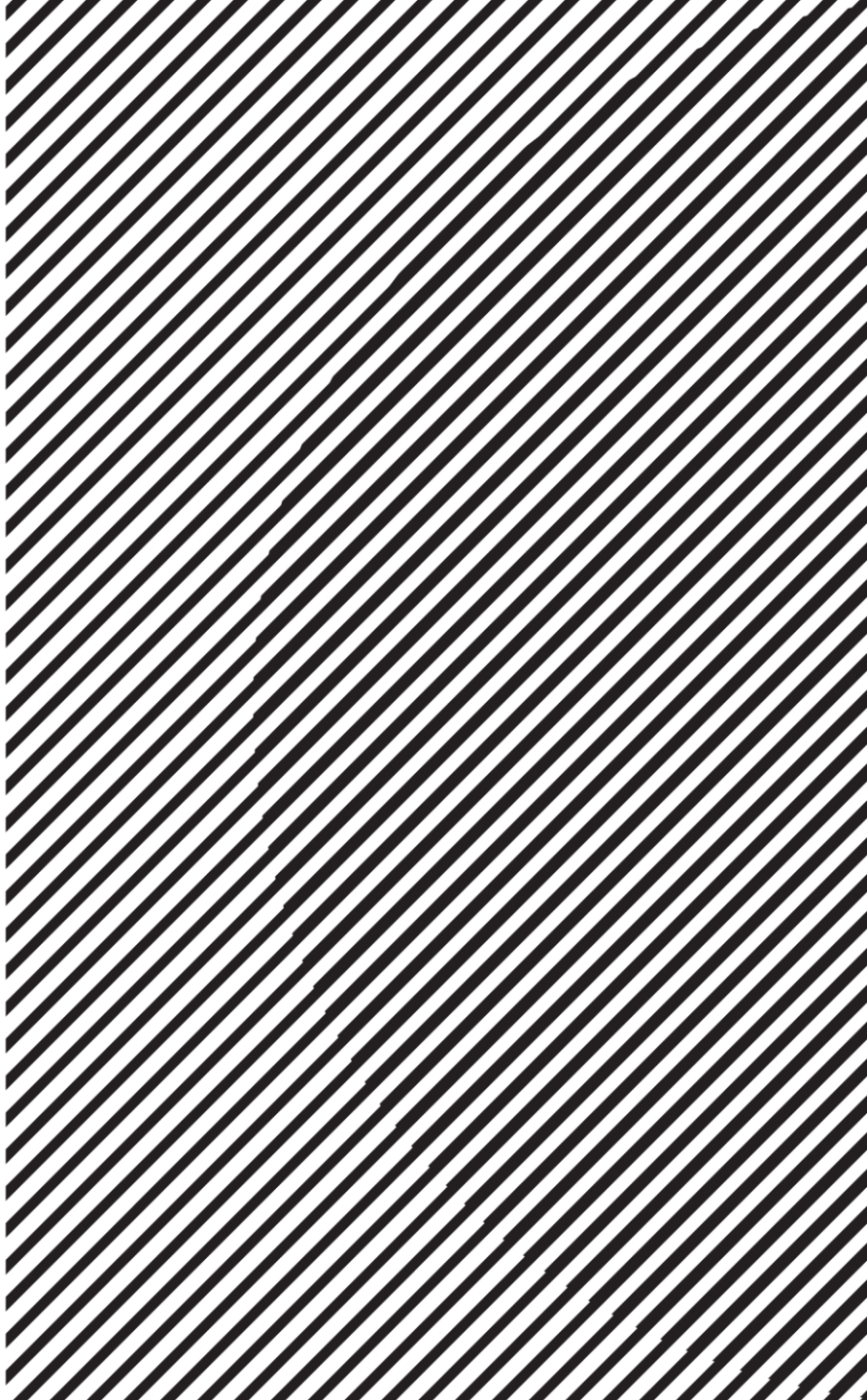
# Linear Algebra

주 재 걸
KAIST 김재철 AI대학원



# References

- Main textbook
  - Lay et al. Linear Algebra and Its Applications, 5th edition, 2015
    - https://www.amazon.com/Linear-Algebra-Its-Applications-5th/dp/032198238X
- Other textbook
  - Gilbert Strang, Introduction to Linear Algebra, 5th edition, 2016
  - Gilbert Strang, Linear Algebra and Its Applications, 4th edition, 2016
- Online lecture
  - Gilbert Strang's MIT Lecture
    - https://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/

# Lecture Overview

- Elements in linear algebra
- Linear system
- Linear combination, vector equation,
  Four views of matrix multiplication
- Linear independence, span, and subspace
- Linear transformation
- Least squares
- Eigendecomposition
- Singular value decomposition
- Principal component analysis

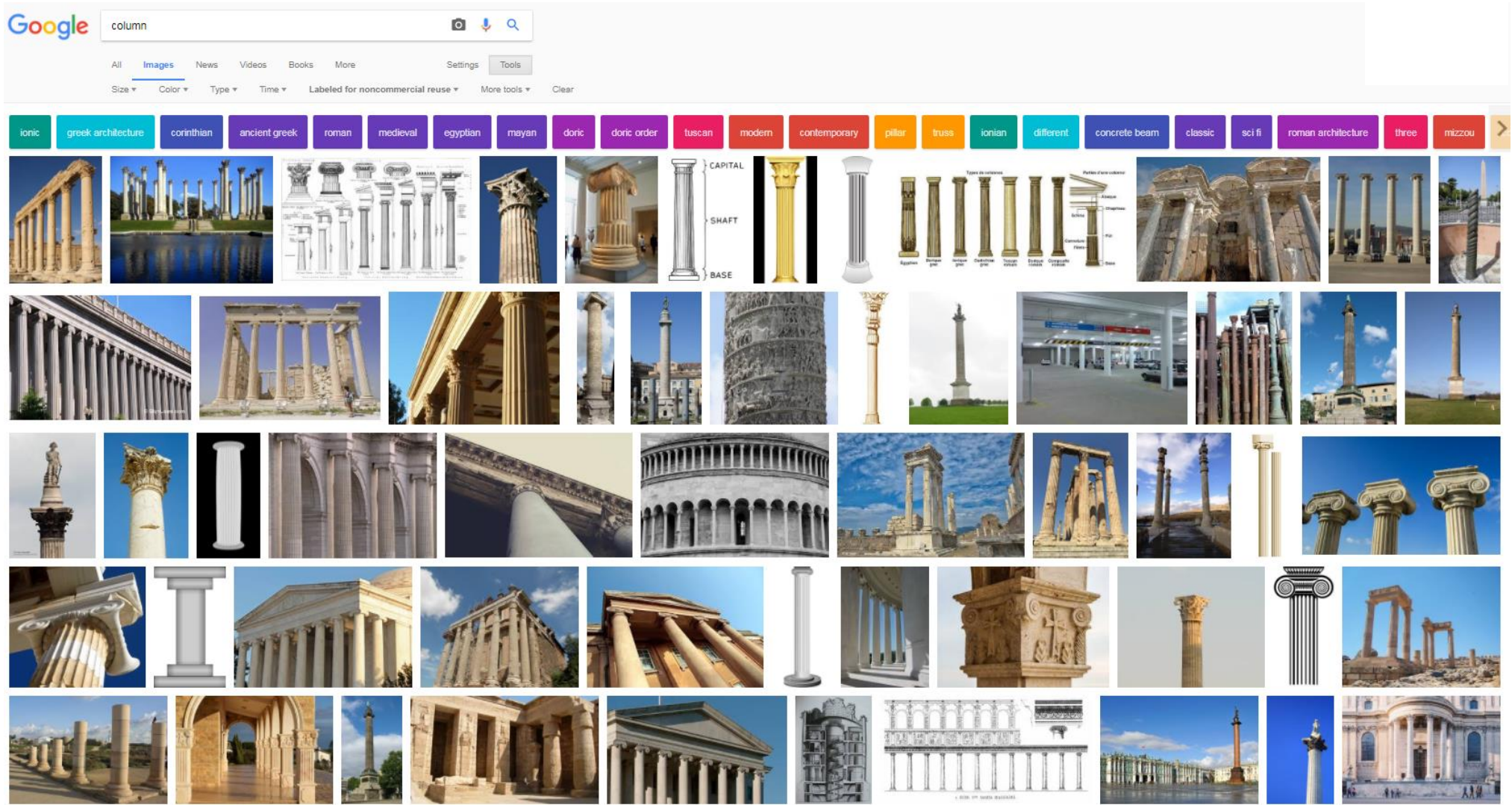
# Lecture Overview

- Elements in linear algebra
- Linear system
- Linear combination, vector equation,
  Four views of matrix multiplication
- Linear independence, span, and subspace
- Linear transformation
- Least squares
- Eigendecomposition
- Singular value decomposition
- Principal component analysis

# Scalar, Vector, and Matrix

- Scalar: a single number $s \in \mathbb{R}$ (lower case), e.g., 3.8
- Vector: an ordered list of numbers, e.g. $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$ (boldface, lower-case), e.g., $\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \in \mathbb{R}^3$
- Matrix: a two-dimensional array of numbers, e.g. $A = \begin{bmatrix} 1 & 6 \\ 3 & 4 \\ 5 & 2 \end{bmatrix} \in \mathbb{R}^{3\times 2}$ (capital letter)
  - Matrix size: $3 \times 2$ means 3 rows and 2 columns
  - Row vector: a horizontal vector
  - Column vector: a vertical vector

# Column is Vertical Vector
# (Don’t be Confused!)



# Column Vector and Row Vector

- A vector of $n$-dimension is usually a column vector, i.e., a matrix of the size $n \times 1$
  - $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n = \mathbb{R}^{n \times 1}$
- Thus, a row vector is usually written as its transpose, i.e.,
  - $\mathbf{x}^T = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}^T = [x_1 \quad x_2 \quad \cdots \quad x_n] \in \mathbb{R}^{1 \times n}$

# Matrix Notations

- $A \in \mathbb{R}^{n \times n}$ : Square matrix (#rows = #columns)
  - e.g., $B = \begin{bmatrix} 1 & 6 \\ 3 & 4 \end{bmatrix}$
- $A \in \mathbb{R}^{m \times n}$ : Rectangular matrix (possible: #rows ≠ #columns)
  - e.g., $A = \begin{bmatrix} 1 & 6 \\ 3 & 4 \\ 5 & 2 \end{bmatrix}$
- $A^T$: Transpose of matrix (mirroring across the main diagonal)
  - e.g., $A^T = \begin{bmatrix} 1 & 3 & 5 \\ 6 & 4 & 2 \end{bmatrix}$
- $A_{ij}$ : $(i,j)$-th component of $A$, e.g., $A_{2,1} = 3$
- $A_{i,:}$ : $i$-th row vector of $A$, e.g., $A_{2,:} = \begin{bmatrix} 3 & 4 \end{bmatrix}$
- $A_{:,j}$ : $j$-th column vector of $A$, e.g., $A_{:,2} = \begin{bmatrix} 6 \\ 4 \\ 2 \end{bmatrix}$

# Vector/Matrix Additions and Multiplications

- $C = A + B$ : Element-wise addition, i.e., $C_{ij} = A_{ij} + B_{ij}$
  - $A, B, C$ should have the same size, i.e., $A, B, C \in \mathbb{R}^{m \times n}$
- $c\mathbf{a}$, $cA$ : Scalar multiple of vector/matrix
  - e.g., $2\begin{bmatrix}3\\2\\1\end{bmatrix} = \begin{bmatrix}6\\4\\2\end{bmatrix}$, $2\begin{bmatrix}1 & 6\\3 & 4\\5 & 2\end{bmatrix} = \begin{bmatrix}2 & 12\\6 & 8\\10 & 4\end{bmatrix}$
- $C = AB$ : Matrix-matrix multiplication, i.e., $C_{ij} = \sum_k A_{i,k} B_{k,j}$
  - e.g., $\begin{bmatrix}1 & 6\\3 & 4\\5 & 2\end{bmatrix}\begin{bmatrix}1 & -1\\2 & 1\end{bmatrix} = \begin{bmatrix}13 & 5\\11 & 1\\9 & -3\end{bmatrix}$, $\begin{bmatrix}3 & 2 & 1\end{bmatrix}\begin{bmatrix}1\\3\\5\end{bmatrix} = [14]$, $\begin{bmatrix}1\\3\\5\end{bmatrix}\begin{bmatrix}1 & 2\end{bmatrix} = \begin{bmatrix}1 & 2\\3 & 6\\5 & 10\end{bmatrix}$

Size: $(3 \times 2)(2 \times 2) = 3 \times 2$, $(1 \times 3)(3 \times 1) = 1 \times 1$, $(3 \times 1)(1 \times 2) = 3 \times 2$

# Matrix multiplication is NOT commutative

$AB \neq BA$: Matrix multiplication is NOT commutative.

- e.g., Given $A \in \mathbb{R}^{2\times 3}$ and $B \in \mathbb{R}^{3\times 5}$, $AB$ is defined, but $BA$ is not even defined.
- What if $BA$ is defined, e.g., $A \in \mathbb{R}^{2\times 3}$ and $B \in \mathbb{R}^{3\times 2}$? Still, the sizes of $AB \in \mathbb{R}^{2\times 2}$and $BA \in \mathbb{R}^{3\times 3}$ does not match, so $AB \neq BA$.
- What if the sizes of $AB$ and $BA$ match, e.g., $A \in \mathbb{R}^{2\times 2}$ and $B \in \mathbb{R}^{2\times 2}$? Still in this case, generally, $AB \neq BA$.
- E.g., $\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}\begin{bmatrix}5 & 6\\ 7 & 8\end{bmatrix} = \begin{bmatrix}19 & 22\\ 43 & 50\end{bmatrix}$

  $\begin{bmatrix}5 & 6\\ 7 & 8\end{bmatrix}\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix} = \begin{bmatrix}23 & 34\\ 31 & 46\end{bmatrix}$

# Other Properties

- $A(B+C) = AB + AC$ : Distributive

$Left$: $\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}\left(\begin{bmatrix}4 & 3\\ 2 & 1\end{bmatrix}+\begin{bmatrix}1 & 3\\ 5 & 7\end{bmatrix}\right) = \begin{bmatrix}19 & 22\\ 43 & 50\end{bmatrix}$ $Right$: $\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}\begin{bmatrix}4 & 3\\ 2 & 1\end{bmatrix}+\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}\begin{bmatrix}1 & 3\\ 5 & 7\end{bmatrix} = \begin{bmatrix}19 & 22\\ 43 & 50\end{bmatrix}$

- $A(BC) = (AB)C$ : Associative

$\begin{bmatrix}1 & 6\\ 3 & 4\\ 5 & 2\end{bmatrix}\begin{bmatrix}1\\ 2\end{bmatrix}\begin{bmatrix}3 & 2 & 1\end{bmatrix} \Rightarrow$ $Left$: $\begin{bmatrix}1 & 6\\ 3 & 4\\ 5 & 2\end{bmatrix}\begin{bmatrix}3 & 2 & 1\\ 6 & 4 & 2\end{bmatrix} = \begin{bmatrix}39 & 26 & 13\\ 33 & 22 & 11\\ 27 & 18 & 9\end{bmatrix}$ $Right$: $\begin{bmatrix}13\\ 11\\ 9\end{bmatrix}\begin{bmatrix}3 & 2 & 1\end{bmatrix} = \begin{bmatrix}39 & 26 & 13\\ 33 & 22 & 11\\ 27 & 18 & 9\end{bmatrix}$

$(3\times 2)\,(2\times 1)\,(1\times 3)$ $(3\times 2)\ (2\times 3)$ $(3\times 3)$ $(3\times 1)\,(1\times 3)$ $(3\times 3)$

- $(AB)^T = B^T A^T$ : Property of transpose

$Left$: $\left(\begin{bmatrix}1 & 6\\ 3 & 4\\ 5 & 2\end{bmatrix}\begin{bmatrix}1\\ 2\end{bmatrix}\begin{bmatrix}3 & 2 & 1\end{bmatrix}\right)^T = \begin{bmatrix}39 & 26 & 13\\ 33 & 22 & 11\\ 27 & 18 & 9\end{bmatrix}^T$ $Right$: $\begin{bmatrix}3\\ 2\\ 1\end{bmatrix}\begin{bmatrix}1 & 2\end{bmatrix}\begin{bmatrix}1 & 3 & 5\\ 6 & 4 & 2\end{bmatrix} = \begin{bmatrix}39 & 33 & 27\\ 26 & 22 & 18\\ 13 & 11 & 9\end{bmatrix}$

$(3\times 2)(2\times 1)(1\times 3)$ $(3\times 3)$ $(3\times 1)(1\times 2)$ $(2\times 3)$ $(3\times 3)$

# Lecture Overview

- Elements in linear algebra
- Linear system
- Linear combination, vector equation,
  Four views of matrix multiplication
- Linear independence, span, and subspace
- Linear transformation
- Least squares
- Eigendecomposition
- Singular value decomposition
- Principal component analysis

# Linear Equation

- A **linear equation** in the variables $x_1, \cdots, x_n$ is an equation that can be written in the form

$$a_1x_1+a_2x_2+\cdots+a_nx_n = b,$$

where $b$ and the coefficients $a_1, \cdots, a_n$ are real or complex numbers that are usually known in advance.

- The above equation can be written as

$$\mathbf{a}^T\mathbf{x} = b$$

where $\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$ and $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$.

ex) $3x + 5y = 4$

$$a = \begin{bmatrix} 3 \\ 5 \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\mathbf{a}^T\mathbf{x} = \begin{bmatrix} 3 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= 3x_1 + 5x_2 = \mathbf{x}^T\mathbf{a}$$

$x + 2y + 3xy = 5$ is linear equation?

→ No, $3xy$ is a quadratic term.

(All terms need to be linear!)

# Linear System: Set of Equations

- A **system of linear equations** (or a **linear system**) is a collection of one or more linear equations involving the same variables - say, $x_1, \cdots, x_n$.

# Linear System Example

- Suppose we collected persons' weight, height, and life-span (e.g., how long s/he lived)

| Person ID | Weight | Height | Is_smoking | Life-span |
|---|---|---|---|---|
| 1 | 60kg | 5.5ft | Yes (=1) | 66 |
| 2 | 65kg | 5.0ft | No (=0) | 74 |
| 3 | 55kg | 6.0ft | Yes (=1) | 78 |

- We want to set up the following linear system:

$$60x_1 + 5.5x_2 + 1 \cdot x_3 = 66$$
$$65x_1 + 5.0x_2 + 0 \cdot x_3 = 74$$
$$55x_1 + 6.0x_2 + 1 \cdot x_3 = 78$$

coefficient →

- Once we solve for $x_1$, $x_2$, and $x_3$, given a new person with his/her weight, height, and is_smoking, we can predict his/her life-span.

# Linear System Example

- The essential information of a linear system can be written compactly using a **matrix**.
- In the following set of equations,

$$60x_1 + 5.5x_2 + 1 \cdot x_3 = 66$$
$$65x_1 + 5.0x_2 + 0 \cdot x_3 = 74$$
$$55x_1 + 6.0x_2 + 1 \cdot x_3 = 78$$

- Let's collect all the coefficients on the left and form a matrix

$$A = \begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \end{bmatrix}$$

- Also, let's form two vectors: $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 66 \\ 74 \\ 78 \end{bmatrix}$

# From Multiple Equations to Single Matrix Equation

- Multiple equations can be converted into a **single** matrix equations

$$\begin{aligned} 60x_1+5.5x_2+1\cdot x_3&=66\\ 65x_1+5.0x_2+0\cdot x_3&=74\\ 55x_1+6.0x_2+1\cdot x_3&=78 \end{aligned} \quad\Rightarrow\quad \begin{bmatrix} 60 & 5.5 & 1\\ 65 & 5.0 & 0\\ 55 & 6.0 & 1 \end{bmatrix}\begin{bmatrix} x_1\\ x_2\\ x_3 \end{bmatrix}=\begin{bmatrix} 66\\ 74\\ 78 \end{bmatrix} \quad\Leftarrow\quad \begin{aligned} \mathbf{a}_1^T\mathbf{x}&=66\\ \mathbf{a}_2^T\mathbf{x}&=74\\ \mathbf{a}_3^T\mathbf{x}&=78 \end{aligned}$$

$$A \qquad \mathbf{x} = \mathbf{b}$$

- How can we solve for $\mathbf{x}$?

In other words,

$$\begin{bmatrix} 60 & 5.5 & 1 \end{bmatrix}\begin{bmatrix} x_1\\ x_2\\ x_3 \end{bmatrix}=66=\mathbf{a}_1^T\mathbf{x}$$

$$\begin{bmatrix} 65 & 5.0 & 0 \end{bmatrix}\begin{bmatrix} x_1\\ x_2\\ x_3 \end{bmatrix}=74=\mathbf{a}_2^T\mathbf{x}$$

$$\begin{bmatrix} 55 & 6.0 & 1 \end{bmatrix}\begin{bmatrix} x_1\\ x_2\\ x_3 \end{bmatrix}=78=\mathbf{a}_3^T\mathbf{x}$$

# Identity Matrix

- **Definition**: An identity matrix is a square matrix whose diagonal entries are all 1's, and all the other entries are zeros. Often, we denote it as $I_n \in \mathbb{R}^{n \times n}$.
  - e.g., $I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
- An identity matrix $I_n$ preserves any vector $\mathbf{x} \in \mathbb{R}^n$ after multiplying $\mathbf{x}$ by $I_n$ :

$$\forall \mathbf{x} \in \mathbb{R}^n, \qquad I_n \mathbf{x} = \mathbf{x}$$

# Inverse Matrix

- **Definition**: For a square matrix $A \in \mathbb{R}^{n \times n}$, its inverse matrix $A^{-1}$ is defined such that

$$A^{-1}A = AA^{-1} = I_n.$$

- For a $2 \times 2$ matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, its inverse matrix $A^{-1}$ is defined as

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

# Inverse Matrix

For matrix $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \in \mathbb{R}^{2\times 2}$, its inverse matrix $A^{-1}$ is defined as

$$A^{-1} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}^{-1} = \frac{1}{(1)(4)-(2)(3)} \begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$$

$$= -\frac{1}{2} \begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$$

Both equations hold true:

- $AA^{-1} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2$
- $A^{-1}A = \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2$

# Solving Linear System via Inverse Matrix

- We can now solve $A\mathbf{x} = \mathbf{b}$ as follows:

$$A\mathbf{x} = \mathbf{b}$$
$$A^{-1}A\mathbf{x} = A^{-1}\mathbf{b}$$
$$I_n\mathbf{x} = A^{-1}\mathbf{b}$$
$$\mathbf{x} = A^{-1}\mathbf{b}$$

# Solving Linear System via Inverse Matrix

- **Example:**

$$\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 66 \\ 74 \\ 78 \end{bmatrix} \Rightarrow A^{-1} = \begin{bmatrix} 0.0870 & 0.0087 & -0.0870 \\ -1.1304 & 0.0870 & 1.1314 \\ 2.0000 & -1.0000 & -1.0000 \end{bmatrix}$$

$A$ $\quad\quad$ $\mathbf{x} = \mathbf{b}$

- One can verify

$$A^{-1}A = AA^{-1} = I_n.$$

- $\mathbf{x} = A^{-1}\mathbf{b} = \begin{bmatrix} 0.0870 & 0.0087 & -0.0870 \\ -1.1304 & 0.0870 & 1.1314 \\ 2.0000 & -1.0000 & -1.0000 \end{bmatrix} \begin{bmatrix} 66 \\ 74 \\ 78 \end{bmatrix} = \begin{bmatrix} -0.4 \\ 20 \\ -20 \end{bmatrix}$

# Solving Linear System via Inverse Matrix

- Now, the life-span can be written as

$$\text{(life-span)} = -0.4 \times \text{(weight)} + 20 \times \text{(height)} - 20 \times \text{(is\_smoking)}.$$

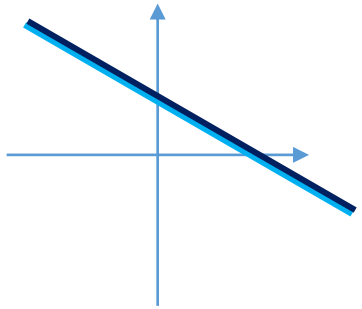
# Non-Invertible Matrix $A$ for $A\mathbf{x} = \mathbf{b}$

- Note that **if** $A$ is invertible, the solution is uniquely obtained as
$$\mathbf{x} = A^{-1}\mathbf{b}.$$
- **What if** $A$ is non-invertible, i.e., the inverse does not exist?
  - E.g., For $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$, in $A^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$, the denominator $ad - bc = 0$, so $A$ is not invertible.

- For $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, $ad - bc$ is called the determinant of $A$, or $\det A$.

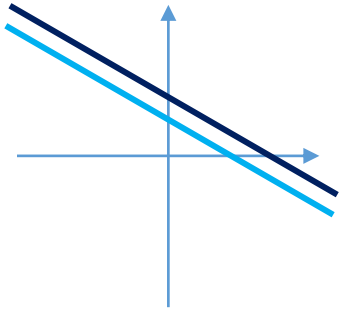
# Non-Invertible Matrix $A$ for $A\mathbf{x} = \mathbf{b}$

- e.g., $A\mathbf{x} = \mathbf{b}$, $\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 12 \end{bmatrix} = \begin{cases} x + 2y = 4 \\ 3x + 6y = 12 \end{cases}$



Overlapping
(Infinitely Many Solutions)

- e.g., $A\mathbf{x} = \mathbf{b}$, $\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 13 \end{bmatrix} = \begin{cases} x + 2y = 4 \\ 3x + 6y = 13 \end{cases}$



Parallel
(No Solution)

# Non-Invertible Matrix $A$ for $A\mathbf{x} = \mathbf{b}$

- Thus, if $A$ is non-invertible, $A\mathbf{x} = \mathbf{b}$ will have either no solution or infinitely many solutions.

# Does a Matrix Have an Inverse Matrix?

- For $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \in \mathbb{R}^{2\times 2}$, $ad - bc$ determines whether the given matrix is invertible or now, and this number is called the determinant of $A$, or $\det A$.

- That is, $\det A$ determines whether $A$ is invertible (when $\det A \neq 0$) or not (when $\det A = 0$).

- For more details on how to compute the determinant of a matrix $A \in \mathbb{R}^{n\times n}$ where $n \geq 3$, you can study the following:
  - https://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/video-lectures/lecture-18-properties-of-determinants/
  - https://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/video-lectures/lecture-19-determinant-formulas-and-cofactors/

# Does a Matrix Have an Inverse Matrix?

- For matrix $A \in \mathbb{R}^{3\times 3}$,

$$\det\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = \begin{vmatrix} a & b & c & a & b \\ d & e & f & d & e \\ g & h & i & g & h \end{vmatrix}$$

$$= aei + bfg + cdh - ceg - afh - bdi$$

- Alternatively, it can be rewritten as

$$\det\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix}$$

# Obtaining Inverse Matrix

- If invertible, is there any formula for computing an inverse matrix of a matrix $A \in \mathbb{R}^{n \times n}$ where $n \geq 3$?

- The closed-form solution does not exist, but one can compute it.
- How to compute it can be found at Gaussian elimination in Lay Ch1.2 and Lay Ch2.2.

or

# Rectangular Matrix $A$ in $A\mathbf{x} = \mathbf{b}$

- **What if** $A$ is a rectangular matrix, e.g., $A \in \mathbb{R}^{m \times n}$, where $m \neq n$?

| Person ID | Weight | Height | Is_smoking | Life-span |
|---|---|---|---|---|
| 1 | 60kg | 5.5ft | Yes (=1) | 66 |
| 2 | 65kg | 5.0ft | No (=0) | 74 |
| 3 | 55kg | 6.0ft | Yes (=1) | 78 |

$$\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 66 \\ 74 \\ 78 \end{bmatrix}$$

- Recall $m$ = #equations and $n$ = #variables. $A$ $\mathbf{x} = \mathbf{b}$
- $m < n$: more variables than equations
  - Usually infinitely many solutions exist (under-determined system).
- $m > n$: more equations than variables
  - Usually no solution exists (over-determined system).
- To study how to compute the solution in these general cases, check out Lay Ch1.2 and Lay Ch1.5.

# Rectangular Matrix $A$ in $A\mathbf{x} = \mathbf{b}$

- For under-determined system, our prediction model can be written in many different forms:
  - Solution 1:

$$(\text{life-span}) = -0.4 \times (\text{weight}) + 20 \times (\text{height}) - 20 \times (\text{is\_smoking})$$

  - Solution 2:

$$(\text{life-span}) = -1.2 \times (\text{weight}) + 25 \times (\text{height}) - 15 \times (\text{is\_smoking})$$

  - Solution 3:

$$(\text{life-span}) = -1.2 \times (\text{weight}) + 16 \times (\text{height}) - 25 \times (\text{is\_smoking})$$

- In such a case where we have infinitely many solutions, we often utilize regularization, e.g., L1 or L2 regularization, which work as a method for risk management.

# Lecture Overview

- Elements in linear algebra
- Linear system
- Linear combination, vector equation,
  Four views of matrix multiplication
- Linear independence, span, and subspace
- Linear transformation
- Least squares
- Eigendecomposition
- Singular value decomposition
- Principal component analysis

# Linear Combinations

- Given vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p$ in $\mathbb{R}^n$ and given scalars $c_1, c_2, \cdots, c_p$,

$$c_1\mathbf{v}_1 + \cdots + c_p\mathbf{v}_p$$

  is called a **linear combination** of $\mathbf{v}_1, \cdots, \mathbf{v}_p$ with **weights or coefficients** $c_1, \cdots, c_p$.

- The weights in a linear combination can be any real numbers, including zero.

# From Matrix Equation to Vector Equation

- Recall the matrix equation of a linear system:

| Person ID | Weight | Height | Is_smoking | Life-span |
|---|---|---|---|---|
| 1 | 60kg | 5.5ft | Yes (=1) | 66 |
| 2 | 65kg | 5.0ft | No (=0) | 74 |
| 3 | 55kg | 6.0ft | Yes (=1) | 78 |

$$\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 66 \\ 74 \\ 78 \end{bmatrix}$$

$$A \qquad \mathbf{x} = \mathbf{b}$$

- A matrix equation can be converted into a vector equation:

$$\begin{bmatrix} 60 \\ 65 \\ 55 \end{bmatrix} x_1 + \begin{bmatrix} 5.5 \\ 5.0 \\ 6.0 \end{bmatrix} x_2 + \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} x_3 = \begin{bmatrix} 66 \\ 74 \\ 78 \end{bmatrix}$$

$$\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \mathbf{a}_3 x_3 = \mathbf{b}$$

# Existence of Solution for $A\mathbf{x} = \mathbf{b}$

- Consider its vector equation:

$$\begin{bmatrix}60\\65\\55\end{bmatrix}x_1 + \begin{bmatrix}5.5\\5.0\\6.0\end{bmatrix}x_2 + \begin{bmatrix}1\\0\\1\end{bmatrix}x_3 = \begin{bmatrix}66\\74\\78\end{bmatrix}$$

$$\mathbf{a}_1x_1 + \mathbf{a}_2x_2 + \mathbf{a}_3x_3 = \mathbf{b}$$

- When does the solution exist for $A\mathbf{x} = \mathbf{b}$?

# Span

- **Definition**: Given a set of vectors $\mathbf{v}_1, \cdots, \mathbf{v}_p \in \mathbb{R}^n$, Span $\{\mathbf{v}_1, \cdots, \mathbf{v}_p\}$ is defined as the set of all linear combinations of $\mathbf{v}_1, \cdots, \mathbf{v}_p$.
- That is, Span $\{\mathbf{v}_1, \cdots, \mathbf{v}_p\}$ is the collection of all vectors that can be written in the form

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 \cdots + c_p\mathbf{v}_p$$

with arbitrary scalars $c_1, \cdots, c_p$.

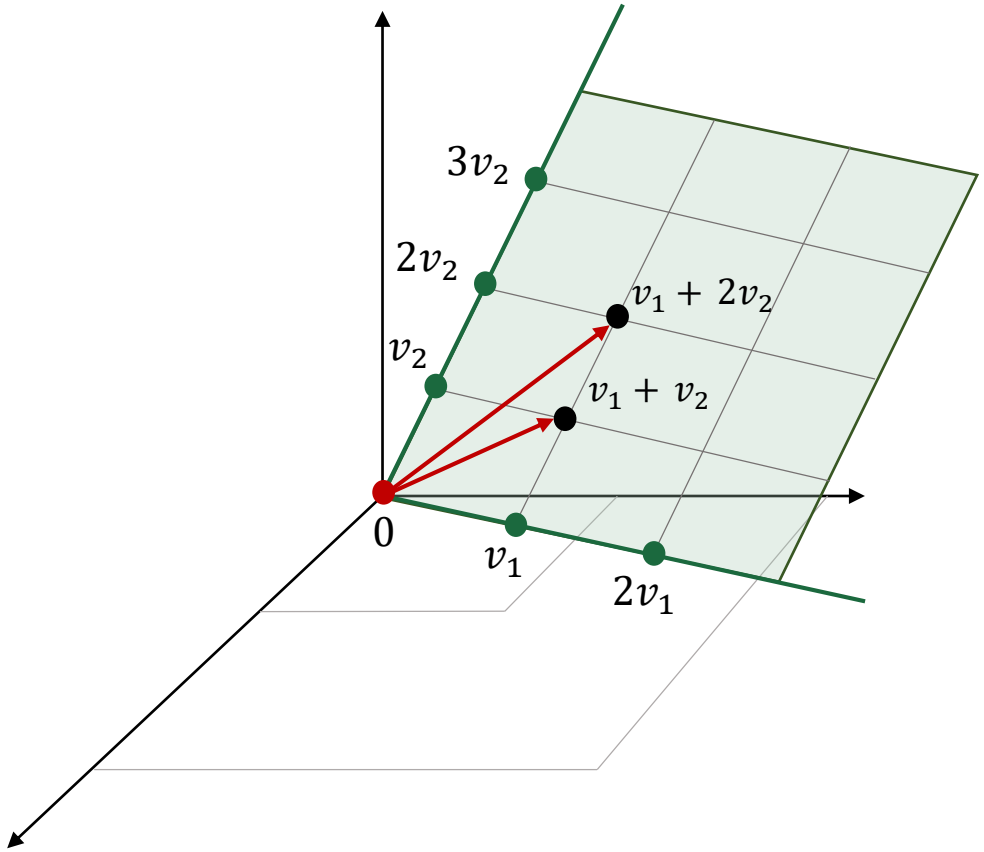
- Span $\{\mathbf{v}_1, \cdots, \mathbf{v}_p\}$ is also called the **subset of** $\mathbb{R}^n$ **spanned** (or **generated**) by $\mathbf{v}_1, \cdots, \mathbf{v}_p$.

# Span

- e.g., For $c_1 = 1, c_2 = 1, c_1 \begin{bmatrix}1\\2\\3\end{bmatrix} + c_2 \begin{bmatrix}2\\3\\4\end{bmatrix} = \begin{bmatrix}\mathbf{3}\\\mathbf{5}\\\mathbf{7}\end{bmatrix}$.
- For $c_1 = 1, c_2 = 0, c_1 \begin{bmatrix}1\\2\\3\end{bmatrix} + c_2 \begin{bmatrix}2\\3\\4\end{bmatrix} = \begin{bmatrix}\mathbf{1}\\\mathbf{2}\\\mathbf{3}\end{bmatrix}$.
- For $c_1 = 0, c_2 = 1, c_1 \begin{bmatrix}1\\2\\3\end{bmatrix} + c_2 \begin{bmatrix}2\\3\\4\end{bmatrix} = \begin{bmatrix}\mathbf{2}\\\mathbf{3}\\\mathbf{4}\end{bmatrix}$.
- For $c_1 = 0, c_2 = 0, c_1 \begin{bmatrix}1\\2\\3\end{bmatrix} + c_2 \begin{bmatrix}2\\3\\4\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\\mathbf{0}\\\mathbf{0}\end{bmatrix}$.
- Therefore, $\text{Span}\left\{\begin{bmatrix}1\\2\\3\end{bmatrix}, \begin{bmatrix}2\\3\\4\end{bmatrix}\right\} = \left\{\begin{bmatrix}3\\5\\7\end{bmatrix}, \begin{bmatrix}1\\2\\3\end{bmatrix}, \begin{bmatrix}2\\3\\4\end{bmatrix}, \begin{bmatrix}0\\0\\0\end{bmatrix}, \ldots\right\}$

# Geometric Description of Span

- If $\mathbf{v}_1$ are $\mathbf{v}_2$ nonzero vectors in $\mathbb{R}^3$, with $\mathbf{v}_2$ not a multiple of $\mathbf{v}_1$, then Span $\{\mathbf{v}_1, \mathbf{v}_2\}$ is the plane in $\mathbb{R}^3$ that contains $\mathbf{v}_1, \mathbf{v}_2$ and $\mathbf{0}$.
- In particular, Span $\{\mathbf{v}_1, \mathbf{v}_2\}$ contains the line in $\mathbb{R}^3$ through $\mathbf{v}_1$ and $\mathbf{0}$ and the line through $\mathbf{v}_2$ and $\mathbf{0}$.



# Geometric Interpretation of Vector Equation

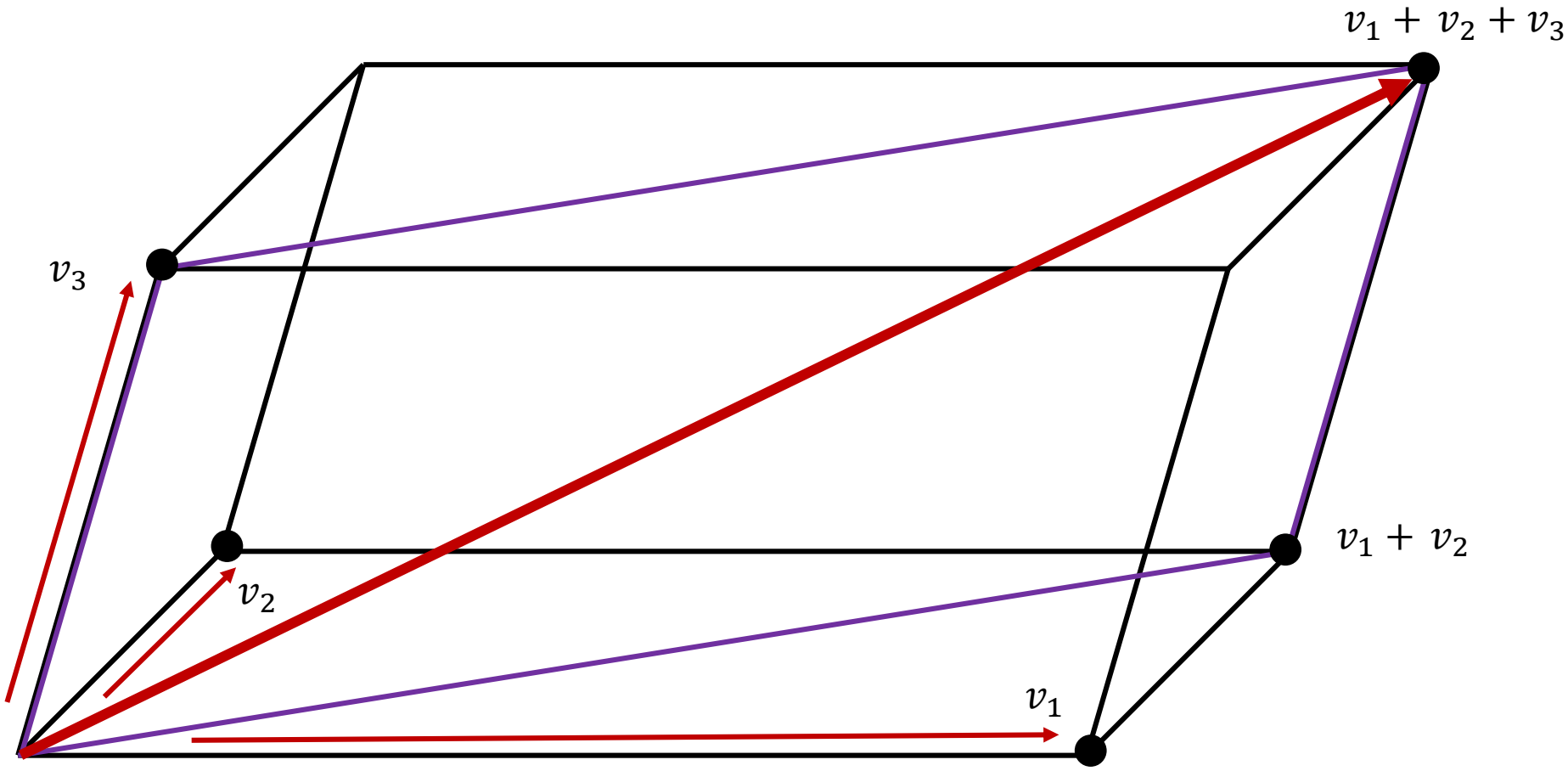
- Finding a linear combination of given vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ to be equal to $\mathbf{b}$:

$$\begin{bmatrix}60\\65\\55\end{bmatrix}x_1 + \begin{bmatrix}5.5\\5.0\\6.0\end{bmatrix}x_2 + \begin{bmatrix}1\\0\\1\end{bmatrix}x_3 = \begin{bmatrix}66\\74\\78\end{bmatrix}$$

$$\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \mathbf{a}_3 x_3 = \mathbf{b}$$

- The solution exists only when $\mathbf{b} \in$ Span $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$.

# Parallelepiped



# Matrix Multiplications as Linear Combinations of Vectors

- **Recall**: we defined matrix-matrix multiplications as the inner product between the row on the left and the column on the right:
  - e.g., $\begin{bmatrix} 1 & 6 \\ 3 & 4 \\ 5 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 13 & 5 \\ 11 & 1 \\ 9 & -3 \end{bmatrix}$
- Inspired by the vector equation, we can view $A\mathbf{x}$ as a linear combination of columns of the left matrix:
  - $\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = A\mathbf{x} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \mathbf{a}_3 x_3$

# Matrix Multiplications as Column Combinations

- Linear combinations of columns
  - Left matrix: bases, right matrix: coefficients

***One column on the right***

$$\begin{bmatrix}1 & 1 & 0\\ 1 & 0 & 1\\ 1 & -1 & 1\end{bmatrix}\begin{bmatrix}1\\2\\3\end{bmatrix}=\begin{bmatrix}1\\1\\1\end{bmatrix}1+\begin{bmatrix}1\\0\\-1\end{bmatrix}2+\begin{bmatrix}0\\1\\1\end{bmatrix}3$$

***Multi-columns on the right***

$$\begin{bmatrix}1 & 1 & 0\\ 1 & 0 & 1\\ 1 & -1 & 1\end{bmatrix}\begin{bmatrix}1 & -1\\2 & 0\\3 & 1\end{bmatrix}=\begin{bmatrix}x_1\ y_1\\ x_2\ y_2\\ x_3\ y_3\end{bmatrix}=[\mathbf{x}\ \mathbf{y}]$$

$$\mathbf{x}=\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}1\\1\\1\end{bmatrix}1+\begin{bmatrix}1\\0\\-1\end{bmatrix}2+\begin{bmatrix}0\\1\\1\end{bmatrix}3$$

$$\mathbf{y}=\begin{bmatrix}y_1\\y_2\\y_3\end{bmatrix}=\begin{bmatrix}1\\1\\1\end{bmatrix}(-1)+\begin{bmatrix}1\\0\\-1\end{bmatrix}0+\begin{bmatrix}0\\1\\1\end{bmatrix}1$$

# Matrix Multiplications as Row Combinations

- Linear combinations of rows of the right matrix
  - Right matrix: bases, left matrix: coefficients

***One row on the left***

$$\begin{bmatrix}1 & 2 & 3\end{bmatrix}\begin{bmatrix}1 & 1 & 0\\ 1 & 0 & 1\\ 1 & -1 & 1\end{bmatrix} = \begin{matrix}1\times[1 & 1 & 0]\\ +2\times[1 & 0 & 1]\\ +3\times[1 & -1 & 1]\end{matrix}$$

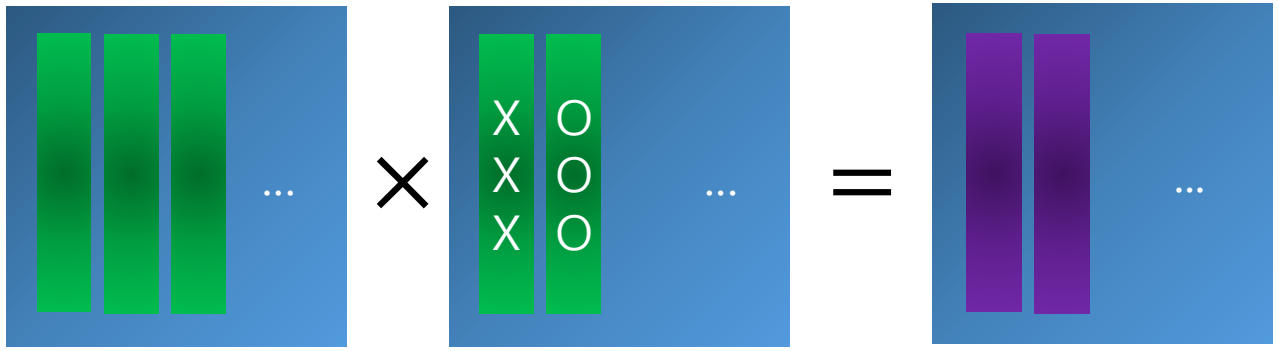
***Multiple rows on the left***

$$\begin{bmatrix}1 & 2 & 3\\ 1 & 0 & -1\end{bmatrix}\begin{bmatrix}1 & 1 & 0\\ 1 & 0 & 1\\ 1 & -1 & 1\end{bmatrix} = \begin{bmatrix}x_1 & x_2 & x_3\\ y_1 & y_2 & y_3\end{bmatrix} = \begin{bmatrix}\mathbf{x}^{\mathrm{T}}\\ \mathbf{y}^{\mathrm{T}}\end{bmatrix}$$

$$\mathbf{x}^{\mathrm{T}} = [x_1 \quad x_2 \quad x_3] = 1[1 \quad 1 \quad 0] + 2[1 \quad 0 \quad 1] + \quad 3[1 \quad -1 \quad 1]$$

$$\mathbf{y}^{\mathrm{T}} = [y_1 \quad y_2 \quad y_3] = 1[1 \quad 1 \quad 0] + 0[1 \quad 0 \quad 1] + (-1)[1 \quad -1 \quad 1]$$

# Matrix Multiplications as Row Combinations




# Matrix Multiplications as Sum of (Rank-1) Outer Products

- (Rank-1) outer product

$$\begin{bmatrix}1\\1\\1\end{bmatrix}\begin{bmatrix}1 & 2 & 3\end{bmatrix}=\begin{bmatrix}1 & 2 & 3\\1 & 2 & 3\\1 & 2 & 3\end{bmatrix}$$

- Sum of (Rank-1) outer products

$$\begin{bmatrix}1 & 1\\1 & -1\\1 & 1\end{bmatrix}\begin{bmatrix}1 & 2 & 3\\4 & 5 & 6\end{bmatrix}=\begin{bmatrix}1\\1\\1\end{bmatrix}\begin{bmatrix}1 & 2 & 3\end{bmatrix}+\begin{bmatrix}1\\-1\\1\end{bmatrix}\begin{bmatrix}4 & 5 & 6\end{bmatrix}$$

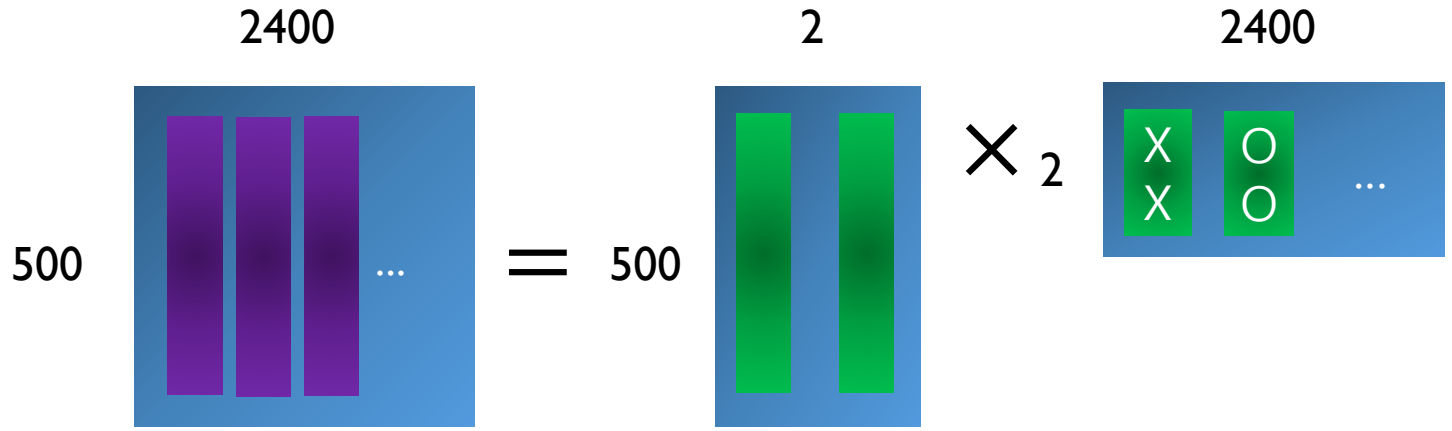
$$=\begin{bmatrix}1 & 2 & 3\\1 & 2 & 3\\1 & 2 & 3\end{bmatrix}+\begin{bmatrix}4 & 5 & 6\\-4 & -5 & -6\\4 & 5 & 6\end{bmatrix}$$

# Matrix Multiplications as Sum of (Rank-1) Outer Products

- Sum of (Rank-1) outer products is widely used in machine learning
  - Covariance matrix in multivariate Gaussian
  - Gram matrix in style transfer

# Matrix Multiplications as Sum of (Rank-1) Outer Products

- Low-rank matrix factorization




e.g. $x^2 - 2x - 3 = (x-3)(x+1)$

# Lecture Overview

- Elements in linear algebra
- Linear system
- Linear combination, vector equation,
  Four views of matrix multiplication
- Linear independence, span, and subspace
- Linear transformation
- Least squares
- Eigendecomposition
- Singular value decomposition
- Principal component analysis

# Recall: Linear System

- Recall the matrix equation of a linear system:

| Person ID | Weight | Height | Is_smoking | Life-span |
|---|---|---|---|---|
| 1 | 60kg | 5.5ft | Yes (=1) | 66 |
| 2 | 65kg | 5.0ft | No (=0) | 74 |
| 3 | 55kg | 6.0ft | Yes (=1) | 78 |

$$\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 66 \\ 74 \\ 78 \end{bmatrix}$$

$$A \qquad \mathbf{x} = \mathbf{b}$$

- Or, a vector equation is written as

$$\begin{bmatrix} 60 \\ 65 \\ 55 \end{bmatrix} x_1 + \begin{bmatrix} 5.5 \\ 5.0 \\ 6.0 \end{bmatrix} x_2 + \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} x_3 = \begin{bmatrix} 66 \\ 74 \\ 78 \end{bmatrix}$$

$$\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \mathbf{a}_3 x_3 = \mathbf{b}$$

# Uniqueness of Solution for $A\mathbf{x} = \mathbf{b}$

- The solution exists only when $\mathbf{b} \in$ Span $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$.

$$\begin{bmatrix} 60 \\ 65 \\ 55 \end{bmatrix} x_1 + \begin{bmatrix} 5.5 \\ 5.0 \\ 6.0 \end{bmatrix} x_2 + \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} x_3 = \begin{bmatrix} 66 \\ 74 \\ 78 \end{bmatrix}$$

$$\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \mathbf{a}_3 x_3 = \mathbf{b}$$

- If the solution exists for $A\mathbf{x} = \mathbf{b}$, when is it unique?
- It is unique when $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ are linearly independent.
- Infinitely many solutions exist when $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ are linearly dependent.

# Linear Independence

**(Practical) Definition:**

- Given a set of vectors $\mathbf{v}_1, \cdots, \mathbf{v}_p \in \mathbb{R}^n$, check if $\mathbf{v}_j$ can be represented as a linear combination of the previous vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{j-1}\}$ for $j = 1, \ldots, p$, e.g.,

$$\mathbf{v}_j \in \text{Span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{j-1}\} \text{ for some } j = 1, \ldots, p?$$

- If at least one such $\mathbf{v}_j$ is found, then $\{\mathbf{v}_1, \cdots, \mathbf{v}_p\}$ is linearly dependent.
- If no such $\mathbf{v}_j$ is found, then $\{\mathbf{v}_1, \cdots, \mathbf{v}_p\}$ is linearly independent.

# Linear Independence

**(Formal) Definition**:

- Consider $x_1\mathbf{v}_1 + x_2\mathbf{v}_2 \cdots + x_p\mathbf{v}_p = \mathbf{0}$.
- Obviously, one solution is $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$,

  which we call a trivial solution.

- $\mathbf{v}_1, \cdots, \mathbf{v}_p$ are linearly independent if this is the only solution.
- $\mathbf{v}_1, \cdots, \mathbf{v}_p$ are linearly dependent if this system also has other nontrivial solutions, e.g., at least one $x_i$ being nonzero.

# Two Definitions are Equivalent

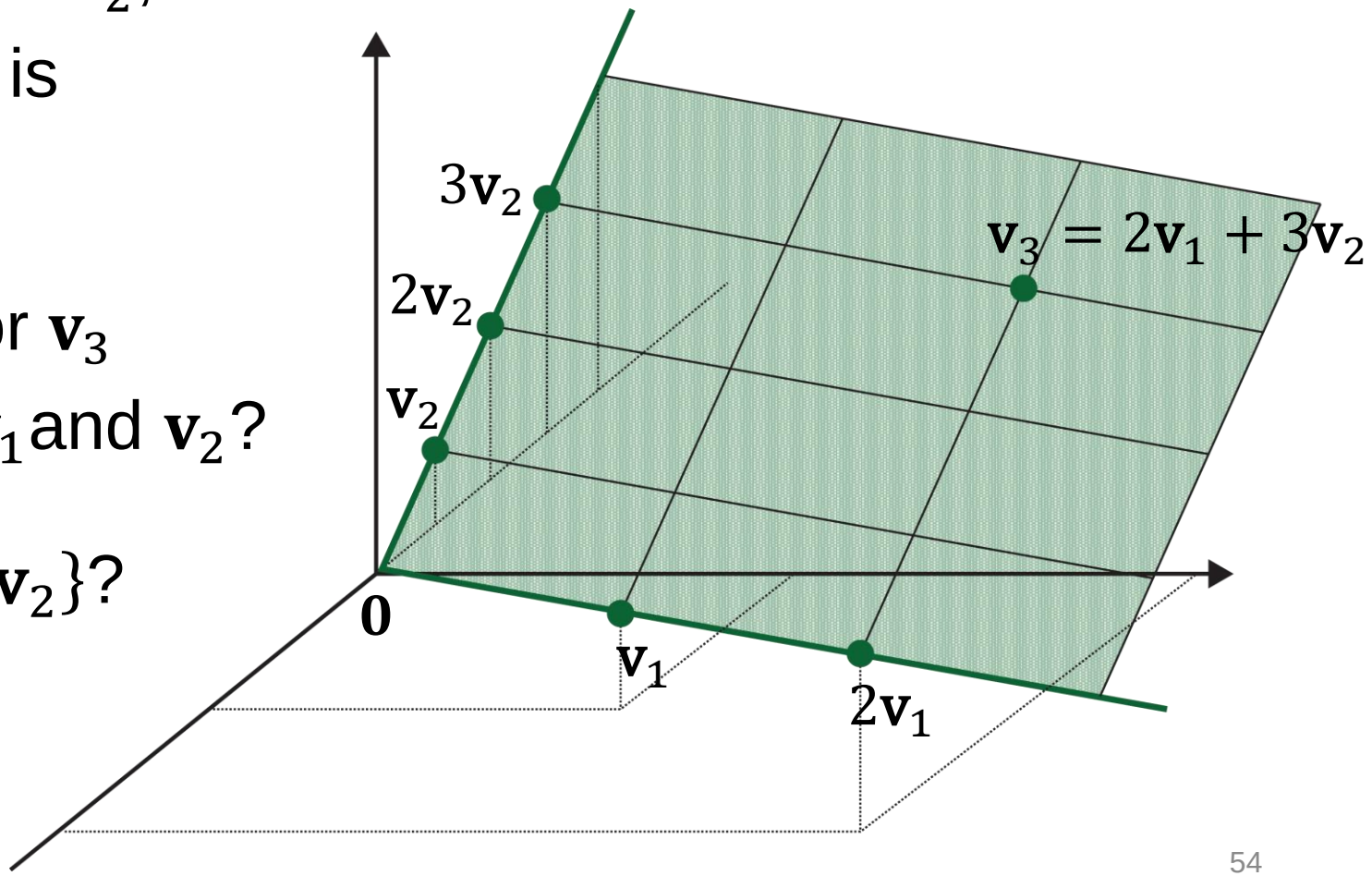
- If $\mathbf{v}_1, \cdots, \mathbf{v}_p$ are linearly dependent, consider a nontrivial solution.
- In the solution, let's denote $j$ as the last index such that $x_j \neq 0$.
- Then, one can write $x_j \mathbf{v}_j = -x_1 \mathbf{v}_1 - \cdots - x_{j-1} \mathbf{v}_{j-1}$, and safely divide it by $x_j$, resulting in

$$\mathbf{v}_j = -\frac{x_1}{x_j}\mathbf{v}_1 - \cdots - \frac{x_{j-1}}{x_j}\mathbf{v}_{j-1} \in \text{Span}\left\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{j-1}\right\}$$

which means $\mathbf{v}_j$ can be represented as a linear combination of the previous vectors.

# Geometric Understanding of Linear Dependence

- Given two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, Suppose Span $\{\mathbf{v}_1, \mathbf{v}_2\}$ is the plane on the right.
- When is the third vector $\mathbf{v}_3$ linearly dependent of $\mathbf{v}_1$ and $\mathbf{v}_2$?
- That is, $\mathbf{v}_3 \in$ Span $\{\mathbf{v}_1, \mathbf{v}_2\}$?



# Linear Dependence

- A linearly dependent vector does not increase Span!
- If $\mathbf{v}_3 \in$ Span $\{\mathbf{v}_1, \mathbf{v}_2\}$, then

$$\text{Span } \{\mathbf{v}_1, \mathbf{v}_2\} = \text{Span } \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\},$$

- Why?
- Suppose $\mathbf{v}_3 = d_1\mathbf{v}_1 + d_2\mathbf{v}_2$, then the linear combination of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ can be written as

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = (c_1 + d_1)\mathbf{v}_1 + (c_1 + d_1)\mathbf{v}_2$$

which is also a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$.

# Linear Dependence and Linear System Solution

- Also, a linearly dependent set produces multiple possible linear combinations of a given vector.
- Given a vector equation $x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + x_3\mathbf{v}_3 = \mathbf{b}$, suppose the solution is $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$, i.e., $3\mathbf{v}_1 + 2\mathbf{v}_2 + 1\mathbf{v}_3 = \mathbf{b}$.
- Suppose also $\mathbf{v}_3 = 2\mathbf{v}_1 + 3\mathbf{v}_2$, a linearly dependent case.
- Then, $3\mathbf{v}_1 + 2\mathbf{v}_2 + 1\mathbf{v}_3 = 3\mathbf{v}_1 + 2\mathbf{v}_2 + (2\mathbf{v}_1 + 3\mathbf{v}_2) = 5\mathbf{v}_1 + 5\mathbf{v}_2$, so $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 5 \\ 0 \end{bmatrix}$ is another solution. Many other solutions exist.

# Linear Dependence and Linear System Solution

- Actually, many more solutions exist.
- e.g., $3\mathbf{v}_1 + 2\mathbf{v}_2 + 1\mathbf{v}_3 = 3\mathbf{v}_1 + 2\mathbf{v}_2 + (2\mathbf{v}_3 - 1\mathbf{v}_3)$
  $= 3\mathbf{v}_1 + 2\mathbf{v}_2 + 2(2\mathbf{v}_1 + 3\mathbf{v}_2) - 1\mathbf{v}_3 = 7\mathbf{v}_1 + 8\mathbf{v}_2 - 1\mathbf{v}_3$,

thus $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ 8 \\ -1 \end{bmatrix}$ is another solution.

# Uniqueness of Solution for $A\mathbf{x} = \mathbf{b}$

- The solution exists only when $\mathbf{b} \in$ Span $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$.

$$\begin{bmatrix} 60 \\ 65 \\ 55 \end{bmatrix} x_1 + \begin{bmatrix} 5.5 \\ 5.0 \\ 6.0 \end{bmatrix} x_2 + \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} x_3 = \begin{bmatrix} 66 \\ 74 \\ 78 \end{bmatrix}$$

$$\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \mathbf{a}_3 x_3 = \mathbf{b}$$

- If the solution exists for $A\mathbf{x} = \mathbf{b}$, when is it unique?
- It is unique when $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ are linearly independent.
- Infinitely many solutions exist when $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ are linearly dependent.

# Span and Subspace

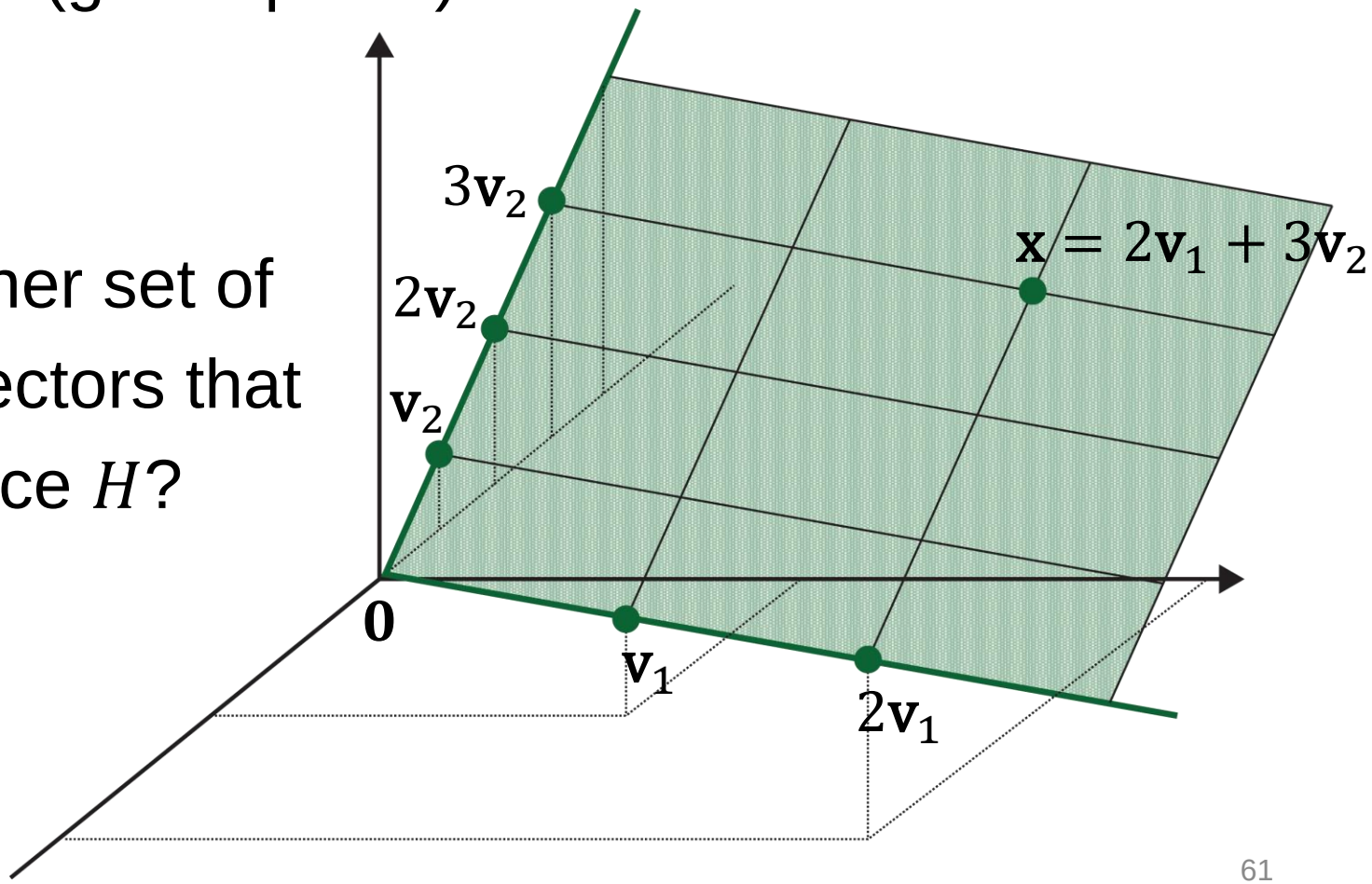
- **Definition**: A **subspace** $H$ is defined as a subset of $\mathbb{R}^n$ closed under linear combination:
  - For any two vectors, $\mathbf{u}_1, \mathbf{u}_2 \in H$, and any two scalars $c$ and $d$, $c\mathbf{u}_1 + d\mathbf{u}_2 \in H$.
- Span $\{\mathbf{v}_1, \cdots, \mathbf{v}_p\}$ is always a subspace. Why?
  - $\mathbf{u}_1 = a_1\mathbf{v}_1 + \cdots + a_p\mathbf{v}_p$, $\mathbf{u}_2 = b_1\mathbf{v}_1 + \cdots + b_p\mathbf{v}_p$
  - $$c\mathbf{u}_1 + d\mathbf{u}_2 = c\left(a_1\mathbf{v}_1 + \cdots + a_p\mathbf{v}_p\right) + d\left(b_1\mathbf{v}_1 + \cdots + b_p\mathbf{v}_p\right)$$
    $$= (ca_1 + db_1)\mathbf{v}_1 + \cdots + \left(ca_p + db_p\right)\mathbf{v}_p$$
- In fact, a subspace is always represented as Span $\{\mathbf{v}_1, \cdots, \mathbf{v}_p\}$.

# Basis of a Subspace

- **Definition**: A **basis** of a subspace $H$ is a set of vectors that satisfies both of the following:
  - Fully spans the given subspace $H$
  - Linearly independent (i.e., no redundancy)
- In the previous example, where $H = \text{Span}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, $\text{Span}\{\mathbf{v}_1, \mathbf{v}_2\}$ forms a plane, but $\mathbf{v}_3 = 2\mathbf{v}_1 + 3\mathbf{v}_2 \in \text{Span}\{\mathbf{v}_1, \mathbf{v}_2\}$, $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a basis of $H$, but not $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ nor $\{\mathbf{v}_1\}$ is a basis.

# Non-Uniqueness of Basis

- Consider a subspace $H$ (green plane).
- Is a basis unique?
- That is, is there any other set of linearly independent vectors that span the same subspace $H$?



# Dimension of Subspace

- What is then unique, given a particular subspace $H$?
- Even though different bases exist for $H$, the number of vectors in any basis for $H$ will be unique.
- We call this number as the dimension of $H$, denoted as dim $H$.
- In the previous example, the dimension of the plane is 2, meaning any basis for this subspace contains exactly two vectors.

# Column Space of Matrix

- **Definition**: The **column space** of a matrix $A$ is the subspace spanned by the columns of $A$.
  We call the column space of $A$ as Col $A$.

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \Rightarrow \quad \text{Col } A = \text{Span} \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \right\}$$

- What is dim Col $A$?

# Matrix with Linearly Dependent Columns

- Given $A = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$, note that $\begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$,

  i.e., the third column is a linear combination of the first two.

$$\text{Col } A = \text{Span}\left\{\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}\right\} \quad \Rightarrow \quad \text{Col } A = \text{Span}\left\{\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}\right\}$$

- What is dim Col $A$?

# Rank of Matrix

- **Definition**: The **rank** of a matrix $A$, denoted by rank $A$, is the dimension of the column space of $A$:
  - rank $A$ = dim Col $A$

# Null Space of a Matrix

- **Definition**: The **null space** of a matrix $A \in \mathbb{R}^{m \times n}$ is the set of all solutions of a homogeneous linear system, $A\mathbf{x} = \mathbf{0}$.

  We denote the null space of $A$ as Nul $A$.

- For $A = \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_m^T \end{bmatrix}$, $\mathbf{x}$ should satisfy the following:

$$\mathbf{a}_1^T\mathbf{x} = 0, \mathbf{a}_2^T\mathbf{x} = 0, \dots, \mathbf{a}_m^T\mathbf{x} = 0$$

- That is, $\mathbf{x}$ should be orthogonal to every row vector in $A$.

# Null Space is a Subspace

- **Theorem**: The null space of a matrix $A \in \mathbb{R}^{m \times n}$ is a subspace of $\mathbb{R}^n$. In other words, the set of all the solutions of a system $A\mathbf{x} = \mathbf{0}$ is a subspace of $\mathbb{R}^n$.

$$\text{Nul } A = (\text{Row } A)^{\perp}$$
$$\text{Nul } A^{T} = (\text{Col } A)^{\perp}$$

- **Note**: An eigenspace thus have a set of basis vectors with a particular dimension.

# Orthogonal Complement

- If a vector $\mathbf{z}$ is orthogonal to every vector in a subspace $W$ of $\mathbb{R}^n$, then $\mathbf{z}$ is said to be **orthogonal to** $W$.
- The set of all vectors $\mathbf{z}$ that are orthogonal to $W$ is called the **orthogonal complement** of $W$ and is denoted by $W^{\perp}$ (and read as "$W$ perpendicular" or simply "$W$ perp").
- A vector $\mathbf{x} \in \mathbb{R}^n$ is in $W^{\perp}$ if and only if $\mathbf{x}$ is orthogonal to every vector in a set that spans $W$.
- $W^{\perp}$ is a subspace of $\mathbb{R}^n$.
- Nul $A = (\text{Row } A)^{\perp}$. equivalent to 여집합
- Likewise, Nul $A^T = (\text{Col } A)^{\perp}$.

ex) $R^{2x3}$, Row A and Col A have linearly independence

dim Row A = 2
dim Col A = 3

dim Nul A = 1 dim Nul $A^T$ = 0

dim span {Nul A, Row A} = 3 - > orthogonal Complement
dim span {Nul $A^T$, Col A} = 3 - > orthogonal Complement

Four

# Fundamental Subspaces Given by $A$

Nul A, Nul $A^T$, Col A, Row A,

- Nul $A = (\text{Row } A)^{\perp}$.
- Nul $A^T = (\text{Col } A)^{\perp}$.

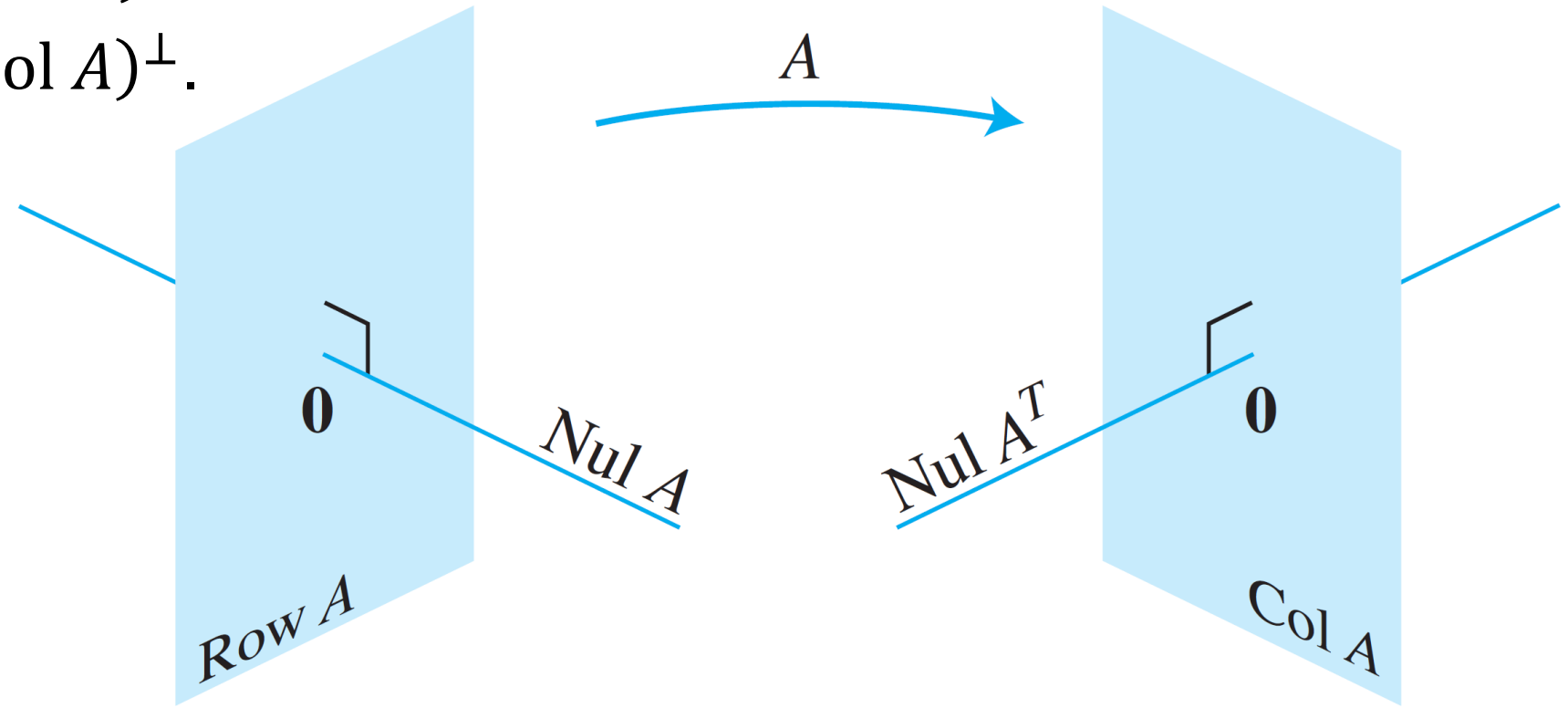



**FIGURE 8** The fundamental subspaces determined by an $m \times n$ matrix $A$.

# Basis Theorem

define dim as p - > satisfying only one codition for basis is enough
ex) Linear indp - > automatically fully span subspace
Fully span subspace - > automatically linealy indp

- Let $H$ be a $p$-dimensional subspace of $\mathbb{R}^n$.
- Any linearly independent set of exactly $p$ elements in $H$ is automatically a basis for $H$.
- Also, any set of $p$ elements of $H$ that spans $H$ is automatically a basis for $H$.

# Lecture Overview

- Elements in linear algebra
- Linear system
- Linear combination, vector equation,
  Four views of matrix multiplication
- Linear independence, span, and subspace
- Linear transformation
- Least squares
- Eigendecomposition
- Singular value decomposition
- Principal component analysis

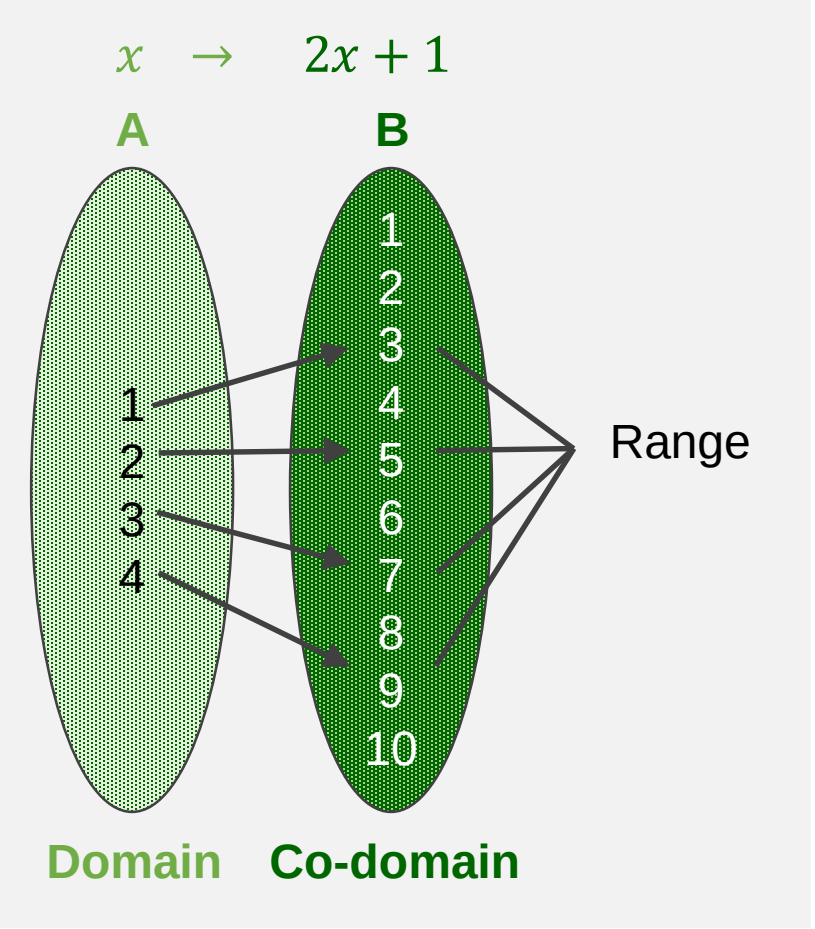
# Summary So Far

- Scalars, vectors, matrices, and their operations such as addition, scalar multiple, matrix multiplication, transpose
- Linear system: solving using inverse matrix
- Matrix equation and vector equation
- Linear combination and Span
  - When does the solution of a linear system exist?
- Four views of matrix multiplication: inner product, column combination, row combination, sum of rank-1 outer products
- Linear independence
  - If the solution of a linear system exists, when is it unique or many?
- Subspace
  - Subset of vectors in $\mathbb{R}^n$ closed under linear combination
  - Basis and dimension
  - Column space and rank of a matrix

# Transformation

- A **transformation**, **function**, or **mapping**, $T$ maps an input $x$ to an output $y$
  - Mathematical notation: $T$: $x \mapsto y$
- **Domain**: Set of all the possible values of $x$ (정의역)
- **Co-domain**: Set of all the possible values of $y$ (공역)
- **Image**: a mapped output $y$, given $x$ (상, 함수값)
- **Range**: Set of all the output values mapped by each $x$ in the domain (치역)

- **Note**: the output mapped by a particular $x$ is uniquely determined.



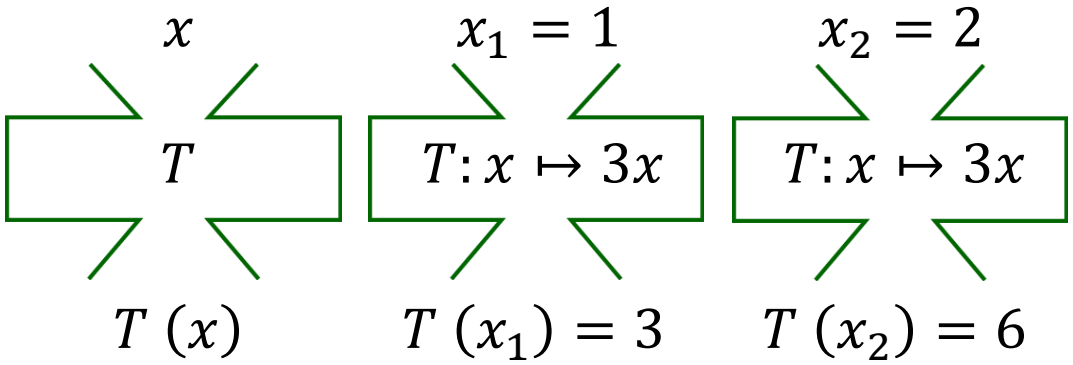
# Linear Transformation

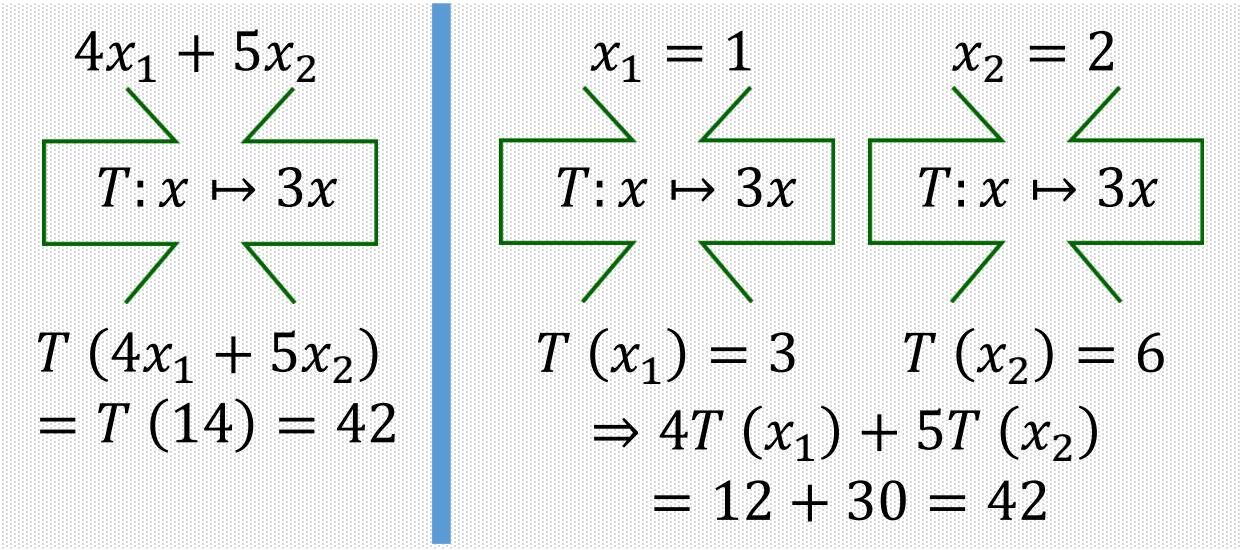
교환법칙이 성립하면 linear T

- **Definition**: A transformation (or mapping) $T$ is **linear** if:

$I.$ $T\,(c\mathbf{u} + d\mathbf{v}) = cT\,(\mathbf{u}) + dT\,(\mathbf{v})$ for all $\mathbf{u}, \mathbf{v}$ in the domain of $T$ and for all scalars $c$ and $d$

- Simple example: $T\colon\ x \mapsto y,\ T(x) = y\ = 3x$



# Transformations between Vectors

- $T: \mathbf{x} \in \mathbb{R}^n \mapsto \mathbf{y} \in \mathbb{R}^m$: Mapping $n$-dim vector to $m$-dim vector
- Example:

$$T: \mathbf{x} \in \mathbb{R}^3 \mapsto \mathbf{y} \in \mathbb{R}^2 \quad \mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \in \mathbb{R}^3 \quad \mapsto \quad \mathbf{y} = T(\mathbf{x}) = \begin{bmatrix} 4 \\ 5 \end{bmatrix} \in \mathbb{R}^2$$

# Matrix of Linear Transformation

- Example: Suppose $T$ is a linear transformation from $\mathbb{R}^2$ to $\mathbb{R}^3$ such that $T\left(\begin{bmatrix}1\\0\end{bmatrix}\right) = \begin{bmatrix}2\\-1\\1\end{bmatrix}$ and $T\left(\begin{bmatrix}0\\1\end{bmatrix}\right) = \begin{bmatrix}0\\1\\2\end{bmatrix}$. With no additional information, find a formula for the image of an arbitrary **x** in $\mathbb{R}^2$

$$\mathbf{x} = \begin{bmatrix}x_1\\x_2\end{bmatrix} = x_1\begin{bmatrix}1\\0\end{bmatrix} + x_2\begin{bmatrix}0\\1\end{bmatrix}$$

$$\Rightarrow T(\mathbf{x}) = T\left(x_1\begin{bmatrix}1\\0\end{bmatrix} + x_2\begin{bmatrix}0\\1\end{bmatrix}\right) = x_1 T\left(\begin{bmatrix}1\\0\end{bmatrix}\right) + x_2 T\left(\begin{bmatrix}0\\1\end{bmatrix}\right)$$

$$= x_1\begin{bmatrix}2\\-1\\1\end{bmatrix} + x_2\begin{bmatrix}0\\1\\2\end{bmatrix} = \begin{bmatrix}2 & 0\\-1 & 1\\1 & 2\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}$$

Column - Column perspective

# Matrix of Linear Transformation

cf.) Affine Trans: y=Ax+b - > 선형성 만족 X (교환법칙 X)
but, [A,b][[x],[1]] - > A'x'로 선형형태로 변환 가능

- In general, let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. Then $T$ is always written as a matrix-vector multiplication, i.e.,

$$T(\mathbf{x}) = A\mathbf{x} \text{ for all } \mathbf{x} \in \mathbb{R}^n$$

- In fact, the $j$-th column of $A \in \mathbb{R}^{m \times n}$ is equal to the vector $T(\mathbf{e}_j)$, where $\mathbf{e}_j$ is the $j$-th column of the identity matrix in $\mathbb{R}^{n \times n}$:

e == standard basis vector

$$A = [T(\mathbf{e}_1) \quad \cdots \quad T(\mathbf{e}_n)]$$

좌표계의 변환

- Here, the matrix $A$ is called the **standard matrix** of the linear transformation $T$

# Matrix of Linear Transformation

- Example: Find the standard matrix $A$ of a linear transformation from $\mathbb{R}^3$ to $\mathbb{R}^2$ such that

$$T\left(\begin{bmatrix}1\\0\\0\end{bmatrix}\right)=\begin{bmatrix}1\\2\end{bmatrix}, T\left(\begin{bmatrix}0\\1\\0\end{bmatrix}\right)=\begin{bmatrix}4\\3\end{bmatrix}, \text{ and } T\left(\begin{bmatrix}0\\0\\1\end{bmatrix}\right)=\begin{bmatrix}5\\6\end{bmatrix}.$$

정의역 $$\mathbf{x}=\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=x_1\begin{bmatrix}1\\0\\0\end{bmatrix}+x_2\begin{bmatrix}0\\1\\0\end{bmatrix}+x_3\begin{bmatrix}0\\0\\1\end{bmatrix}$$

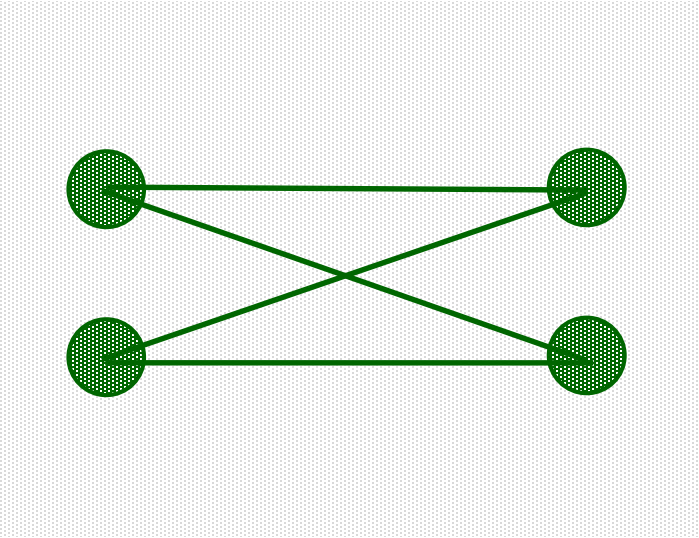
$$\Rightarrow T(\mathbf{x})=T\left(x_1\begin{bmatrix}1\\0\\0\end{bmatrix}+x_2\begin{bmatrix}0\\1\\0\end{bmatrix}+x_3\begin{bmatrix}0\\0\\1\end{bmatrix}\right)=x_1T\left(\begin{bmatrix}1\\0\\0\end{bmatrix}\right)+x_2T\left(\begin{bmatrix}0\\1\\0\end{bmatrix}\right)+x_3T\left(\begin{bmatrix}0\\0\\1\end{bmatrix}\right)$$

$$=x_1\begin{bmatrix}1\\2\end{bmatrix}+x_2\begin{bmatrix}4\\3\end{bmatrix}+x_2\begin{bmatrix}5\\6\end{bmatrix}=\begin{bmatrix}1&4&5\\2&3&6\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=A\mathbf{x}$$

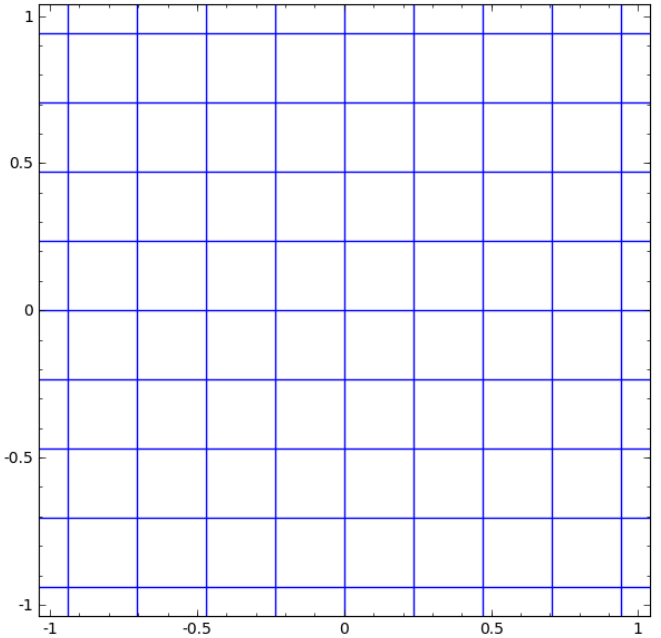
# Linear Transformation in Neural Networks

- Fully-connected layers (linear layer)

Linear 로만 model이 잘 안만들어질때,
- 차원을 높이는 거는 도찐개찐 (subspace dim이 그대로니까)
- non - linear를 feature로 추가하여 model upgrade하는것이 타당



$$\mathbf{x} \xrightarrow[T_1]{} \mathbf{y}$$

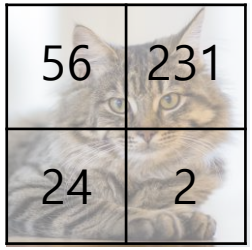


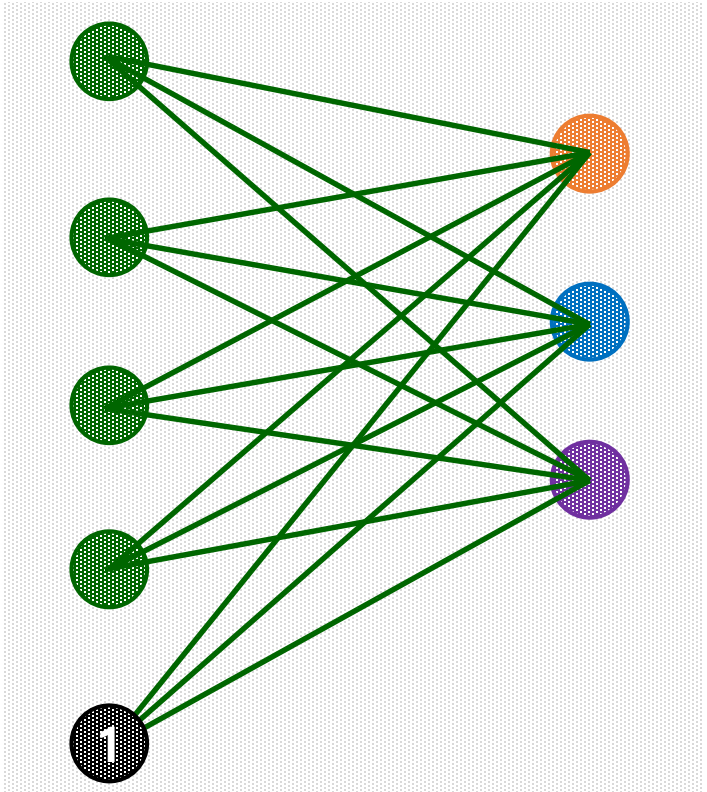

https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

# Affine Layer in Neural Networks

cf.) Affine Trans: y=Ax+b - > 선형성 만족 X (교환법칙 X)
but, [A,b][x,[1]] - > A'x'로 선형형태로 변환 가능

- Fully-connected layers usually involve a bias term. That is why we call it an affine layer, but not a linear layer.
- Example: Image with 4 pixels and 3 classes (cat/dog/ship)




| | | | |
|---|---|---|---|
| 0.2 | -0.5 | 0.1 | 2 |
| 1.5 | 1.3 | 2.1 | 1 |
| -.2 | 0.3 | 0.7 | -1.3 |

| |
|---|
| 56 |
| 231 |
| 24 |
| 2 |

+

| |
|---|
| 1.1 |
| 3.2 |
| -1.2 |

=

| |
|---|
| -96.8 |
| 439.9 |
| 711 |

=56

| | | | |
|---|---|---|---|
| 0.2 | | | |
| 1.5 | | | |
| -.2 | | | |

+231

| | | | |
|---|---|---|---|
| -0.5 | | | |
| 1.3 | | | |
| 0.3 | | | |

+24

| | | | |
|---|---|---|---|
| 0.1 | | | |
| 2.1 | | | |
| 0.7 | | | |

+2

| | | | |
|---|---|---|---|
| 2 | | | |
| 1 | | | |
| -1.3 | | | |

+1

| | | | |
|---|---|---|---|
| 1.1 | | | |
| 3.2 | | | |
| -1.2 | | | |

=

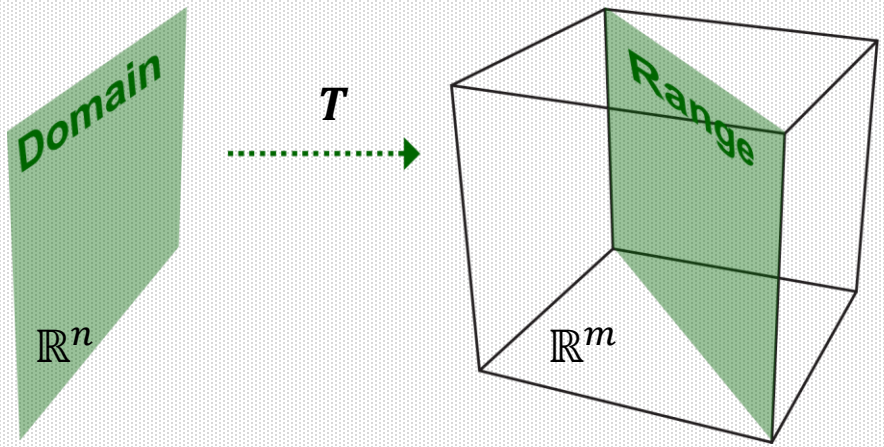
| | | | | |
|---|---|---|---|---|
| 0.2 | -0.5 | 0.1 | 2 | 1.1 |
| 1.5 | 1.3 | 2.1 | 1 | 3.2 |
| -.2 | 0.3 | 0.7 | -1.3 | -1.2 |

| |
|---|
| 56 |
| 231 |
| 24 |
| 2 |
| 1 |

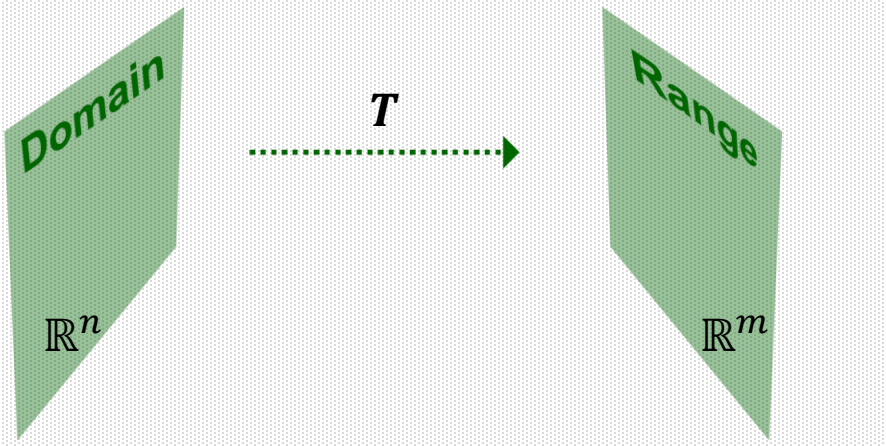
# ONTO and ONE-TO-ONE

전사, projection < - > 공역 == 치역, 정의역 원소 갯수 >= 공역 원소 갯수

- **Definition:** A mapping $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is said to be onto $\mathbb{R}^m$ if each $\mathbf{b} \in \mathbb{R}^m$ is the image of at least one $\mathbf{x} \in \mathbb{R}^n$.
  That is, the range is equal to the co-domain.




$T$ is NOT onto $\mathbb{R}^m$

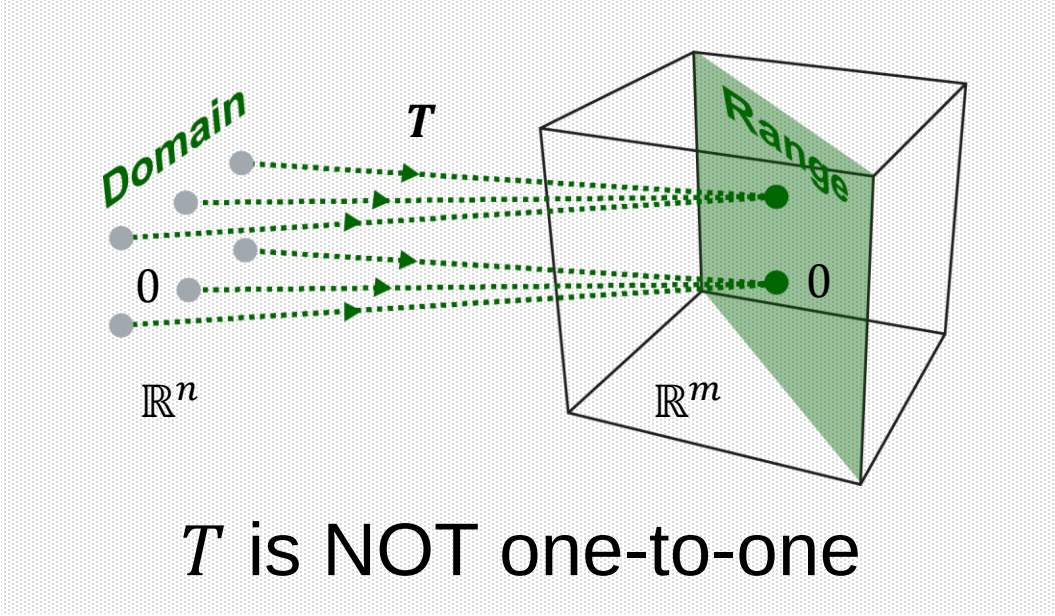



$T$ is onto $\mathbb{R}^m$

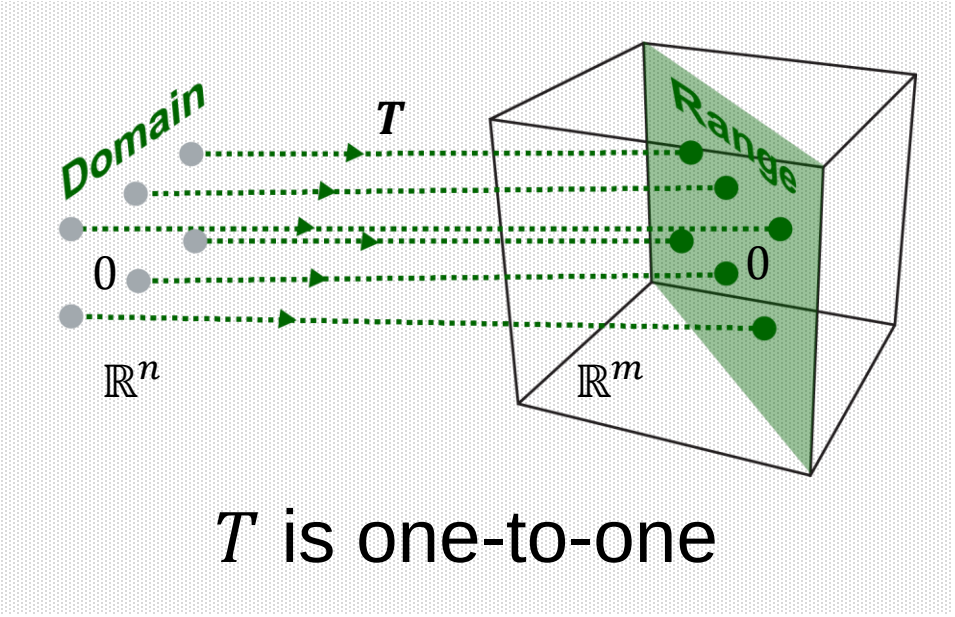
# ONTO and ONE-TO-ONE

일대일 함수,

- **Definition:** A mapping $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is said to be one-to-one if each $\mathbf{b} \in \mathbb{R}^m$ is the image of at most one $\mathbf{x} \in \mathbb{R}^n$.
  That is, each output vector in the range is mapped by only one input vector, no more than that.



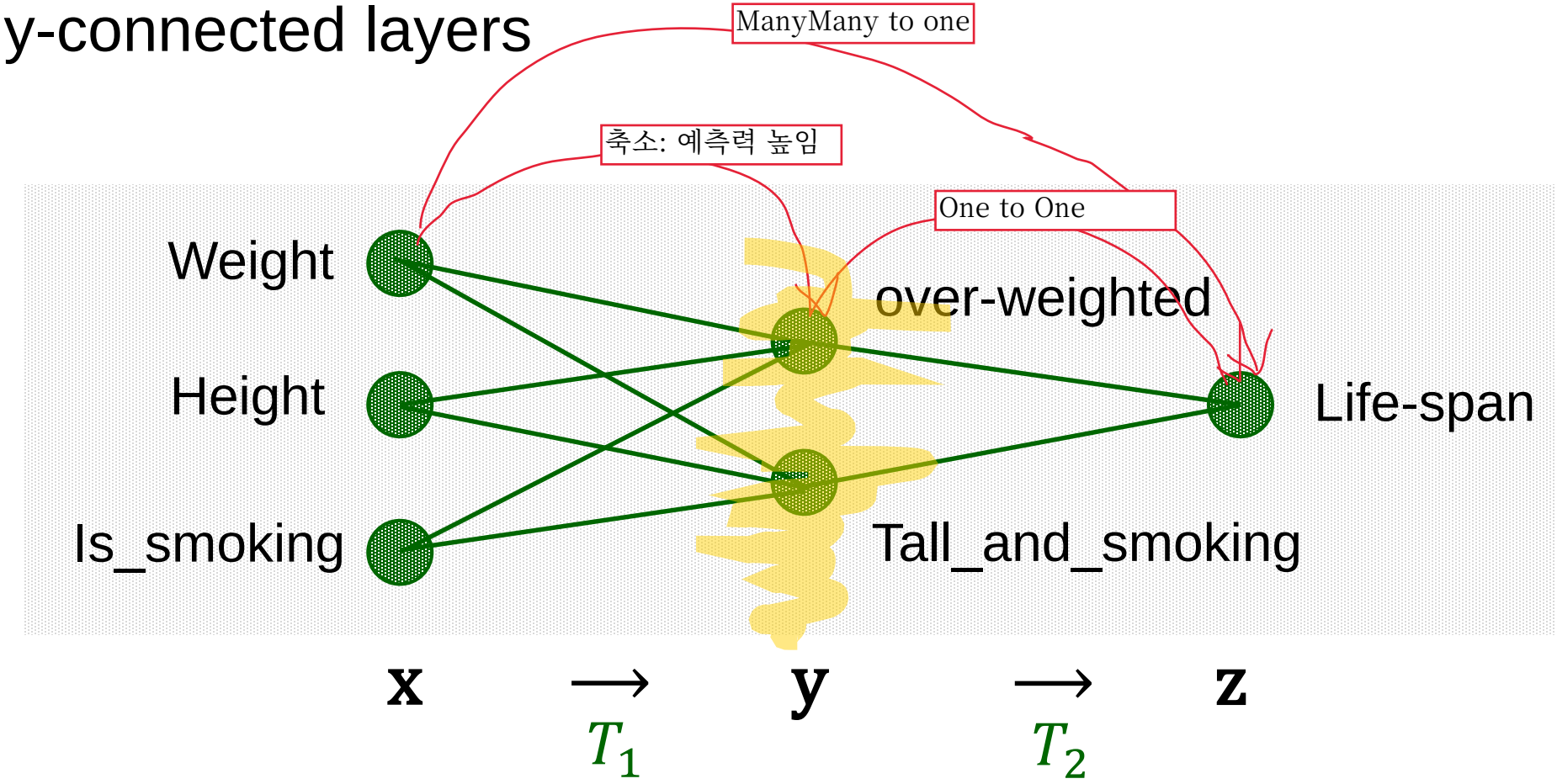

$T$ is NOT one-to-one




$T$ is one-to-one

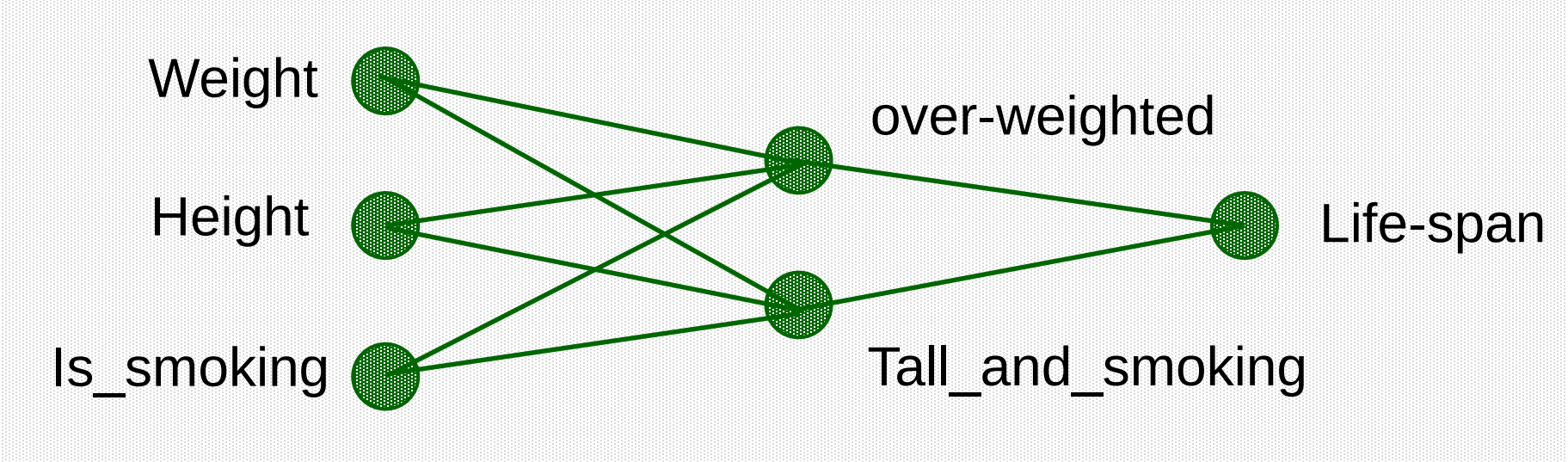
# Neural Network Example

- Fully-connected layers



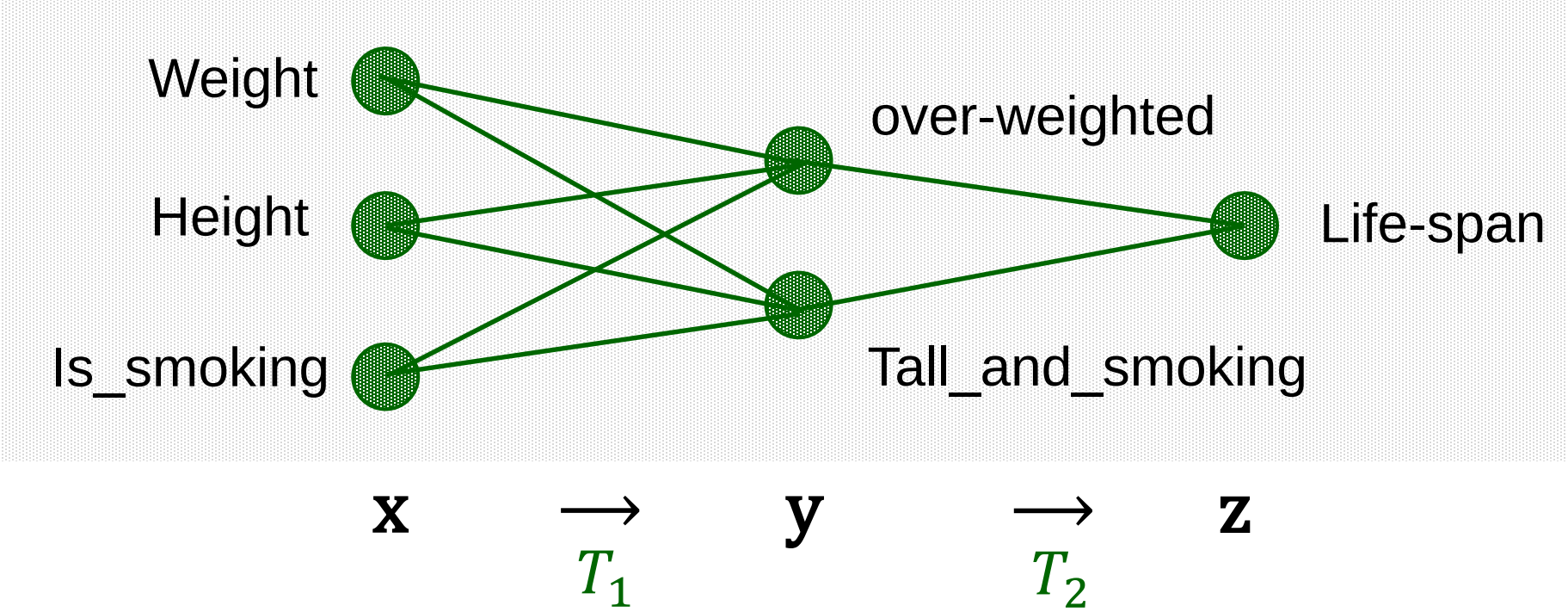
# Neural Network Example: ONE-TO-ONE

- Will there be many (or unique) people mapped to the same (over_weighted, tall_and_smoking)?



# Neural Network Example: ONTO

- Is there any (over_weighted, tall_and_smoking) that does not exist at all?



# ONTO and ONE-TO-ONE

- Let $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a linear transformation, i.e.,

$$T(\mathbf{x}) = A\mathbf{x} \text{ for all } \mathbf{x} \in \mathbb{R}^n.$$

- $T$ is one-to-one if and only if the columns of $A$ are linearly independent. Ax=b의 해가 하나다 - > Linearly ind

- $T$ maps $\mathbb{R}^n$ onto $\mathbb{R}^m$ if and only if the columns of $A$ span $\mathbb{R}^m$.

공역 == 치역 ,
Col A == $R^m$

# ONTO and ONE-TO-ONE

- **Example:**

$$\text{Let } T(\mathbf{x}) = A\mathbf{x} = \begin{bmatrix} 2 & 0 \\ -1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Is $T$ one-to-one? yes, Col A linealy ind
- Does $T$ map $\mathbb{R}^2$ onto $\mathbb{R}^3$? No, dim Col A = = 2,

# ONTO and ONE-TO-ONE

- **Example:**

$$\text{Let } T(\mathbf{x}) = A\mathbf{x} = \begin{bmatrix} 1 & 4 & 5 \\ 2 & 3 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

- Is $T$ one-to-one? No, linearly dep
- Does $T$ map $\mathbb{R}^3$ onto $\mathbb{R}^2$? Yes, dim Col A == 2

# Further Study

- Gaussian elimination, row reduction, echelon form
  - Lay Ch1.2,
- LU factorization: efficiently solving linear systems
  - Lay Ch2.5
- Computing invertible matrices
  - Lay Ch2.2
- Invertible matrix theorem for square matrices
  - Lay Ch2.3, Ch2.9

# Lecture Overview

- Elements in linear algebra
- Linear system
- Linear combination, vector equation,
  Four views of matrix multiplication
- Linear independence, span, and subspace
- Linear transformation
- Least squares
- Eigendecomposition
- Singular value decomposition
- Principal component analysis

# Over-determined Linear Systems (#equations ≫ #variables)

- Recall a linear system:

| Person ID | Weight | Height | Is_smoking | Life-span |
|---|---|---|---|---|
| 1 | 60kg | 5.5ft | Yes (=1) | 66 |
| 2 | 65kg | 5.0ft | No (=0) | 74 |
| 3 | 55kg | 6.0ft | Yes (=1) | 78 |

$$60x_1+5.5x_2+1\cdot x_3=66$$
$$65x_1+5.0x_2+0\cdot x_3=74$$
$$55x_1+6.0x_2+1\cdot x_3=78$$

# Over-determined Linear Systems (#equations ≫ #variables)

- Recall a linear system:
- What if we have much more data examples?

| Person ID | Weight | Height | Is_smoking | Life-span |
|---|---|---|---|---|
| 1 | 60kg | 5.5ft | Yes (=1) | 66 |
| 2 | 65kg | 5.0ft | No (=0) | 74 |
| 3 | 55kg | 6.0ft | Yes (=1) | 78 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

$$\begin{aligned} 60x_1 + 5.5x_2 + 1 \cdot x_3 &= 66 \\ 65x_1 + 5.0x_2 + 0 \cdot x_3 &= 74 \\ 55x_1 + 6.0x_2 + 1 \cdot x_3 &= 78 \\ \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad &\vdots \end{aligned}$$

- Matrix equation:

$$\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 66 \\ 74 \\ 78 \\ \vdots \end{bmatrix}$$

$$A \qquad \mathbf{x} = \mathbf{b}$$

$m \gg n$: more equations than variables

➡ Usually no solution exists

# Vector Equation Perspective

- Vector equation form: $\begin{bmatrix}60\\65\\55\\\vdots\end{bmatrix}x_1+\begin{bmatrix}5.5\\5.0\\6.0\\\vdots\end{bmatrix}x_2+\begin{bmatrix}1\\0\\1\\\vdots\end{bmatrix}x_3=\begin{bmatrix}66\\74\\78\\\vdots\end{bmatrix}$

$$\mathbf{a}_1x_1+\ \mathbf{a}_2x_2\ +\ \mathbf{a}_3x_3=\ \mathbf{b}$$

- Compared to the original space $\mathbb{R}^n$, where $\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3,\mathbf{b}\in\mathbb{R}^n$, Span $\{\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3\}$ will be a thin hyperplane, so it is likely that $\mathbf{b}\notin$ Span $\{\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3\}$

  ➡ No solution exists.

# Motivation for Least Squares

- Even if no solution exists, we want to approximately obtain the solution for an over-determined system.

- Then, how can we define the best approximate solution for our purpose?

# Inner Product

- Given $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, we can consider $\mathbf{u}$ and $\mathbf{v}$ as $n \times 1$ matrices.
- The transpose $\mathbf{u}^T$ is a $1 \times n$ matrix, and the matrix product $\mathbf{u}^T\mathbf{v}$ is a $1 \times 1$ matrix, which we write as a scalar without brackets.
- The number $\mathbf{u}^T\mathbf{v}$ is called the **inner product** or **dot product** of $\mathbf{u}$ and $\mathbf{v}$, and it is written as $\mathbf{u} \cdot \mathbf{v}$.
- For $\mathbf{u} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}, \mathbf{v} = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}, \mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T\mathbf{v} = \begin{bmatrix} 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} = [14]$

$(1 \times 3)(3 \times 1) = 1 \times 1$

# Properties of Inner Product

- **Theorem:** Let $\mathbf{u}, \mathbf{v},$ and $\mathbf{w}$ be vectors in $\mathbb{R}^n$, and let $c$ be a scalar. Then
  - **a)** $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$
  - **b)** $(\mathbf{u} + \mathbf{v}) \cdot \mathbf{w} = \mathbf{u} \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{w}$
  - **c)** $(c\mathbf{u}) \cdot \mathbf{v} = c(\mathbf{u} \cdot \mathbf{v}) = \mathbf{u} \cdot (c\mathbf{v})$
  - **d)** $\mathbf{u} \cdot \mathbf{u} \geq 0$, and $\mathbf{u} \cdot \mathbf{u} = 0$ if and only if $\mathbf{u} = \mathbf{0}$
- Properties (b) and (c) can be combined to produce the following useful rule:

$$(c_1\mathbf{u}_1 + \cdots + c_p\mathbf{u}_p) \cdot \mathbf{w} = c_1(\mathbf{u}_1 \cdot \mathbf{w}) + \cdots + c_p(\mathbf{u}_p \cdot \mathbf{w})$$

내적은 선형성을 가진다

# Vector Norm

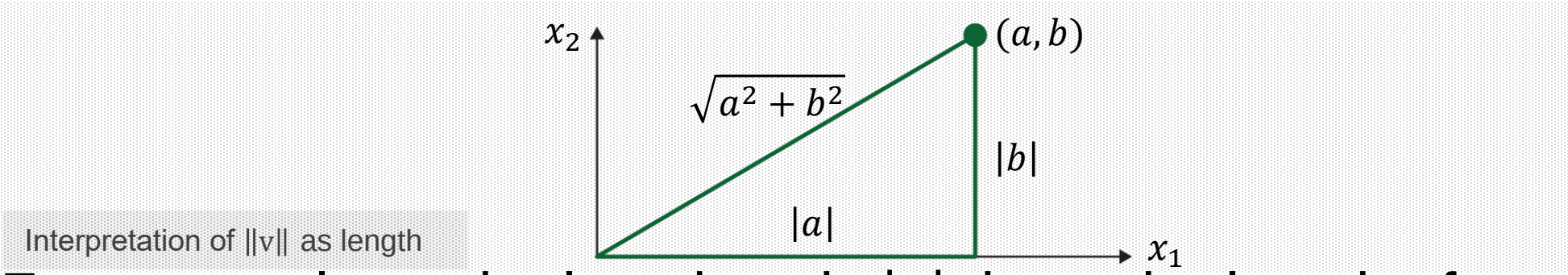
- For $\mathbf{v} \in \mathbb{R}^n$, with entries $v_1, \ldots, v_n$, the square root of $\mathbf{v} \cdot \mathbf{v}$ is defined because $\mathbf{v} \cdot \mathbf{v}$ is nonnegative.

- **Definition**: The length (or norm) of $\mathbf{v}$ is the non-negative scalar $\|\mathbf{v}\|$ defined as the square root of $\mathbf{v} \cdot \mathbf{v}$ :

$$\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} \text{ and } \|\mathbf{v}\|^2 = \mathbf{v} \cdot \mathbf{v}$$

# Geometric Meaning of Vector Norm

- Suppose $\mathbf{v} \in \mathbb{R}^2$, say, $\mathbf{v} = \begin{bmatrix} a \\ b \end{bmatrix}$.
- $\|\mathbf{v}\|$ is the length of the line segment from the origin to $\mathbf{v}$.
- This follows from Pythagorean Theorem applied to a triangle such as the one shown in the following figure:




Interpretation of ||v|| as length

- For any scalar $c$, the length $c\mathbf{v}$ is $|c|$ times the length of $\mathbf{v}$ That is,

$$\|c\mathbf{v}\| = |c|\|\mathbf{v}\|$$

# Unit Vector

- A vector whose length is $1$ is called a unit vector.
- Normalizing a vector: Given a nonzero vector $\mathbf{v}$, if we divide it by its length, we obtain a unit vector $\mathbf{u} = \frac{1}{\|\mathbf{v}\|}\mathbf{v}$.
- $\mathbf{u}$ is in the same direction as $\mathbf{v}$, but its length is 1.

# Distance between Vectors in $\mathbb{R}^n$

- **Definition:** For $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, the **distance between $\mathbf{u}$ and $\mathbf{v}$**, written as dist $(\mathbf{u}, \mathbf{v})$, is the length of the vector $\mathbf{u} - \mathbf{v}$ That is,

$$\text{dist}\,(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|$$

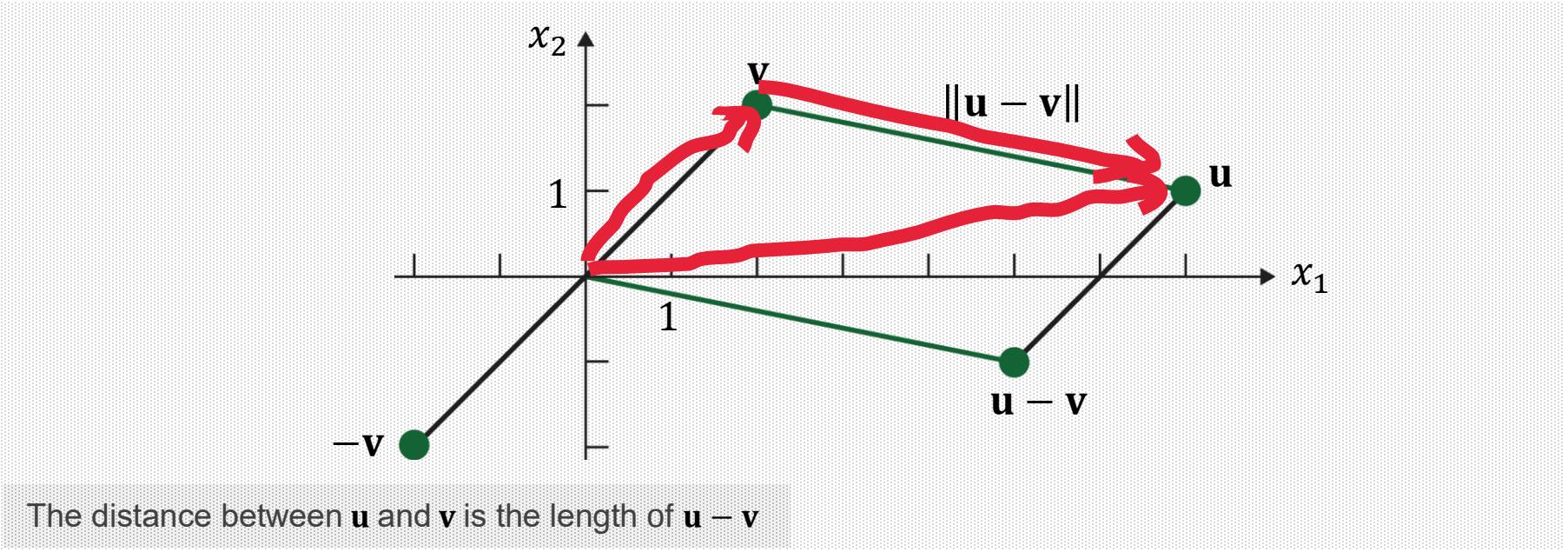
- **Example**: Compute the distance between the vector $\mathbf{u} = \begin{bmatrix} 6 \\ 1 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$.
- **Solution**: Calculate

$$\mathbf{u} - \mathbf{v} = \begin{bmatrix} 6 \\ 1 \end{bmatrix} - \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

$$\|\mathbf{u} - \mathbf{v}\| = \sqrt{3^2 + (-1)^2} = \sqrt{10}$$

# Distance between Vectors in $\mathbb{R}^n$

- The distance from $\mathbf{u}$ to $\mathbf{v}$ is the same as the distance from $\mathbf{u} - \mathbf{v}$ to $0$.



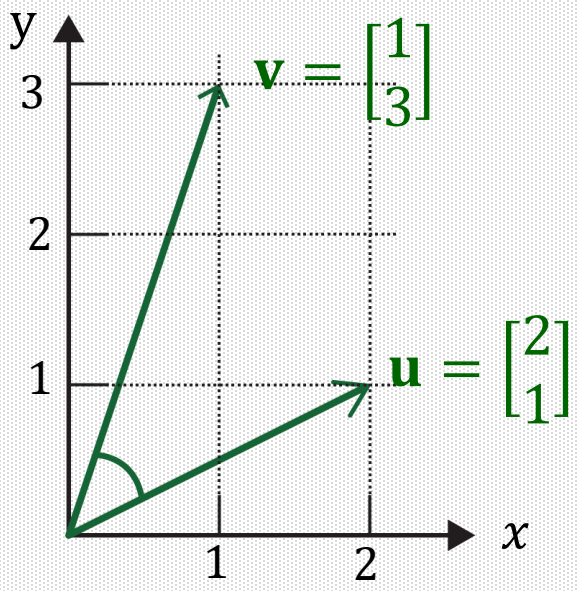

The distance between $\mathbf{u}$ and $\mathbf{v}$ is the length of $\mathbf{u} - \mathbf{v}$

# Inner Product and Angle Between Vectors

- Inner product between $\mathbf{u}$ and $\mathbf{v}$ can be rewritten using their norms and angle:

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\| \cos\theta$$

- **Example**:




$$\mathbf{u} \cdot \mathbf{v} = \begin{bmatrix}2\\1\end{bmatrix} \cdot \begin{bmatrix}1\\3\end{bmatrix} = [2 \quad 1] \begin{bmatrix}1\\3\end{bmatrix} = 5$$

$$\|\mathbf{u}\| = \sqrt{2^2 + 1^2} = \sqrt{5} \qquad \|\mathbf{v}\| = \sqrt{1^2 + 3^2} = \sqrt{10}$$

$$\mathbf{u} \cdot \mathbf{v} = 5 = \|\mathbf{u}\|\|\mathbf{v}\| \cos\theta = \sqrt{5} \cdot \sqrt{10} \cos\theta$$

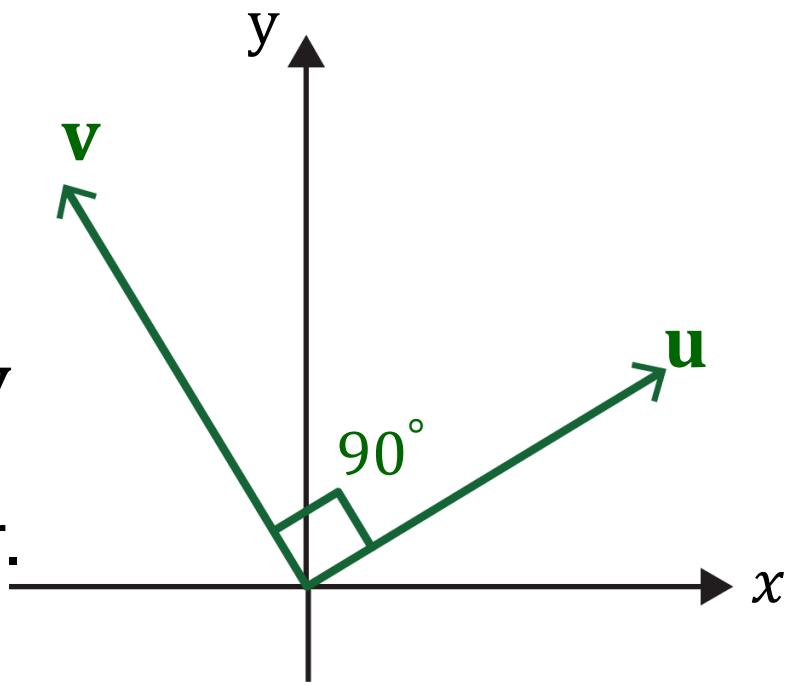
$$\Rightarrow \cos\theta = \frac{5}{\sqrt{50}} = \frac{1}{\sqrt{2}}$$

$$\Rightarrow \theta = 45^\circ$$

# Orthogonal Vectors

- **Definition:** $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^n$ are **orthogonal** (to each other) if $\mathbf{u} \cdot \mathbf{v} = 0$

  That is,
  $\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\| \cos\theta = 0.$
  - $\cos\theta = 0$ for nonzero vectors $\mathbf{u}$ and $\mathbf{v}$
  - $\theta = 90°$ ($\mathbf{u} \perp \mathbf{v}$).
  - $\mathbf{u}$ and $\mathbf{v}$ are perpendicular each other.



# Back to Over-Determined System

- Let's start with the original problem:

| Person ID | Weight | Height | Is_smoking | Life-span |
|---|---|---|---|---|
| 1 | 60kg | 5.5ft | Yes (=1) | 66 |
| 2 | 65kg | 5.0ft | No (=0) | 74 |
| 3 | 55kg | 6.0ft | Yes (=1) | 78 |

$$\underset{A}{\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \end{bmatrix}} \underset{\mathbf{x}}{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}} = \underset{\mathbf{b}}{\begin{bmatrix} 66 \\ 74 \\ 78 \end{bmatrix}}$$

- Using the inverse matrix, the solution is $\mathbf{x} = \begin{bmatrix} -0.4 \\ 20 \\ -20 \end{bmatrix}$.

# Back to Over-Determined System

- Let's add an additional example:

| Person ID | Weight | Height | Is_smoking | Life-span |
|---|---|---|---|---|
| 1 | 60kg | 5.5ft | Yes (=1) | 66 |
| 2 | 65kg | 5.0ft | No (=0) | 74 |
| 3 | 55kg | 6.0ft | Yes (=1) | 78 |
| 4 | 50kg | 5.0ft | Yes (=1) | 72 |

$$\underset{A}{\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ 50 & 5.0 & 1 \end{bmatrix}} \underset{\mathbf{x}}{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}} = \underset{\mathbf{b}}{\begin{bmatrix} 66 \\ 74 \\ 78 \\ 72 \end{bmatrix}}$$

- Now, let's use the previous solution $\mathbf{x} = \begin{bmatrix} -0.4 \\ 20 \\ -20 \end{bmatrix}$

$$\underset{A}{\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ 50 & 5.0 & 1 \end{bmatrix}} \underset{\mathbf{x}}{\begin{bmatrix} -0.4 \\ 20 \\ -20 \end{bmatrix}} = \begin{bmatrix} 66 \\ 74 \\ 78 \\ 60 \end{bmatrix} \neq \underset{\mathbf{b}}{\begin{bmatrix} 66 \\ 74 \\ 78 \\ 72 \end{bmatrix}} \quad \underset{\text{Errors}}{\begin{vmatrix} (\mathbf{b} - A\mathbf{x}) \\ 0 \\ 0 \\ 0 \\ 12 \end{vmatrix}}$$

# Back to Over-Determined System

- How about using slightly different solution $\mathbf{x} = \begin{bmatrix} -0.12 \\ 16 \\ -9.5 \end{bmatrix}$?

$$A \qquad \mathbf{x} \qquad\qquad \neq \mathbf{b}$$

$$\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ 50 & 5.0 & 1 \end{bmatrix} \begin{bmatrix} -0.12 \\ 16 \\ -9.5 \end{bmatrix} = \begin{bmatrix} 71.3 \\ 72.2 \\ 79.9 \\ 64.5 \end{bmatrix} \neq \begin{bmatrix} 66 \\ 74 \\ 78 \\ 72 \end{bmatrix}$$

Errors

$$(\mathbf{b} - A\mathbf{x}) = \begin{bmatrix} -5.3 \\ 1.8 \\ -1.9 \\ 7.5 \end{bmatrix}$$

# Which One is Better Solution?

$$\begin{matrix} & A & & \mathbf{x} & & & \neq & \mathbf{b} \\ \begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ 50 & 5.0 & 1 \end{bmatrix} & & \begin{bmatrix} -0.12 \\ 16 \\ -9.5 \end{bmatrix} & = & \begin{bmatrix} 71.3 \\ 72.2 \\ 79.9 \\ 64.5 \end{bmatrix} & \neq & \begin{bmatrix} 66 \\ 74 \\ 78 \\ 72 \end{bmatrix} \end{matrix}$$

Errors

$$\left| \begin{matrix} (\mathbf{b} - A\mathbf{x}) \\ -5.3 \\ 1.8 \\ -1.9 \\ 7.5 \end{matrix} \right.$$

$$\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ 50 & 5.0 & 1 \end{bmatrix} \begin{bmatrix} -0.4 \\ 20 \\ -20 \end{bmatrix} = \begin{bmatrix} 66 \\ 74 \\ 78 \\ 60 \end{bmatrix} \neq \begin{bmatrix} 66 \\ 74 \\ 78 \\ 72 \end{bmatrix} \left| \begin{matrix} 0 \\ 0 \\ 0 \\ 12 \end{matrix} \right.$$

# Least Squares: Best Approximation Criterion

- Let's use the squared sum of errors:

| | $A$ | | $\mathbf{x}$ | | $\neq$ | $\mathbf{b}$ | Errors $(\mathbf{b}-A\mathbf{x})$ | Sum of squared errors |
|---|---|---|---|---|---|---|---|---|
| $\begin{bmatrix}60\\65\\55\\50\end{bmatrix}$ | $\begin{matrix}5.5\\5.0\\6.0\\5.0\end{matrix}$ | $\begin{matrix}1\\0\\1\\1\end{matrix}$ | $\begin{bmatrix}-0.12\\16\\-9.5\end{bmatrix}$ | $=$ | $\begin{bmatrix}71.3\\69\\79.9\\64.5\end{bmatrix}$ | $\neq \begin{bmatrix}66\\74\\78\\72\end{bmatrix}$ | $\begin{matrix}-5.3\\1.8\\-1.9\\7.5\end{matrix}$ | $\left((-5.3)^2+1.8^2+(-1.9)^2+7.5^2\right)^{0.5}$ $=9.55$ *Better solution* |
| $\begin{bmatrix}60\\65\\55\\50\end{bmatrix}$ | $\begin{matrix}5.5\\5.0\\6.0\\5.0\end{matrix}$ | $\begin{matrix}1\\0\\1\\1\end{matrix}$ | $\begin{bmatrix}-0.4\\20\\-20\end{bmatrix}$ | $=$ | $\begin{bmatrix}66\\74\\78\\60\end{bmatrix}$ | $\neq \begin{bmatrix}66\\74\\78\\72\end{bmatrix}$ | $\begin{matrix}0\\0\\0\\12\end{matrix}$ | $(0^2+0^2+0^2+12^2)^{0.5}$ $=12$ |

# Least Squares Problem

- Now, the sum of squared errors can be represented as $\|\mathbf{b} - A\mathbf{x}\|$.
- **Definition**: Given an overdetermined system $A\mathbf{x} \simeq \mathbf{b}$ where $A \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^n$, and $m \gg n$, a least squares solution $\hat{\mathbf{x}}$ is defined as

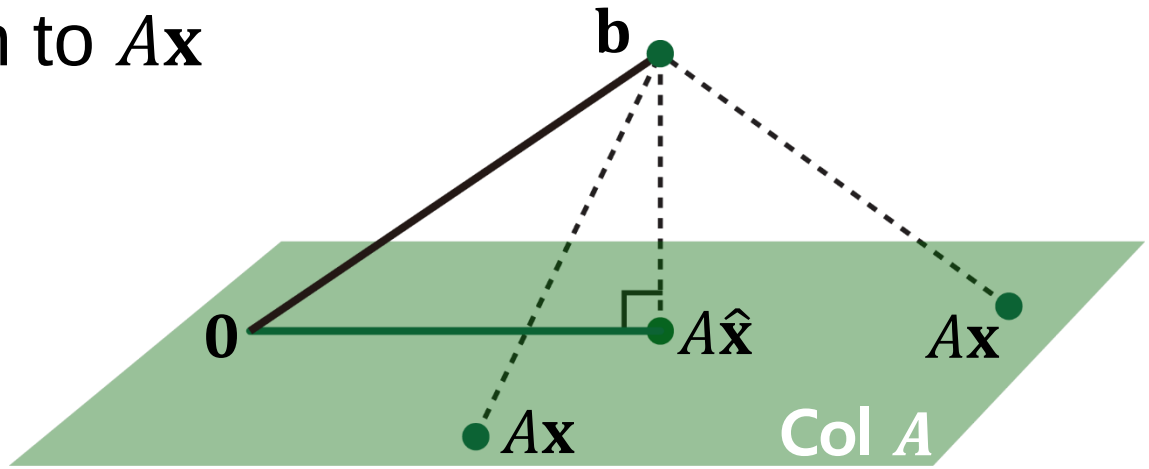
$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{b} - A\mathbf{x}\|$$

- The most important aspect of the least-squares problem is that no matter what **x** we select, the vector $A\mathbf{x}$ will necessarily be in the column space Col $A$.
- Thus, we seek for **x** that makes $A\mathbf{x}$ as the closest point in Col $A$ to $\mathbf{b}$.

# Geometric Interpretation of Least Squares

- The vector $\mathbf{b}$ is closer to $A\hat{\mathbf{x}}$ than to $A\mathbf{x}$ for other $\mathbf{x}$.

- To satisfy this, the vector $\mathbf{b} - A\hat{\mathbf{x}}$ should be orthogonal to Col $A$.



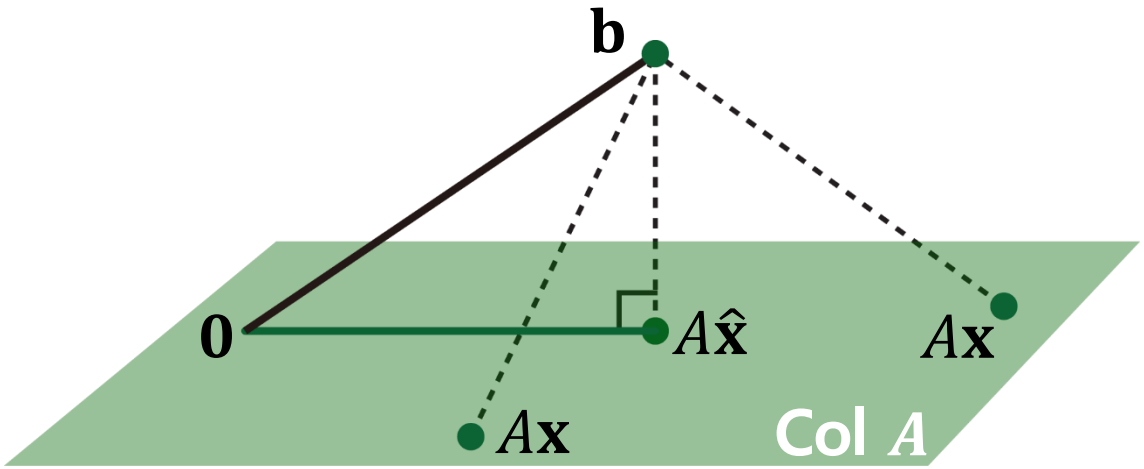

- This means $\mathbf{b} - A\hat{\mathbf{x}}$ should be orthogonal to any vector in Col $A$:

$$\mathbf{b} - A\hat{\mathbf{x}} \perp (x_1\mathbf{a}_1 + x_2\mathbf{a}_2 \cdots + x_n\mathbf{a}_n) \text{ for any vector } \mathbf{x}$$

x_hat: solution

# Geometric Interpretation of Least Squares

- $\mathbf{b} - A\hat{\mathbf{x}} \perp (x_1\mathbf{a}_1 + x_2\mathbf{a}_2 \cdots + x_n\mathbf{a}_n)$ for any vector $\mathbf{x}$




- Or equivalently, A의 모든 basis와 수직

$$\begin{aligned}(\mathbf{b} - A\hat{\mathbf{x}}) &\perp \mathbf{a}_1 \\ (\mathbf{b} - A\hat{\mathbf{x}}) &\perp \mathbf{a}_2 \\ &\vdots \\ (\mathbf{b} - A\hat{\mathbf{x}}) &\perp \mathbf{a}_n\end{aligned} \quad \Rightarrow \quad \begin{aligned}\mathbf{a}_1^T(\mathbf{b} - A\hat{\mathbf{x}}) &= 0 \\ \mathbf{a}_2^T(\mathbf{b} - A\hat{\mathbf{x}}) &= 0 \\ &\vdots \\ \mathbf{a}_n^T(\mathbf{b} - A\hat{\mathbf{x}}) &= 0\end{aligned} \quad \Rightarrow \quad \boxed{A^T(\mathbf{b} - A\hat{\mathbf{x}}) = \mathbf{0}}$$

# Normal Equation

- Finally, given a least squares problem, $A\mathbf{x} \simeq \mathbf{b}$, we obtain

$$A^T A\hat{\mathbf{x}} = A^T\mathbf{b},$$

  which is called a normal equation.

- This can be viewed as a new linear system, $C\mathbf{x} = \mathbf{d}$, where a square matrix $C = A^T A \in \mathbb{R}^{n\times n}$, and $\mathbf{d} = A^T\mathbf{b} \in \mathbb{R}^n$.

- If $C = A^T A$ is invertible, then the solution is computed as

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

# Another Derivation of Normal Equation

arg문제에 있어 단조증가 함수를 이용한 transfomation은 문제 없다.
likelyhood 문제 같은 경우 자주 쓰이는 트릭

- $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{b} - A\mathbf{x}\| = \arg\min_{\mathbf{x}} \|\mathbf{b} - A\mathbf{x}\|^2$

$$= \arg\min_{\mathbf{x}} (\mathbf{b} - A\mathbf{x})^T(\mathbf{b} - A\mathbf{x}) = \mathbf{b}^T\mathbf{b} - \mathbf{x}^T A^T\mathbf{b} - \mathbf{b}^T A\mathbf{x} + \mathbf{x}^T A^T A\mathbf{x}$$

scalar

- > 미분
$0 - A^Tb - (b^TA)^T + A^TAx + (x^TA^TA)^T$

$f(x) = a^Tx = x^Ta = (a^Tx)^T$
$f'(x) = a$

- Computing derivatives w.r.t. $\mathbf{x}$, we obtain

vector calculus

$$-A^T\mathbf{b} - A^T\mathbf{b} + 2A^TA\mathbf{x} = \mathbf{0} \quad \Leftrightarrow \quad A^TA\mathbf{x} = A^T\mathbf{b}$$

w/ rank x

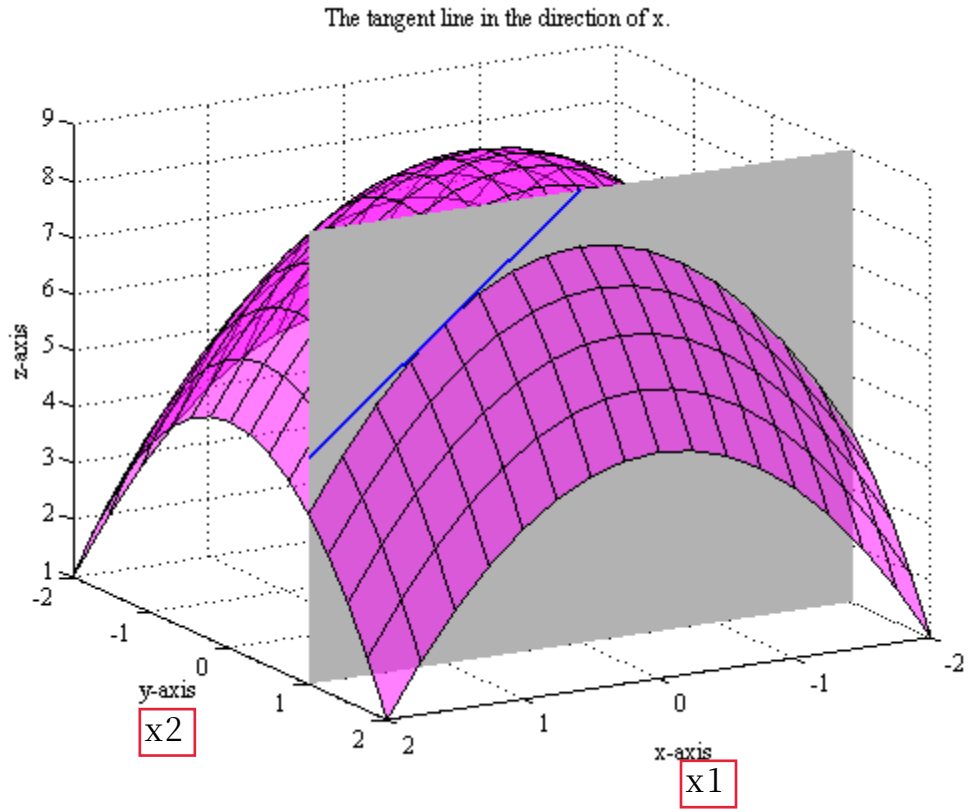
- Thus, if $C = A^TA$ is invertible, then the solution is computed as

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

역행렬 cost가 비쌈

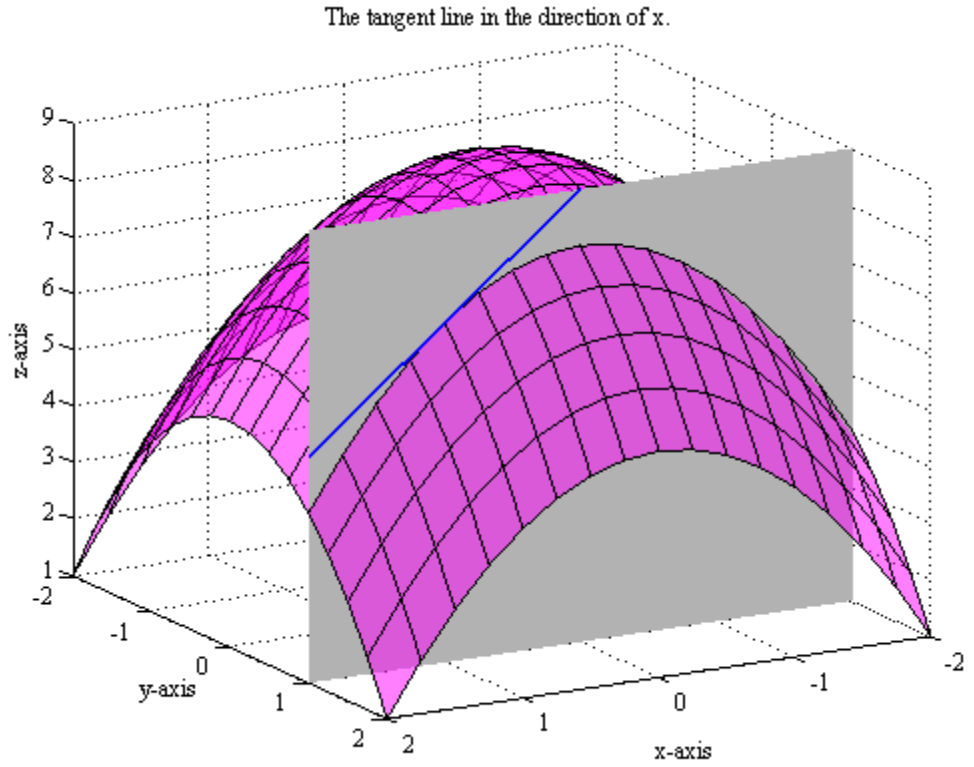
# Partial Derivative

- For a multi-variate function, e.g., $f(x, y)$, one can consider a univariate function by assigning particular values to all other variables, e.g., $g(x) = f(x, y = 1)$. Then, one can consider a partial derivative $\frac{d}{dx} g(x)$ with respect to $x$.



# Partial Derivative

- For a multi-variate function, e.g., $f(x, y)$, one can consider a univariate function by assigning particular values to all other variables, e.g., $g(x) = f(x, y = 1)$. Then, one can consider a partial derivative $\frac{d}{dx} g(x)$ with respect to $x$.



# Life-Span Example

| Person ID | Weight | Height | Is_smoking | Life-span |
|---|---|---|---|---|
| 1 | 60kg | 5.5ft | Yes (=1) | 66 |
| 2 | 65kg | 5.0ft | No (=0) | 74 |
| 3 | 55kg | 6.0ft | Yes (=1) | 78 |
| 4 | 50kg | 5.0ft | Yes (=1) | 72 |

$$\underset{A}{\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ 50 & 5.0 & 1 \end{bmatrix}} \underset{\mathbf{x}}{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}} \simeq \underset{\mathbf{b}}{\begin{bmatrix} 66 \\ 74 \\ 78 \\ 72 \end{bmatrix}}$$

- The normal equation $A^T A\hat{\mathbf{x}} = A^T\mathbf{b}$ is

Least squares solution
= minimum of sum of squares of error

$$\begin{bmatrix} 60 & 65 & 55 & 50 \\ 5.5 & 5.0 & 6.0 & 5.0 \\ 1 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ 50 & 5.0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 60 & 65 & 55 & 50 \\ 5.5 & 5.0 & 6.0 & 5.0 \\ 1 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 66 \\ 74 \\ 78 \\ 72 \end{bmatrix}$$

$$\begin{bmatrix} 13350 & 1235 & 165 \\ 1235 & 116.25 & 16.5 \\ 165 & 16.5 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 16600 \\ 1561 \\ 216 \end{bmatrix}$$

# What If $C = A^T A$ is NOT Invertible?

- Given $A^{\mathrm{T}}A\mathbf{x} = A^{\mathrm{T}}\mathbf{b}$, what if $C = A^T A$ is NOT invertible?
- Remember that in this case, the system has either no solution or infinitely many solutions.
- However, the solution always exist for this “normal” equation, and thus infinitely many solutions exist.

$A^{\mathrm{T}}A$의 해가 없는 경우는 없다,
수선의 발이 없는 경우는 없다

- When $C = A^T A$ is NOT invertible?
  If and only if the columns of $A$ are linearly dependent. Why?

해가 많을땐 risk mamagement를 위해 가중치 norm이 작은걸 택한다: regularization

해가 무수히 많다. 수선의 발은 있지만 발을 만들기 위한
x1, ..., xn set가 무수히 많다

- However, $C = A^T A$ is usually invertible. Why?

# Orthogonal Projection Perspective

- Back to the case of invertible $C = A^T A$, consider the orthogonal projection of $\mathbf{b}$ onto Col $A$ as

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

  where $C = A(A^T A)^{-1} A^T$.

- One can see that the orthogonal projection is actually a **linear transformation** $f(\mathbf{b}) = C\mathbf{b}$ where the standard matrix is defined as $C = A(A^T A)^{-1} A^T$.

- What if $A$ has orthonormal columns? (More in the next slides.)

# Orthogonal and Orthonormal Sets

- **Definition**: A set of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ in $\mathbb{R}^n$ is an **orthogonal set** if each pair of distinct vectors from the set is orthogonal That is, if $\mathbf{u}_i \cdot \mathbf{u}_j = 0$ whenever $i \neq j$.

- **Definition**: A set of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ in $\mathbb{R}^n$ is an **orthonormal set** if it is an orthogonal set of unit vectors.

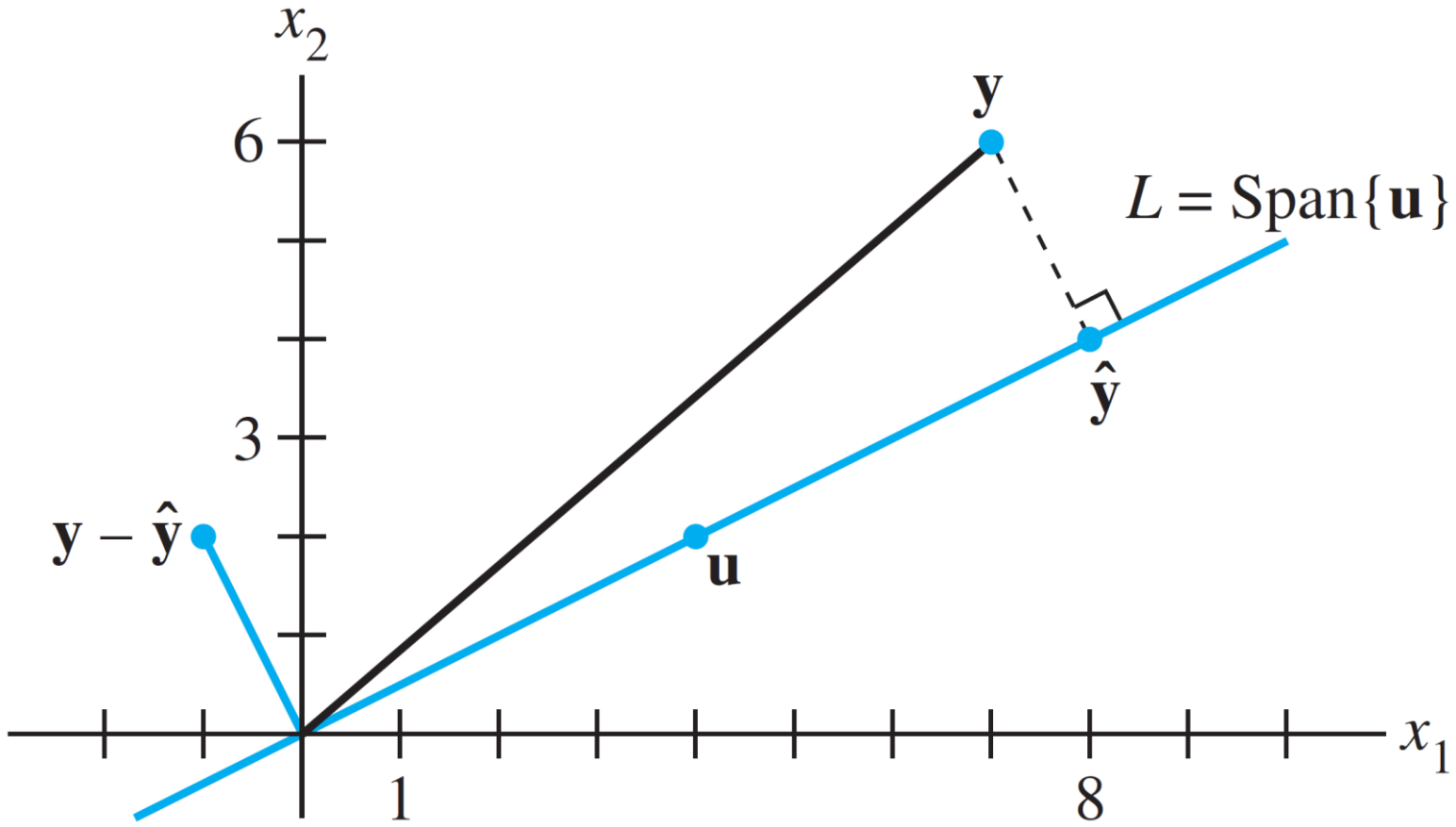
- Is an orthogonal (or orthonormal) set also a linearly independent set? What about its converse?

# Orthogonal and Orthonormal Basis

- Consider basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ of a $p$-dimensional subspace $W$ in $\mathbb{R}^n$.
- Can we make it as an orthogonal (or orthonormal) basis?
  - Yes, it can be done by Gram–Schmidt process. → QR factorization.
- Given the orthogonal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ of $W$,
  let's compute the orthogonal projection of $\mathbf{y} \in \mathbb{R}^n$ onto $W$.

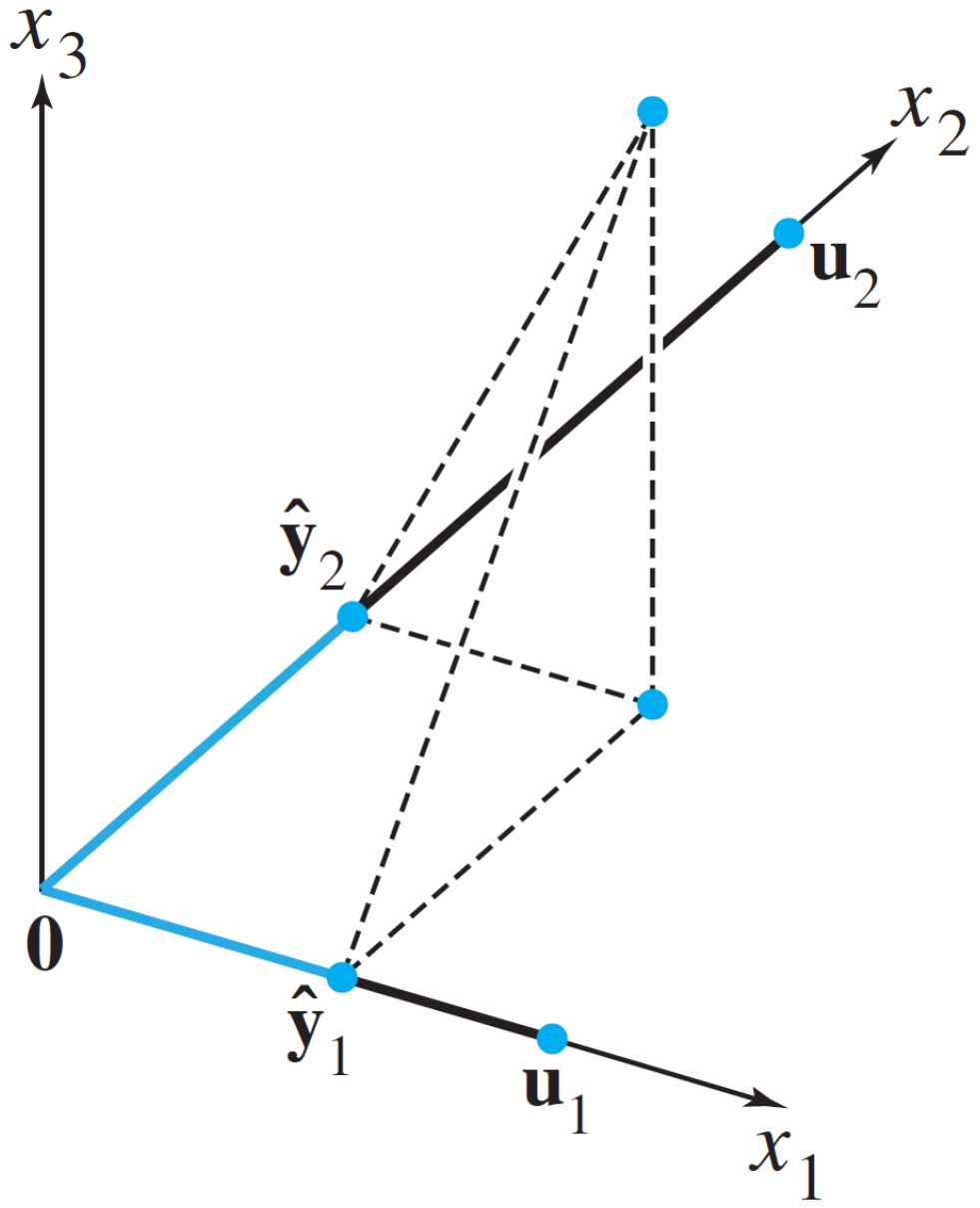
# Orthogonal Projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ onto Line

- Consider the orthogonal projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ onto one-dimensional subspace $L$.
- $\hat{\mathbf{y}} = \text{proj}_L\mathbf{y} = \frac{\mathbf{y}\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u}$
- If $\mathbf{u}$ is a unit vector, $\hat{\mathbf{y}} = \text{proj}_L\mathbf{y} = (\mathbf{y}\cdot\mathbf{u})\mathbf{u}$



# Orthogonal Projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ onto Plane

- Consider the orthogonal projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ onto two-dimensional subspace *W*
- $\hat{\mathbf{y}} = \text{proj}_L\mathbf{y} = \frac{\mathbf{y}\cdot\mathbf{u}_1}{\mathbf{u}_1\cdot\mathbf{u}_1}\mathbf{u}_1 + \frac{\mathbf{y}\cdot\mathbf{u}_2}{\mathbf{u}_2\cdot\mathbf{u}_2}\mathbf{u}_2$
- If $\mathbf{u}_1$ and $\mathbf{u}_2$ are unit vectors, $\hat{\mathbf{y}} = \text{proj}_L\mathbf{y} = (\mathbf{y}\cdot\mathbf{u}_1)\mathbf{u}_1 + (\mathbf{y}\cdot\mathbf{u}_2)\mathbf{u}_2$
- Projection is done independently on each orthogonal basis vector.



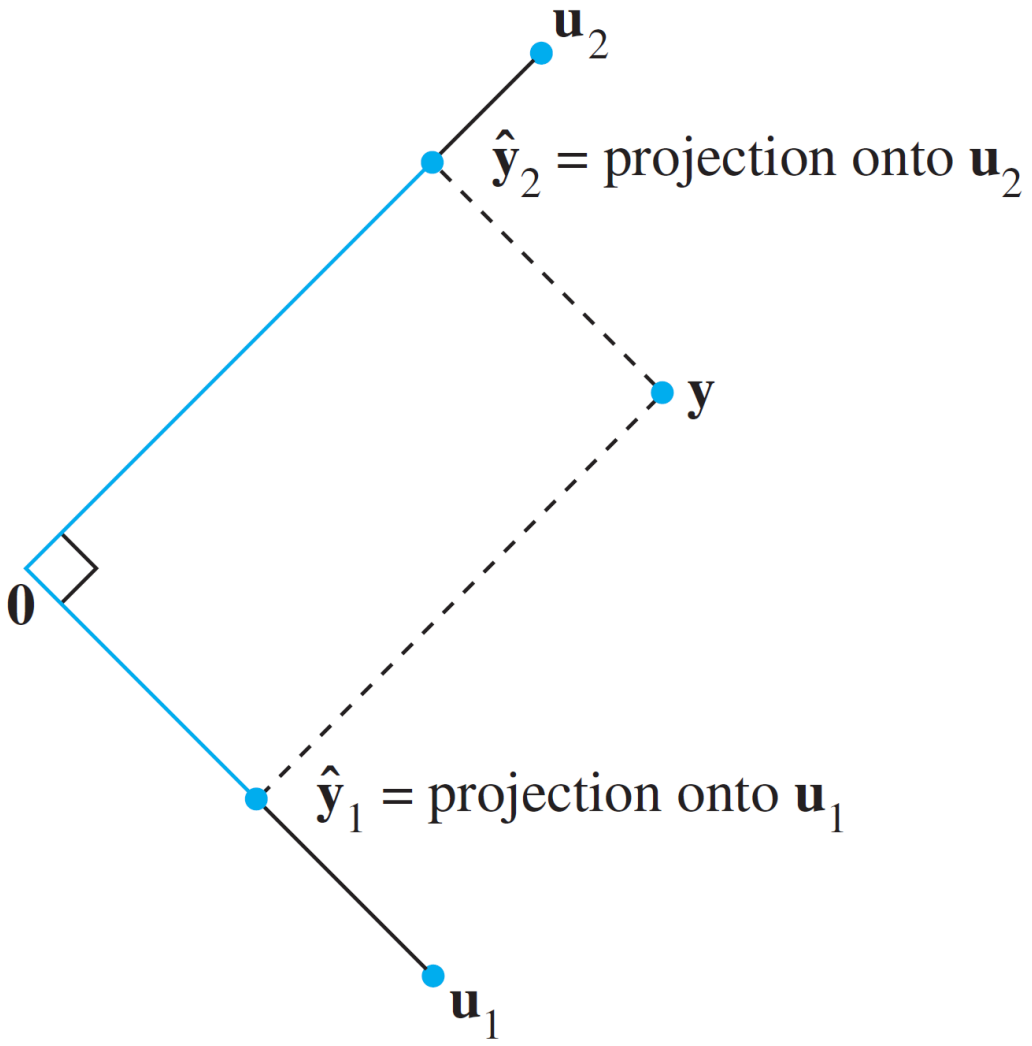
# Orthogonal Projection when $\mathbf{y} \in W$

- Consider the orthogonal projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ onto two-dimensional subspace $W$, where $\mathbf{y} \in W$

- $\hat{\mathbf{y}} = \text{proj}_L \mathbf{y} = \mathbf{y} = \frac{\mathbf{y} \cdot \mathbf{u}_1}{\mathbf{u}_1 \cdot \mathbf{u}_1} \mathbf{u}_1 + \frac{\mathbf{y} \cdot \mathbf{u}_2}{\mathbf{u}_2 \cdot \mathbf{u}_2} \mathbf{u}_2$

- If $\mathbf{u}_1$ and $\mathbf{u}_2$ are unit vectors, $\hat{\mathbf{y}} = \mathbf{y} = (\mathbf{y} \cdot \mathbf{u}_1)\mathbf{u}_1 + (\mathbf{y} \cdot \mathbf{u}_2)\mathbf{u}_2$

- The solution is the same as before. Why?



# Transformation: Orthogonal Projection

- Consider a transformation of orthogonal projection $\hat{\mathbf{b}}$ of $\mathbf{b}$, given orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2\}$ of a subspace $W$:

$$\hat{\mathbf{b}} = f(\mathbf{b}) = (\mathbf{b} \cdot \mathbf{u}_1)\mathbf{u}_1 + (\mathbf{b} \cdot \mathbf{u}_2)\mathbf{u}_2$$
$$= (\mathbf{u}_1^T\mathbf{b})\mathbf{u}_1 + (\mathbf{u}_2^T\mathbf{b})\mathbf{u}_2$$
$$= \mathbf{u}_1(\mathbf{u}_1^T\mathbf{b}) + \mathbf{u}_2(\mathbf{u}_2^T\mathbf{b})$$
$$= (\mathbf{u}_1\mathbf{u}_1^T)\mathbf{b} + (\mathbf{u}_2\mathbf{u}_2^T)\mathbf{b}$$
$$= (\mathbf{u}_1\mathbf{u}_1^T + \mathbf{u}_2\mathbf{u}_2^T)\mathbf{b}$$
$$= [\mathbf{u}_1 \quad \mathbf{u}_2]\begin{bmatrix}\mathbf{u}_1^T \\ \mathbf{u}_2^T\end{bmatrix}\mathbf{b} = UU^T\mathbf{b} = C\mathbf{b} \Rightarrow \text{linear transformation!}$$

# Orthogonal Projection Perspective

- Let's verify the following, when $A = U = [\mathbf{u}_1 \quad \mathbf{u}_2]$ has orthonormal columns:
  Back to the case of invertible $C = A^T A$, consider the orthogonal projection of $\mathbf{b}$ onto Col $A$ as

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

- $C = A^T A = \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \end{bmatrix} [\mathbf{u}_1 \quad \mathbf{u}_2] = I$. Thus,

$$\hat{\mathbf{b}} = A\hat{\mathbf{x}} = A(A^T A)^{-1} A^T \mathbf{b} = A(I)^{-1} A^T \mathbf{b} = AA^T \mathbf{b} = UU^T \mathbf{b}$$

# Further Study

- Least-squares derivation from maximum likelihood perspective (via Gaussian distribution)
  - Kevin Murphy, “Machine Learning: A Probabilistic Perspective,” Ch7.2
- Orthogonal projection and QR decomposition
  - Lay Ch6.2, Ch.6.3, Ch6.4

# Gram-Schmidt Orthogonalization

- **Example 1**: Let $W = \text{Span}\{\mathbf{x}_1, \mathbf{x}_2\}$, where $\mathbf{x}_1 = \begin{bmatrix} 3 \\ 6 \\ 0 \end{bmatrix}$ and $\mathbf{x}_2 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$.

  Construct an orthogonal basis $\{\mathbf{v}_1, \mathbf{v}_2\}$ for $W$.

- **Solution:** Let $\mathbf{v}_1 = \mathbf{x}_1$. Next, Let $\mathbf{v}_2$ the component of $\mathbf{x}_2$ orthogonal to $\mathbf{x}_1$, i.e.,

$$\mathbf{v}_2 = \mathbf{x}_2 - \frac{\mathbf{x}_2 \cdot \mathbf{x}_1}{\mathbf{x}_1 \cdot \mathbf{x}_1}\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} - \frac{15}{45}\begin{bmatrix} 3 \\ 6 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}.$$

- The set $\{\mathbf{v}_1, \mathbf{v}_2\}$ is an orthogonal basis for $W$.

# Gram-Schmidt Orthogonalization

- **Example 2**: Let $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$, $\mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$, and $\mathbf{x}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$. Then $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is clearly linearly independent and thus a basis for a subspace $W$ of $\mathbb{R}^4$. Construct an orthogonal basis for $W$.

# Gram-Schmidt Orthogonalization

- **Solution:**
- **Step 1.** Let $\mathbf{v}_1 = \mathbf{x}_1$ and $W_1 = \text{Span}\{\mathbf{x}_1\} = \text{Span}\{\mathbf{v}_1\}$.
- **Step 2.** Let $\mathbf{v}_2$ be the vector produced by subtracting from $\mathbf{x}_2$ its projection onto the subspace $W_1$. That is, let

$$\mathbf{v}_2 = \mathbf{x}_2 - \text{proj}_{W_1}\mathbf{x}_2 = \mathbf{x}_2 - \frac{\mathbf{x}_2 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 = \begin{bmatrix} -3/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$

- $\mathbf{v}_2$ is the component of $\mathbf{x}_2$ orthogonal to $\mathbf{x}_1$, and $\{\mathbf{v}_1, \mathbf{v}_2\}$ is an orthogonal basis for the subspace $W_2$ spanned by $\mathbf{x}_1$ and $\mathbf{x}_2$.

# Gram-Schmidt Orthogonalization

- **Step 2' (optional).** If appropriate, scale $\mathbf{v}_2$ to simplify later computations, e.g.,

$$\mathbf{v}_2 = \begin{bmatrix} -3/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} \longrightarrow \mathbf{v}_2' = \begin{bmatrix} -3 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

# Gram-Schmidt Orthogonalization

- **Step 3.** Let $\mathbf{v}_3$ be the vector produced by subtracting from $\mathbf{x}_3$ its projection onto the subspace $W_2$. Use the orthogonal basis $\{\mathbf{v}_1, \mathbf{v}_2'\}$ to compute this projection onto $W_2$:

$$\text{proj}_{W_2}\mathbf{x}_3 = \frac{\mathbf{x}_3 \cdot \mathbf{v}_1}{\mathbf{v}_3 \cdot \mathbf{v}_1}\mathbf{v}_1 + \frac{\mathbf{x}_3 \cdot \mathbf{v}_2'}{\mathbf{v}_3 \cdot \mathbf{v}_2'}\mathbf{v}_2' = \begin{bmatrix} 0 \\ 2/3 \\ 2/3 \\ 2/3 \end{bmatrix}$$

- Then $\mathbf{v}_3$ is the component of $\mathbf{x}_3$ orthogonal to $W_2$, namely,

$$\mathbf{v}_3 = \mathbf{x}_3 - \text{proj}_{W_2}\mathbf{x}_3 = \begin{bmatrix} 0 \\ -2/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

# Gram-Schmidt Orthogonalization

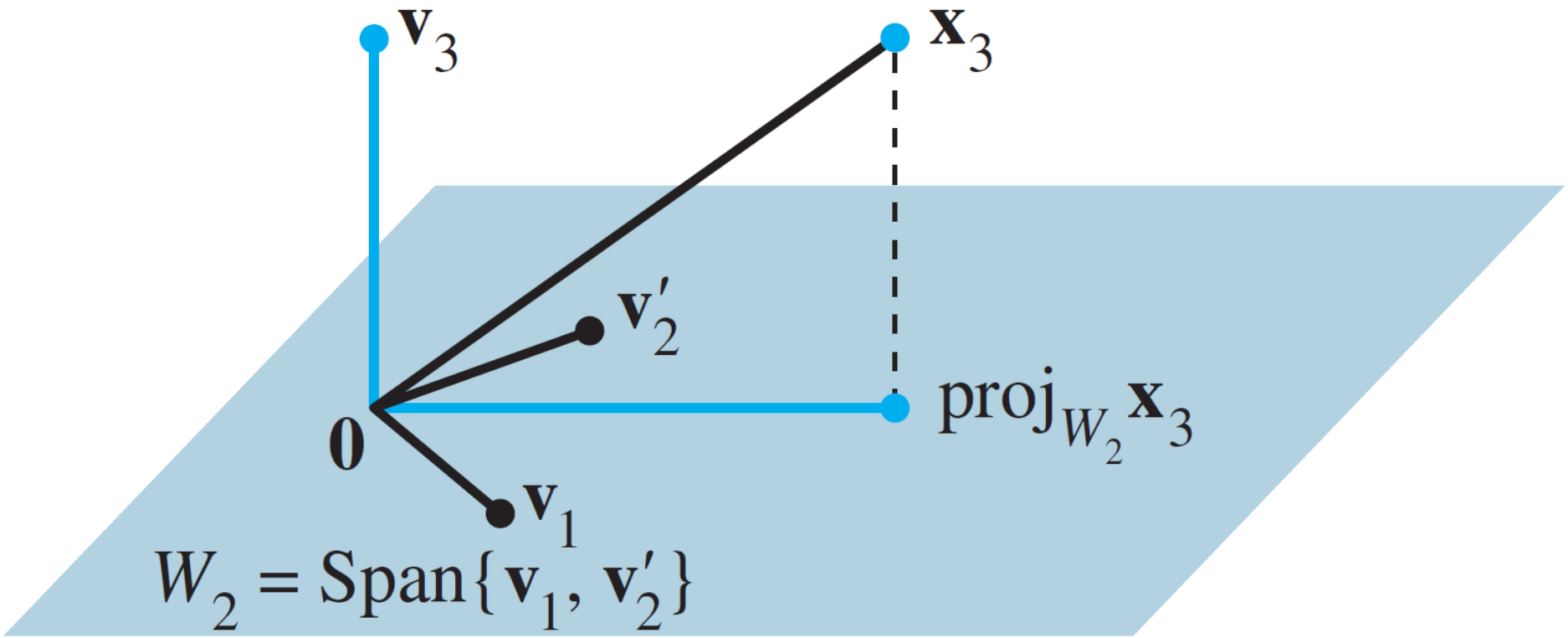



**FIGURE 2** The construction of $\mathbf{v}_3$ from $\mathbf{x}_3$ and $W_2$.

Figure from Lay Ch6.4

# QR Factorization

- If $A$ is an $m \times n$ matrix with linearly independent columns, then $A$ can be factored as $A = QR$, where $Q$ is an $m \times n$ matrix whose columns form an orthonormal basis for $\mathrm{Col}\, A$ and $R$ is an $n \times n$ upper triangular invertible matrix with positive entries on its diagonal.

# Computing QR Factorization

- **Step 1 (Construction of $Q$):** The columns of $A$ form a basis for $\mathrm{Col}\, A$ since they are linearly independent. Let these columns be $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Then, we can construct the orthonormal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ for $\mathrm{Col}\, A$ by the Gram-Schmidt process described by Theorem 11. Using this basis, we can construct $Q$ as

$$Q = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_n]$$

# Computing QR Factorization

- **Step 2 (Construction of $R$):** From (1) in Theorem 11, for $k = 1, \dots, n$, $\mathbf{x}_k$ is in $\text{Span}\{\mathbf{x}_1, \dots, \mathbf{x}_k\} = \text{Span}\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$. Therefore, there exist constants $r_{1k}, \dots, r_{kk}$ such that

$$\mathbf{x}_k = r_{1k}\mathbf{u}_1 + \cdots + r_{kk}\mathbf{u}_k + 0 \cdot \mathbf{u}_{k+1} + \cdots + 0 \cdot \mathbf{u}_n$$

- We can always make $r_{kk} \geq 0$ because if $r_{kk} < 0$, then we can multiply both $r_{kk}$ and $\mathbf{u}_k$ by -1. Using this linear combination representation, we can construct $\mathbf{r}_k$, the $k$-th column of $R$, as

$$\mathbf{r}_k = \begin{bmatrix} r_{1k} \\ \vdots \\ r_{kk} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

# Computing QR Factorization

- That is, $\mathbf{x}_k = Q\mathbf{r}_k$ for $k = 1, \ldots, n$. Let $R = [\mathbf{r}_1 \quad \cdots \quad \mathbf{r}_n]$. Then,
$$A = [\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_n] = [Q\mathbf{r}_1 \quad \cdots \quad Q\mathbf{r}_n] = QR$$

- The fact that $R$ is invertible follows easily from the fact that the columns of $A$ are linearly independent (Exercise 19). Since $R$ is clearly upper triangular (from the previous slide) and invertible, the diagonal entries $r_{kk}$'s should be nonzero. By combining this with the fact that $r_{kk} \geq 0$, $r_{kk}$'s must be positive.

# Example: QR Factorization

- **Example 4:** Find a QR factorization of $A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$.
- **Solution:** Let $A = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3]$. We first obtain $\mathbf{v}_1 = \mathbf{x}_1$ and its normalized vector is $\mathbf{u}_1 = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}$.
- Thus, $\mathbf{x}_1 = 2\mathbf{u}_1$, which gives us $\mathbf{r}_{11} = 2$, i.e., $\mathbf{r}_1 = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$.

# Example: QR Factorization

- Next, we obtain $\mathbf{v}_3$ as $\mathbf{v}_3 = \mathbf{x}_3 - \text{proj}_{W_2}\mathbf{x}_3 = \mathbf{x}_3 - \frac{\mathbf{x}_3 \cdot \mathbf{u}_1}{\mathbf{u}_1 \cdot \mathbf{u}_1}\mathbf{u}_1 - \frac{\mathbf{x}_3 \cdot \mathbf{u}_2}{\mathbf{u}_2 \cdot \mathbf{u}_2}\mathbf{u}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} - 1\begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix} - \frac{2}{\sqrt{12}}\begin{bmatrix} -3/\sqrt{12} \\ 1/\sqrt{12} \\ 1/\sqrt{12} \\ 1/\sqrt{12} \end{bmatrix} = \begin{bmatrix} 0 \\ -2/3 \\ 1/3 \\ 1/3 \end{bmatrix}$ and its normalized vector $\mathbf{u}_2$ as $\mathbf{u}_2 = \begin{bmatrix} 0 \\ -2/\sqrt{6} \\ 1/\sqrt{6} \\ 1/\sqrt{6} \end{bmatrix}$.
- Thus, $\mathbf{x}_3 = 1\mathbf{u}_1 + \frac{2}{\sqrt{12}}\mathbf{u}_2 + \frac{2}{\sqrt{6}}\mathbf{u}_3$, i.e., $\mathbf{r}_3 = \begin{bmatrix} 1 \\ 2/\sqrt{12} \\ 2/\sqrt{6} \end{bmatrix}$.

# Example: QR Factorization

- In conclusion, $Q = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \mathbf{u_3}] = \begin{bmatrix} 1/2 & -3/\sqrt{12} & 0 \\ 1/2 & 1/\sqrt{12} & -2/\sqrt{6} \\ 1/2 & 1/\sqrt{12} & 1/\sqrt{6} \\ 1/2 & 1/\sqrt{12} & 1/\sqrt{6} \end{bmatrix}$

and $R = [\mathbf{r}_1 \quad \mathbf{r}_2 \quad \mathbf{r}_3] = \begin{bmatrix} 2 & -3/2 & 1 \\ 0 & -3/\sqrt{12} & 2/\sqrt{12} \\ 0 & 0 & 2/\sqrt{6} \end{bmatrix}$.

# Lecture Overview

- Elements in linear algebra
- Linear system
- Linear combination, vector equation,
  Four views of matrix multiplication
- Linear independence, span, and subspace
- Linear transformation
- Least squares
- Eigendecomposition
- Singular value decomposition
- Principal component analysis

# Eigenvectors and Eigenvalues

- **Definition**: An **eigenvector** of a **square** matrix $A \in \mathbb{R}^{n \times n}$ is a **nonzero** vector $\mathbf{x} \in \mathbb{R}^n$ such that $A\mathbf{x} = \lambda\mathbf{x}$ for some scalar $\lambda$
  In this case, $\lambda$ is called an **eigenvalue** of $A$, and
  such an $\mathbf{x}$ is called an ***eigenvector corresponding to $\lambda$***.

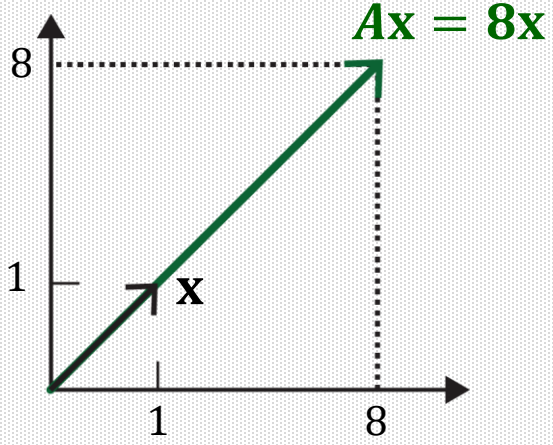
# Transformation Perspective

- Consider a linear transformation $T(\mathbf{x}) = A\mathbf{x}$.
- If $\mathbf{x}$ is an eigenvector, then $T(\mathbf{x}) = A\mathbf{x} = \lambda\mathbf{x}$, which means the output vector has **the same direction** as $\mathbf{x}$, but the length is scaled by a factor of $\lambda$.

- **Example:** For $A = \begin{bmatrix} 2 & 6 \\ 5 & 3 \end{bmatrix}$, an eigenvector is $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ since

$$T(\mathbf{x}) = A\mathbf{x} = \begin{bmatrix} 2 & 6 \\ 5 & 3 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 8 \\ 8 \end{bmatrix} = 8\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$A \quad \mathbf{x} \quad = \quad 8\,\mathbf{x}$$



# Computational Advantage

- Which computation is faster between $\begin{bmatrix} 2 & 6 \\ 5 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $8 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$?

# Eigenvectors and Eigenvalues

- The equation $A\mathbf{x} = \lambda\mathbf{x}$ can be re-written as

$$(A - \lambda I)\mathbf{x} = \mathbf{0}$$

- $\lambda$ is an eigenvalue of an $n \times n$ matrix $A$ if and only if this equation has a **nontrivial** solution (since $\mathbf{x}$ should be a nonzero vector).

# Eigenvectors and Eigenvalues

$$(A - \lambda I)\mathbf{x} = \mathbf{0}$$

- The set of *all* solutions of the above equation is the **null space** of the matrix $(A - \lambda I)$, which we call the **eigenspace** of $A$ **corresponding to $\lambda$**.
- The eigenspace consists of the zero vector and all the eigenvectors corresponding to $\lambda$, satisfying the above equation.

# Example: Eigenvalues and Eigenvectors

- **Example**: Show that $8$ is an eigenvalue of a matrix

  $A = \begin{bmatrix} 2 & 6 \\ 5 & 3 \end{bmatrix}$ and find the corresponding eigenvectors.

- **Solution**: The scalar 8 is an eigenvalue of $A$ if and only if the equation $(A - 8I)\mathbf{x} = \mathbf{0}$ has a nontrivial solution:

$$(A - 8I)\mathbf{x} = \begin{bmatrix} -6 & 6 \\ 5 & -5 \end{bmatrix}\mathbf{x} = \mathbf{0}$$

- The solution is $\mathbf{x} = c\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ for any nonzero scalar $c$, which is Span $\left\{\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right\}$.

# Example: Eigenvalues and Eigenvectors

- In the previous example, $-3$ is also an eigenvalue:

$$(A + 3I)\mathbf{x} = \begin{bmatrix} 5 & 6 \\ 5 & 6 \end{bmatrix} \mathbf{x} = \mathbf{0}$$

- The solution is $\mathbf{x} = c \begin{bmatrix} 1 \\ -5/6 \end{bmatrix}$ for any nonzero scalar $c$, which is Span $\left\{ \begin{bmatrix} 1 \\ -5/6 \end{bmatrix} \right\}$.

# Characteristic Equation

- How can we find the eigenvalues such as $8$ and $-3$?
- If $(A-\lambda I)\mathbf{x}=\mathbf{0}$ has a nontrivial solution, then the columns of $(A-\lambda I)$ should be noninvertible.
- If it is invertible, $\mathbf{x}$ cannot be a nonzero vector since

$$(A-\lambda I)^{-1}(A-\lambda I)\mathbf{x}=(A-\lambda I)^{-1}\mathbf{0}\Longrightarrow\mathbf{x}=\mathbf{0}$$

- Thus, we can obtain eigenvalues by solving

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

called a **characteristic equation**.

- Also, the solution is not unique, and thus $A-\lambda I$ has linearly dependent columns.

# Example: Characteristic Equation

- In the previous example, $A = \begin{bmatrix} 2 & 6 \\ 5 & 3 \end{bmatrix}$ is originally invertible since

$$\det(A) = \det\begin{bmatrix} 2 & 6 \\ 5 & 3 \end{bmatrix} = 6 - 30 = -24 \neq 0.$$

- By solving the characteristic equation, we want to find $\lambda$ that makes $A - \lambda I$ non-invertible:

$$\det(A - \lambda I) = \det\begin{bmatrix} 2-\lambda & 6 \\ 5 & 3-\lambda \end{bmatrix}$$

$$= (2-\lambda)(3-\lambda) - 30$$

$$= -\lambda^2 - 5\lambda - 25 = (8-\lambda)(-3-\lambda) = 0$$

$$\lambda = -3 \text{ or } 8$$

# Example: Characteristic Equation

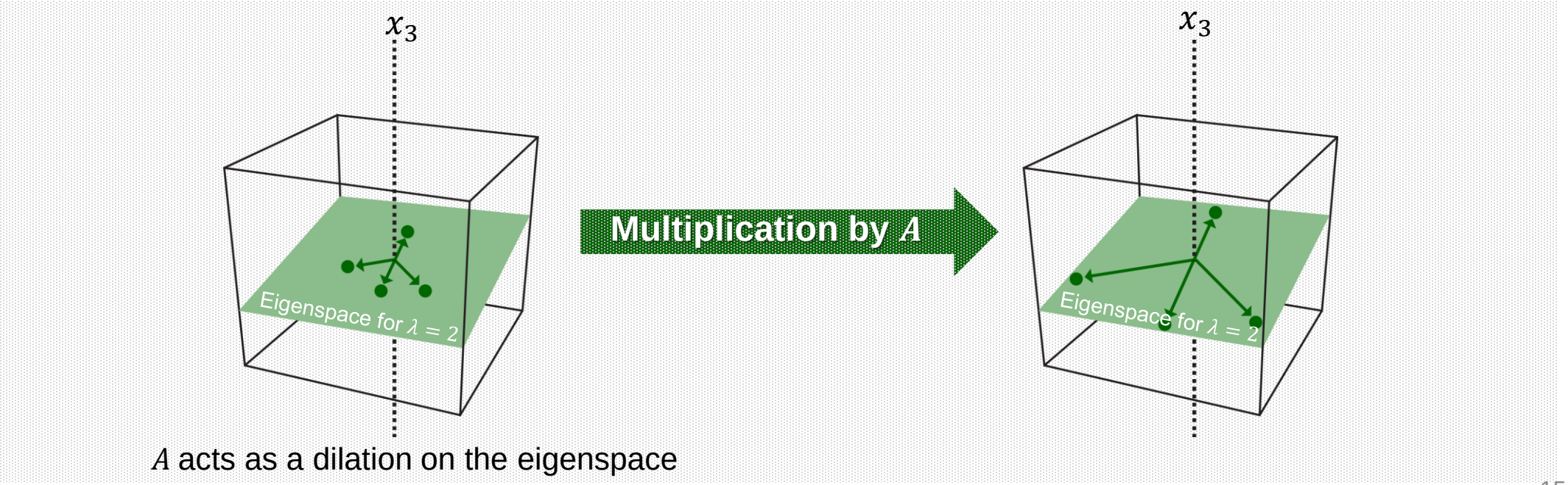
- Once obtaining eigenvalues, we compute the eigenvectors for each $\lambda$ by solving

$$(A - \lambda I)\mathbf{x} = \mathbf{0}$$

# Eigenspace

- Note that the dimension of the eigenspace (corresponding to a particular $\lambda$) can be larger than one. In this case, any vector in the eigenspace satisfies

$$T(\mathbf{x}) = A\mathbf{x} = \lambda\mathbf{x}$$




$A$ acts as a dilation on the eigenspace

# Finding All Eigenvalues and Eigenvectors

- In summary, we can find all the possible eigenvalues and eigenvectors, as follows.
- First, find all the eigenvalue by solving the **characteristic equation**:

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

- Second, for each eigenvalue $\lambda$, solve for $(A - \lambda I)\mathbf{x} = \mathbf{0}$ and obtain the set of basis vectors of the corresponding eigenspace.

# Diagonalization

- We want to change a given square matrix $A \in \mathbb{R}^{n\times n}$ into a diagonal matrix via the following form:

$$D = P^{-1}AP$$

where $P \in \mathbb{R}^{n\times n}$ is an invertible matrix and $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix. This is called a diagonalization of $A$.

- It is not always possible to diagonalize $A$. For $A$ to be diagonalizable, an invertible $P$ should exist such that $P^{-1}AP$ becomes a diagonal matrix.

# Finding $P$ and $D$

- How can we find an invertible $P$ and the resulting diagonal matrix $D = P^{-1}AP$?
- $D = P^{-1}AP \Longrightarrow PD = AP$
- Let us represent the following:
- $P = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_n]$ where $\mathbf{v}_i$'s are column vectors of $P$
- $D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}$

# Finding $P$ and $D$

- $AP = A[\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_n] = [A\mathbf{v}_1 \quad A\mathbf{v}_2 \quad \cdots \quad A\mathbf{v}_n]$
- $PD = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_n] \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}$

  $= [\lambda_1\mathbf{v}_1 \quad \lambda_2\mathbf{v}_2 \quad \cdots \quad \lambda_n\mathbf{v}_n]$
- $PD = AP \iff [A\mathbf{v}_1 \quad A\mathbf{v}_2 \quad \cdots \quad A\mathbf{v}_n] = [\lambda_1\mathbf{v}_1 \quad \lambda_2\mathbf{v}_2 \quad \cdots \quad \lambda_n\mathbf{v}_n]$

# Finding $P$ and $D$

- Equating columns, we obtain

$$A\mathbf{v}_1 = \lambda_1\mathbf{v}_1,\ A\mathbf{v}_2 = \lambda_2\mathbf{v}_2,\ldots,\ A\mathbf{v}_n = \lambda_n\mathbf{v}_n$$

- Thus, $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ should be eigenvectors and $\lambda_1, \lambda_2, \ldots, \lambda_n$ should be eigenvalues.

- Then, for $PD = AP \Longrightarrow D = P^{-1}AP$ to be true, $P$ should invertible.

- In this case, the resulting diagonal matrix $D$ has eigenvalues as diagonal entries.

# Diagonalizable Matrix

- For $P$ to be invertible,
  $P$ should be a square matrix in $\mathbb{R}^{n \times n}$, and
  $P$ should have $n$ linearly independent columns.
- Recall columns of $P$ are eigenvectors.
  Hence, $A$ should have $n$ linearly independent eigenvectors.
- It is not always the case, but if it is, $A$ is diagonalizable.

# Eigendecomposition

- If $A$ is diagonalizable, we can write $D = P^{-1}AP$.

- We can also write $A = PDP^{-1}$.
  which we call eigendecomposition of $A$.

- $A$ being diagonalizable is equivalent to
  $A$ having eigendecomposition.

# Linear Transformation via Eigendecomposition

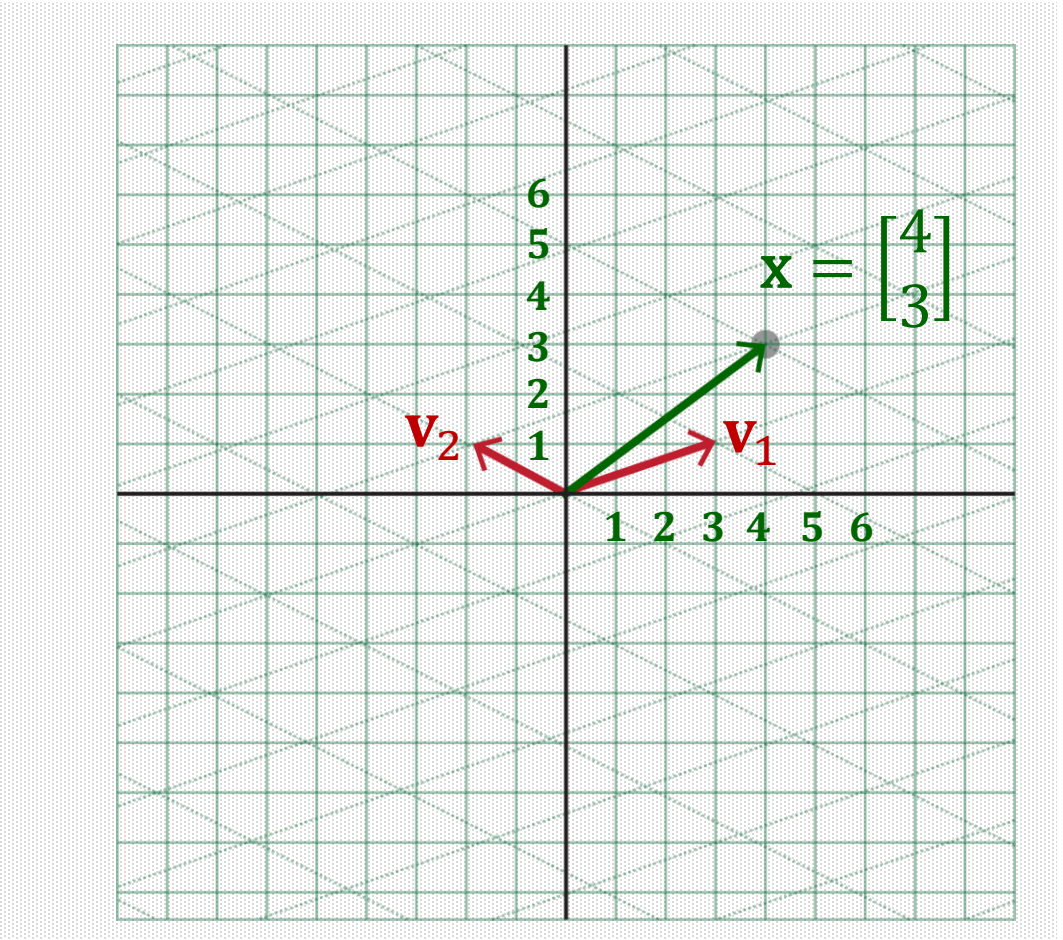
- Suppose $A$ is diagonalizable, thus having eigendecomposition

$$A = PDP^{-1}$$

- Consider the linear transformation $T(\mathbf{x}) = A\mathbf{x}$.
- $T(\mathbf{x}) = A\mathbf{x} = PDP^{-1}\mathbf{x} = P\big(D(P^{-1}\mathbf{x})\big)$.

# Change of Basis

- Suppose $A\mathbf{v}_1 = -1\mathbf{v}_1$ and $A\mathbf{v}_2 = 2\mathbf{v}_2$.
- $T(\mathbf{x}) = A\mathbf{x} = PDP^{-1}\mathbf{x} = P\big(D(P^{-1}\mathbf{x})\big)$
- Let $\mathbf{y} = P^{-1}\mathbf{x}$. Then,
$$P\mathbf{y} = \mathbf{x}$$
- $\mathbf{y}$ is a new coordinate of $\mathbf{x}$ with respect to a new basis of eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2\}$.



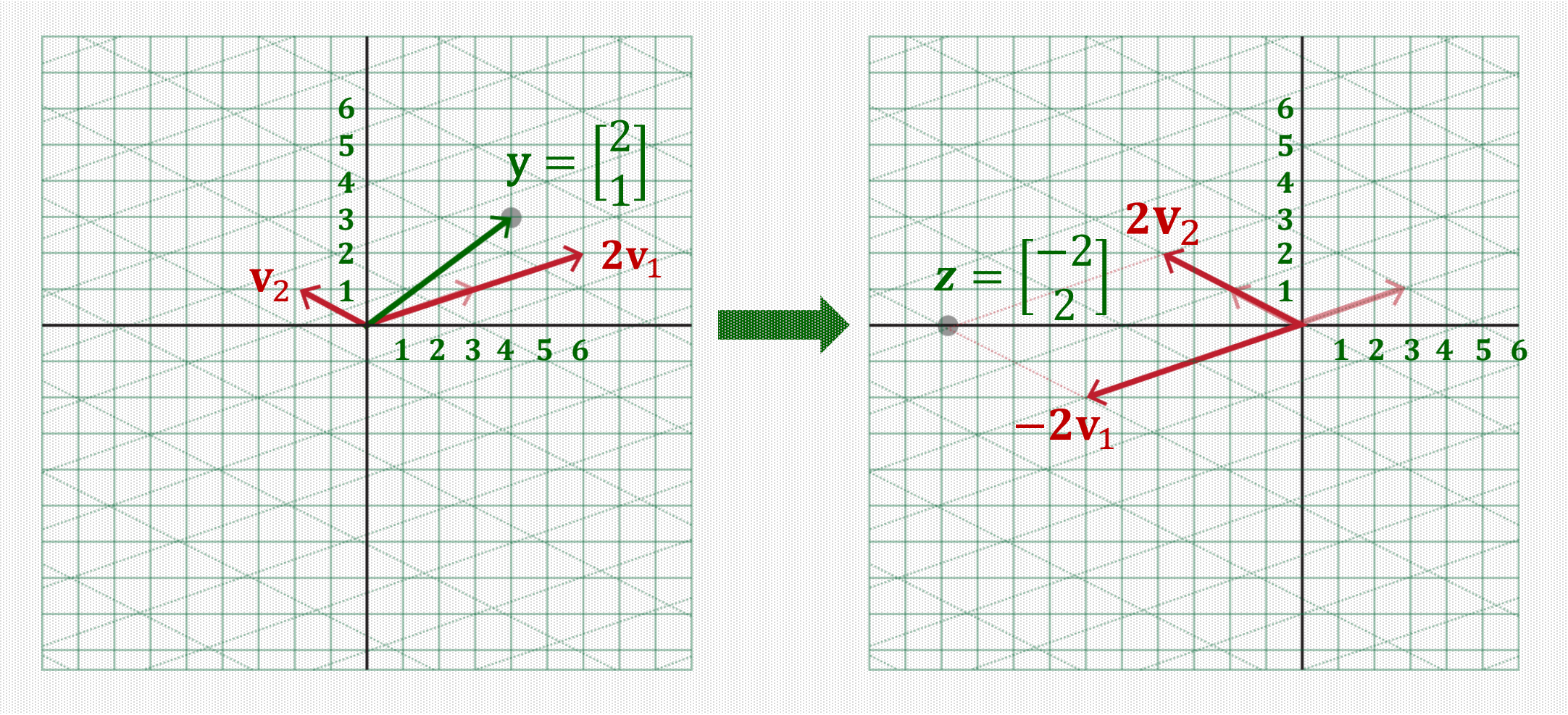

$$\mathbf{x} = \begin{bmatrix} 4 \\ 3 \end{bmatrix} = 4\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 3\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 4 \\ 3 \end{bmatrix} = P\mathbf{y} = [\mathbf{v}_1 \quad \mathbf{v}_2]\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = 2\mathbf{v}_1 + 1\mathbf{v}_2 \Longrightarrow \mathbf{y} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

# Element-wise Scaling

- $T(\mathbf{x}) = P\left(D\left(P^{-1}\mathbf{x}\right)\right) = P(D\mathbf{y})$
- Let $z = D\mathbf{y}$. This computation is a simple element-wise scaling of $\mathbf{y}$.
- **Example**: Suppose $D = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix}$. Then

$$z = D\mathbf{y} = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} (-1) \times 2 \\ 2 \times 1 \end{bmatrix} = \begin{bmatrix} -2 \\ 2 \end{bmatrix}$$

# Dimension-wise Scaling



# Back to Original Basis

- $T(\mathbf{x}) = P(D\mathbf{y}) = Pz$
- $z$ is still a coordinate based on the new basis $\{\mathbf{v}_1, \mathbf{v}_2\}$.
- $Pz$ converts $z$ to another coordinates based on the original standard basis.
- That is, $Pz$ is a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$ using the coefficient vector $z$.
- That is,

$$Pz = [\mathbf{v}_1 \quad \mathbf{v}_2] \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \mathbf{v}_1 z_1 + \mathbf{v}_2 z_2$$

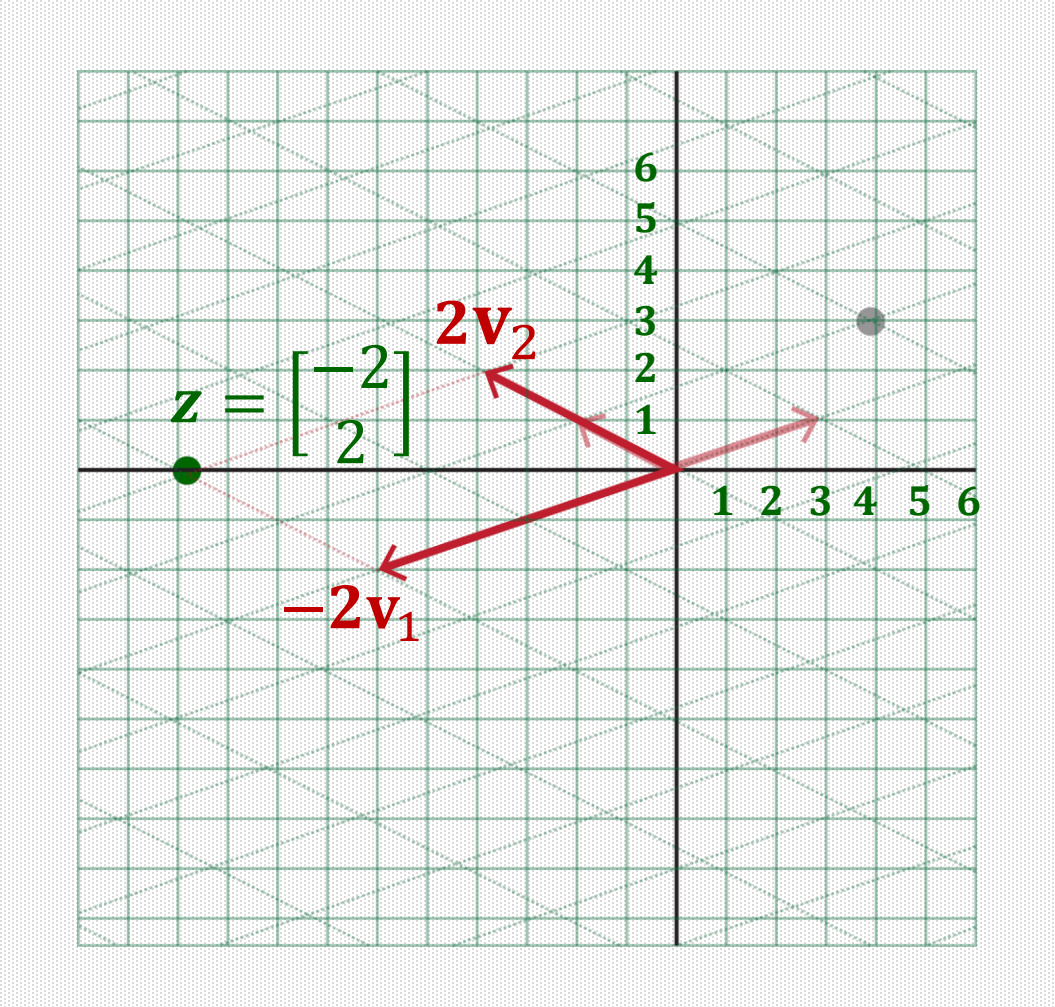
# Back to Original Basis

- $T(\mathbf{x}) = Pz = [\mathbf{v}_1 \quad \mathbf{v}_2]\begin{bmatrix}-2\\2\end{bmatrix}$

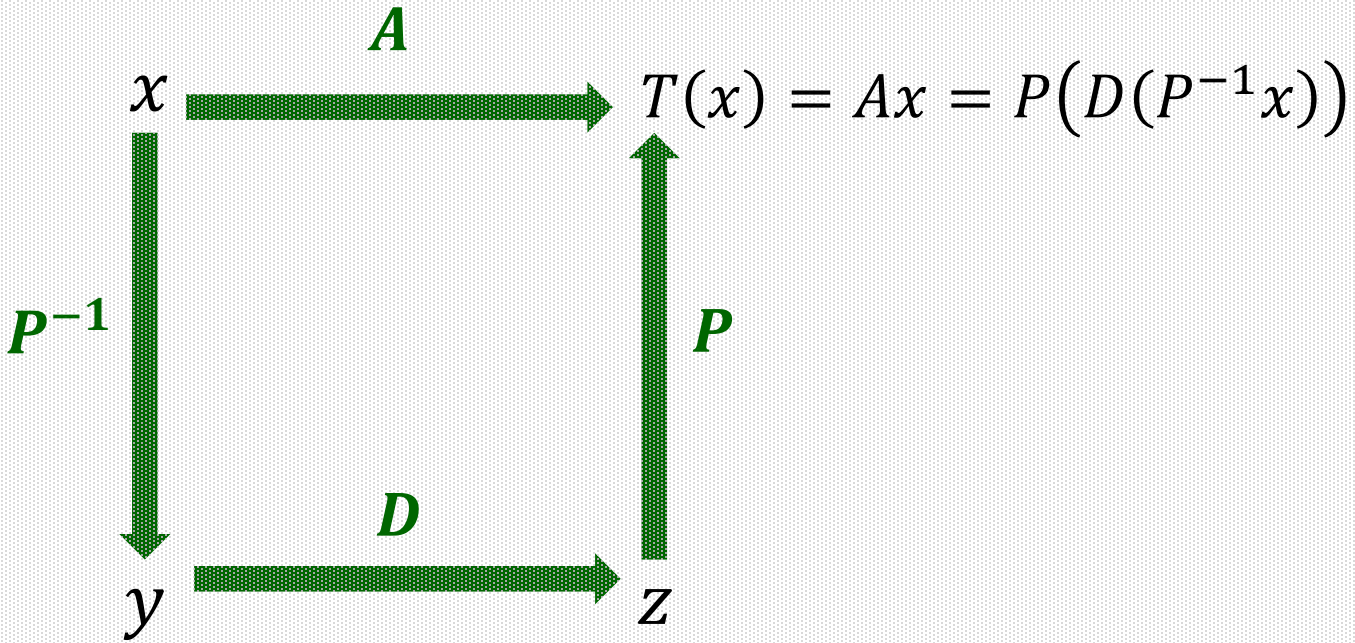
$$= -2\mathbf{v}_1 + 2\mathbf{v}_2$$

$$= -2\begin{bmatrix}3\\1\end{bmatrix} + 2\begin{bmatrix}-2\\1\end{bmatrix}$$

$$= \begin{bmatrix}-10\\0\end{bmatrix}$$



# Overview of Transformation using Eigendecomposition



# Linear Transformation via $A^k$

- Now, consider recursive transformation $A \times A \times \cdots \times A\mathbf{x} = A^k\mathbf{x}$.
- If $A$ is diagonalizable, $A$ has eigendecomposition

$$A = PDP^{-1}$$

- $A^k = (PDP^{-1})(PDP^{-1}) \cdots (PDP^{-1}) = PD^kP^{-1}$
- $D^k$ is simply computed as

$$D^k = \begin{bmatrix} \lambda_1^k & 0 & \cdots & 0 \\ 0 & \lambda_2^k & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n^k \end{bmatrix}$$

# Linear Transformation via $A^k$

- $A^k\mathbf{x} = PD^kP^{-1}\mathbf{x}$ can be computed in the similar manner to the previous example.
- It is much faster to compute $P\left(D^k(P^{-1}\mathbf{x})\right)$ than to compute $A^k\mathbf{x}$.

# Lecture Overview

- Elements in linear algebra
- Linear system
- Linear combination, vector equation,
  Four views of matrix multiplication
- Linear independence, span and subspace
- Linear transformation
- Least squares
- Eigendecomposition
- Singular value decomposition
- Principal component analysis

# Singular Value Decomposition (SVD)

- Given a rectangular matrix $A \in \mathbb{R}^{m \times n}$,
  its singular value decomposition is written as

$$A = U \Sigma V^T$$

  where

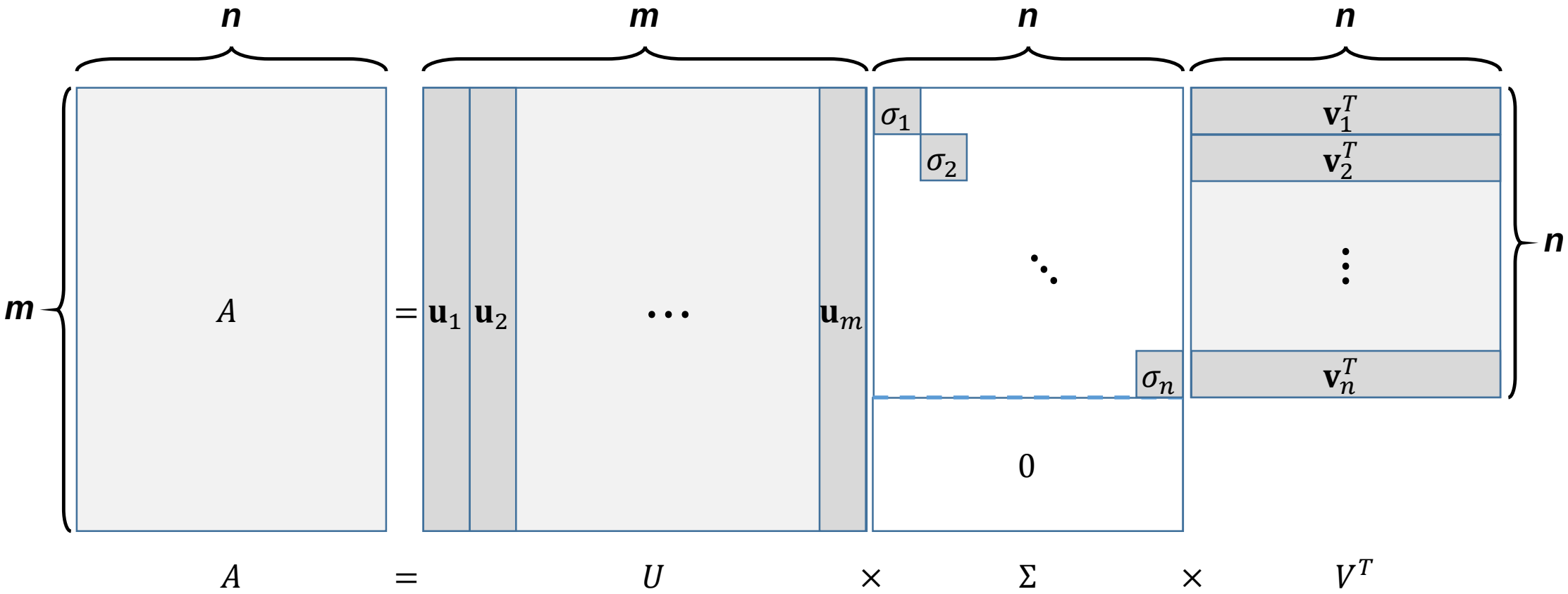
- $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$: matrices with orthonormal columns, providing an orthonormal basis of Col $A$ and Row $A$, respectively

- $\Sigma \in \mathbb{R}^{m \times n}$: a diagonal matrix whose entries are in a decreasing order, i.e., $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)}$

# Basic Form of SVD

- Given a matrix $A \in \mathbb{R}^{m \times n}$ where $m > n$, SVD gives

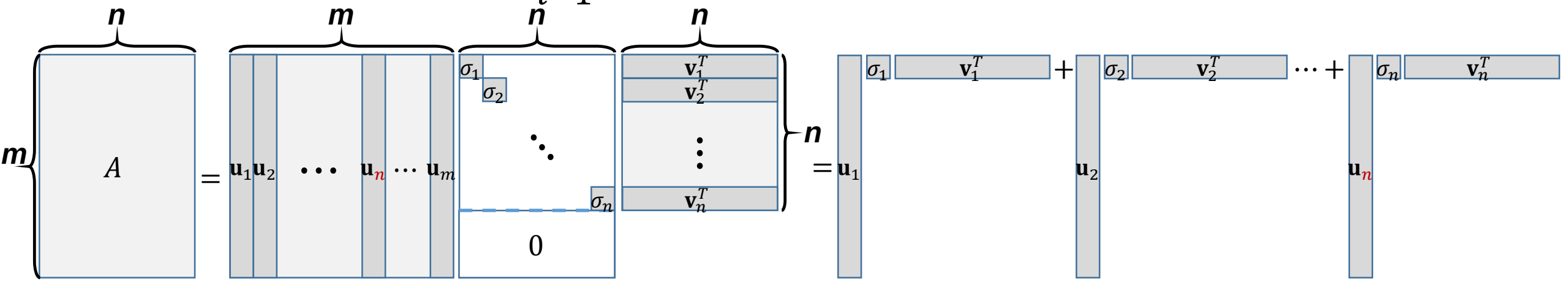
$$A = U\Sigma V^T$$



# SVD as Sum of Outer Products

- $A$ can also be represented as the sum of outer products

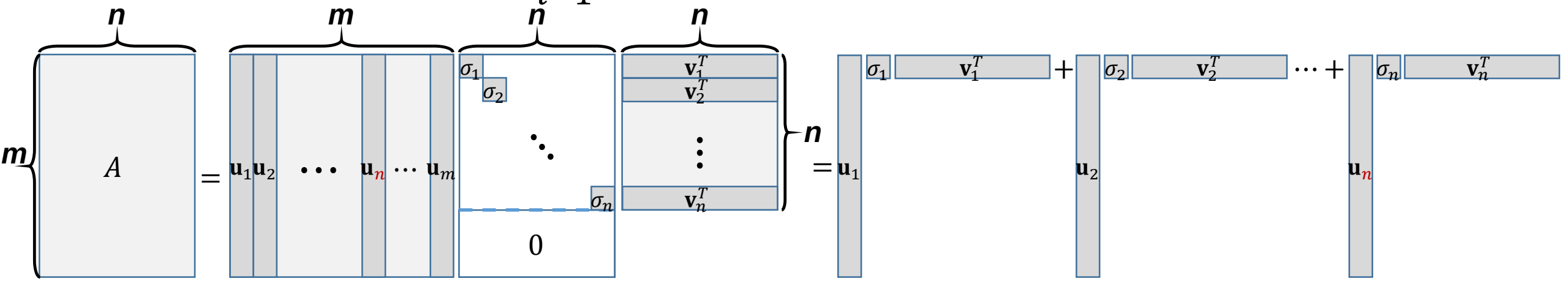
$$A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \qquad \text{where } \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$$

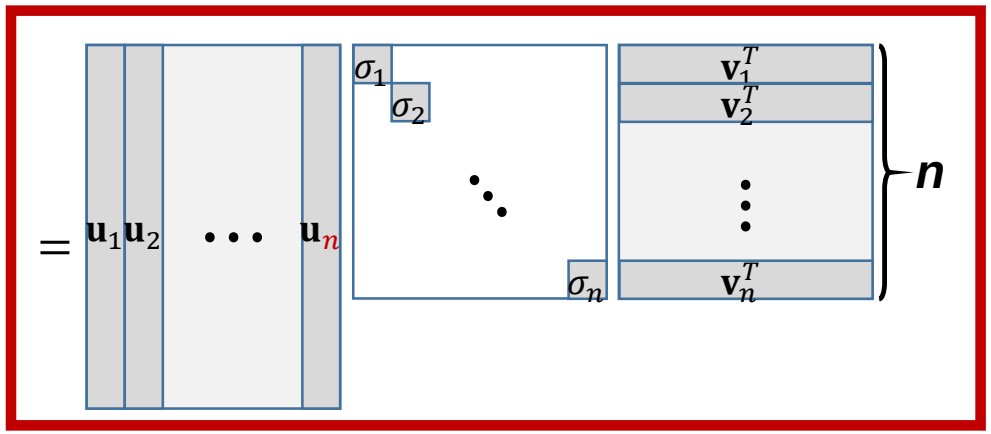


# Reduced Form of SVD

- $A$ can also be represented as the sum of outer products

$$A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \qquad \text{where } \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$$



# Another Perspective of SVD

- We can easily find two orthonormal basis sets, $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ for Col $A$ and $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ for Row $A$, by using, say, Gram–Schmidt orthogonalization.
- Are these unique orthonormal basis sets?

- No. Then, can we jointly find them such that

$$A\mathbf{v}_i = \sigma_i \mathbf{u}_i, \qquad \forall\, i = 1, \dots, n$$

$n$

# Another Perspective of SVD

- Let us denote $U = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_n] \in \mathbb{R}^{m\times n}$, $V = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_n] \in \mathbb{R}^{n\times n}$, and $\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n \end{bmatrix} \in \mathbb{R}^{n\times n}$
- Consider $AV = A[\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_n] = [A\mathbf{v}_1 \quad A\mathbf{v}_2 \quad \cdots \quad A\mathbf{v}_n]$ and

$$U\Sigma = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_n] \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n \end{bmatrix}$$

$$= [\sigma_1\mathbf{u}_1 \quad \sigma_2\mathbf{u}_2 \quad \cdots \quad \sigma_n\mathbf{u}_n]$$

- $AV = U\Sigma \Leftrightarrow [A\mathbf{v}_1 \quad A\mathbf{v}_2 \quad \cdots \quad A\mathbf{v}_n] = [\sigma_1\mathbf{u}_1 \quad \sigma_2\mathbf{u}_2 \quad \cdots \quad \sigma_n\mathbf{u}_n]$
- $V^{-1} = V^T$ since $V \in \mathbb{R}^{n\times n}$ has orthonormal columns.
- Thus $AV = U\Sigma \Leftrightarrow A = U\Sigma V^T$

# Computing SVD

- First, we form $AA^T$ and $A^TA$ and compute eigendecomposition of each:

$$AA^T = U\Sigma V^T \mathrm{V}\Sigma^T U^T = U\Sigma\Sigma^T U^T = U\Sigma^2 U^T$$
$$A^TA = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T\Sigma V^T = V\Sigma^2 V^T$$

- Can we find the following?
  1. Orthogonal eigenvector matrices $U$ and $V$
  2. Eigenvalues in $\Sigma^2$ that are all positive
  3. Eigenvalues in $\Sigma^2$ that are shared by $AA^T$ and $A^TA$
- Yes, since $AA^T$ and $A^TA$ are symmetric positive (semi-)definite.
  - More details in the next slides.

# Diagonalization of Symmetric Matrices

- In general, $A \in \mathbb{R}^{n \times n}$ is diagonalizable if and only if $n$ linearly independent eigenvectors exist.

- How about a symmetric matrix $S \in \mathbb{R}^{n \times n}$, where $S^T = S$?

- $S$ is always diagonalizable.
- Furthermore, $S$ is orthogonally diagonalizable, meaning that their eigenvectors are not only linearly independent, but also orthogonal to each other.

# Spectral Theorem of Symmetric Matrices

Consider a symmetric matrix $S \in \mathbb{R}^{n \times n}$, where $S^T = S$.

- $A$ has $n$ real eigenvalues, counting multiplicities.
- The dimension of the eigenspace for each eigenvalue equals the multiplicity of $\lambda$ as a root of the characteristic equation.
- The eigenspaces are mutually orthogonal. That is, eigenvectors corresponding to different eigenvalues are orthogonal.
- To sum up, $A$ is orthogonally diagonalizable.

- Proofs in Lay Ch7.1

# Spectral Decomposition

Eigendecomposition of a symmetric matrix,
also known as spectral decomposition, is represented as

- $$S = UDU^{-1} = UDU^{T} = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_n] \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \vdots \\ \mathbf{u}_n^T \end{bmatrix}$$

$$= [\lambda_1\mathbf{u}_1 \quad \lambda_2\mathbf{u}_2 \quad \cdots \quad \lambda_n\mathbf{u}_n] \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \vdots \\ \mathbf{u}_n^T \end{bmatrix}$$

$$= \lambda_1\mathbf{u}_1\mathbf{u}_1^T + \lambda_2\mathbf{u}_2\mathbf{u}_2^T + \cdots + \lambda_n\mathbf{u}_n\mathbf{u}_n^T$$

- Each term, $\lambda_i\mathbf{u}_i\mathbf{u}_i^T$ can be viewed as a projection matrix onto the subspace spanned by $\mathbf{u}_i$, scaled by its eigenvalue $\lambda_i$.

# Positive Definite Matrices

- **Definition**: $A \in \mathbb{R}^{n \times n}$ is positive definite if $\mathbf{x}^T A\mathbf{x} > 0, \quad \forall \mathbf{x} \neq \mathbf{0}$.
- **Definition**: $A \in \mathbb{R}^{n \times n}$ is positive semi-definite if $\mathbf{x}^T A\mathbf{x} \geq 0, \quad \forall \mathbf{x} \neq \mathbf{0}$.

- **Theorem**: $A \in \mathbb{R}^{n \times n}$ is positive definite if and only if the eigenvalues of $A$ are all positive.
- Proofs in Lay Ch7.2

# Symmetric Positive Definite Matrices

- If $S \in \mathbb{R}^{n \times n}$ is symmetric and positive-definite, then the spectral decomposition will have all positive eigvenvalues:

- $$S = UDU^T = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_n] \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \vdots \\ \mathbf{u}_n^T \end{bmatrix}$$

$$= \lambda_1 \mathbf{u}_1 \mathbf{u}_1^T + \lambda_2 \mathbf{u}_2 \mathbf{u}_2^T + \cdots + \lambda_n \mathbf{u}_n \mathbf{u}_n^T$$

where $\lambda_j > 0, \forall j = 1, \cdots, n$

# Back to Computing SVD

- In the following,

$$AA^T = U\Sigma V^T \mathrm{V}\Sigma^T U^T = U\Sigma\Sigma^T U^T = U\Sigma^2 U^T$$
$$A^T A = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T \Sigma U^T = V\Sigma^2 V^T$$

- Can we prove that both $AA^T$ and $A^T A$ are symmetric positive-(semi-)definite?
- Symmetric: $(AA^T)^T = AA^T$ and $(A^T A)^T = A^T A$
- Positive-(semi-)definite
  - $\mathbf{x}^T AA^T \mathbf{x} = (A^T \mathbf{x})^T (A^T \mathbf{x}) = \|A^T \mathbf{x}\|^2 \geq 0$
  - $\mathbf{x}^T A^T A\mathbf{x} = (A\mathbf{x})^T (A\mathbf{x}) = \|A\mathbf{x}\|^2 \geq 0$
- Thus, we can find
  1. Orthogonal eigenvector matrices $U$ and $V$
  2. Eigenvalues in $\Sigma^2$ that are all positive

# Things to Note

- Given any rectangular matrix $A \in \mathbb{R}^{m \times n}$, its SVD always exists.
- Given a square matrix $A \in \mathbb{R}^{n \times n}$, its eigendecomposition does not always exist, but its SVD always exists.
- Given a square, symmetric positive (semi-)definite matrix $S \in \mathbb{R}^{n \times n}$, its eigendecomposition always exists, and it is actually the same as its SVD.

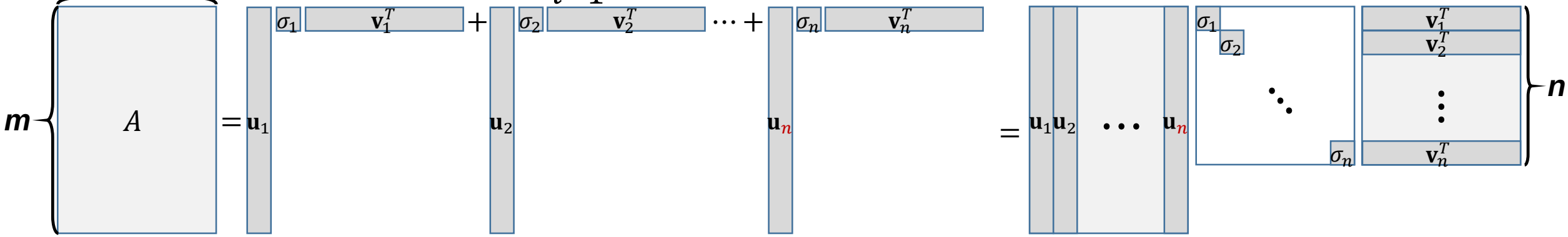
# Eigendecomposition in Machine Learning

- In machine learning, we often handle symmetric positive (semi-)definite matrix.
- Given a (feature-by-data item) matrix $A \in \mathbb{R}^{m \times n}$,
- $A^T A$ represents a (data item-by-data item) similarity matrix between all pairs of data items, where the similarity is computed as an inner product.
- Likewise, $AA^T$ represents a (feature-by-feature) similarity matrix between all pairs of features, indicating a kind of correlations between features.
  - Covariance matrix in principal component analysis
  - Gram matrix in style transfer

# Low-Rank Approximation of a Matrix

- Recall a rectangular matrix $A \in \mathbb{R}^{m \times n}$, its SVD can be represented as the sum of outer products

$$A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \qquad \text{where } \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$$




- Consider the problem of the best low-rank approximation of $A$:

$$\hat{A}_r = \arg\min_{A_r} \|A - A_r\|_F \text{ subject to } \text{rank}(A_r) \leq r$$

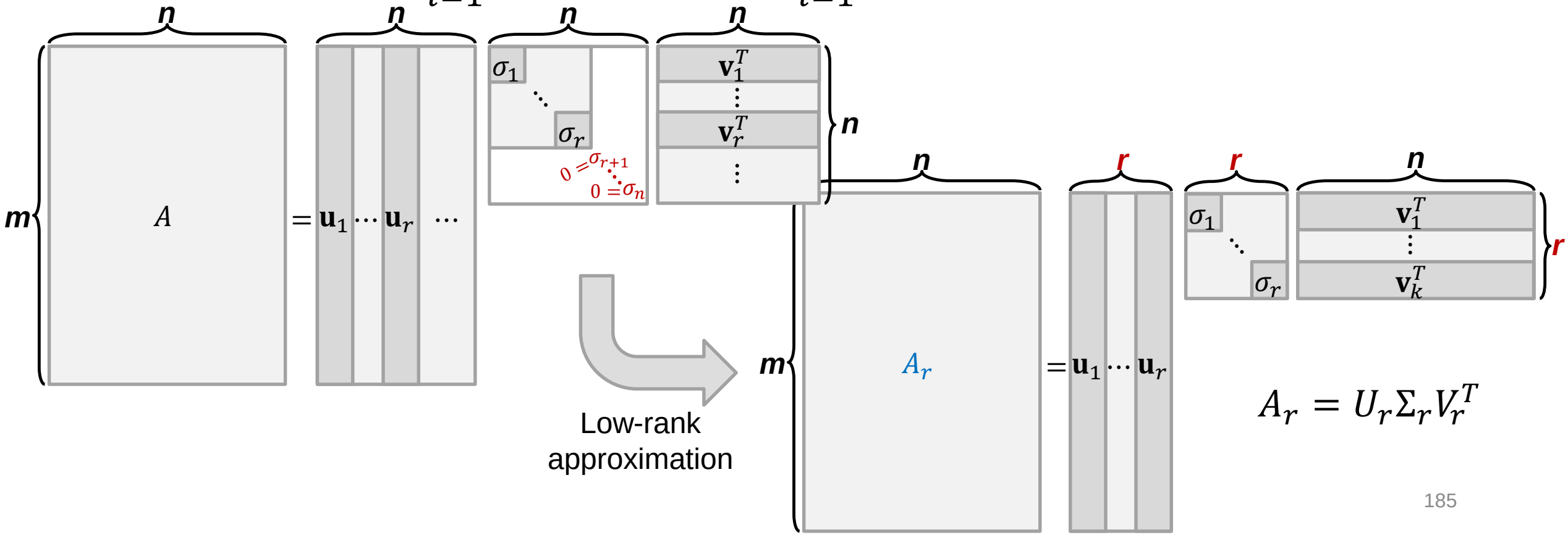
- The optimal solution is given as

$$\hat{A}_r = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \qquad \text{where } \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$$

# Low-Rank Approximation of a Matrix

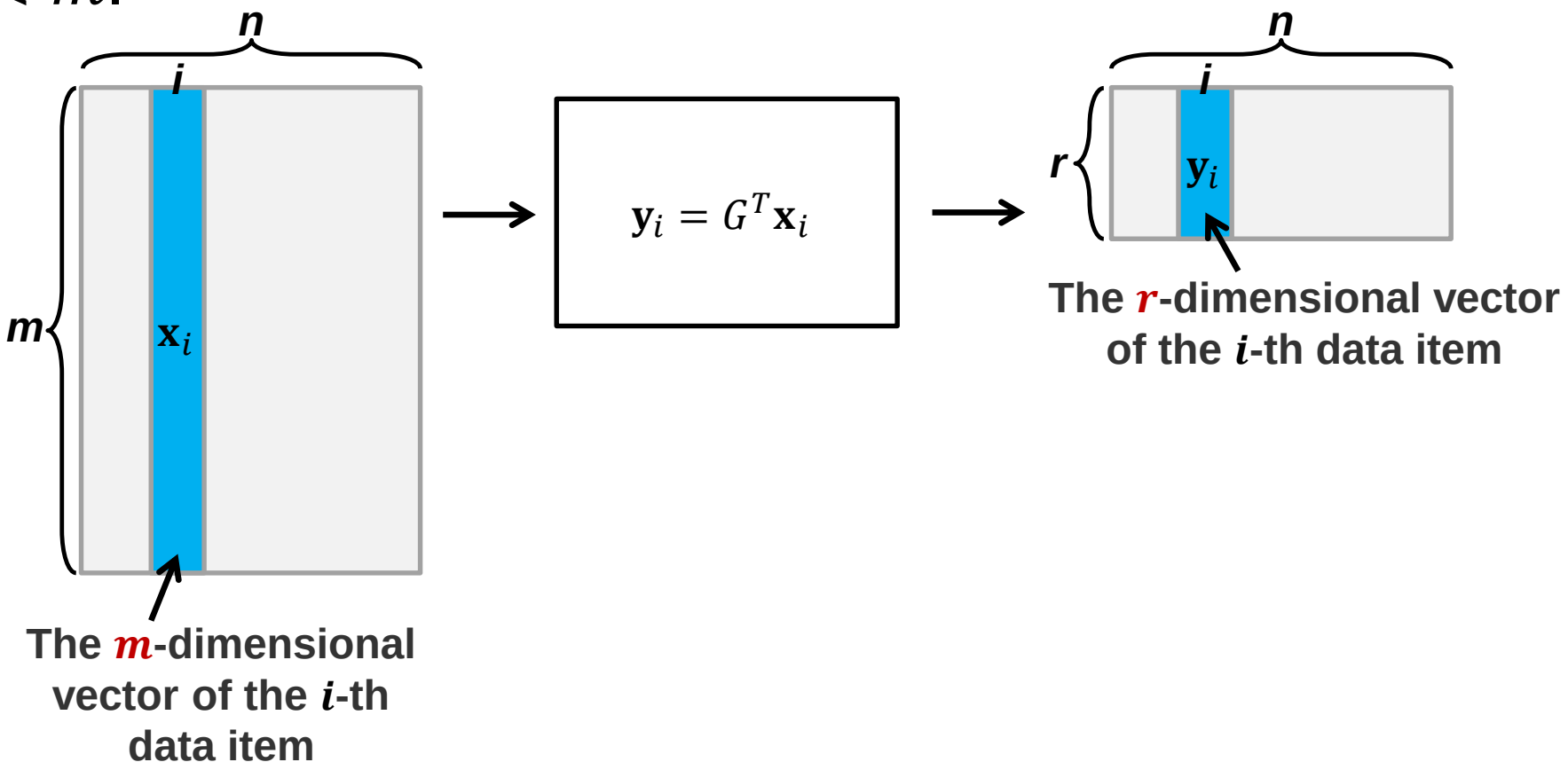
- We approximate $A$ as $A_r$ by setting $\sigma_i = 0$ for $\forall i \geq (r+1)$

$$A = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \simeq A_r = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = U_r \Sigma_r V_r^T$$



# Dimension-Reducing Transformation

- Given a (feature-by-data item) matrix $X \in \mathbb{R}^{m \times n}$, consider the linear transformation, $G^T : \mathbf{x} \in \mathbb{R}^m \mapsto \mathbf{y} \in \mathbb{R}^r$, where $G \in \mathbb{R}^{m \times r}$ and $r < m$.



# Dimension-Reducing Transformation

- Can we find the linear transformation, $\mathbf{y}_i = G^T \mathbf{x}_i$, where the columns of $G \in \mathbb{R}^{m \times r}$ are orthonormal, that best preserves the pairwise similarity between data items, $S = X^T X$?
- $Y = G^T X$, and their pairwise similarity is written as

$$Y^T Y = (G^T X)^T G^T X = X^T G G^T X$$

- Then, the above problem is written as

$$\hat{G} = \arg\min_G \|S - X^T G G^T X\|_F \text{ subject to } G^T G = I_k$$

- Given $X = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$, the optimal solution is given as

$$\hat{G} = U_r = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_r]$$

# Dimension-Reducing Transformation

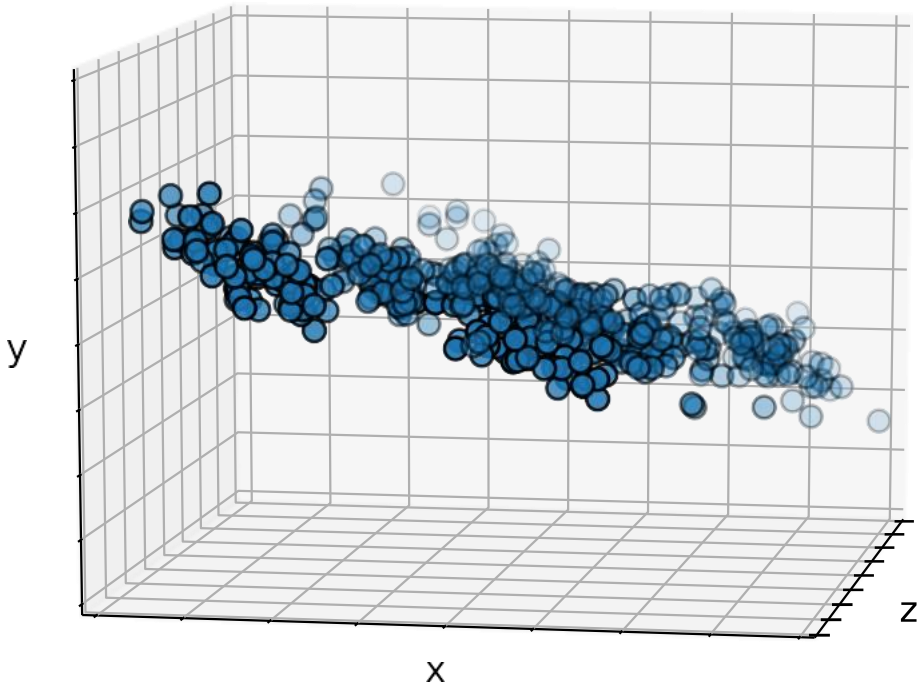
- In this case, $Y = \hat{G}^T X = U_r^T U \Sigma V^T = \Sigma_r V_r^T$.
- We can show that this generates the best solution for the best rank-$r$ approximation of $S$.

# Lecture Overview

- Elements in linear algebra
- Linear system
- Linear combination, vector equation,
  Four views of matrix multiplication
- Linear independence, span and subspace
- Linear transformation
- Least squares
- Eigendecomposition
- Singular value decomposition
- Principal component analysis

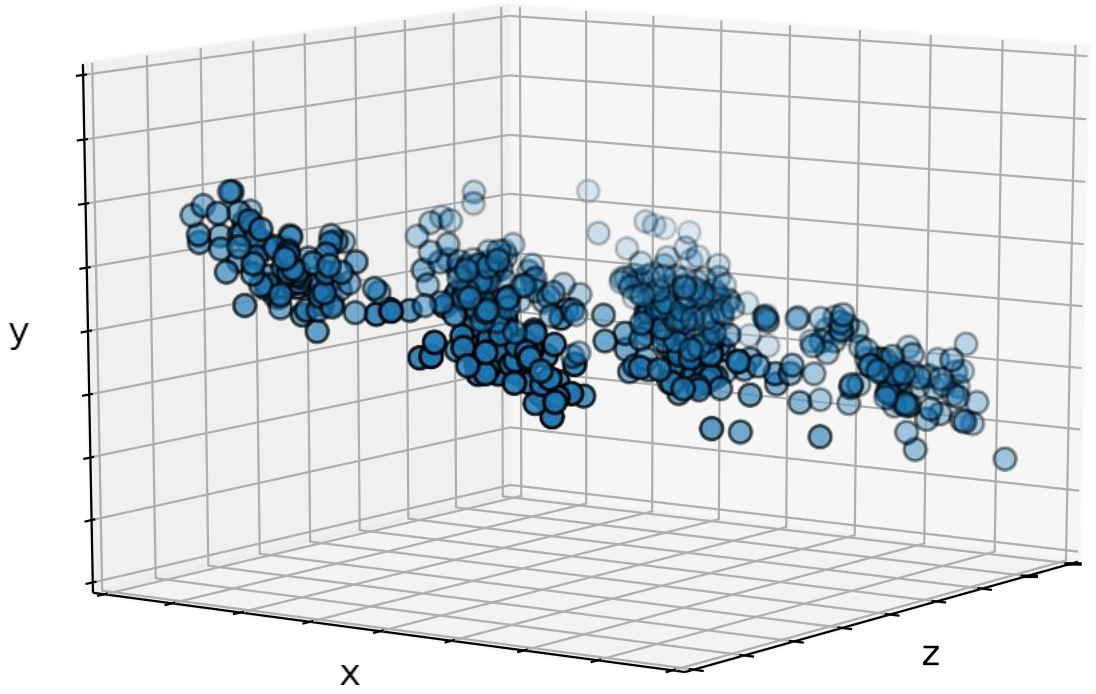
# Intuition

- What can we do to get a **better view** of this data set?



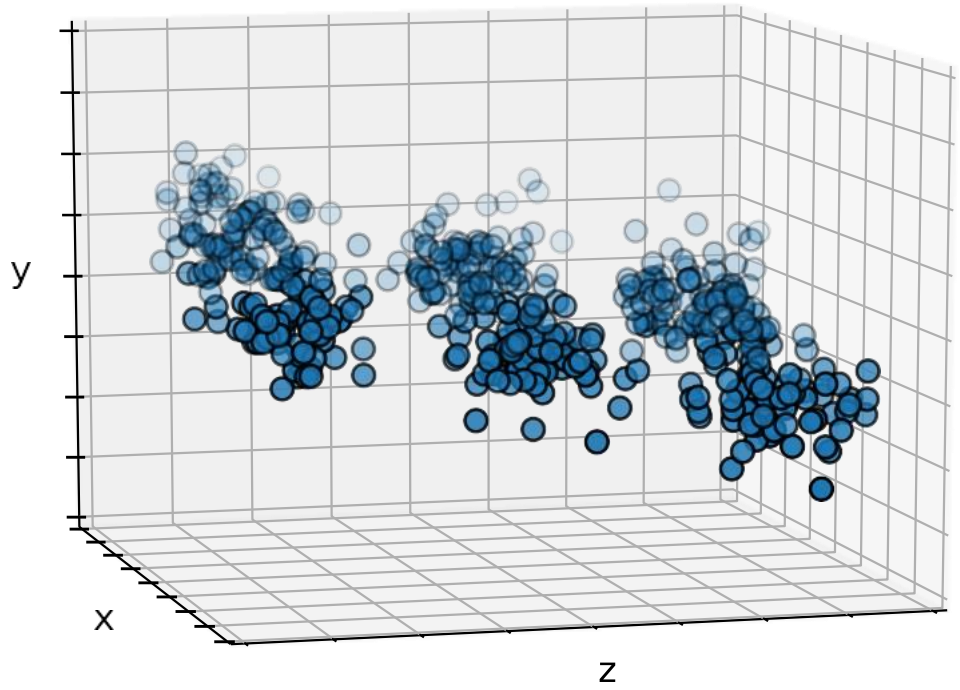
# Intuition

- What can we do to get a **better view** of this data set?
- We can try to rotate the axes until we find the **best angle** to view the data



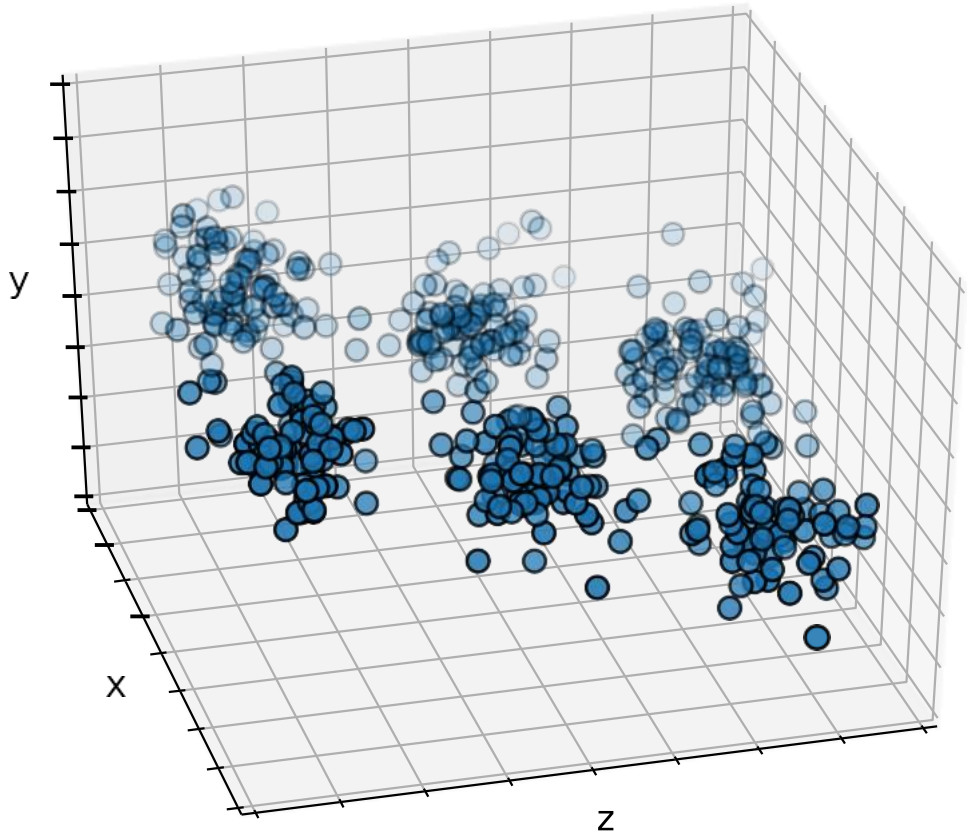
# Intuition

- What can we do to get a **better view** of this data set?
- We can try to rotate the axes until we find the **best angle** to view the data



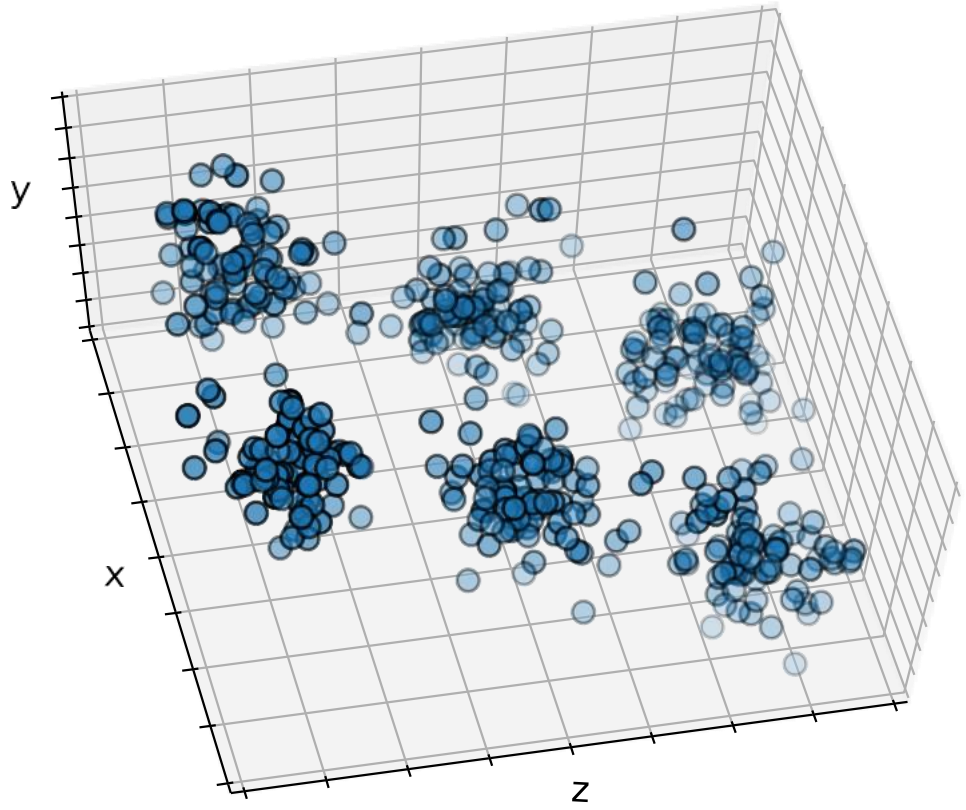
# Intuition

- What can we do to get a **better view** of this data set?
- We can try to rotate the axes until we find the **best angle** to view the data



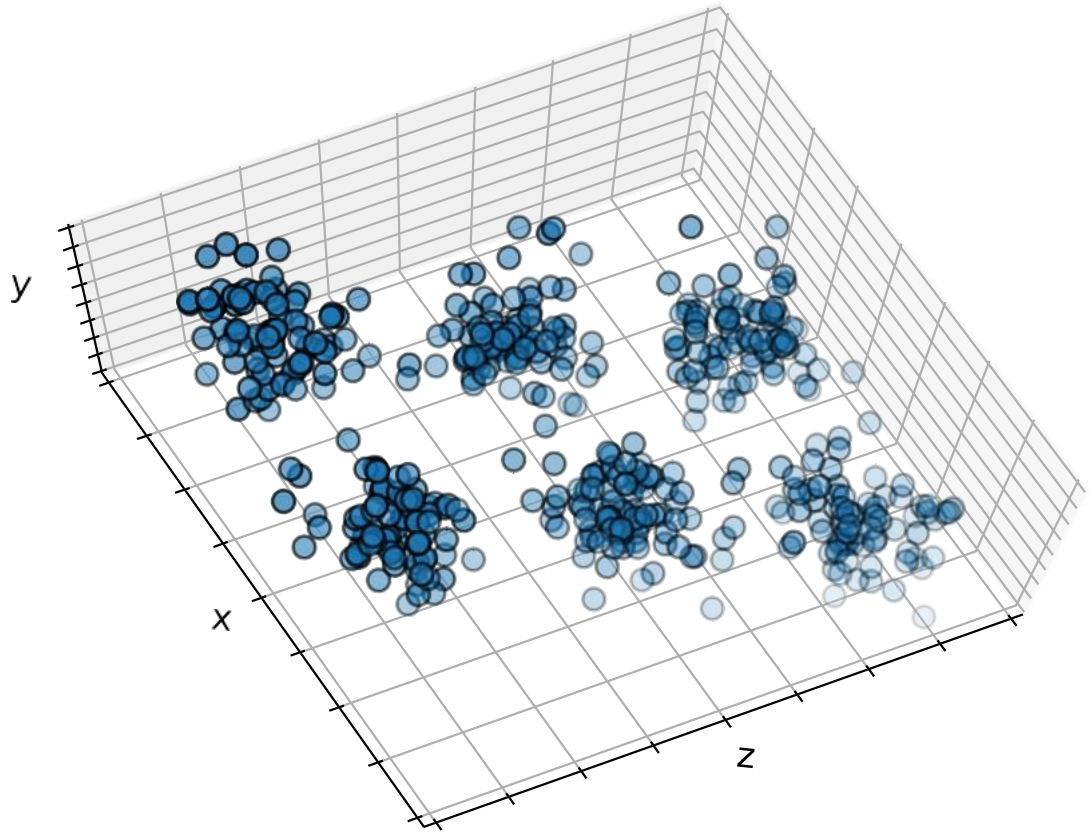
# Intuition

- What can we do to get a **better view** of this data set?
- We can try to rotate the axes until we find the **best angle** to view the data



# Intuition

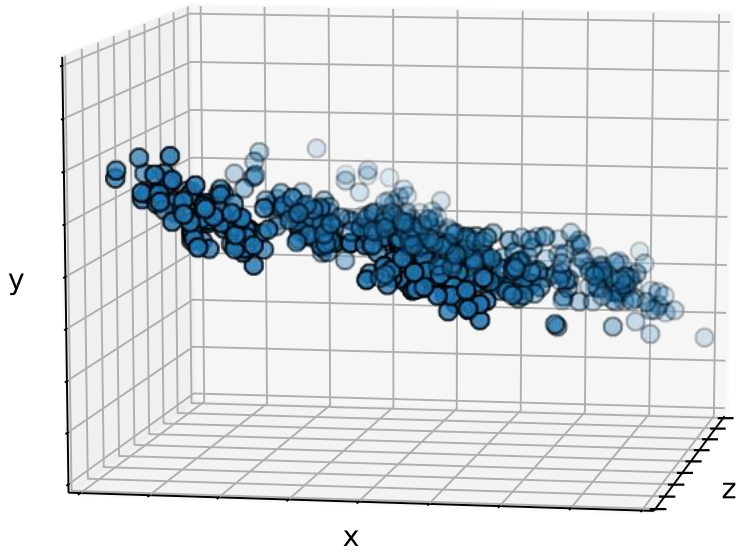
- What can we do to get a **better view** of this data set?
- We can try to rotate the axes until we find the **best angle** to view the data



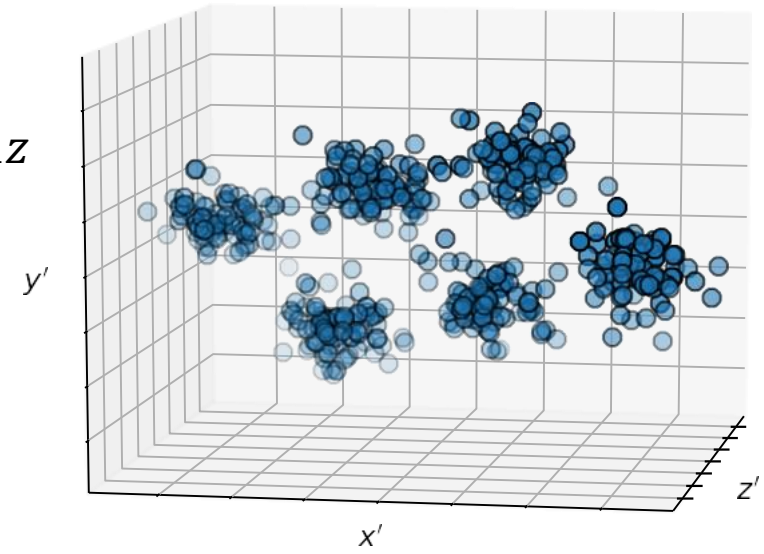

We get a better understanding of the data by looking from this angle than from our original perspective.

# Intuition

- Rotating the axes is equivalent to changing the coordinate system (or rotating the data)
- The axes that give the best view of the data are called **principal components**
  - These are the directions in which the data varies the most!



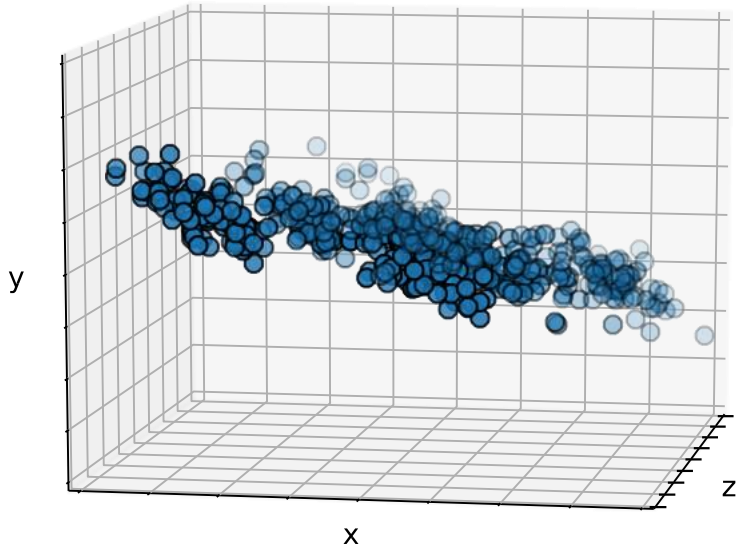

$$x' = 0.56x - 0.25y - 0.79z$$
$$y' = -0.75x + 0.25y - 0.61z$$
$$z' = 0.35x + 0.94y - 0.04z$$



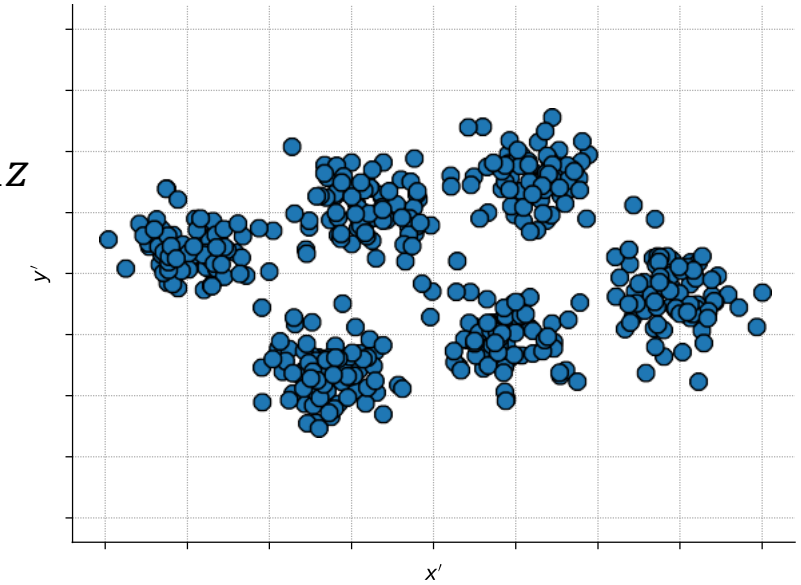

- We also compute how much variance is explained by each component:
  - $x'$ explains ~75%, $y'$ explains ~25%, $z'$ explains <1%

# Intuition

- Rotating the axes is equivalent to changing the coordinate system (or rotating the data)
- The axes that give the best view of the data are called **principal components**
  - These are the directions in which the data varies the most!




$$x' = 0.56x - 0.25y - 0.79z$$
$$y' = -0.75x + 0.25y - 0.61z$$
$$z' = 0.35x + 0.94y - 0.04z$$




- We also compute how much variance is explained by each component:
  - $x'$ explains ~75%, $y'$ explains ~25%, ~~$z'$ explains <1%~~
  - We can discard components that explain too little variance!

# Mathematical Preliminaries

# Random Variable

- A random variable is a variable whose possible values are outcomes of a random phenomenon, e.g., rolling a die.
  - The source of uncertainty in a random variable can be either:
    - objective – the result of a random process
    - subjective – the results of incomplete knowledge
  - A random variable has a probability distribution that specifies the probability of its value falling in any given interval.
- A random variate is a particular outcome of a random variable.
  - It is the result of sampling from the probability distribution.
- The expected value $E$ (or mean) of a random variable is the long-run average of its random variates.
  - For the discrete case, it is the probability-weighted average of all possible outcomes:

$$E[\ ] = 3.5$$

- The variance $\sigma^2$ is a measure of the dispersion of random variates around the expected value.
  - $\sigma$ is the standard deviation.

# Random Variable

- Let $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ be two sets of random variates sampled together from two random variables, i.e., $x_i$ is sampled at the same time as $y_i$
- $X$ and $Y$ are just subsets of the whole (potentially infinite) population
  - Then, the sample mean and the sample variance are just estimates of the true mean and variance values:

$$\text{Mean}(X) = E[X] = \mu_X = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{Var}(X) = \sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_X)^2$$

$n-1$ instead of $n$ is Bessel's correction that accounts for the biased estimation of the mean.

# Random Variable

- Covariance measures the joint variability of two random variables, e.g., if larger values of $X$ correspond to larger values of $Y$ and smaller values of $X$ to smaller values of $Y$:

$$\mathrm{Cov}(X,Y) = \sigma_{X,Y}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_X)\cdot(y_i - \mu_Y)$$

$n-1$ instead of $n$ is Bessel's correction that accounts for the biased estimation of the mean.

# Features as Random Variables

- Consider a data set $X \in \mathbb{R}^{m \times n}$ with $m$ examples and $n$ features:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix}$$

  - Each row represents an example (vector of $n$ features).
  - Each column represents the same feature of all examples.
- We can regard each feature (column) as a **random variable**.
  - This means that $x_{i,j}$ (feature $j$ of example $i$) is a **random variate**.
  - Examples are made up of $n$ random variables that are sampled together.
- This makes $X$ a set of $m$ sampling events of $n$ random variables.
  - We can compute the sample mean, variance, and covariance .

# Centering Data Set

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix}$$

$$\mu_X^T = [\mu_1 \quad \mu_2 \quad \cdots \quad \mu_n] = \frac{1}{m}\left[\sum_{i=1}^{m} x_{i,1} \quad \sum_{i=1}^{m} x_{i,2} \quad \cdots \quad \sum_{i=1}^{m} x_{i,n}\right]$$

$$\bar{X} = \begin{bmatrix} x_{1,1} - \mu_1 & x_{1,2} - \mu_2 & \cdots & x_{1,n} - \mu_n \\ x_{2,1} - \mu_1 & x_{2,2} - \mu_2 & \cdots & x_{2,n} - \mu_n \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} - \mu_1 & x_{m,2} - \mu_2 & \cdots & x_{m,n} - \mu_n \end{bmatrix} = X - 1_{m\times 1}\mu_X^T$$

Centered version of $X$

# Computing Covariance of Data Set

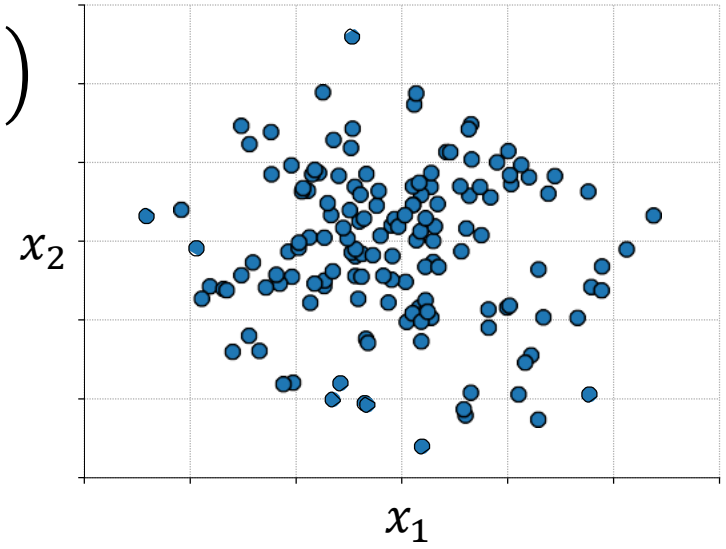
$$\bar{X} = \begin{bmatrix} x_{1,1} - \mu_1 & x_{1,2} - \mu_2 & \cdots & x_{1,n} - \mu_n \\ x_{2,1} - \mu_1 & x_{2,2} - \mu_2 & \cdots & x_{2,n} - \mu_n \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} - \mu_1 & x_{m,2} - \mu_2 & \cdots & x_{m,n} - \mu_n \end{bmatrix} \qquad \bar{X}^T = \begin{bmatrix} x_{1,1} - \mu_1 & x_{2,1} - \mu_1 & \cdots & x_{m,1} - \mu_1 \\ x_{1,2} - \mu_2 & x_{2,2} - \mu_2 & \cdots & x_{m,2} - \mu_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,n} - \mu_n & x_{2,n} - \mu_n & \cdots & x_{m,n} - \mu_n \end{bmatrix}$$

$$\frac{1}{m-1}\bar{X}^T\bar{X} = \frac{1}{m-1}\begin{bmatrix} \sum_{i=1}^{m}(x_{i,1} - \mu_1)^2 & \sum_{i=1}^{m}(x_{i,1} - \mu_1)\cdot(x_{i,2} - \mu_2) & \cdots & \sum_{i=1}^{m}(x_{i,1} - \mu_1)\cdot(x_{i,n} - \mu_n) \\ \sum_{i=1}^{m}(x_{i,2} - \mu_2)\cdot(x_{i,1} - \mu_1) & \sum_{i=1}^{m}(x_{i,2} - \mu_2)^2 & \cdots & \sum_{i=1}^{m}(x_{i,2} - \mu_2)\cdot(x_{i,n} - \mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{m}(x_{i,n} - \mu_n)\cdot(x_{i,1} - \mu_1) & \sum_{i=1}^{m}(x_{i,n} - \mu_n)\cdot(x_{i,2} - \mu_2) & \cdots & \sum_{i=1}^{m}(x_{i,n} - \mu_n)^2 \end{bmatrix}$$

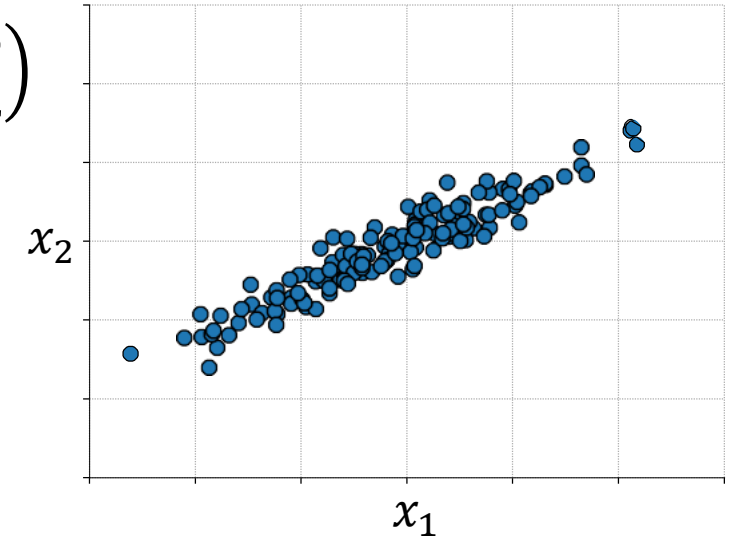
$$= \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Var}(X_n) \end{bmatrix} \stackrel{\text{def}}{=} \mathrm{Cov}(X) = \Sigma_X$$

# Covariance of Data Set

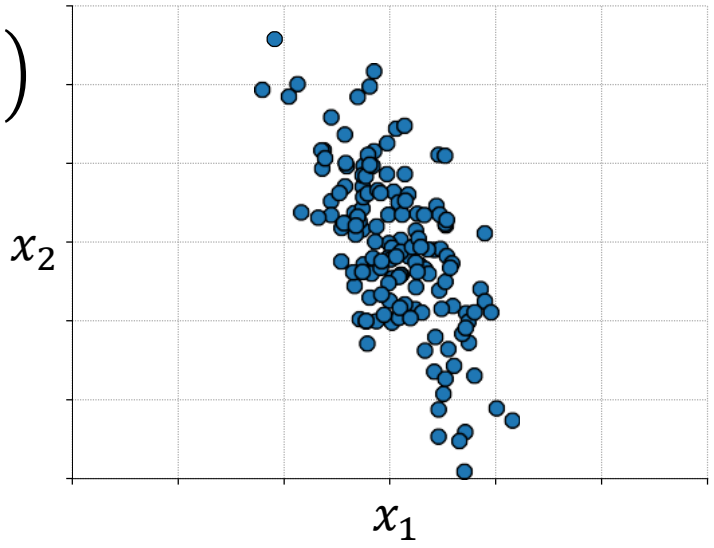
$$\Sigma_X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



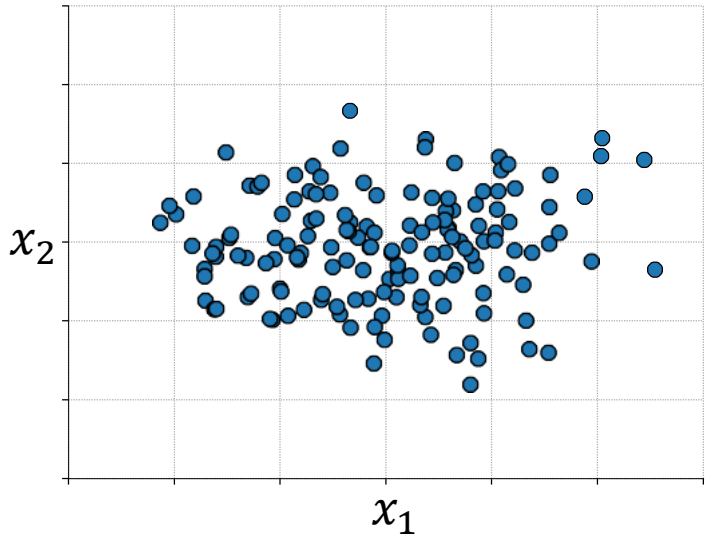

$$\Sigma_X = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 0.4 \end{pmatrix}$$




$$\Sigma_X = \begin{pmatrix} 0.2 & -0.3 \\ -0.3 & 1 \end{pmatrix}$$




$$\Sigma_X = \begin{pmatrix} 1 & 0 \\ 0 & 0.5 \end{pmatrix}$$



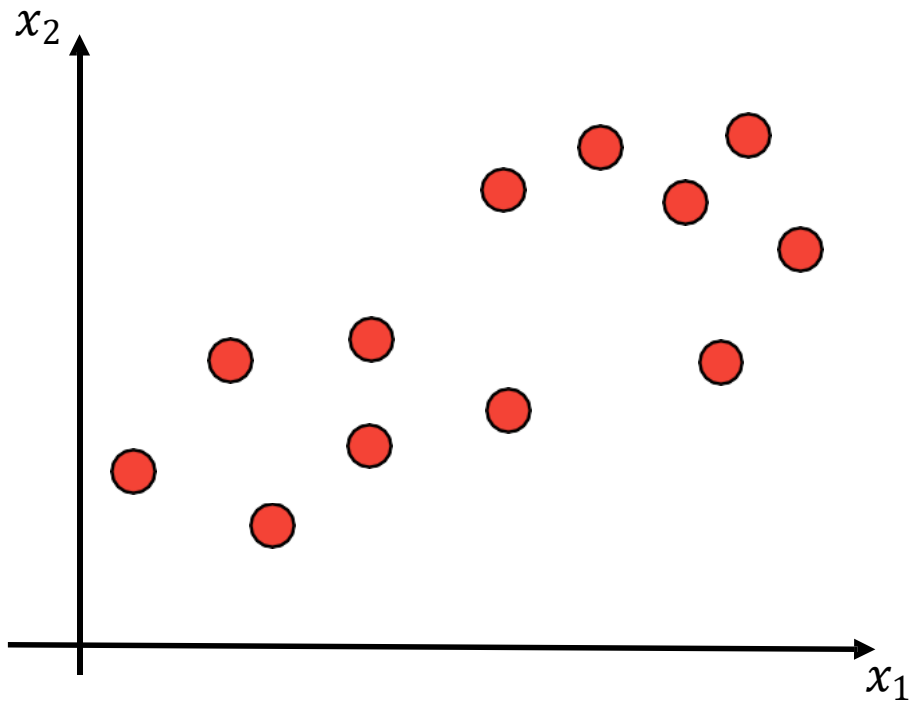
# Finding Principal Components

# Principal Component Analysis

- Principal Component Analysis (PCA) transforms a data set into a new **orthogonal coordinate system** in which the data are centered and the features are completely uncorrelated.
  - The mean of the new data set is 0.
  - The covariance of any pair of distinct features is 0.
- The features of the transformed data set are called principal components.
- Principal components are sorted in a descending order by variance, i.e., the first component has the largest variance.
  - Components with low variance can be discarded, making PCA a method of dimensionality reduction.

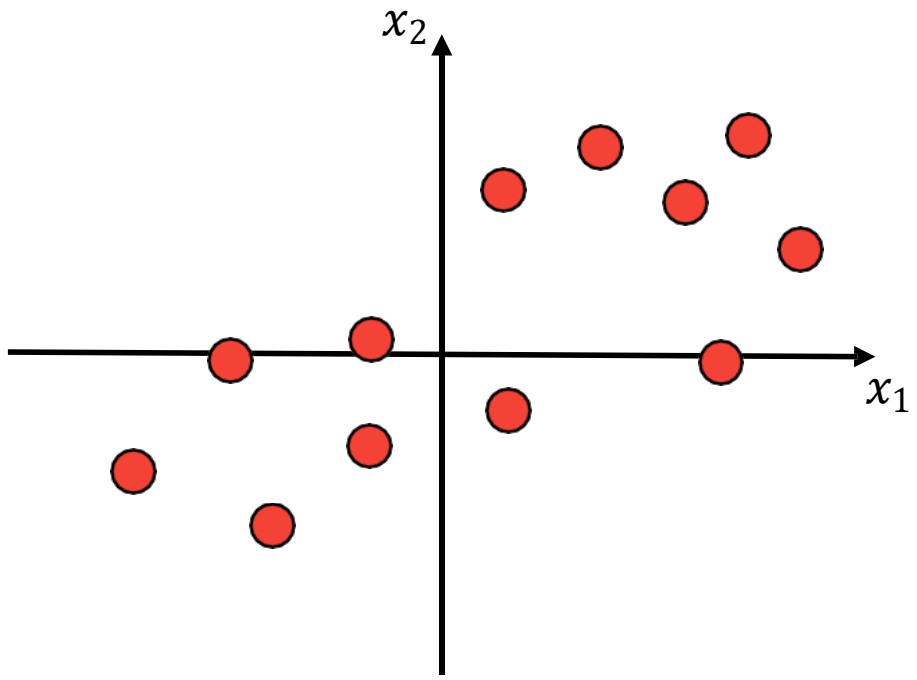
# Finding First Principal Component

- The first step is centering the data (subtracting the mean from all data points).



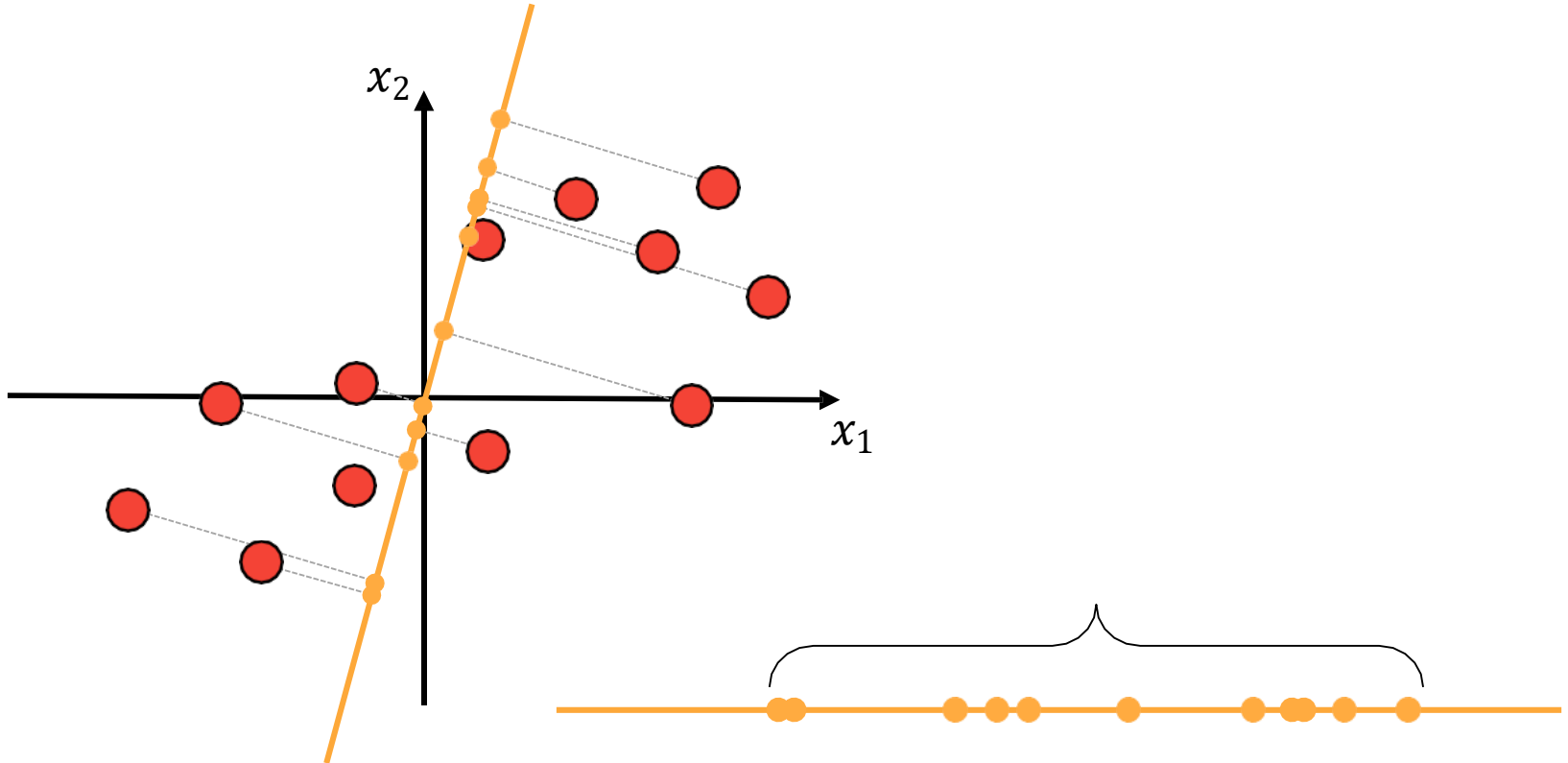
# Finding First Principal Component

- The first step is centering the data (subtracting the mean from all data points).



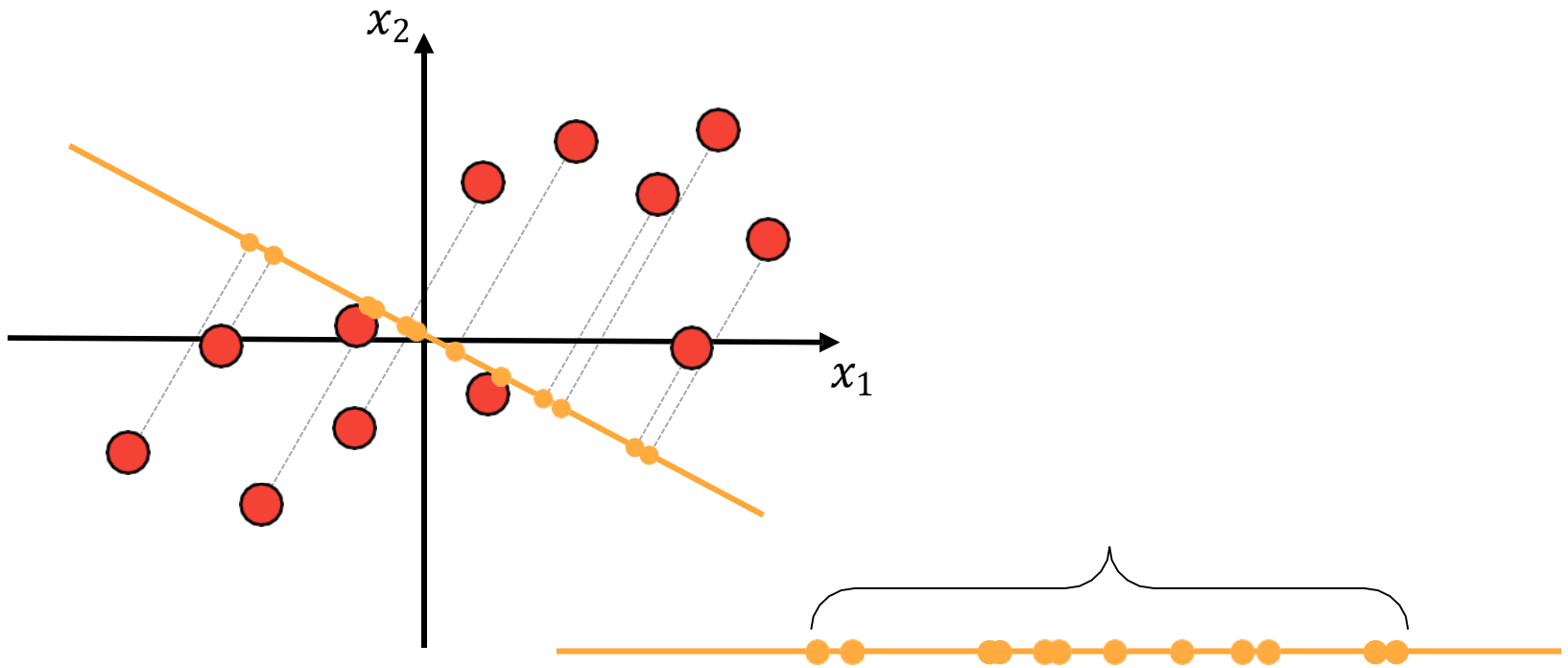
# Finding First Principal Component

- The first step is centering the data (subtracting the mean from all data points).
- The first principal component is the direction on which the data has the largest variance.
  - We are looking for a line on which the projection of data points is as spread out as possible.



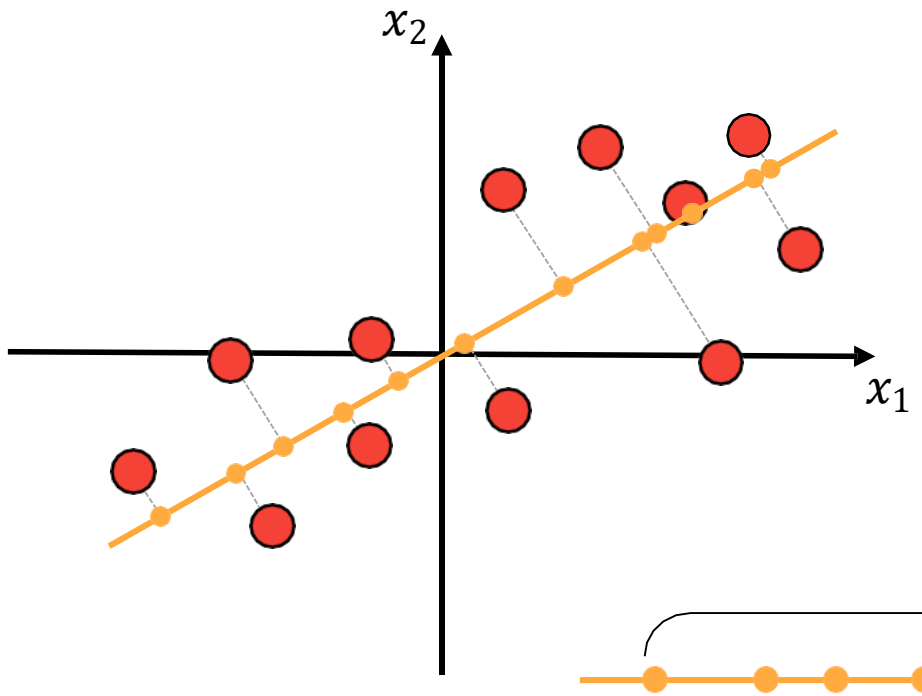
# Finding First Principal Component

- The first step is centering the data (subtracting the mean from all data points).
- The first principal component is the direction on which the data has the largest variance.
  - We are looking for a line on which the projection of data points is as spread out as possible.

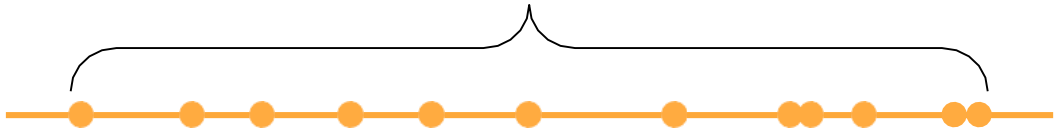


# Finding First Principal Component

- The first step is centering the data (subtracting the mean from all data points).
- The first principal component is the direction on which the data has the largest variance.
  - We are looking for a line on which the projection of data points is as spread out as possible.

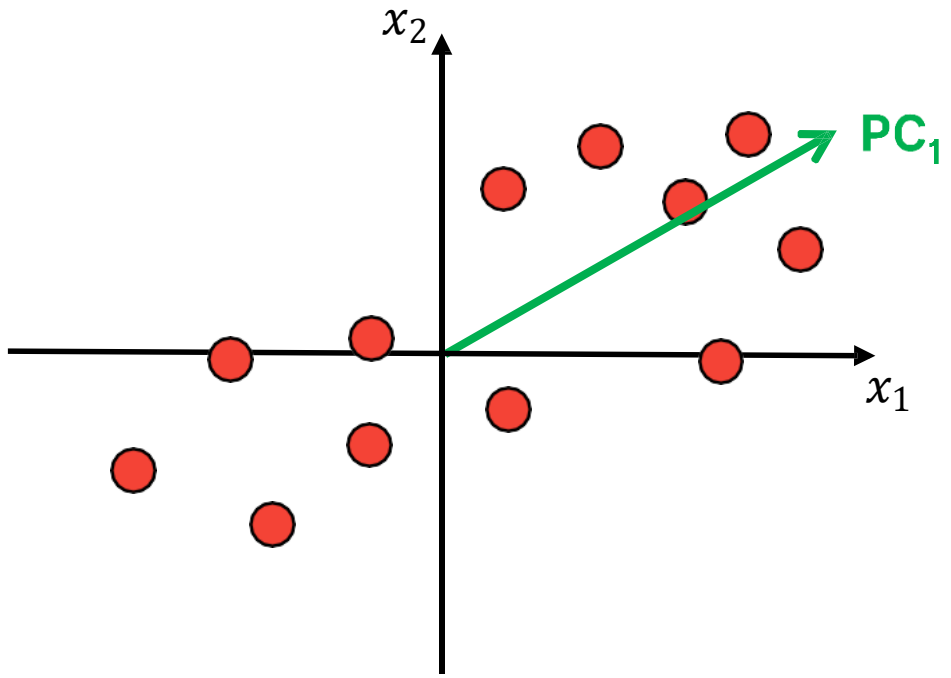



It can be proven that this is equivalent to finding the line that minimizes the sum of distances to the points



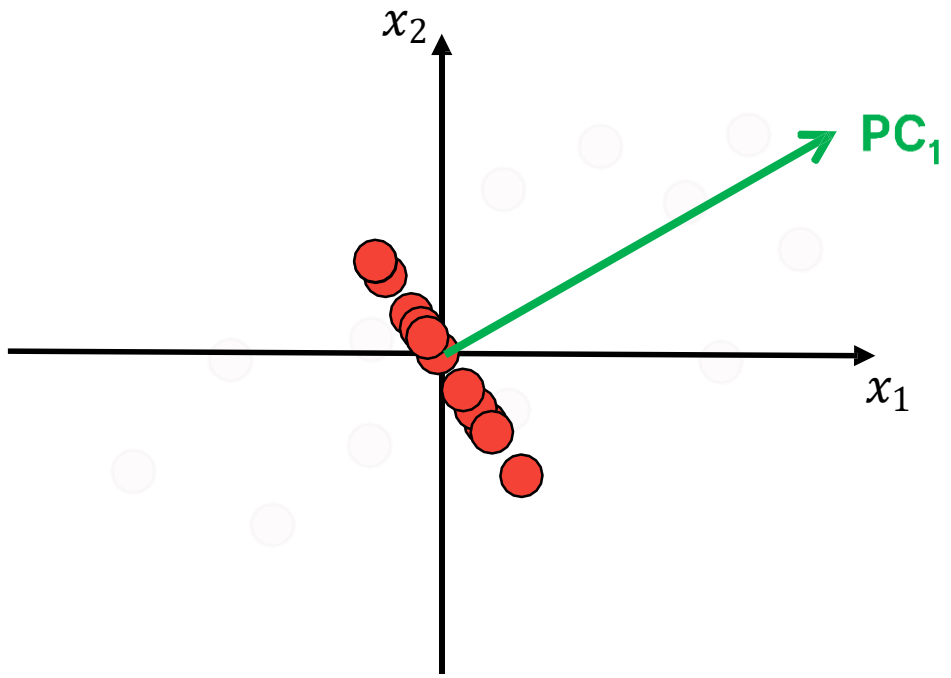
# Finding Second Principal Component

- If we subtract the projection of points on the first PC from all the points, we obtain a data set that has zero variance on that direction.



# Finding Second Principal Component

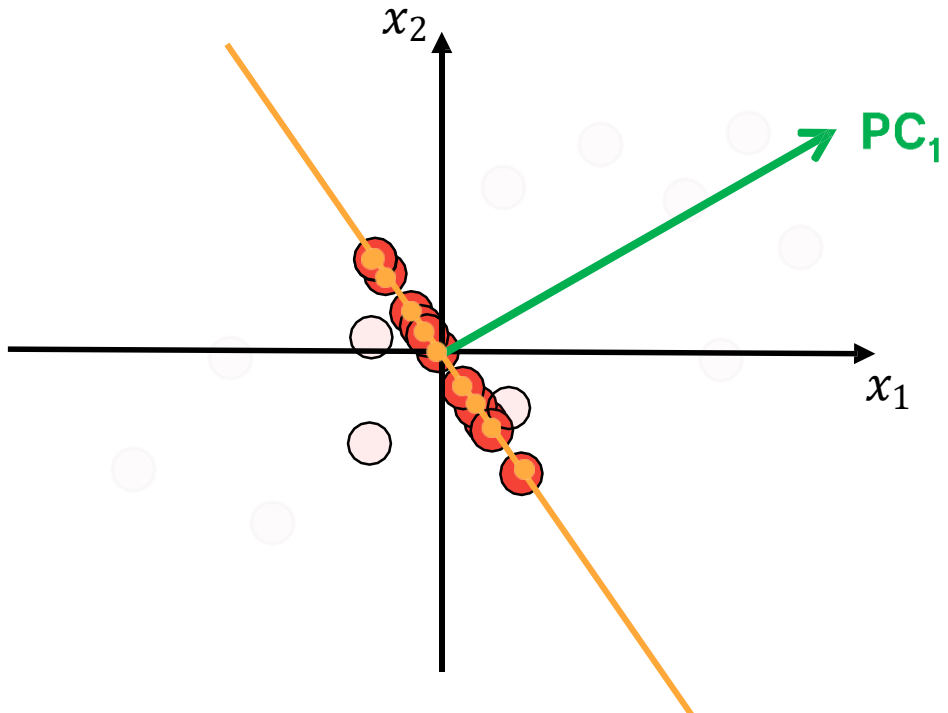
- If we subtract the projection of points on the first PC from all the points, we obtain a data set that has zero variance on that direction.



# Finding Second Principal Component

- If we subtract the projection of points on the first PC from all the points, we obtain a data set that has zero variance on that direction.
- We now look for a line on which the projection of data points from the new data set is as spread out as possible.
  - The result will be the second principal component

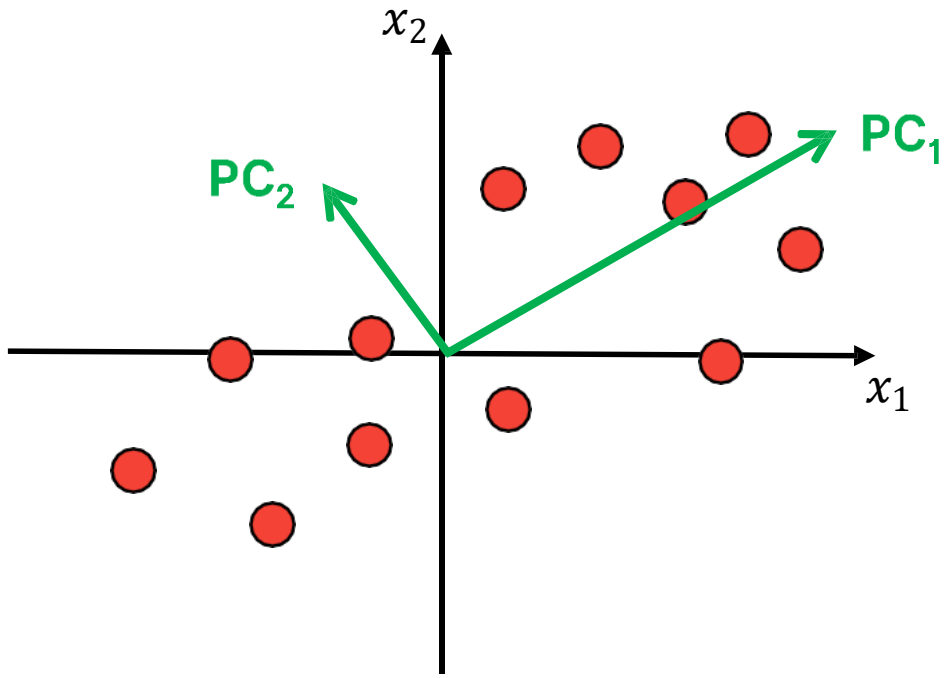
Trivial in two dimensions




The second PC is the direction in which the data varies the most, after eliminating the variance on the first PC

# Finding Further Principal Component

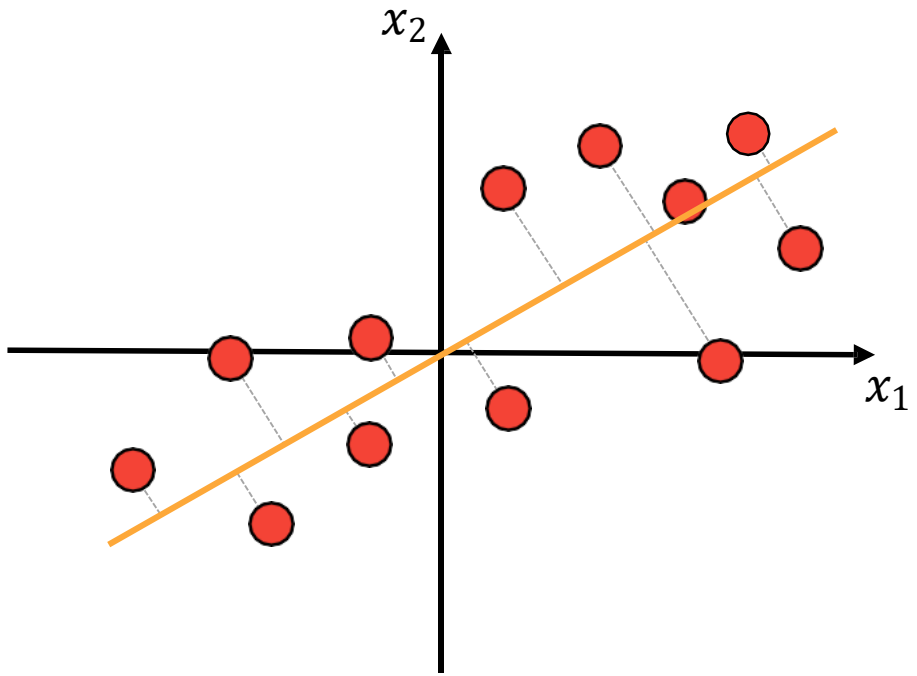
- If we subtract the projection of points on the first PC from all the points, we obtain a data set that has zero variance on that direction.
- We now look for a line on which the projection of data points from the new data set is as spread out as possible.
  - The result will be the second principal component.




The process would go on for multiple dimensions

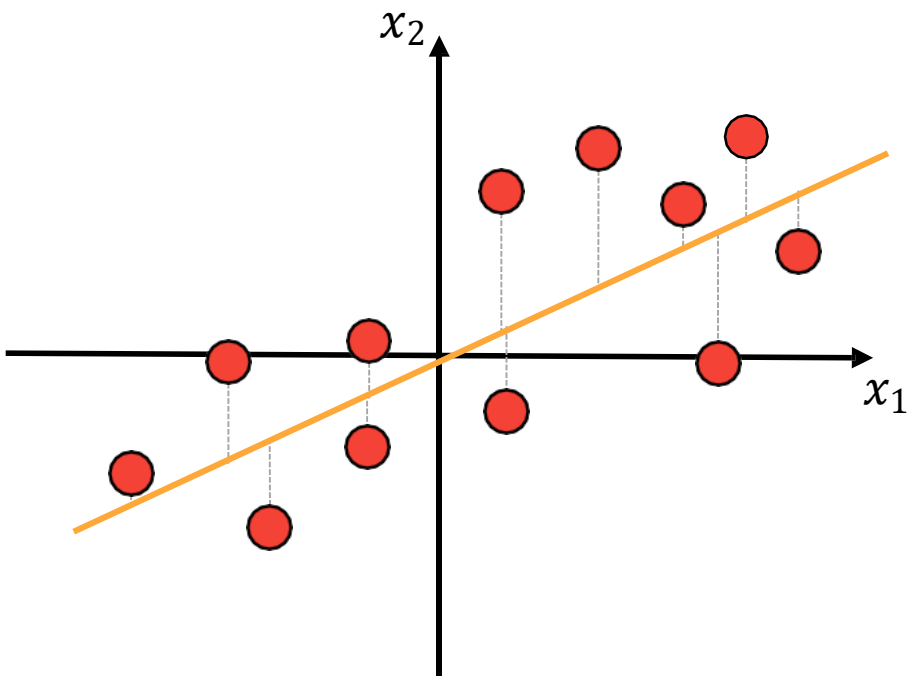
# PCA versus Linear Regression

PCA finds the line that minimizes the **sum of distances** to the data points.




$x_1$ and $x_2$ are both features, i.e., independent variables.

Linear Regression finds the line that minimizes the **sum of squared distances** to the labels.

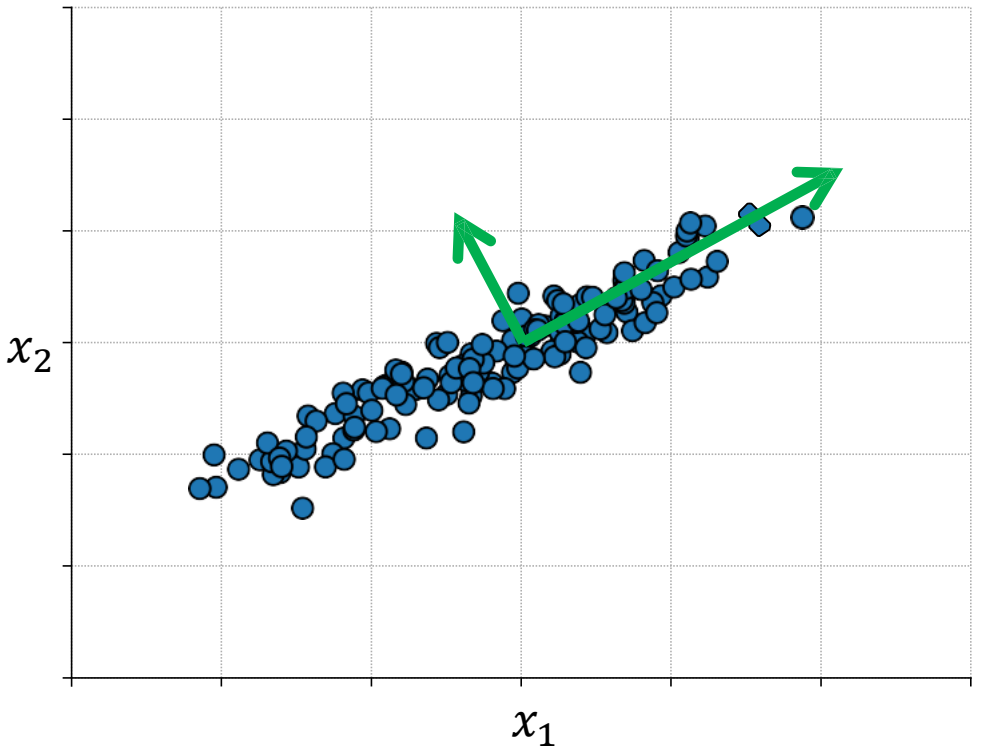



$x_2$ is the label, i.e., dependent variable.

# PCA as Eigendecomposition

# PCA as Eigendecomposition

- Let $X \in \mathbb{R}^{m \times n}$ be a data set with covariance $\Sigma_X$.
- Q: How do the eigenvectors of $\Sigma_X$ look like?
  - Intuitively, consider $\Sigma_X$ responsible for the shape of $X$.
  - Recall that the eigenvectors are the vectors that do not change direction when multiplied by $\Sigma_X$.
- A: The eigenvectors of the covariance matrix are in fact the principal components of matrix $X$.
  - The corresponding eigenvalues are equal to the amount of variance explained by each component.



# PCA as Eigendecomposition (Python)

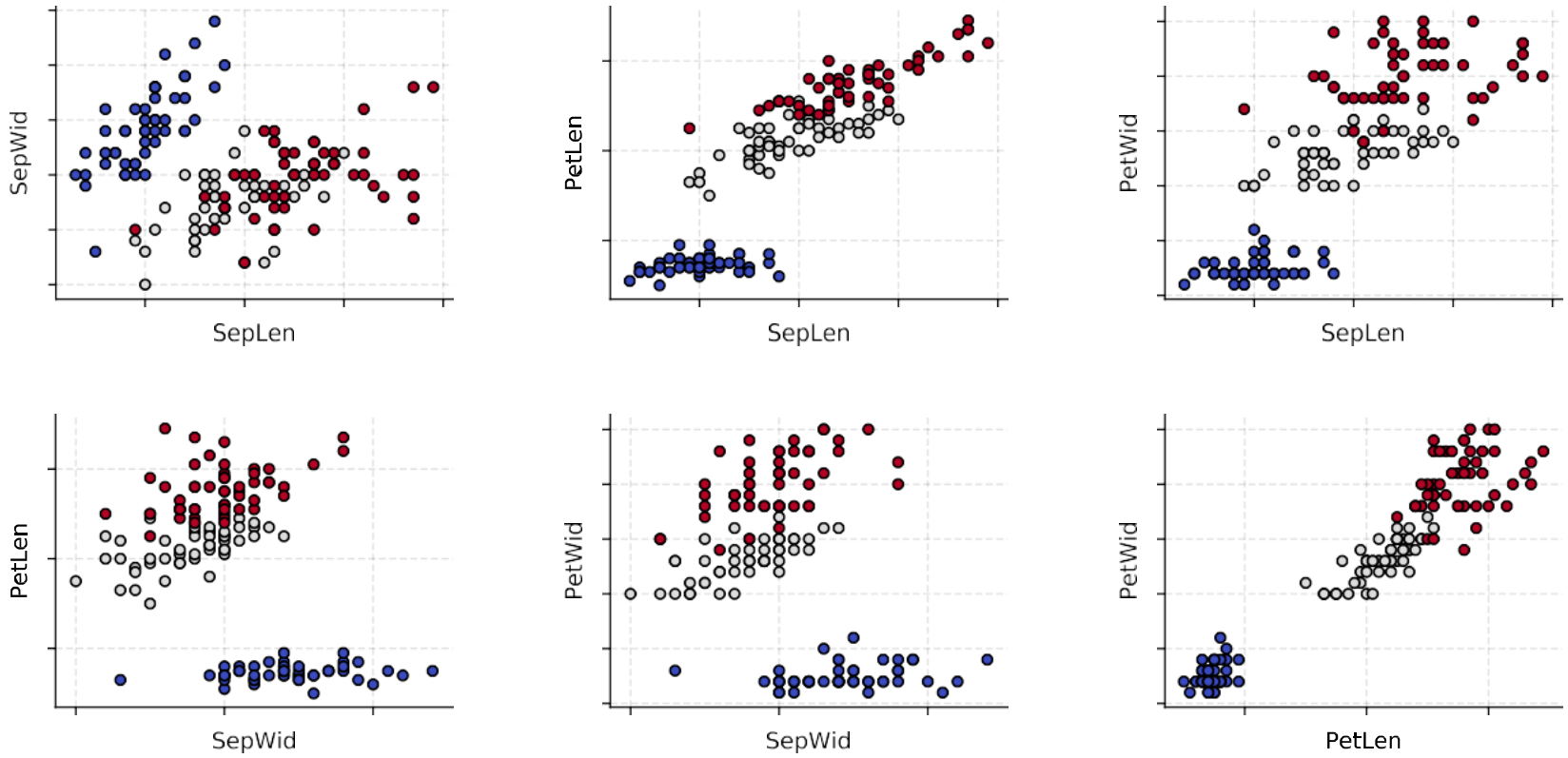
```
def PCA(X,k):
    [m,n] = X.shape
    miu = np.mean(X, axis=0) # compute mean of each feature
    X_bar = X - miu # center X
    cov_X = 1 / (m-1) * np.matmul(X_bar.T, X_bar) # compute covariance
    V, lambdas = np.linalg.eig(cov_X) # apply eigen decomposition
   idx = np.argsort(np.diag(-lambdas)) # sort lambdas descdending
   V_star = V[:,idx[:k]] # permute and keep only first k PCs
   X_star = np.matmul(X,V_star) # apply transformation to X
   return X_star, V_star
```

# Application of PCA (Python)

```
from sklearn.decomposition import PCA

pca = PCA(n_components = 2)
# number of components to keep, if ”None” it keeps all components

# find the principal components:
pca.fit(X)

# rotate X into the new coordinate system:
X_pca = pca.transform(X)

# coeficients of the original components to produce the new components:
pca.components_

# variance of projections per component:
pca.explained_variance_

# ratio of explained variance per component:
pca.explained_variance_ratio_
```

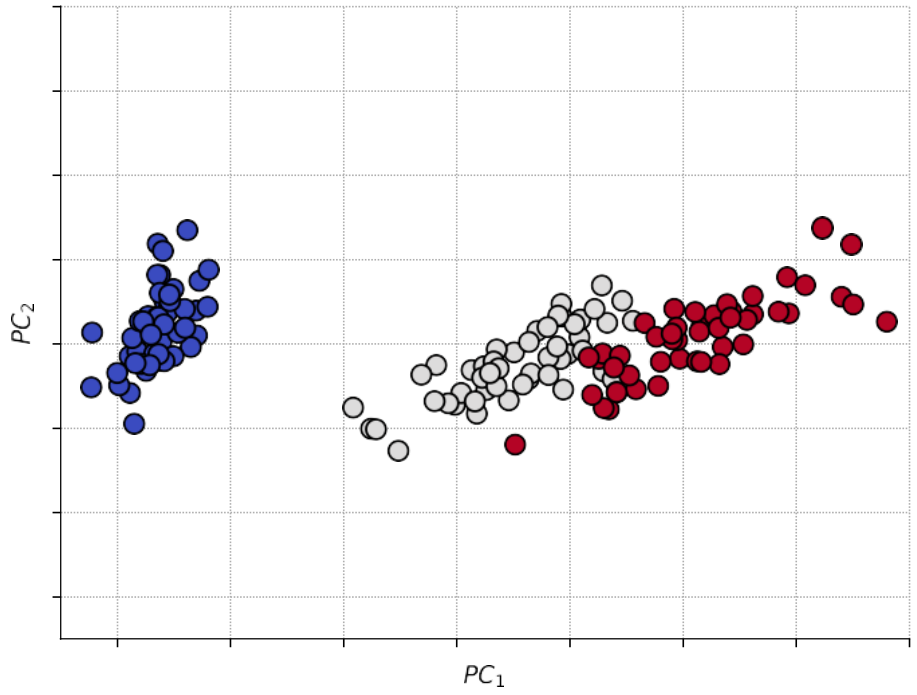
# Example

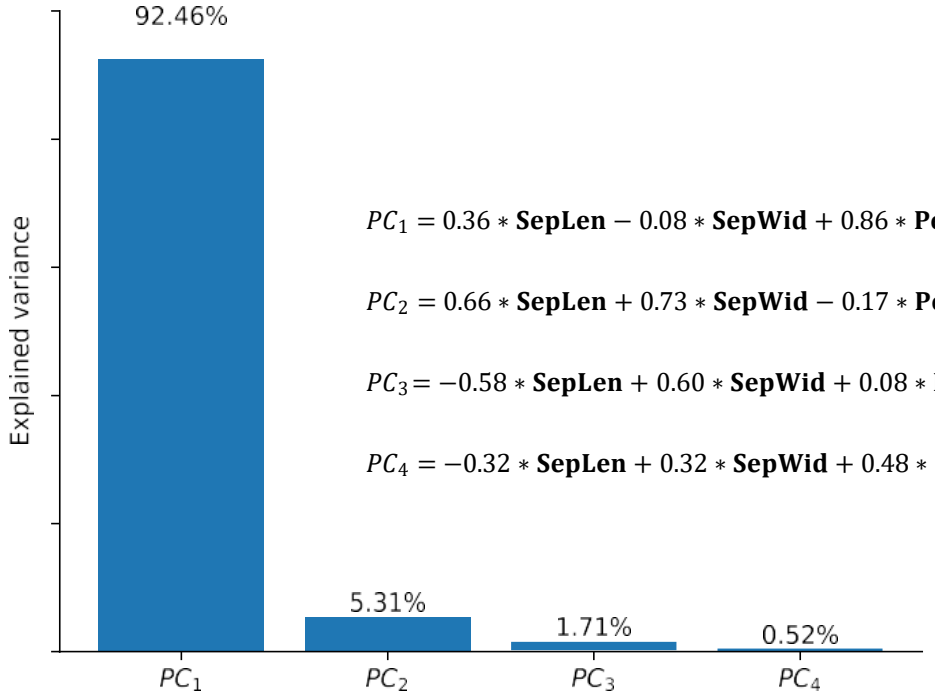
- Iris data set: 150 data samples of flowers with 4 features and 3 classes
  - $X \in \mathbb{R}^{150\times4}$
  - Features: sepal length (cm), sepal width (cm), petal length (cm), petal width (cm)



# Example

- Iris data set: 150 data samples of flowers with 4 features and 3 classes
  - $X \in \mathbb{R}^{150\times4}$
  - Features: sepal length (cm), sepal width (cm), petal length (cm), petal width (cm)




$PC_1 = 0.36 * \mathbf{SepLen} - 0.08 * \mathbf{SepWid} + 0.86 * \mathbf{PetLen} + 0.36 * \mathbf{PetWid}$

$PC_2 = 0.66 * \mathbf{SepLen} + 0.73 * \mathbf{SepWid} - 0.17 * \mathbf{PetLen} - 0.08 * \mathbf{PetWid}$

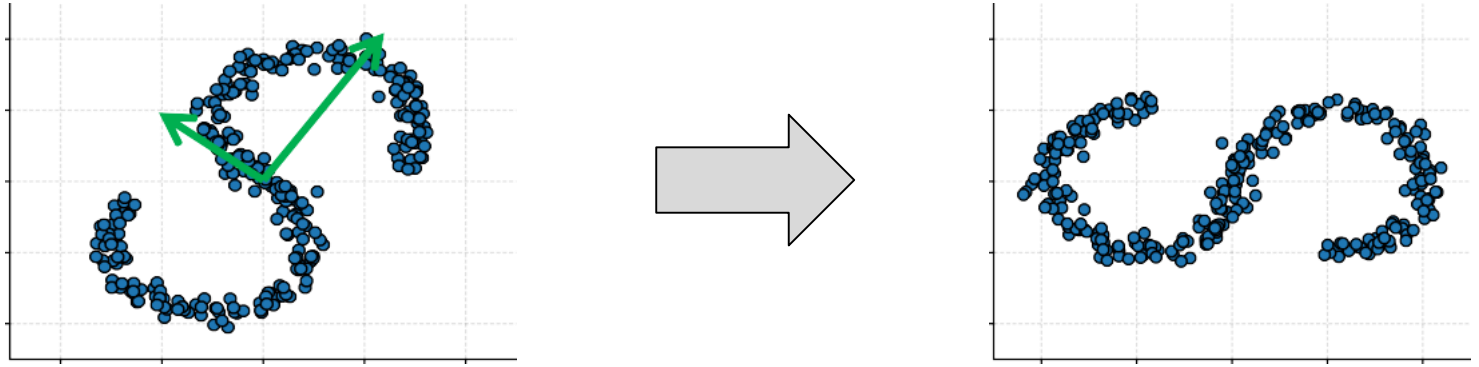
$PC_3 = -0.58 * \mathbf{SepLen} + 0.60 * \mathbf{SepWid} + 0.08 * \mathbf{PetLen} + 0.55 * \mathbf{PetWid}$

$PC_4 = -0.32 * \mathbf{SepLen} + 0.32 * \mathbf{SepWid} + 0.48 * \mathbf{PetLen} - 0.75 * \mathbf{PetWid}$

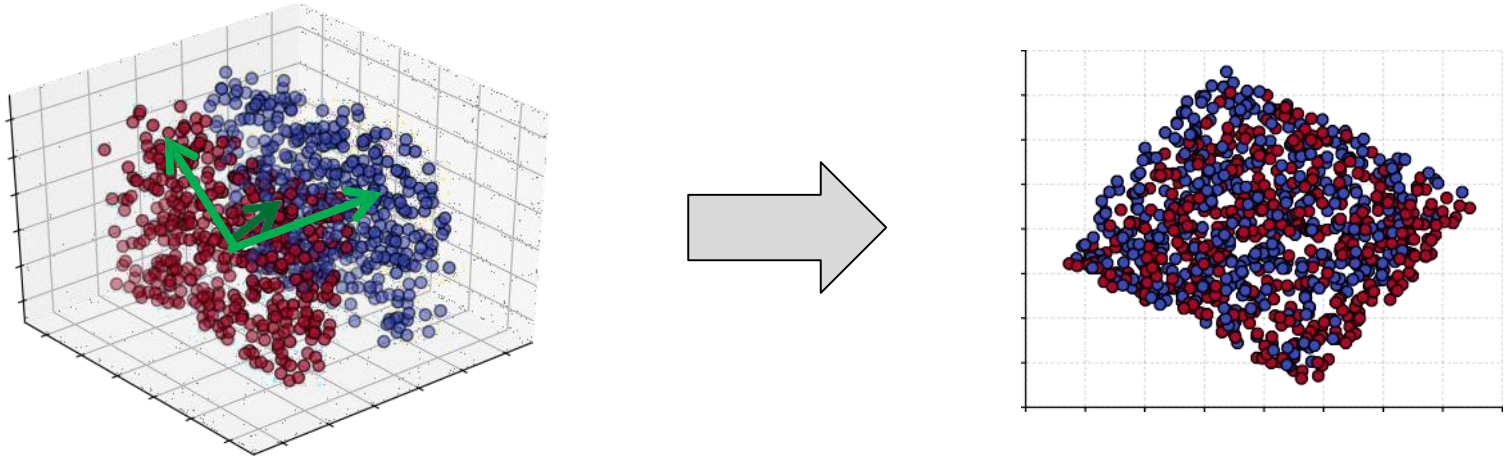
First two components explain almost 98% of the total variance.

# Limitations

- PCA does not consider non-linear correlations:

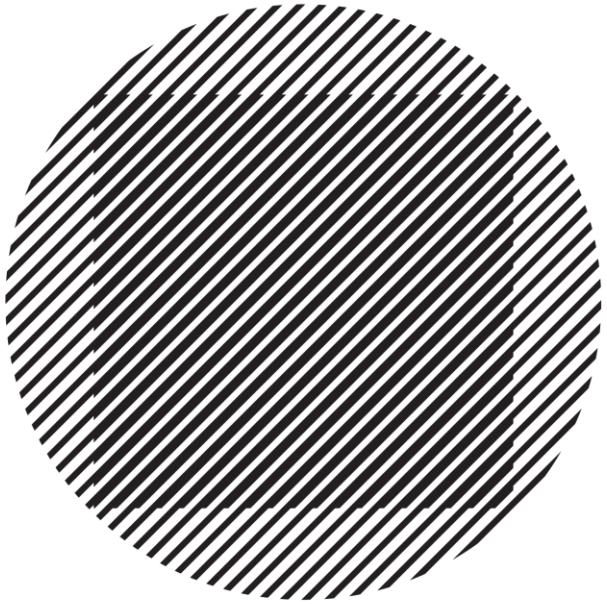


- PCA does not take labels into account, only variance of features:



# Summary

- Principal Component Analysis transforms a data set into a new orthogonal coordinate system in which the data is centered and the features are completely uncorrelated.
- The directions of the new coordinate system are called principal components.
  - PCs are sorted by variance, such that the first PC has the largest variance (it explains most of the data variance).
- Components with low variance can be removed, making PCA a method of dimensionality reduction.
- PCA does not treat non-linear correlations, and it does not take labels into account.



# THANK YOU