2\,G\right]X_2^TX_2$$

$$= (X_1^TX_1)^{-1}X_1^TX_2 + \left[(X_1^TX_1)^{-1}X_1^TX_2\right]G\left[X_2^TX_1(X_1^TX_1)^{-1}X_1^TX_2 - X_2^TX_2\right]$$

$$= (X_1^TX_1)^{-1}X_1^TX_2 - \left[(X_1^TX_1)^{-1}X_1^TX_2\right]G\,G^{-1}$$

$$= 0$$

③ Similarly $M_{21}=0$

④ $M_{22} = -G\,X_2^TX_1(X_1^TX_1)^{-1}X_1^TX_2 + G\,X_2^TX_2 = G\,X_2^T\left[I - X_1(X_1^TX_1)^{-1}X_1^T\right]X_2$

$$= G\,G^{-1} = I.$$

Determinant

- The determinant of a square matrix A, denoted $\det(A)$ or $|A|$, is a scalar value
- Geometric interpretation: Scaling factor of the linear transformation
- For 2×2 matrix: $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$
- Properties:
  - $\det(AB) = \det(A)\det(B)$
  - $\det(A^{-1}) = 1/\det(A)$
  - $\det(A^\top) = \det(A)$
- Application: Testing invertibility, change of variables in integration

Determinant

- The determinant of a square matrix A, denoted $\det(A)$ or $|A|$, is a scalar value
- Geometric interpretation: Scaling factor of the linear transformation
- For 2×2 matrix: $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$
- Properties:
  - $\det(AB) = \det(A)\det(B)$
  - $\det(A^{-1}) = 1/\det(A)$
  - $\det(A^\top) = \det(A)$
- Application: Testing invertibility, change of variables in integration

Determinant

- The determinant of a square matrix A, denoted $\det(A)$ or $|A|$, is a scalar value
- Geometric interpretation: Scaling factor of the linear transformation
- For 2×2 matrix: $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$
- Properties:
  - $\det(AB) = \det(A)\det(B)$
  - $\det(A^{-1}) = 1/\det(A)$ $\det(AA^{-1}) = \det(I) = 1 = \det(A)\det(A^{-1})$
  - $\det(A^\top) = \det(A)$
- Application: Testing invertibility, change of variables in integration

Determinant

- The determinant of a square matrix A, denoted $\det(A)$ or $|A|$, is a scalar value
- Geometric interpretation: Scaling factor of the linear transformation
- For 2×2 matrix: $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$
- Properties:
  - $\det(AB) = \det(A)\det(B)$
  - $\det(A^{-1}) = 1/\det(A)$
  - $\det(A^\top) = \det(A)$
- Application: Testing invertibility, change of variables in integration

Contents

Matrix Derivatives

(Matrix Cookbook)

Let $\boldsymbol{A}$ be a $k \times k$ matrix of constants, $\boldsymbol{a}$ be a $k \times 1$ vector of constants, and $\boldsymbol{y}$ be a $k \times 1$ vector of variables.

$$\frac{\partial z}{\partial y} = \left(\frac{\partial z}{\partial y_i}\right)_i = \left(\frac{\partial}{\partial y_i} \sum_j a_j y_j\right)_i = \left(\frac{\partial}{\partial y_i} a_i y_i\right)_i = (a_i)_i = a$$

1. If $z = \boldsymbol{a}^\top \boldsymbol{y}$, then
$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial(\boldsymbol{a}^\top \boldsymbol{y})}{\partial \boldsymbol{y}} = \boldsymbol{a}$$

2. If $z = \boldsymbol{y}^\top \boldsymbol{y}$, then
$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial(\boldsymbol{y}^\top \boldsymbol{y})}{\partial \boldsymbol{y}} = 2\boldsymbol{y}$$

3. If $z = \boldsymbol{a}^\top \boldsymbol{A}\boldsymbol{y}$, then
$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial(\boldsymbol{a}^\top \boldsymbol{A}\boldsymbol{y})}{\partial \boldsymbol{y}} = \boldsymbol{A}^\top \boldsymbol{a}$$

4. If $z = \boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}$ and $\boldsymbol{A}$ is symmetric, then
$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y})}{\partial \boldsymbol{y}} = 2\boldsymbol{A}\boldsymbol{y}$$

Matrix Derivatives

Let $\boldsymbol{A}$ be a $k \times k$ matrix of constants, $\boldsymbol{a}$ be a $k \times 1$ vector of constants, and $\boldsymbol{y}$ be a $k \times 1$ vector of variables.

1. If $z = \boldsymbol{a}^\top \boldsymbol{y}$, then
$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial (\boldsymbol{a}^\top \boldsymbol{y})}{\partial \boldsymbol{y}} = \boldsymbol{a}$$
2. If $z = \boldsymbol{y}^\top \boldsymbol{y}$, then
$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial (\boldsymbol{y}^\top \boldsymbol{y})}{\partial \boldsymbol{y}} = 2\boldsymbol{y}$$
3. If $z = \boldsymbol{a}^\top \boldsymbol{A}\boldsymbol{y}$, then
$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial (\boldsymbol{a}^\top \boldsymbol{A}\boldsymbol{y})}{\partial \boldsymbol{y}} = \boldsymbol{A}^\top \boldsymbol{a}$$
4. If $z = \boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}$ and $\boldsymbol{A}$ is symmetric, then
$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial (\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y})}{\partial \boldsymbol{y}} = 2\boldsymbol{A}\boldsymbol{y}$$

Matrix Derivatives

Let $\boldsymbol{A}$ be a $k \times k$ matrix of constants, $\boldsymbol{a}$ be a $k \times 1$ vector of constants, and $\boldsymbol{y}$ be a $k \times 1$ vector of variables.

1. If $z = \boldsymbol{a}^\top \boldsymbol{y}$, then

$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial (\boldsymbol{a}^\top \boldsymbol{y})}{\partial \boldsymbol{y}} = \boldsymbol{a}$$

2. If $z = \boldsymbol{y}^\top \boldsymbol{y}$, then

$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial (\boldsymbol{y}^\top \boldsymbol{y})}{\partial \boldsymbol{y}} = 2\boldsymbol{y}$$

3. If $z = \boldsymbol{a}^\top \boldsymbol{A} \boldsymbol{y}$, then

$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial (\boldsymbol{a}^\top \boldsymbol{A} \boldsymbol{y})}{\partial \boldsymbol{y}} = \boldsymbol{A}^\top \boldsymbol{a}$$

$z = b^\top y \qquad b = A^\top a$

$\Rightarrow \frac{\partial z}{\partial y} = b = A^\top a,$

4. If $z = \boldsymbol{y}^\top \boldsymbol{A} \boldsymbol{y}$ and $\boldsymbol{A}$ is symmetric, then

$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial (\boldsymbol{y}^\top \boldsymbol{A} \boldsymbol{y})}{\partial \boldsymbol{y}} = 2\boldsymbol{A}\boldsymbol{y}$$

Matrix Derivatives

Let $\boldsymbol{A}$ be a $k \times k$ matrix of constants, $\boldsymbol{a}$ be a $k \times 1$ vector of constants, and $\boldsymbol{y}$ be a $k \times 1$ vector of variables.

1. If $z = \boldsymbol{a}^\top \boldsymbol{y}$, then
$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial(\boldsymbol{a}^\top \boldsymbol{y})}{\partial \boldsymbol{y}} = \boldsymbol{a}$$
2. If $z = \boldsymbol{y}^\top \boldsymbol{y}$, then
$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial(\boldsymbol{y}^\top \boldsymbol{y})}{\partial \boldsymbol{y}} = 2\boldsymbol{y}$$
3. If $z = \boldsymbol{a}^\top \boldsymbol{A}\boldsymbol{y}$, then
$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial(\boldsymbol{a}^\top \boldsymbol{A}\boldsymbol{y})}{\partial \boldsymbol{y}} = \boldsymbol{A}^\top \boldsymbol{a}$$
4. If $z = \boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}$ and $\boldsymbol{A}$ is symmetric, then
$$\frac{\partial z}{\partial \boldsymbol{y}} = \frac{\partial(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y})}{\partial \boldsymbol{y}} = 2\boldsymbol{A}\boldsymbol{y}$$

Let $z = y^T A y$

$$\frac{\partial z}{\partial y_\ell} = \frac{\partial}{\partial y_\ell} \sum_{(i,j)} a_{i,j}\, y_i y_j$$

$$= \frac{\partial}{\partial y_\ell} \left[\sum_{i=j=\ell} a_{\ell\ell} y_\ell^2 + \sum_{\substack{i=\ell \\ j\neq\ell}} a_{i,j} y_i y_j + \sum_{\substack{j=\ell \\ i\neq\ell}} a_{i,j} y_i y_j \right]$$

$$= 2 a_{\ell\ell} y_\ell + \sum_{j\neq\ell} a_{\ell,j} y_j + \sum_{i\neq\ell} a_{i\ell} y_i$$

$$= \sum_j a_{\ell,j} y_j + \sum_i a_{i,\ell} y_i$$

$$\Rightarrow \frac{\partial z}{\partial y} = \Big(\sum_j a_{\ell,j} y_j \Big)_\ell + \Big(\sum_i a_{i,\ell} y_i \Big)_\ell$$

$$= Ay + A^T y = 2Ay$$

if assume A symmetric

More Derivative Rules

- Application: Gradient descent optimization

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla f(\mathbf{w}_t)$$

  where $\nabla f(\mathbf{w})$ is the gradient of the objective function

- Example: For linear regression with loss $L(\mathbf{w}) = \|\mathbf{y} - X\mathbf{w}\|^2$, the gradient is

$$\nabla L(\mathbf{w}) = -2X^\top(\mathbf{y} - X\mathbf{w})$$

- Chain rule for matrix derivatives: If $z = f(\mathbf{y})$ and $\mathbf{y} = g(\mathbf{x})$, then

$$\frac{\partial z}{\partial \mathbf{x}} = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)^\top \frac{\partial z}{\partial \mathbf{y}}$$

Expectations of Random Vectors

Let $\boldsymbol{A}$ be a $k \times k$ matrix of constants, $\boldsymbol{a}$ be a $k \times 1$ vector of constants, and $\boldsymbol{y}$ be a $k \times 1$ random vector with mean $\boldsymbol{\mu}$ and nonsingular variance–covariance matrix V.

1. $\mathbb{E}(\boldsymbol{a}^\top \boldsymbol{y}) = \boldsymbol{a}^\top \boldsymbol{\mu}$
2. $\mathbb{E}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{\mu}$
3. $\text{Var}(\boldsymbol{a}^\top \boldsymbol{y}) = \boldsymbol{a}^\top V \boldsymbol{a}$
4. $\text{Var}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A} V \boldsymbol{A}^\top$

   Note: If $V = \sigma^2 I$, then $\text{Var}(\boldsymbol{A}\boldsymbol{y}) = \sigma^2 \boldsymbol{A}\boldsymbol{A}^\top$
5. $\mathbb{E}(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}) = \text{trace}(AV) + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}$

   Note: If $V = \sigma^2 I$, then $\mathbb{E}(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}) = \sigma^2 \text{trace}(A) + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}$

Expectations of Random Vectors

Let $\boldsymbol{A}$ be a $k \times k$ matrix of constants, $\boldsymbol{a}$ be a $k \times 1$ vector of constants, and $\boldsymbol{y}$ be a $k \times 1$ random vector with mean $\boldsymbol{\mu}$ and nonsingular variance–covariance matrix V.

1. $\mathbb{E}(\boldsymbol{a}^\top \boldsymbol{y}) = \boldsymbol{a}^\top \boldsymbol{\mu}$
2. $\mathbb{E}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{\mu}$
3. $\mathrm{Var}(\boldsymbol{a}^\top \boldsymbol{y}) = \boldsymbol{a}^\top V \boldsymbol{a}$
4. $\mathrm{Var}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A} V \boldsymbol{A}^\top$

   Note: If $V = \sigma^2 I$, then $\mathrm{Var}(\boldsymbol{A}\boldsymbol{y}) = \sigma^2 \boldsymbol{A}\boldsymbol{A}^\top$
5. $\mathbb{E}(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}) = \mathrm{trace}(AV) + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}$

   Note: If $V = \sigma^2 I$, then $\mathbb{E}(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}) = \sigma^2 \mathrm{trace}(A) + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}$

Expectations of Random Vectors

Let $\boldsymbol{A}$ be a $k \times k$ matrix of constants, $\boldsymbol{a}$ be a $k \times 1$ vector of constants, and $\boldsymbol{y}$ be a $k \times 1$ random vector with mean $\boldsymbol{\mu}$ and nonsingular variance–covariance matrix V.

1. $\mathbb{E}(\boldsymbol{a}^\top \boldsymbol{y}) = \boldsymbol{a}^\top \boldsymbol{\mu}$
2. $\mathbb{E}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{\mu}$
3. $\mathrm{Var}(\boldsymbol{a}^\top \boldsymbol{y}) = \boldsymbol{a}^\top V \boldsymbol{a}$
4. $\mathrm{Var}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A} V \boldsymbol{A}^\top$

   Note: If $V = \sigma^2 I$, then $\mathrm{Var}(\boldsymbol{A}\boldsymbol{y}) = \sigma^2 \boldsymbol{A}\boldsymbol{A}^\top$
5. $\mathbb{E}(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}) = \mathrm{trace}(AV) + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}$

   Note: If $V = \sigma^2 I$, then $\mathbb{E}(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}) = \sigma^2 \mathrm{trace}(A) + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}$

Expectations of Random Vectors

Let $\boldsymbol{A}$ be a $k \times k$ matrix of constants, $\boldsymbol{a}$ be a $k \times 1$ vector of constants, and $\boldsymbol{y}$ be a $k \times 1$ random vector with mean $\boldsymbol{\mu}$ and nonsingular variance–covariance matrix V.

1. $\mathbb{E}(\boldsymbol{a}^\top \boldsymbol{y}) = \boldsymbol{a}^\top \boldsymbol{\mu}$
2. $\mathbb{E}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{\mu}$
3. $\mathrm{Var}(\boldsymbol{a}^\top \boldsymbol{y}) = \boldsymbol{a}^\top V \boldsymbol{a}$
4. $\mathrm{Var}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A} V \boldsymbol{A}^\top$

   Note: If $V = \sigma^2 I$, then $\mathrm{Var}(\boldsymbol{A}\boldsymbol{y}) = \sigma^2 \boldsymbol{A}\boldsymbol{A}^\top$
5. $\mathbb{E}(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}) = \mathrm{trace}(AV) + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}$

   Note: If $V = \sigma^2 I$, then $\mathbb{E}(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}) = \sigma^2 \mathrm{trace}(A) + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}$

Expectations of Random Vectors

Let $\boldsymbol{A}$ be a $k \times k$ matrix of constants, $\boldsymbol{a}$ be a $k \times 1$ vector of constants, and $\boldsymbol{y}$ be a $k \times 1$ random vector with mean $\boldsymbol{\mu}$ and nonsingular variance–covariance matrix V.

$$E[\underbrace{(y-\mu)^T}_{\downarrow} A (y-\mu)]$$

$$= E[Tr(A(y-\mu)(y-\mu)^T)]$$

$$= Tr(A\, E[(y-\mu)(y-\mu)^T])$$

$$= Tr(AV)$$

1. $\mathbb{E}(\boldsymbol{a}^\top \boldsymbol{y}) = \boldsymbol{a}^\top \boldsymbol{\mu}$
2. $\mathbb{E}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{\mu}$
3. $\text{Var}(\boldsymbol{a}^\top \boldsymbol{y}) = \boldsymbol{a}^\top V \boldsymbol{a}$
4. $\text{Var}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A} V \boldsymbol{A}^\top$

   Note: If $V = \sigma^2 I$, then $\text{Var}(\boldsymbol{A}\boldsymbol{y}) = \sigma^2 \boldsymbol{A}\boldsymbol{A}^\top$
5. $\mathbb{E}(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}) = \text{trace}(AV) + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}$

   Note: If $V = \sigma^2 I$, then $\mathbb{E}(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}) = \sigma^2 \text{trace}(A) + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu}$

Applications of Matrix Expectations

- Portfolio variance: For portfolio returns $\mathbf{r}$ with weights $\mathbf{w}$,

  $$\mathrm{Var}(\mathbf{w}^\top \mathbf{r}) = \mathbf{w}^\top \Sigma \mathbf{w}$$

  where Σ is the covariance matrix of returns
- Risk estimation: For quadratic loss functions
- Signal processing: For estimating power in transformed signals
- Econometrics: In GMM and other estimation methods

Applications of Matrix Expectations

- Portfolio variance: For portfolio returns $\mathbf{r}$ with weights $\mathbf{w}$,

$$\mathsf{Var}(\mathbf{w}^\top \mathbf{r}) = \mathbf{w}^\top \Sigma \mathbf{w}$$

  where Σ is the covariance matrix of returns
- Risk estimation: For quadratic loss functions
- Signal processing: For estimating power in transformed signals
- Econometrics: In GMM and other estimation methods

Applications of Matrix Expectations

- Portfolio variance: For portfolio returns $\mathbf{r}$ with weights $\mathbf{w}$,

$$\mathsf{Var}(\mathbf{w}^\top \mathbf{r}) = \mathbf{w}^\top \Sigma \mathbf{w}$$

  where Σ is the covariance matrix of returns

- Risk estimation: For quadratic loss functions
- Signal processing: For estimating power in transformed signals
- Econometrics: In GMM and other estimation methods

Applications of Matrix Expectations

- Portfolio variance: For portfolio returns $\mathbf{r}$ with weights $\mathbf{w}$,

$$\mathsf{Var}(\mathbf{w}^\top \mathbf{r}) = \mathbf{w}^\top \Sigma \mathbf{w}$$

  where Σ is the covariance matrix of returns

- Risk estimation: For quadratic loss functions
- Signal processing: For estimating power in transformed signals
- Econometrics: In GMM and other estimation methods

Contents

Applications in AI

- Neural networks: Weight matrices and activation functions

$$\mathbf{h}^{(l)} = f(W^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)})$$

- Principal Component Analysis (PCA): Eigendecomposition of covariance matrix

$$\Sigma = Q\Lambda Q^\top$$

- Linear regression: Least squares solution

$$\hat{\beta} = (X^\top X)^{-1}X^\top y$$

- Support Vector Machines: Quadratic optimization with linear constraints

Back Propagation: Overview and Motivation

$\ell(\hat{y}, y) = \|y - \hat{y}\|_2^2$

- Loss function: $F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)$
- Goal: Minimize $F(\theta)$ using gradient descent

$$\theta(t+1) = \theta(t) - \alpha_t \nabla F(\theta(t))$$

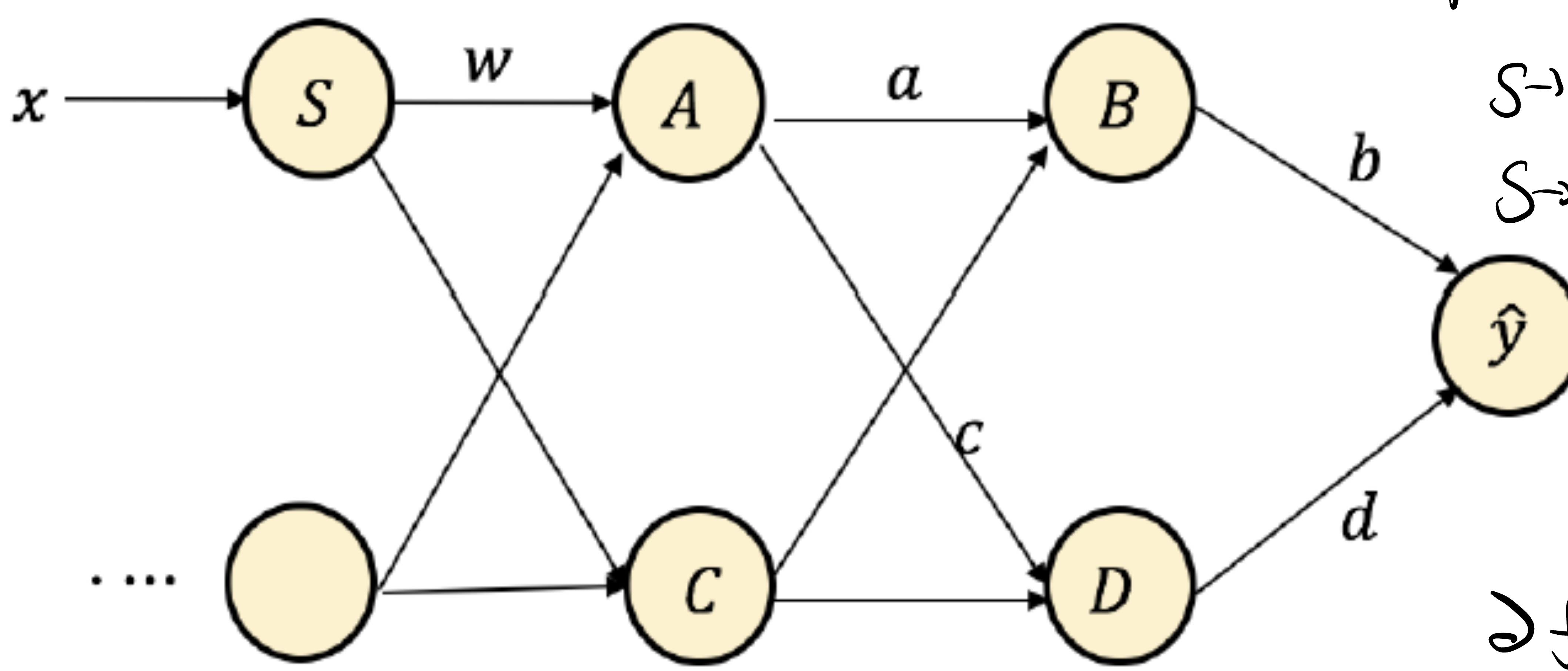
- Back propagation efficiently computes $\nabla F(\theta(t))$ using chain rule

Understanding BP in Level I: Scalar Form of Gradient

- Based on two fundamental rules:
  - Chain Rule: $\frac{df(g(w))}{dw} = \frac{df}{dg}\frac{dg}{dw}$
  - Sum Rule: $\frac{d(f_1(w)+f_2(w))}{dw} = \frac{df_1}{dw} + \frac{df_2}{dw}$
- Practical for coding implementations

Understanding BP in Level I: Scalar Form of Gradient

Example 2.2. Consider a 2-layer neural network with scalar output. We are interested in computing the derivative of this output $\hat{y}$ over a scalar parameter w. This function w.r.t. w can be represented in graph:



path 1:

$S \to A \to B \to \hat{y}$

$S \to A \to D \to \hat{y}$

$f_1(w) = b\phi(a\phi(wx))$

$f_2(w) = d\phi(c\phi(wx))$

$$\frac{\partial\, f_1(w) + f_2(w)}{\partial w} = \frac{\partial f_1(w)}{\partial w} + \frac{\partial f_2(w)}{\partial w}$$

$[b\phi'(a\phi(wx))][a\phi'(wx)][x]$

The computation of $\frac{\partial \hat{y}}{\partial w}$ can be summarized as follows:

BP Level II: Matrix Form Understanding

$x \xrightarrow{V} Vx \xrightarrow{U} UVx$

- Consider a 2-layer linear network (The weight matrices U, V are parameterized by θ) $f_\theta(x) = UVx$.
- Given n data points (x_i, y_i), the goal is to minimize the loss function

$\|A\|_F^2 = \sum_{(i,j)} A_{ij}^2$

$\|A\| = \max_{\|x\|\le 1} \|Ax\|$

$$F \triangleq \frac{1}{n}\sum_{i=1}^{n} \|UVx_i - y_i\|^2,$$

$h = UV - Y$

$\frac{\partial F}{\partial V} = \frac{\partial F}{\partial h}\frac{\partial h}{\partial V} = 2U^T(UV-Y)$

$d_y \times d_x$

with U, V to be determined.

- The question is how to take gradient of F w.r.t. the matrix V?
- Even simpler, how to compute $\frac{\partial F}{\partial V}$ with $F \triangleq \|UV - Y\|_F^2$? Here suppose that $U \in \mathbb{R}^{d_y \times d_1}$, $V \in \mathbb{R}^{d_1 \times d_x}$, $Y \in \mathbb{R}^{d_y \times d_x}$.

$R^{d_1 \times d_x}$

U

$d_y \times d_1$

$2(UV-Y)$

$d_y \times d_x$

BP Level II: Matrix Form Understanding

$$x \in R^{d\times 1} \xrightarrow{V} Vx \in R^{k\times 1} \xrightarrow{\phi} \phi(Vx) \in R$$

$\phi: R^k \to R^1$

- For $g(V) \triangleq \phi(Vx)$ with $x \in \mathbb{R}^{d\times 1}$ and $V \in \mathbb{R}^{k\times d}$, define $h = Vx$.

  Then

$$\underbrace{\frac{\partial g}{\partial V}}_{k\times d} = \frac{\partial \phi}{\partial h} x^\top = \frac{\partial g}{\partial h} \cdot x^T$$

$$= \underbrace{\frac{\partial g}{\partial h}}_{k\times 1} \; \underbrace{\frac{\partial h}{\partial V}}_{x \to d\times 1}$$

- Exercise:

$$\frac{\partial \|AWB + C\|_F^2}{\partial W} = 2A^\top (AWB + C)B^\top$$

$$\frac{\partial g}{\partial h} = \frac{\partial \phi(Vx)}{\partial Vx}$$

$$= \phi'(Vx)$$

BP for General Deep Non-linear Network

$\frac{\partial z^1}{\partial h^1} = D^1$

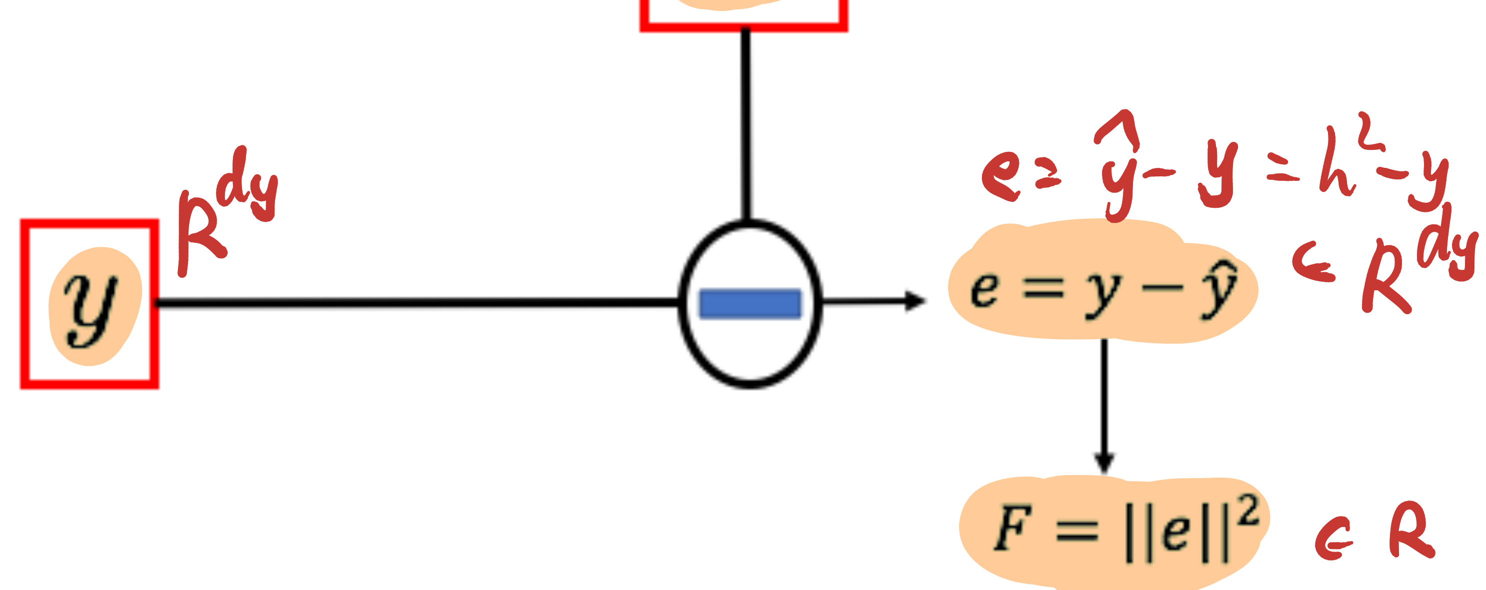
$z^1 = \phi(h^1) \Leftrightarrow \forall i,\ z^1_i = \phi(h^1_i) \Leftrightarrow \frac{\partial z^1_i}{\partial h^1_i} = \phi'(h^1_i)$

Now derive the gradient of fully-connected neural network with quadratic loss. The objective f_θ is defined based on the following diagram:

$d_L \equiv d_y$

$$\underset{\mathbb{R}^{d_x}}{x} \xrightarrow{\overset{\mathbb{R}^{d_1 \times d_x}}{W^1}} \underset{\mathbb{R}^{d_1}}{h^1} \xrightarrow{\phi} \underset{\mathbb{R}^{d_1}}{z^1} \xrightarrow{\overset{\mathbb{R}^{d_2 \times d_1}}{W^2}} \underset{\mathbb{R}^{d_2}}{h^2} \rightarrow \cdots \rightarrow z^{L-1} \xrightarrow{W^L} \boxed{h^L}\ \mathbb{R}^{d_L}$$

$h^1 = W^1 x$



$$\frac{\partial F}{\partial W^1} = \underbrace{\frac{\partial F}{\partial h^1}}_{\in \mathbb{R}^{d_1}} \cdot x^T$$

$$= \left(\frac{\partial e}{\partial h^1}\right)^T \left(\frac{\partial F}{\partial e}\right) x^T$$

$$= \left(\frac{\partial e}{\partial h^1}\right)^T (2e)\, x^T$$

Figure: Diagram for the operator F $\quad \mathrm{diag}(\phi'(h^{L-1}_i))_{i=1}^{d_{L-1}}$

$$= \left(\frac{\partial e}{\partial h^L} \frac{\partial h^L}{\partial z^{L-1}} \frac{\partial z^{L-1}}{\partial h^{L-1}} \cdots \frac{\partial z^1}{\partial h^1}\right)^T (2e)\, x^T = \left(I \cdot W^L \cdot D^{L-1} W^{L-1} D^{L-2} \cdots W^2 D^1\right)^T (2e)\, x^T$$

BP for General Deep Non-linear Network

- The derivative $\frac{\partial F}{\partial W^1}$ is computed as follows:

$$
\begin{aligned}
\frac{\partial F}{\partial W^1} &= \frac{\partial F}{\partial h^1} x^\top && \text{(1a)} \\
&= \left(\frac{\partial e}{\partial h^1}\right)^\top \left(\frac{\partial F}{\partial e}\right) x^\top && \text{(1b)} \\
&= \left(\frac{\partial e}{\partial h^1}\right)^\top 2e \cdot x^\top \\
&= \left(\frac{\partial e}{\partial h^L} \frac{\partial h^L}{\partial z^{L-1}} \cdots \frac{\partial h^1}{\partial z^1} \frac{\partial z^1}{\partial h^1}\right)^\top 2e \cdot x^\top && \text{(1c)} \\
&= \left(W^L D^{L-1} W^{L-1} D^{L-2} \cdots W^2 D^1\right)^\top 2e \cdot x^\top && \text{(1d)}
\end{aligned}
$$

BP for General Deep Non-linear Network

- The general formula $\frac{\partial F}{\partial W^\ell}$ is left as exercise:

$$\frac{\partial F}{\partial W^\ell} = (W^L D^{L-1} \cdots W^{\ell+1} D^\ell)^\top \cdot 2e \cdot (z^{\ell-1})^\top$$

  This formula can be expressed in a recursive way, which is the mechanism of the BP technique.

- BP is an efficient way to compute all gradients $\frac{\partial F}{\partial W^\ell}$ for $\ell = 1, \ldots, L$. The navie computation complexity is $\mathcal{O}(d^2 L^2)$; while the BP complexity is $\mathcal{O}(d^2 L)$.

Further Reading

- Strang, G. (2016). *Introduction to Linear Algebra*
- Boyd, S. & Vandenberghe, L. (2018). *Introduction to Applied Linear Algebra*
- MIT OpenCourseWare: Linear Algebra