Module 3-1:
Linear Algebra I

國立中山大學電機系
魏家博

Outline

- Linear System of Equations
  - Consistent/Inconsistent Systems
  - Linear Dependence and Linear Independence
  - Rank and Nullity
- Linear Transformations
- Least Squares Problems and Matrix Calculus
  - Matrix Calculus
  - Local Minima/Global Minima
  - Gradient Descent

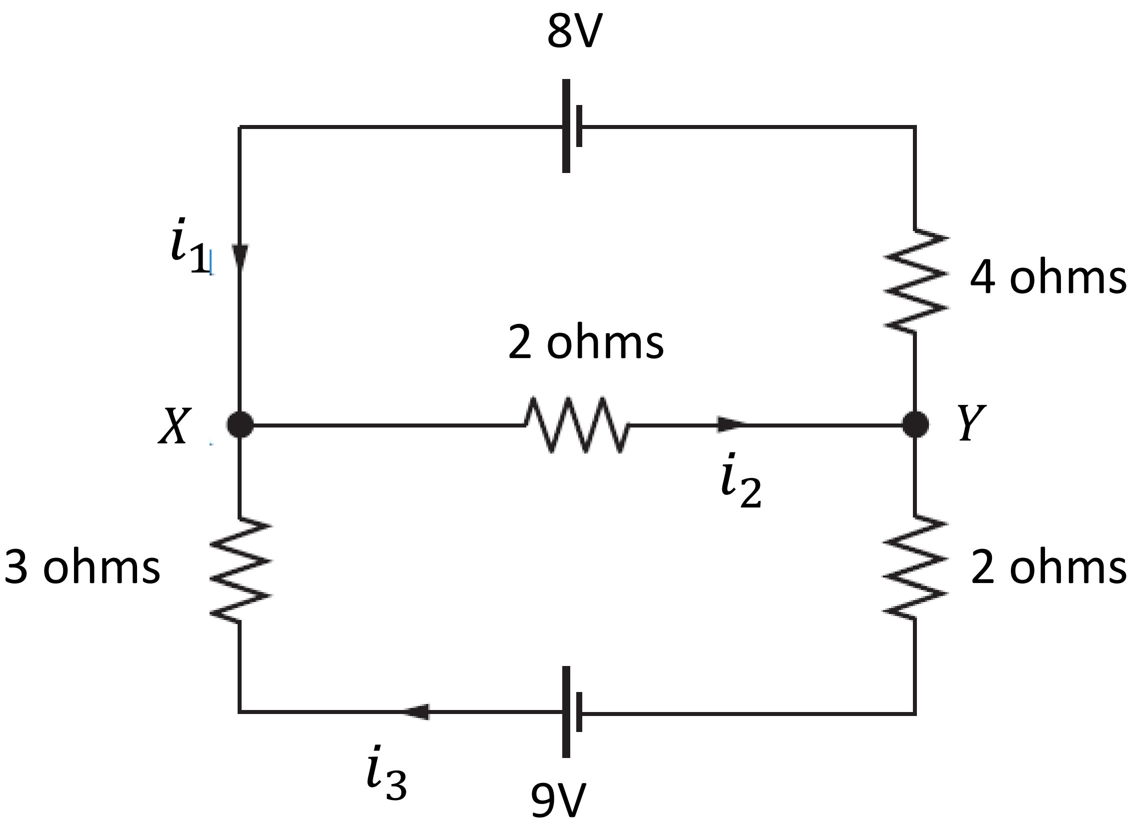
System of Equations

- A linear system of m equations in n unknowns is described as

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\
\vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \qquad & \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m
\end{aligned}
$$

- A linear system is inconsistent if it has no solution.
- A consistent system will have a nonempty solution set.

Electrical Networks

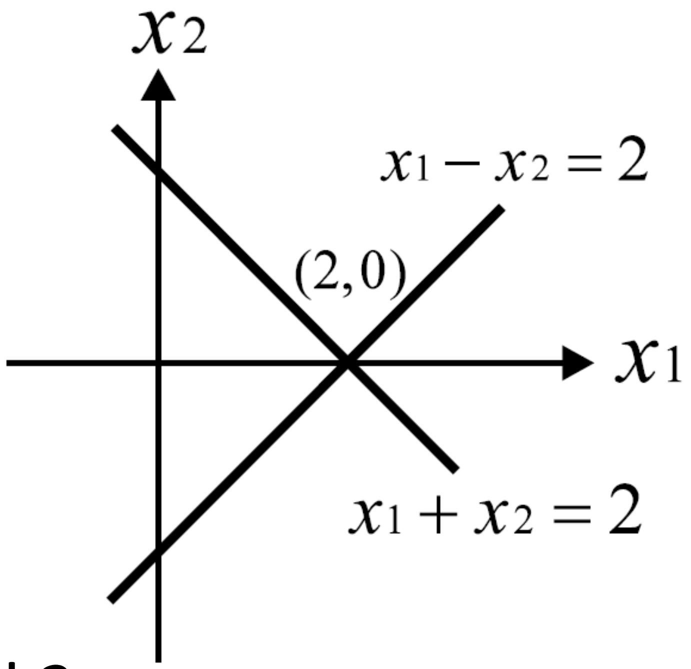


$$
\begin{aligned}
i_1 - i_2 + i_3 &= 0 \quad \text{(node X)} \\
-i_1 + i_2 - i_3 &= 0 \quad \text{(node Y)} \\
4i_1 + 2i_2 &= 8 \quad \text{(top loop)} \\
2i_2 + 5i_3 &= 9 \quad \text{(bottom loop)}
\end{aligned}
$$

$$
\underbrace{\begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 4 & 2 & 0 \\ 0 & 2 & 5 \end{bmatrix}}_{\mathbf{A}}
\underbrace{\begin{bmatrix} i_1 \\ i_2 \\ i_3 \end{bmatrix}}_{\mathbf{x}}
=
\underbrace{\begin{bmatrix} 0 \\ 0 \\ 8 \\ 9 \end{bmatrix}}_{\mathbf{b}}
$$

Consistent System (Exact One Solution)

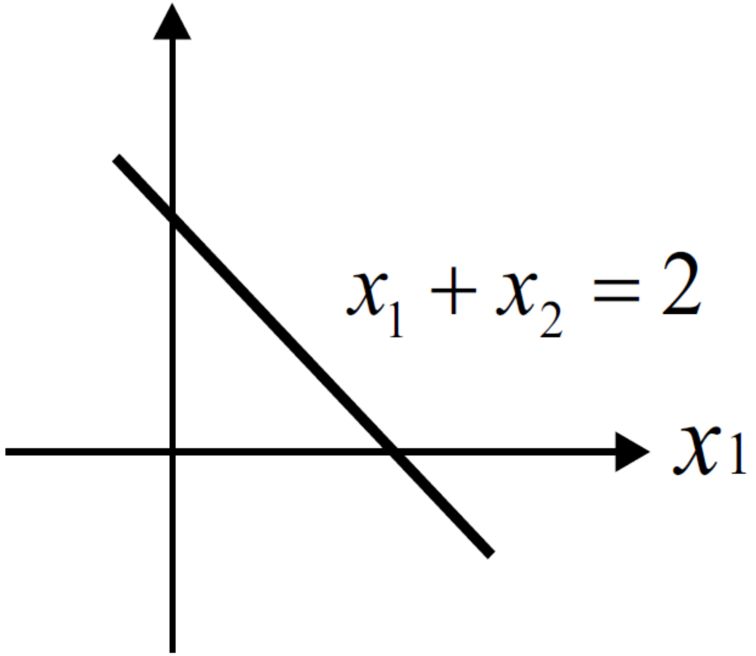
$$\begin{cases} x_1 - x_2 = 2 \\ x_1 + x_2 = 2 \end{cases}$$



- The system contains two lines intersecting at point (2,0).
- The system has 2 equations and 2 unknowns (m=2, n=2).

Consistent System (Infinite Solutions)

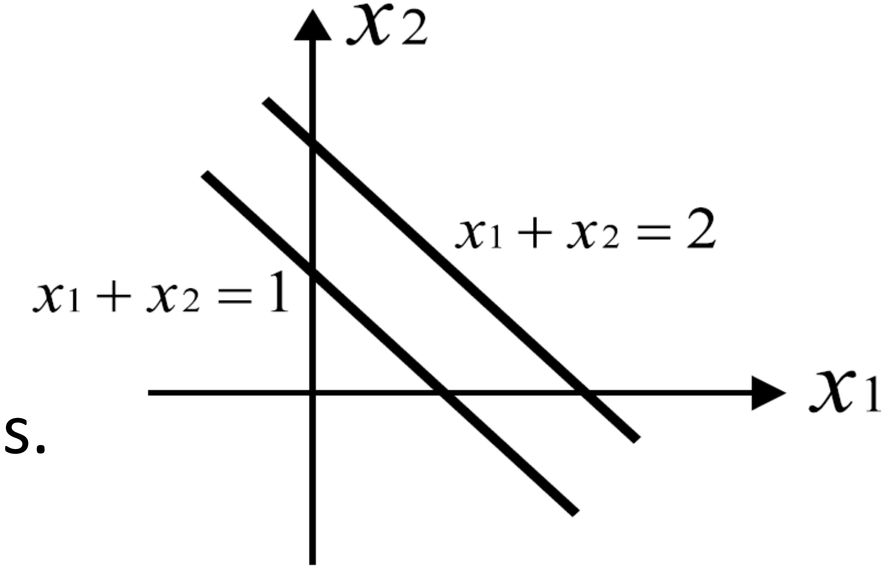
$$\begin{cases} -x_1 - x_2 = -2 \\ \quad x_1 + x_2 = 2 \end{cases}$$



- The system contains a single line.
- The system has 2 equations and 2 unknowns (m=2, n=2).

Inconsistent System (No Solutions)

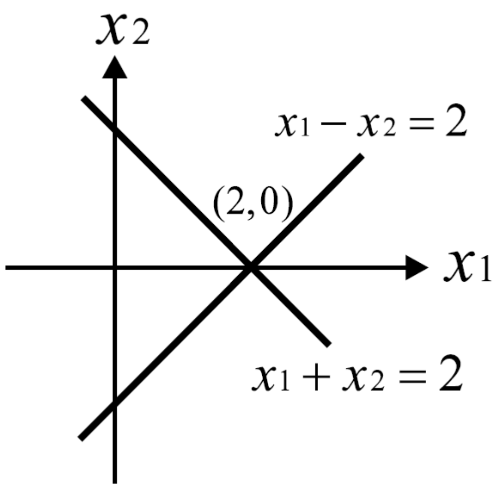
$$\begin{cases} x_1 + x_2 = 1 \\ x_1 + x_2 = 2 \end{cases}$$

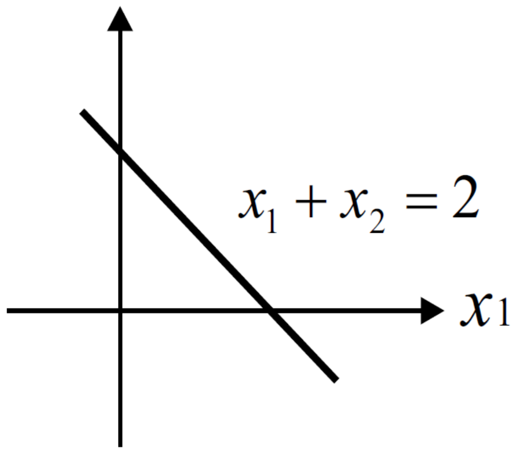


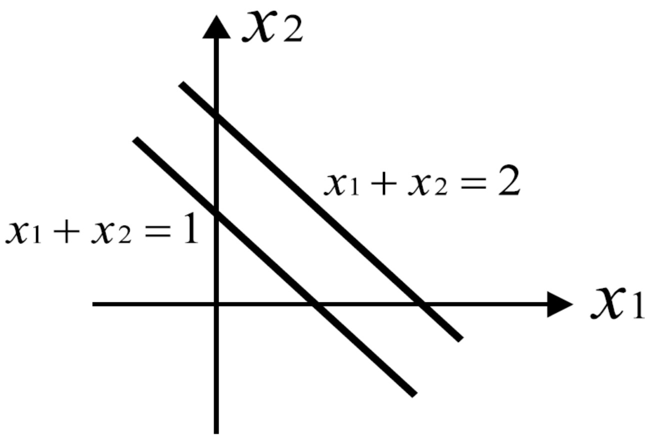
- The system contains two parallel lines.
- The system has 2 equations and 2 unknowns (m=2, n=2).

Solution Sets of Linear Systems

- For a linear system of equations, there are three scenarios for its solutions:
  - Exact one solution (consistent)
  - Infinitely many solutions (consistent)
  - No solutions (inconsistent)

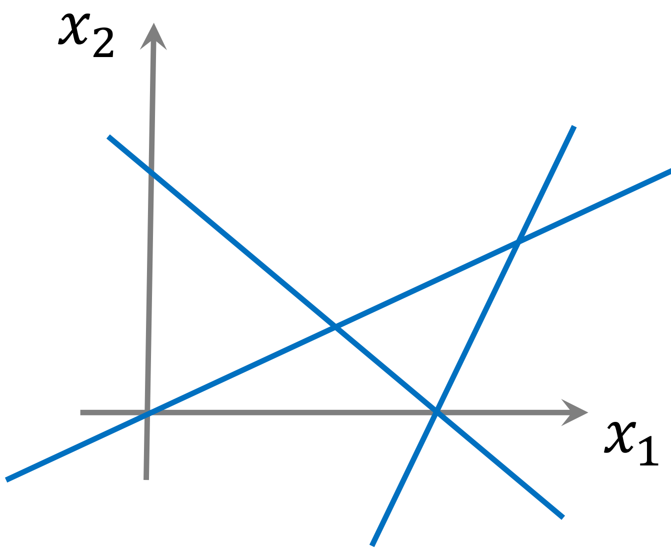






Quiz 1

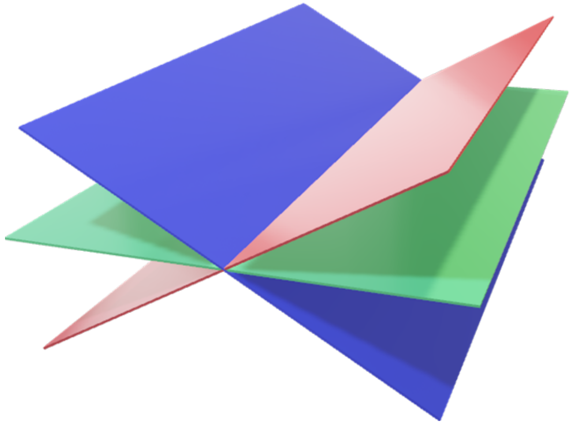
1. For a linear system of equations, is it possible to have a finite number of solutions (more than one solution)?
2. Determine whether the following linear system of equations (m=3, n=2) is consistent or not.



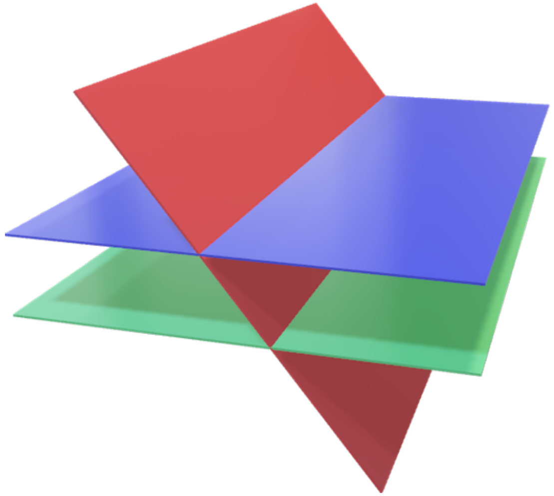
Quiz 2

1. Determine whether the following linear systems of equations (m=3, n=3) is consistent or not.

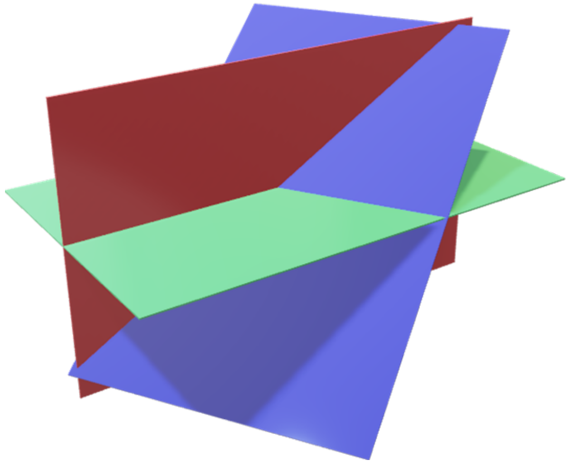
(是否有解？若有解，是唯一解或無限多解?)



(a)



(b)



(c)

Overdetermined/Underdetermined Systems

- A linear system is said to be overdetermined if there are more equations than unknowns ($m > n$).
- A linear system is said to be underdetermined if there are fewer equations than unknowns ($m < n$).

	$m > n$	$m = n$	$m < n$
Consistent (exact one solution)	O	O	X
Consistent (infinitely many solutions)	O	O	O
Inconsistent	O	O	O

Underdetermined Systems

Consider

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 + x_5 &= 2 \\ x_1 + x_2 + x_3 + 2x_4 + 2x_5 &= 3 \\ x_1 + x_2 + x_3 + 2x_4 + 3x_5 &= 2 \end{aligned}$$

Elementary Row Operations

$$\left(\begin{array}{ccccc|c} 1 & 1 & 1 & 1 & 1 & 2 \\ 1 & 1 & 1 & 2 & 2 & 3 \\ 1 & 1 & 1 & 2 & 3 & 2 \end{array}\right) \rightarrow \left(\begin{array}{ccccc|r} 1 & 1 & 1 & 1 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{-1} \end{array}\right) \rightarrow \left(\begin{array}{ccccc|r} 1 & 1 & 1 & 1 & 0 & 3 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{2} \\ 0 & 0 & 0 & 0 & 1 & -1 \end{array}\right)$$

$$\rightarrow \left(\begin{array}{ccccc|r} 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{array}\right) \rightarrow \begin{aligned} x_1 &= 1 - x_2 - x_3 \\ x_4 &= 2 \\ x_5 &= -1 \end{aligned}$$

$\rightarrow$ For any real numbers α and β, $(1 - \alpha - \beta, \alpha, \beta, 2, -1)$ is a solution of the system.

System of Equations in Matrix Form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \quad & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned} \quad \Longleftrightarrow \quad \mathbf{A}\mathbf{x} = \mathbf{b}$$

- Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and vector $\mathbf{b} \in \mathbb{R}^m$ are given, while vector $\mathbf{x} \in \mathbb{R}^n$ is to be solved.

System of Equations in Vector Form

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}}_{\mathbf{b}} \iff \mathbf{A}\mathbf{x} = \mathbf{b}$$

$$x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = \mathbf{b}$$

Linear Combination and Span

$$\mathbf{Ax} = \mathbf{b} \quad \Longleftrightarrow \quad x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = \mathbf{b}$$

- $x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n$ is called a linear combination of $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$.
- The set of all linear combinations of $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$ is called the span of $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$.
- **Theorem**. The system $\mathbf{Ax} = \mathbf{b}$ is consistent if and only if $\mathbf{b}$ can be expressed as a linear combination of $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$.
- Notice that $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n] \in \mathbb{R}^{m\times n}$, $\mathbf{a}_i \in \mathbb{R}^m$, $\mathbf{b} \in \mathbb{R}^m$, and $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$ are the column vectors of $\mathbf{A}$.

Quiz 3

1. Let $\mathbf{A}$ be a 5×3 matrix. If

$$\mathbf{b} = \mathbf{a}_1 + \mathbf{a}_2 = \mathbf{a}_2 + \mathbf{a}_3$$

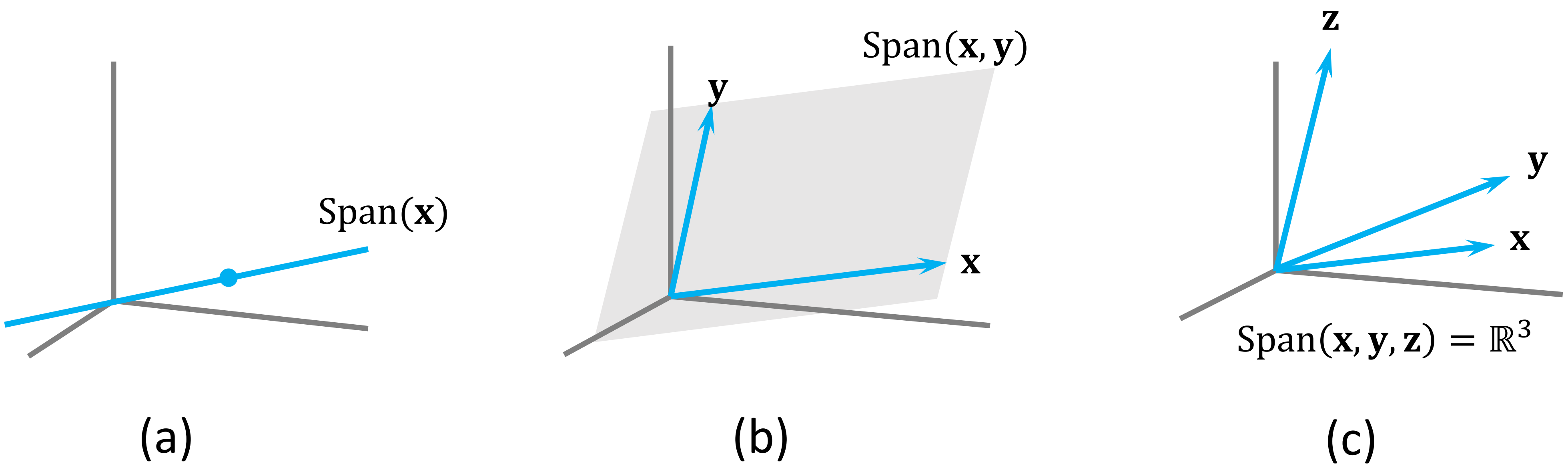
Then what can you conclude about the number of solutions of $\mathbf{Ax} = \mathbf{b}$?
(是否有解？若有解，是唯一解或無限多解?)

2. Let $\mathbf{A}$ be a 3×4 matrix. If

$$\mathbf{b} = \mathbf{a}_1 + \mathbf{a}_2 + \mathbf{a}_3 + \mathbf{a}_4$$

Then what can you conclude about the number of solutions of $\mathbf{Ax} = \mathbf{b}$?

Span Examples in $\mathbb{R}^3$



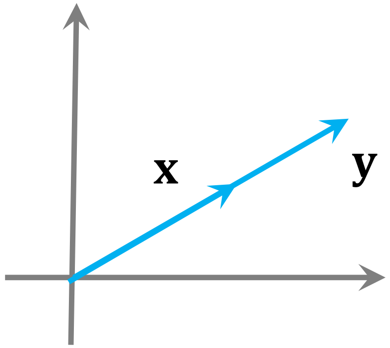
- The span of a vector is a line through the original (one-dimensional).
- The span of two independent vectors is a plane through the origin (two-dimensional).
- The span of three independent vectors is $\mathbb{R}^3$.

Linear Dependence

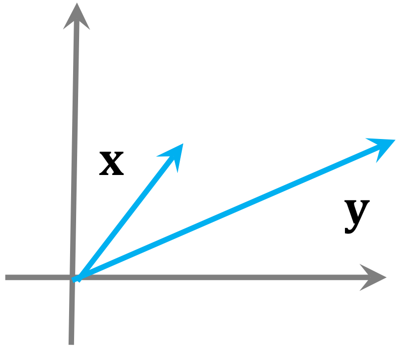
- The vectors $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$ is said to be linearly dependent if there exists scalars $c_1, c_2, \cdots, c_n$, not all zero, such that

$$c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_n\mathbf{a}_n = \mathbf{0}$$

- Geometric Interpretation:
  - If $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent in $\mathbb{R}^2$, than there exist c_1 and c_2 that are not both 0, such that $c_1\mathbf{x} + c_2\mathbf{y} = 0$. Then we have $\mathbf{x} = -\frac{c_2}{c_1}\mathbf{y}$.



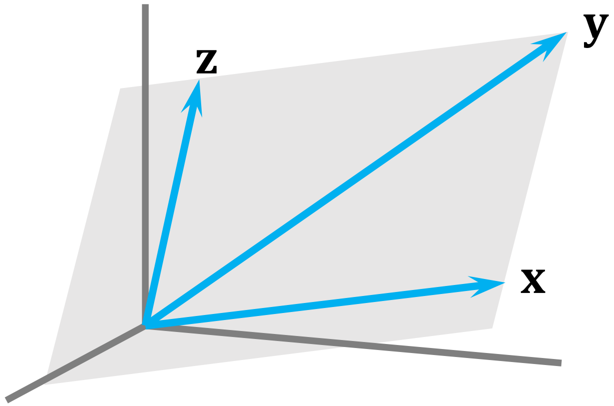
(a) $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent.



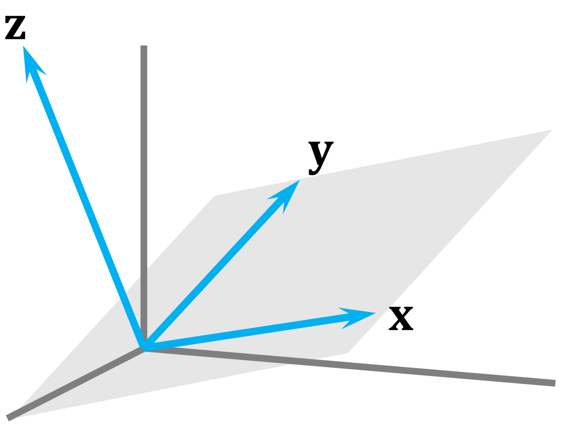
(b) $\mathbf{x}$ and $\mathbf{y}$ are linearly independent.

Linear Dependence (cont.)

- Suppose that $\mathbf{x}$ and $\mathbf{y}$ are linearly independent in $\mathbb{R}^3$.
- Since $\mathbf{x}$, $\mathbf{y}$, and the origin are not collinear, they determine a plane.



(a) $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are linearly dependent.



(b) $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are linearly independent.

Linear Independence

- The vectors $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$ is said to be linearly independent if

$$c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_n\mathbf{a}_n = \mathbf{0}$$

implies that the scalars $c_1, c_2, \cdots, c_n$ must equal zero.

- **Theorem**. The system $\mathbf{Ax} = \mathbf{b}$ has at most one solution for every $\mathbf{b} \in \mathbb{R}^m$ if and only if the column vectors of $\mathbf{A}$ are linearly independent.

Proof. ($\Longleftarrow$) If $\mathbf{x}_1$ and $\mathbf{x}_2$ were both solutions of $\mathbf{Ax} = \mathbf{b}$, then $\mathbf{x}_1 - \mathbf{x}_2$ would be a solution of $\mathbf{Ax} = \mathbf{0}$, because

$$\mathbf{A}(\mathbf{x}_1 - \mathbf{x}_2) = \mathbf{Ax}_1 - \mathbf{Ax}_2 = \mathbf{b} - \mathbf{b} = \mathbf{0}$$

Since the column vectors of $\mathbf{A}$ are linearly independent, we must have $\mathbf{x}_1 - \mathbf{x}_2 = \mathbf{0}$, and hence $\mathbf{x}_1$ must equal $\mathbf{x}_2$.

Example 1.

Let $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\} = \left\{ \begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix} \right\}$. Consider $c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + c_3\mathbf{a}_3 = \mathbf{0}$, i.e.,

$$c_1 \begin{bmatrix}1\\0\\0\end{bmatrix} + c_2 \begin{bmatrix}0\\1\\0\end{bmatrix} + c_3 \begin{bmatrix}0\\0\\1\end{bmatrix} = \mathbf{0} \implies \begin{bmatrix}c_1\\c_2\\c_3\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix}$$

Hence, the vectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ are linearly independent.

Example 2.

Let $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\} = \left\{ \begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}, \begin{bmatrix}2\\1\\0\end{bmatrix} \right\}$. Consider $c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + c_3\mathbf{a}_3 = \mathbf{0}$, i.e.,

$$c_1 \begin{bmatrix}1\\0\\0\end{bmatrix} + c_2 \begin{bmatrix}0\\1\\0\end{bmatrix} + c_3 \begin{bmatrix}2\\1\\0\end{bmatrix} = \mathbf{0} \implies \begin{bmatrix}c_1 + 2c_3\\c_2 + c_3\\0\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix} \implies \begin{bmatrix}c_1\\c_2\\c_3\end{bmatrix} = \begin{bmatrix}-2\alpha\\-\alpha\\\alpha\end{bmatrix}$$

Hence, the vectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ are linearly dependent.

Example 3.

Let $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\} = \left\{\begin{bmatrix}1\\0\\1\end{bmatrix}, \begin{bmatrix}0\\1\\1\end{bmatrix}, \begin{bmatrix}2\\1\\3\end{bmatrix}\right\}$. Consider $c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + c_3\mathbf{a}_3 = \mathbf{0}$, i.e.,

$$c_1\begin{bmatrix}1\\0\\1\end{bmatrix} + c_2\begin{bmatrix}0\\1\\1\end{bmatrix} + c_3\begin{bmatrix}2\\1\\3\end{bmatrix} = \mathbf{0} \implies \begin{bmatrix}c_1 + 2c_3\\c_2 + c_3\\c_1 + c_2 + 3c_3\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix} \implies \begin{bmatrix}c_1\\c_2\\c_3\end{bmatrix} = \begin{bmatrix}-2\alpha\\-\alpha\\\alpha\end{bmatrix}$$

Hence, the vectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ are linearly dependent.

Example 4.

Let $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\} = \{[1,0,2], [0,1,1], [1,1,3]\}$. Consider $c_1\mathbf{b}_1 + c_2\mathbf{b}_2 + c_3\mathbf{b}_3 = 0$, i.e.,

$$c_1[1,0,2] + c_2[0,1,1] + c_3[1,1,3] = \mathbf{0} \implies [c_1 + c_3,\ c_2 + c_3,\ 2c_1 + c_2 + 3c_3] = \mathbf{0}$$

$$[c_1, c_2, c_3] = [-\alpha, -\alpha, \alpha]$$

Hence, the vectors $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ are linearly dependent.

Basis and Dimension

- The vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$ for a basis for a vector space V if
  1. $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$ are linearly independent.
  2. $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$ span V.

- If a vector space V has a basis consisting of n vectors, we say that V has dimension n.

- Example. Below are bases for $\mathbb{R}^3$

$$\left\{\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix}\right\}, \left\{\begin{bmatrix}1\\1\\1\end{bmatrix}, \begin{bmatrix}0\\1\\1\end{bmatrix}, \begin{bmatrix}2\\0\\1\end{bmatrix}\right\}, \left\{\begin{bmatrix}1\\1\\1\end{bmatrix}, \begin{bmatrix}1\\1\\0\end{bmatrix}, \begin{bmatrix}1\\0\\1\end{bmatrix}\right\}$$

Row/Column Spaces and Rank

- The subspace of $\mathbb{R}^{1\times n}$ spanned by the row vectors of $\mathbf{A} \in \mathbb{R}^{m\times n}$ is called the row space of $\mathbf{A}$.
- The subspace of $\mathbb{R}^{m}$ spanned by the column vectors of $\mathbf{A} \in \mathbb{R}^{m\times n}$ is called the column space of $\mathbf{A}$.
- The rank of $\mathbf{A}$ is the dimension of the row space of $\mathbf{A}$.
- An equivalent definition of rank is the maximal number of linearly independent rows of A.
- **Theorem**. The dimension of the row space of $\mathbf{A}$ equals the dimension of the column space of $\mathbf{A}$. That is, $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{A}^T)$.

Example 5.

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & -1 & 1 \\ 2 & 4 & -3 & 1 \\ 1 & 2 & 1 & 3 \end{bmatrix} \xrightarrow{\times -2 \quad \times -1} \begin{bmatrix} 1 & 2 & -1 & 1 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 2 & 2 \end{bmatrix}$$

$$\longrightarrow \begin{bmatrix} 1 & 2 & -1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 2 & 2 \end{bmatrix} \xrightarrow{\times -2 \quad \times 1} \begin{bmatrix} 1 & 2 & 0 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \mathbf{U} \longrightarrow \mathrm{rank}(\mathbf{U}) = \mathrm{rank}(\mathbf{A}) = 2$$

- $\mathbf{U}$ is the reduced row echelon form of $\mathbf{A}$.
- In fact, $\mathbf{A} = \mathbf{EU}$, where $\mathbf{E}$ is a nonsingular matrix.
- Note that $\mathbf{Ax} = \mathbf{0} \;\Leftrightarrow\; \mathbf{EUx} = \mathbf{0} \;\Leftrightarrow\; \mathbf{Ux} = 0$
- Thus, $\mathbf{Ax} = \mathbf{0}$ and $\mathbf{Ux} = \mathbf{0}$ have the same solution set.
  They are equivalent systems.

Null Space and Nullity

- The set of all solutions to the system $\mathbf{Ax} = \mathbf{0}$ is called the null space of $\mathbf{A}$.
- The dimension of the null space of $\mathbf{A}$ is called the nullity of $\mathbf{A}$.
- **Theorem**. If $\mathbf{A}$ is an $m \times n$ matrix, then $\mathrm{rank}(\mathbf{A}) + \mathrm{nullity}(\mathbf{A}) = n$.

In Example 3 and 4:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 1 \\ 1 & 1 & 3 \end{bmatrix} \in \mathbb{R}^{3\times 3}, \mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{A}^T) = 2,\ \mathrm{nullity}(\mathbf{A}) = 1$$

In Example 5:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & -1 & 1 \\ 2 & 4 & -3 & 1 \\ 1 & 2 & 1 & 3 \end{bmatrix} \in \mathbb{R}^{3\times 4},\ \mathrm{rank}(\mathbf{A}) = 2,\ \mathrm{nullity}(\mathbf{A}) = 2$$

Detailed Derivation

- $\mathbf{Ax} = \mathbf{0}$ and $\mathbf{Ux} = \mathbf{0}$ have the same solution set. From Example 5,
- $\mathbf{Ux} = \begin{bmatrix} 1 & 2 & 0 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$
- $\begin{cases} x_1 + 2x_2 + 2x_4 = 0 \\ x_3 + x_4 = 0 \end{cases} \implies \begin{cases} x_1 = -2x_2 - 2x_4 \\ x_3 = -x_4 \end{cases} \implies$ Let $x_2 = \alpha$ and $x_4 = \beta$.
- $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \alpha \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} -2 \\ 0 \\ -1 \\ 1 \end{bmatrix} \implies \text{nullity}(\mathbf{U}) = \text{nullity}(\mathbf{A}) = 2.$

Useful Propositions

Proposition 1. The following three statements are equivalent.

1. The linear system $\mathbf{Ax} = \mathbf{b}$ is consistent.
2. The vector $\mathbf{b}$ can be expressed as a linear combination of $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$.
3. $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}([\mathbf{A} \quad \mathbf{b}])$

Proposition 2. Suppose $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$. The following three statements are equivalent.

1. The linear system $\mathbf{Ax} = \mathbf{b}$ has at most one solution for every $\mathbf{b} \in \mathbb{R}^m$.
2. The column vectors of $\mathbf{A}$ are linearly independent.
3. $\operatorname{rank}(\mathbf{A}) = n$

Useful Propositions (cont.)

Remarks on Proposition 2

- Suppose $\operatorname{rank}(\mathbf{A}) \neq n$. That is, the column vectors of $\mathbf{A}$ are linearly dependent.
- Then, there exists a nonzero vector $\mathbf{z} \in \mathbb{R}^n$ such that $\mathbf{Az} = \mathbf{0}.$
- If $\hat{\mathbf{x}}$ is a solution to $\mathbf{Ax} = \mathbf{b}$, then $\hat{\mathbf{x}} + \boldsymbol{z}$ is also a solution, because
$$\mathbf{A}(\hat{\mathbf{x}} + \boldsymbol{z}) = \mathbf{A}\hat{\mathbf{x}} + \boldsymbol{A}\boldsymbol{z} = \mathbf{A}\hat{\mathbf{x}} = \mathbf{b}$$
- In other words, the linear system $\mathbf{Ax} = \mathbf{b}$ has an infinite number of solutions if $\mathbf{Ax} = \mathbf{b}$ is consistent and $\operatorname{rank}(\mathbf{A}) \neq n$.

Quiz 4

Let $\mathbf{A}$ be an $m \times n$ matrix.

1. What is the minimum value of $\mathrm{rank}(\mathbf{A})$ if $\mathbf{A}$ is not a zero matrix?
2. What is the maximum value of $\mathrm{rank}(\mathbf{A})$?

Let $\mathbf{x}$ and $\mathbf{y}$ be vectors in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively.

3. What is the value of $\mathrm{rank}(\mathbf{x}\mathbf{y}^T)$?

Notice that $\mathbf{x}\mathbf{y}^T = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} [y_1 \quad y_2 \quad \cdots \quad y_m] = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_m \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_m \\ \vdots & \vdots & \ddots & \vdots \\ x_n y_1 & x_n y_2 & \cdots & x_n y_m \end{bmatrix}$

Properties of Square Matrices

- An $n \times n$ matrix $\mathbf{A}$ is said to be nonsingular or invertible if there exits a matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$. The matrix B is said to be a multiplicative inverse of $\mathbf{A}$.
- We refer the multiplicative inverse of a nonsingular matrix $\mathbf{A}$ as the inverse of $\mathbf{A}$ and denote it by $\mathbf{A}^{-1}$.
- An $n \times n$ matrix $\mathbf{A}$ is said to be singular if it does not have a multiplicative inverse.
- **Theorem**. If $\mathbf{A}$ and $\mathbf{B}$ are nonsingular $n \times n$ matrices, then $\mathbf{AB}$ is also nonsingular and $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

Proof.

$$(\mathbf{B}^{-1}\mathbf{A}^{-1})\mathbf{AB} = \mathbf{B}^{-1}(\mathbf{A}^{-1}\mathbf{A})\mathbf{B} = \mathbf{B}^{-1}\mathbf{B} = \mathbf{I}$$
$$(\mathbf{AB})(\mathbf{B}^{-1}\mathbf{A}^{-1}) = \mathbf{A}(\mathbf{BB}^{-1})\mathbf{A}^{-1} = \mathbf{AA}^{-1} = \mathbf{I}$$

Algebraic Rules for Transposes and Inverses

Rules for Transposes

- $(\mathbf{A}^T)^T = \mathbf{A}$
- $(\alpha\mathbf{A})^T = \alpha\mathbf{A}^T$
- $(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$
- $(\mathbf{A}\mathbf{B})^T = \mathbf{B}^T\mathbf{A}^T$

Rules for Inverses

(Only for square and invertible matrices)

- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- $(\alpha\mathbf{A})^{-1} = \alpha^{-1}\mathbf{A}^{-1}$
- $(\mathbf{A}+\mathbf{B})^{-1} = \mathbf{A}^{-1} + \mathbf{B}^{-1}$ (no such equality!)
- $(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$

Example. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$. Then $\mathbf{AB} \in \mathbb{R}^{m\times p}$. Notice that $\mathbf{BA}$ does not exist, because the dimension is not compatible for matrix multiplication.

Algebraic Rules for Determinants

Rules for Determinants

(Only for square matrices)

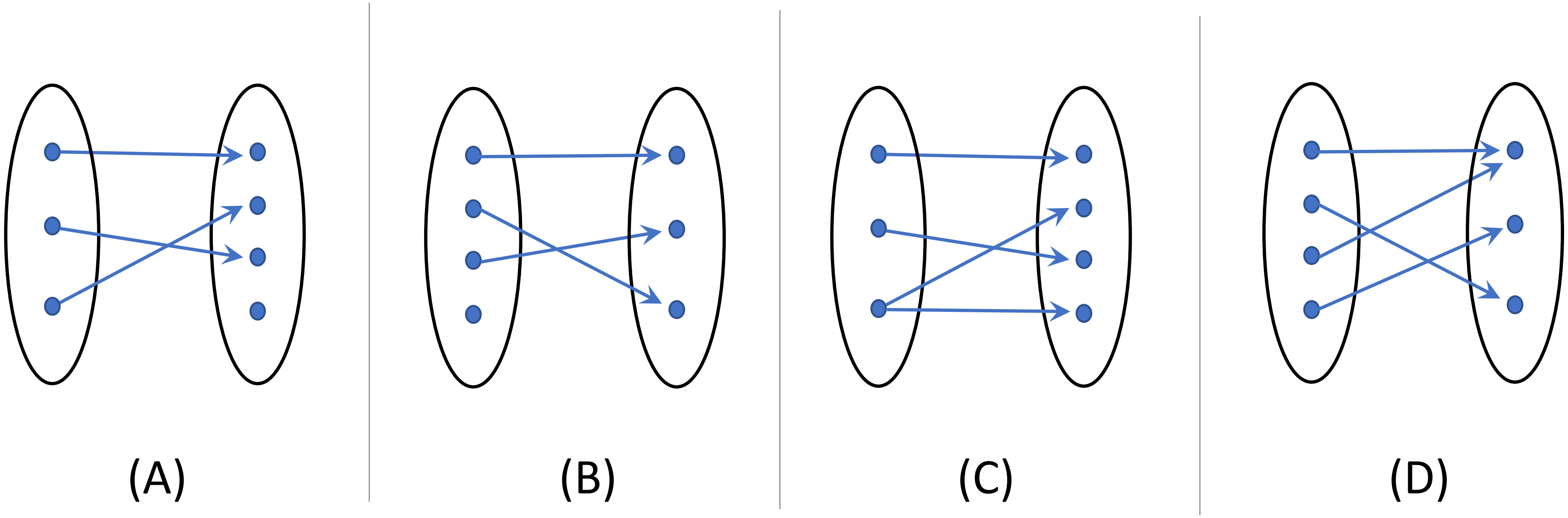
- $\det(\mathbf{A} + \mathbf{B}) = \det(\mathbf{A}) + \det(\mathbf{B})$ (no such equality!)
- $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$
- $\det(\mathbf{A}^T) = \det(\mathbf{A})$
- $\det(\mathbf{A}^{-1}) = \det(\mathbf{A})^{-1}$
- $\det(\alpha\mathbf{A}) = \alpha^n\det(\mathbf{A})$ for $\mathbf{A} \in \mathbb{R}^{n\times n}$

Quiz 5

True-or-False Questions:

1. If $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ matrices, then $(\mathbf{A}+\mathbf{B})(\mathbf{A}-\mathbf{B}) = \mathbf{A}^2 - \mathbf{B}^2$.
2. If $\mathbf{AC} = \mathbf{BC}$ and $\mathbf{C} \neq \mathbf{O}$ (the zero matrix), then $\mathbf{A} = \mathbf{B}$.
3. If $\mathbf{AB} = \mathbf{C}$ and $\mathbf{B}$ is nonsingular, then $\mathbf{A} = \mathbf{B}^{-1}\mathbf{C}$.

What is a Function (Transformation)?



A function L from a vector space V into a vector space W is denoted $L: V \to W$.

Linear Transformations

- A mapping L from a vector space V into a vector space W is said to be a linear transformation if

$$L(\mathbf{v}_1 + \mathbf{v}_2) = L(\mathbf{v}_1) + L(\mathbf{v}_2) \qquad \text{(additivity)}$$
$$L(\alpha\mathbf{v}) = \alpha L(\mathbf{v}) \qquad \text{(homogeneity)}$$

  for all $\mathbf{v}_1, \mathbf{v}_2 \in V$ and for all scalars α.

- An equivalent definition of a linear transformation L is that it satisfies

$$L(\alpha\mathbf{v}_1 + \beta\mathbf{v}_2) = \alpha L(\mathbf{v}_1) + \beta L(\mathbf{v}_2)$$

  for all $\mathbf{v}_1, \mathbf{v}_2 \in V$ and for all scalars α and β.

- In the case that the vector spaces V and W are the same, we refer to a linear transformation $L: V \to V$ as a linear operator on V.

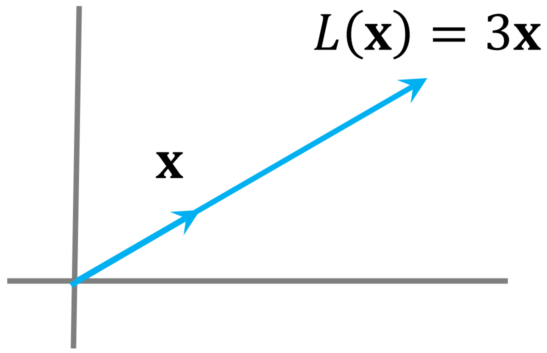
Scaling Operators

Let $L(\mathbf{x}) = 3\mathbf{x}$ for each $\mathbf{x} \in \mathbb{R}^n$. Since

$$L(\alpha\mathbf{x}) = 3(\alpha\mathbf{x}) = \alpha(3\mathbf{x}) = \alpha L(\mathbf{x}),$$

$$L(\mathbf{x}+\mathbf{y}) = 3(\mathbf{x}+\mathbf{y}) = 3\mathbf{x} + 3\mathbf{y} = L(\mathbf{x}) + L(\mathbf{y}),$$

it follows that L is a linear operator.

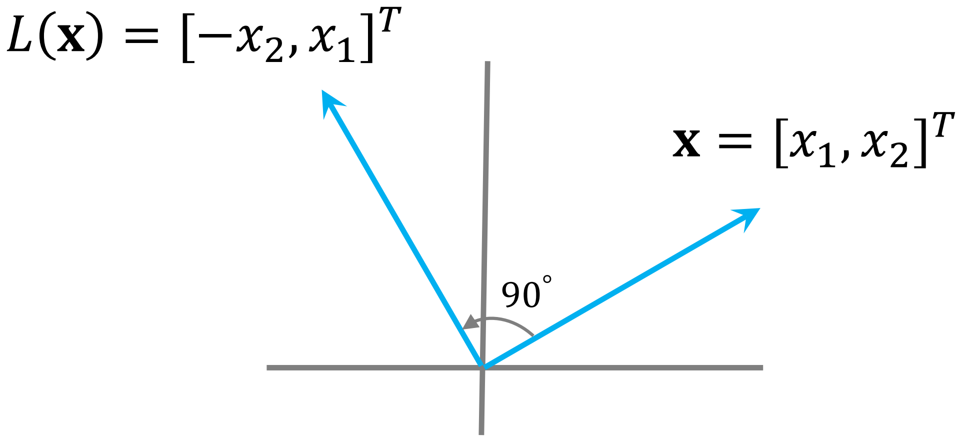


Rotation Operators

Let $L(\mathbf{x}) = [-x_2, x_1]^T$ for each $\mathbf{x} = [x_1, x_2]^T \in \mathbb{R}^2$. Since

$$L(\alpha\mathbf{x} + \beta\mathbf{y}) = \begin{bmatrix} -(\alpha x_2 + \beta y_2) \\ \alpha x_1 + \beta y_1 \end{bmatrix} = \alpha \begin{bmatrix} -x_2 \\ x_1 \end{bmatrix} + \beta \begin{bmatrix} -y_2 \\ y_1 \end{bmatrix} = \alpha L(\mathbf{x}) + \beta L(\mathbf{y}),$$

it follows that L is a linear operator.



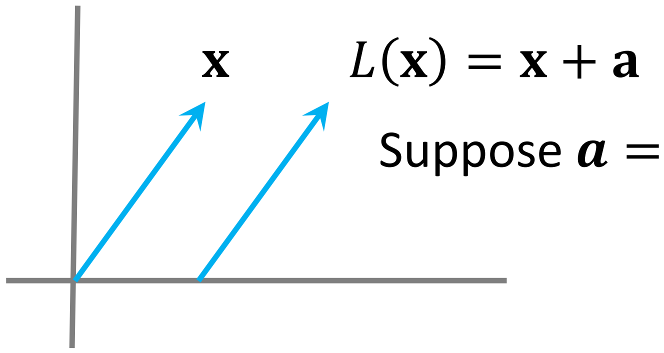
Translation Operators

Let $\mathbf{a}$ be a fixed nonzero vector in $\mathbb{R}^2$. A mapping of the form

$$L(\mathbf{x}) = \mathbf{x} + \mathbf{a}$$

is called a translation. Show that a translation is **not** a linear operator. Illustrate geometrically the effect of a translation.

Solution. Since $L(\mathbf{0}) = \mathbf{0} + \mathbf{a} \neq \mathbf{0}$, L is not a linear operator.



Suppose $\boldsymbol{a} = [1,0]^T$

Examples

5. Determine whether the following are linear transformations from $\mathbb{R}^3$ into $\mathbb{R}^2$.
   a) $L(\mathbf{x}) = (x_2, x_3)^T$ linear
   b) $L(\mathbf{x}) = (0, 0)^T$ linear
   c) $L(\mathbf{x}) = (1 + x_1, x_2)^T$ not linear
   d) $L(\mathbf{x}) = (x_3, x_1 + x_2)^T$ linear

6. Determine whether the following are linear transformations from $\mathbb{R}^2$ into $\mathbb{R}^3$.
   a) $L(\mathbf{x}) = (x_1, x_2, 1)^T$ not linear
   b) $L(\mathbf{x}) = (x_1, x_2, x_1 + 2x_2)^T$ linear
   c) $L(\mathbf{x}) = (x_1, 0, 0)^T$ linear
   d) $L(\mathbf{x}) = (x_1, x_2, x_1^2 + x_2^2)^T$ not linear

Matrix Representation Theorem

Theorem. If $E = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\}$ and $F = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_n\}$ are ordered bases for vector spaces V and W, respectively, then corresponding to each linear transformation $L: V \to W$, there is an $m \times n$ matrix $\mathbf{A}$ such that

$$[L(\mathbf{v})]_F = A[\mathbf{v}]_E \quad \text{for each } \mathbf{v} \in V$$

$$\begin{tikzcd}
\mathbf{v} \in V \arrow[r, "L"] \arrow[d, leftrightarrow] & \mathbf{w} = L(\mathbf{v}) \in W \arrow[d, leftrightarrow] \\
\mathbf{x} = [\mathbf{v}]_E \in \mathbb{R}^n \arrow[r, "\mathbf{A}"] & \mathbf{Ax} = [\mathbf{w}]_F \in \mathbb{R}^m
\end{tikzcd}$$

Matrix Representation of Differentiation

The linear transformation L defined by $L(p) = dp/dx$ maps P_3 into P_2, where the ordered bases $E = [x^2, x, 1]$ and $F = [x, 1]$ for P_3 and P_2.

If $p(x) = ax^2 + bx + c$, then $[p]_E = [a, b, c]^T$.

We know $L(p) = 2ax + b$ and $[L(p)]_F = [2a, b]^T$.

It can be verified that

$$\underbrace{\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} a \\ b \\ c \end{bmatrix}}_{[p]_E} = \underbrace{\begin{bmatrix} 2a \\ b \end{bmatrix}}_{[L(p)]_F}$$

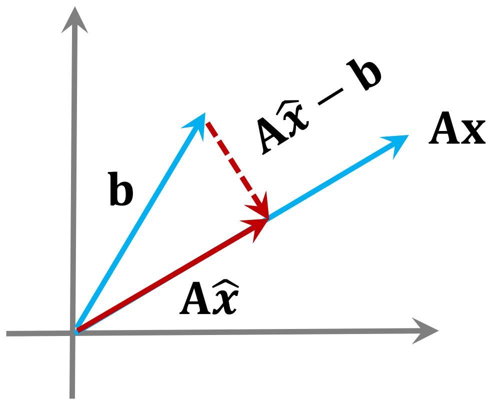
Column vectors of $\mathbf{A}$ can be derived in view of the following

$$L(x^2) = 2x + 0 \cdot 1$$
$$L(x) = 0x + 1 \cdot 1$$
$$L(1) = 0x + 0 \cdot 1$$

Least Squares Problem

- If a linear system $\mathbf{Ax} = \mathbf{b}$ is inconsistent, we can look for a vector $\hat{\mathbf{x}}$ for which is closest to $\mathbf{b}$. Such solution $\hat{\mathbf{x}}$ is called a least squares solution to the linear system.
- Least squares problems refer to the following optimization problem

$$\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2^2 = \min_{\mathbf{x}} (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b})$$



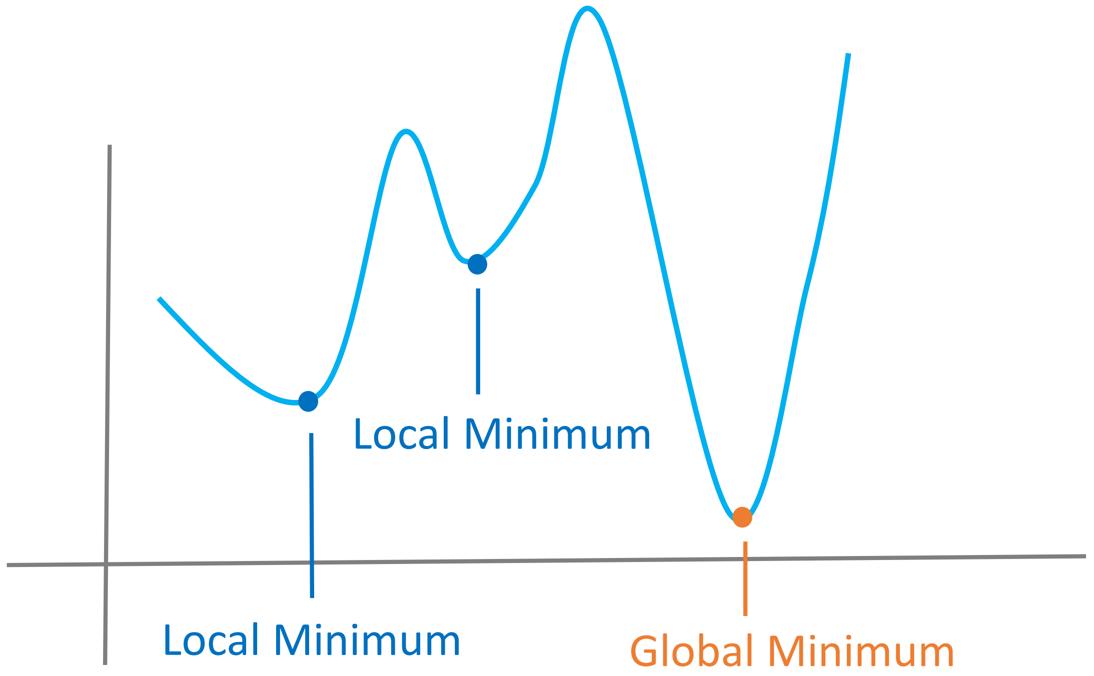
- Notice that the general form of an unconstrained optimization problem is

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$$

where the function $f: \mathbb{R}^n \to \mathbb{R}$ is called an objective function or a loss function.

Global and Local Minima

- We say f has a global minimum at $\mathbf{x}^*$ if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$.
- We say f has a local minimum at $\mathbf{x}^*$ if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x}$ belonging to a neighborhood of $\mathbf{x}^*$.



Global and Local Minimizer

- We say that $\mathbf{x}^*$ is a global minimizer of the optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$$

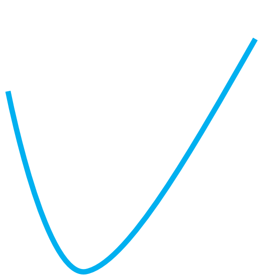
  if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$. The following notation is used to denote $\mathbf{x}^*$.

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$$

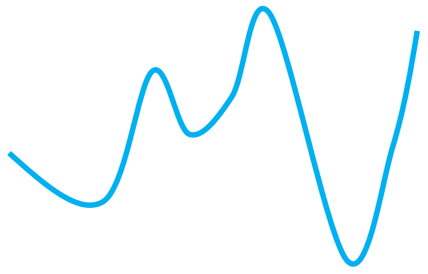
- We say that $\mathbf{x}^*$ is a local minimizer if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x}$ belonging to a neighborhood of $\mathbf{x}^*$.

Convex Objective Functions

- **Theorem**. Any local minimum of a convex objective function is also a global minimum.
- Because the objective function of the least square problem is convex, any local minimum is also a global minimum.
- The objection function of a neural network (**deep learning**) is non-convex.
  - There often exist many local minima that are not global minima.



(a) Convex function



(b) Non-convex function

Single-Variable Calculus

- Least Squares Problem

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \min_{\mathbf{x}} (\mathbf{A}\mathbf{x} - \mathbf{b})^T (\mathbf{A}\mathbf{x} - \mathbf{b})$$

- If $\mathbf{x}$ is a scalar, and let $\mathbf{A} = 2$ and $\mathbf{b} = 3$, then the above becomes

$$\min_{x} (2x - 3)^T (2x - 3) = \min_{x} (2x - 3)^2 = \min_{x} f(x)$$

- The global/local minimum occurs at x satisfying $f'(x) = 0$.

$$f'(x) = 2(2x - 3) \cdot 2 = 0 \quad \Longrightarrow \quad 2x - 3 = 0 \quad \Longrightarrow \quad x = 3/2.$$

Multivariable Calculus

- Least Squares Problem

$$\min_{\mathbf{x}} \|\mathbf{Ax}-\mathbf{b}\|_2^2 = \min_{\mathbf{x}} (\mathbf{Ax}-\mathbf{b})^T(\mathbf{Ax}-\mathbf{b}) = \min_{\mathbf{x}} f(\mathbf{x})$$

- We have

$$f(\mathbf{x}) = (\mathbf{Ax}-\mathbf{b})^T(\mathbf{Ax}-\mathbf{b}) = \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{A}^T\mathbf{b} - \mathbf{b}^T\mathbf{Ax} + \mathbf{b}^T\mathbf{b}$$

- The global/local minimum occurs at $\mathbf{x}$ satisfying $\nabla f(\mathbf{x}) = 0$.
- $\nabla f(\mathbf{x})$ is called the gradient of the function f.
- How can we calculate $\nabla f(\mathbf{x})$?

Matrix Derivatives

- There are six common types

	Scalar	Vector	Matrix
Scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{Y}}{\partial x}$
Vector	$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	
Matrix	$\frac{\partial y}{\partial \mathbf{X}}$		

Derivatives by Vector

$$\mathbf{x} = [x_1 \quad \cdots \quad x_n]^T$$
$$\mathbf{y} = [y_1 \quad \cdots \quad y_m]^T$$

Numerator Layout Notation

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1 \times n}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Denominator Layout Notation (we use this one)

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}^T \in \mathbb{R}^{n}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{n \times m}$$

- $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ is called the Jacobian matrix of $\mathbf{y}$.

Derivatives of Scalars by Vector

$$\mathbf{x} = [x_1 \quad \cdots \quad x_n]^T, \qquad \mathbf{u} = [u_1 \quad \cdots \quad u_n]^T, \qquad \mathbf{u}^T\mathbf{x} = u_1x_1 + \cdots + u_nx_n$$

(C1) If $\mathbf{u}$ is not a function of $\mathbf{x}$:

$$\frac{\partial(\mathbf{u}^T\mathbf{x})}{\partial\mathbf{x}} = \left[\frac{\partial(\mathbf{u}^T\mathbf{x})}{\partial x_1} \quad \cdots \quad \frac{\partial(\mathbf{u}^T\mathbf{x})}{\partial x_n}\right]^T = [u_1 \quad \cdots \quad u_n]^T = \mathbf{u}$$

(C2) If $\mathbf{u}$ is a function of $\mathbf{x}$:

$$\frac{\partial(\mathbf{u}^T\mathbf{x})}{\partial\mathbf{x}} = \frac{\partial\mathbf{u}}{\partial\mathbf{x}}\mathbf{x} + \frac{\partial\mathbf{x}}{\partial\mathbf{x}}\mathbf{u}$$

Computing $\nabla f(\mathbf{x})$

- $f(\mathbf{x}) = (\mathbf{Ax}-\mathbf{b})^T(\mathbf{Ax}-\mathbf{b}) = \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{A}^T\mathbf{b} - \mathbf{b}^T\mathbf{Ax} + \mathbf{b}^T\mathbf{b}$
- The gradient of $f(\mathbf{x})$ is

$$\begin{aligned}\nabla f(\mathbf{x}) &= \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}^T\mathbf{A}^T\mathbf{Ax})}{\partial \mathbf{x}} - \frac{\partial(\mathbf{x}^T\mathbf{A}^T\mathbf{b})}{\partial \mathbf{x}} - \frac{\partial(\mathbf{b}^T\mathbf{Ax})}{\partial \mathbf{x}} \\ &= \frac{\partial(\mathbf{x}^T\mathbf{A}^T\mathbf{Ax})}{\partial \mathbf{x}} - 2\frac{\partial(\mathbf{b}^T\mathbf{Ax})}{\partial \mathbf{x}}\end{aligned}$$

- From (C1), $\frac{\partial(\mathbf{b}^T\mathbf{Ax})}{\partial \mathbf{x}} = \frac{\partial\left(\mathbf{A}^T\mathbf{b}\right)^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^T\mathbf{b}$
- From (C2), $\frac{\partial(\mathbf{x}^T\mathbf{A}^T\mathbf{Ax})}{\partial \mathbf{x}} = \frac{\partial\left(\mathbf{A}^T\mathbf{Ax}\right)^T\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial(\mathbf{A}^T\mathbf{Ax})}{\partial \mathbf{x}}\mathbf{x} + \frac{\partial \mathbf{x}}{\partial \mathbf{x}}(\mathbf{A}^T\mathbf{Ax}) = (\mathbf{A}^T\mathbf{A})^T\mathbf{x} + \mathbf{A}^T\mathbf{Ax} = 2\mathbf{A}^T\mathbf{Ax}$
- Hence, $\nabla f(x) = 2\mathbf{A}^T\mathbf{Ax} - 2\mathbf{A}^T\mathbf{b}$

Analytical Solution to Least Squares Problems

- The objective function can be written as

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T\mathbf{A}^T\mathbf{b} - \mathbf{b}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{b} = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{x}^T\mathbf{A}^T\mathbf{b} + \mathbf{b}^T\mathbf{b}$$

- Setting the gradient of f to zero, we obtain

$$\nabla f(x) = 2\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{A}^T\mathbf{b} = 0 \quad \Longrightarrow \quad \mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b} \quad \text{(Normal Equation)}$$

- If $\mathbf{A}$ has full column rank, then $\mathbf{A}^T\mathbf{A}$ is nonsingular, and we have

$$\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$$

- Notice that
  - The normal equation has at least one solution, because the range space of $\mathbf{A}^T\mathbf{A}$ and the range space of $\mathbf{A}^T$ are the same, i.e., $\mathrm{R}(\mathbf{A}^T\mathbf{A}) = \mathrm{R}(\mathbf{A}^T)$.
  - If the column vectors of $\mathbf{A}$ are linearly independent, then the normal equation has exact one solution.

Multivariable Example

- If $\mathbf{x} = [x_1, x_2]^T$, and let $\mathbf{A} = \begin{bmatrix} 1 & -2 \\ -2 & -1 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, then
- $\mathbf{Ax} - \mathbf{b} = \begin{bmatrix} 1 & -2 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} x_1 - 2x_2 - 1 \\ -2x_1 - x_2 + 1 \end{bmatrix}$
- $\min_{\mathbf{x}} \ (\mathbf{Ax} - \mathbf{b})^T \, (\mathbf{Ax} - \mathbf{b}) = \min_{\mathbf{x}} \, (x_1 - 2x_2 - 1)^2 + (-2x_1 - x_2 + 1)^2$
- Method 1 (manually derived, only works for very simple problems):
- $\begin{cases} x_1 - 2x_2 - 1 = 0 \\ -2x_1 - x_2 + 1 = 0 \end{cases} \implies \begin{cases} x_1 = 0.6 \\ x_2 = -0.2 \end{cases}$

Multivariable Example (cont.)

- Method 2 (normal equation, only works for least squares problems)
- $\nabla f(x) = 2\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{A}^T\mathbf{b} = 0 \quad \Longrightarrow \quad \mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$
- $\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 1 & -2 \\ -2 & -1 \end{bmatrix}^T \begin{bmatrix} 1 & -2 \\ -2 & -1 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$, $\mathbf{A}^T\mathbf{b} = \begin{bmatrix} 1 & -2 \\ -2 & -1 \end{bmatrix}^T \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$
- $\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}^{-1} \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \begin{bmatrix} 0.6 \\ -0.2 \end{bmatrix}$

Multivariable Example (cont.)

- Method 3 (gradient descent, works for general optimization problems)
- Consider the following sequence

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \alpha \,\nabla f(\mathbf{x}_n), n \geq 0$$

- Since $-\nabla f(\mathbf{x}_n)$ is the negative gradient at $\mathbf{x}_n$, we have

$$f(\mathbf{x}_0) \geq f(\mathbf{x}_1) \geq f(\mathbf{x}_2) \geq \cdots$$

- In particular,

$$\begin{aligned}\mathbf{x}_{n+1} &= \mathbf{x}_n - \alpha \,\nabla f(x) \\ &= \mathbf{x}_n - \alpha(2\mathbf{A}^T\mathbf{A}\mathbf{x}_n - 2\mathbf{A}^T\mathbf{b}) \\ &= (\mathbf{I} - 2\alpha\mathbf{A}^T\mathbf{A})\mathbf{x}_n + 2\alpha\mathbf{A}^T\mathbf{b}\end{aligned}$$

- The scalar α is called the learning rate.

Gradient Descent

Pseudo code

for k in range(max_iteration):

    $x^{k+1} = x^k - \alpha\nabla f(x^k)$

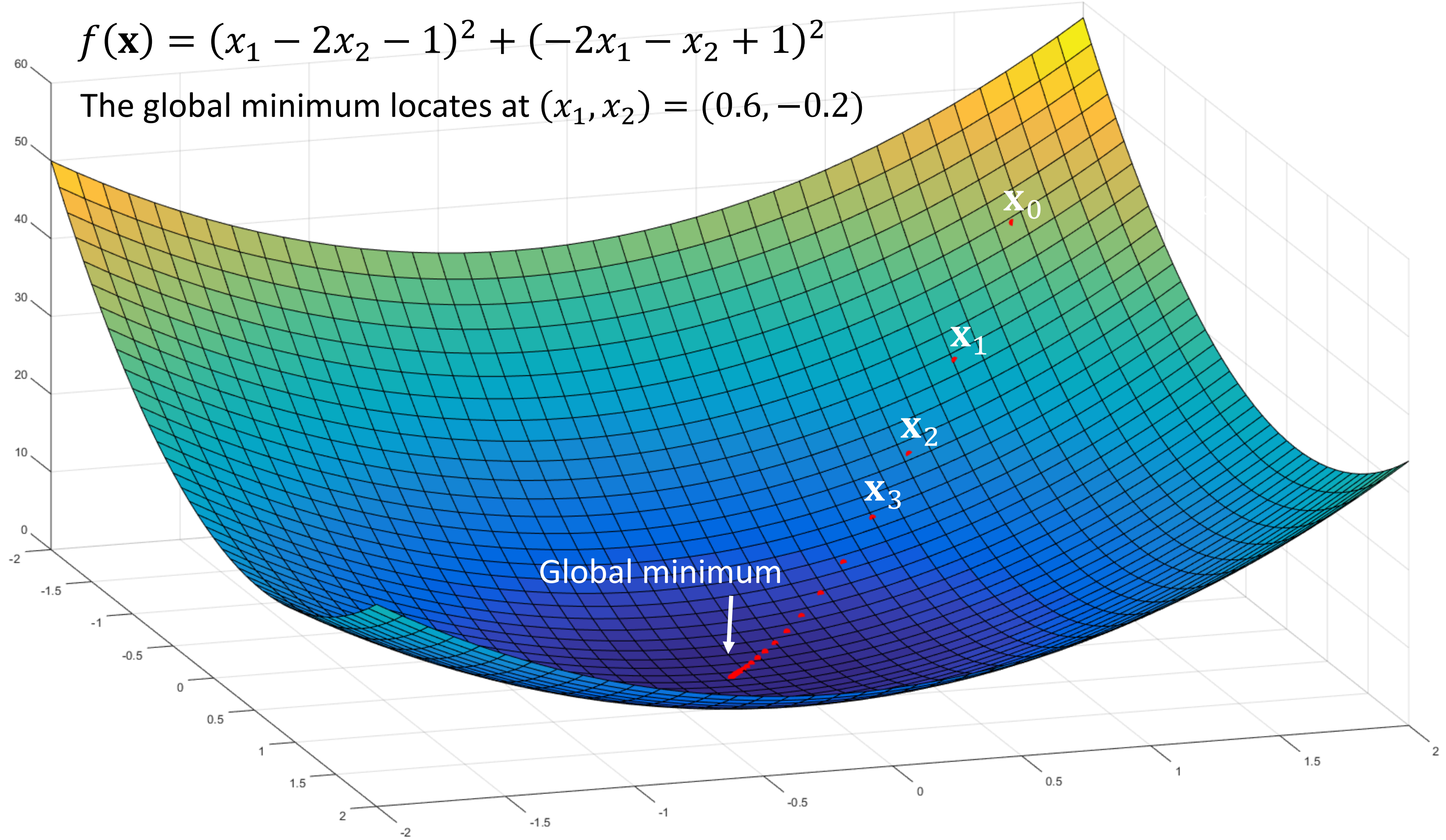
    if $\|x^{k+1} - x^k\|_2 < 10^{-8}$: #stopping criterion

        break

Multivariable Example (cont.)

$$f(\mathbf{x}) = (x_1 - 2x_2 - 1)^2 + (-2x_1 - x_2 + 1)^2$$

The global minimum locates at $(x_1, x_2) = (0.6, -0.2)$



Multivariable Example (cont.)

$\alpha = 0.02$

k	(x_1, x_2)	$f(\mathbf{x})$
0	(-2.00, 2.00)	58.00
1	(-1.48, 1.56)	37.12
2	(-1.06, 1.21)	23.76
3	(-0.73, 0.93)	15.20
4	(-0.46, 0.70)	9.73
5	(-0.25, 0.52)	6.23
⋮	⋮	⋮
29	(0.60,-0.20)	0

$\alpha = 0.01$

k	(x_1, x_2)	$f(\mathbf{x})$
0	(-2.00, 2.00)	58.00
1	(-1.74, 1.78)	46.98
2	(-1.51, 1.58)	38.05
3	(-1.30, 1.40)	30.82
4	(-1.11, 1.24)	24.97
5	(-0.94, 1.10)	20.22
⋮	⋮	⋮
60	(0.60,-0.20)	0

- If the learning rate is too small, then the convergence speed might be slow.
- If the learning rate is too large, then the update sequence cannot converge.

References

[1] Steven J. Leon, *Linear Algebra with Applications*, Pearson, 2015.

[2] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press (http://web.stanford.edu/~boyd/cvxbook/)

[3] Kaare B. Petersen and Michael S. Pedersen, *The Matrix Cookbook* (https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf)