Linear Algebra

주재걸
고려대학교 컴퓨터학과

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

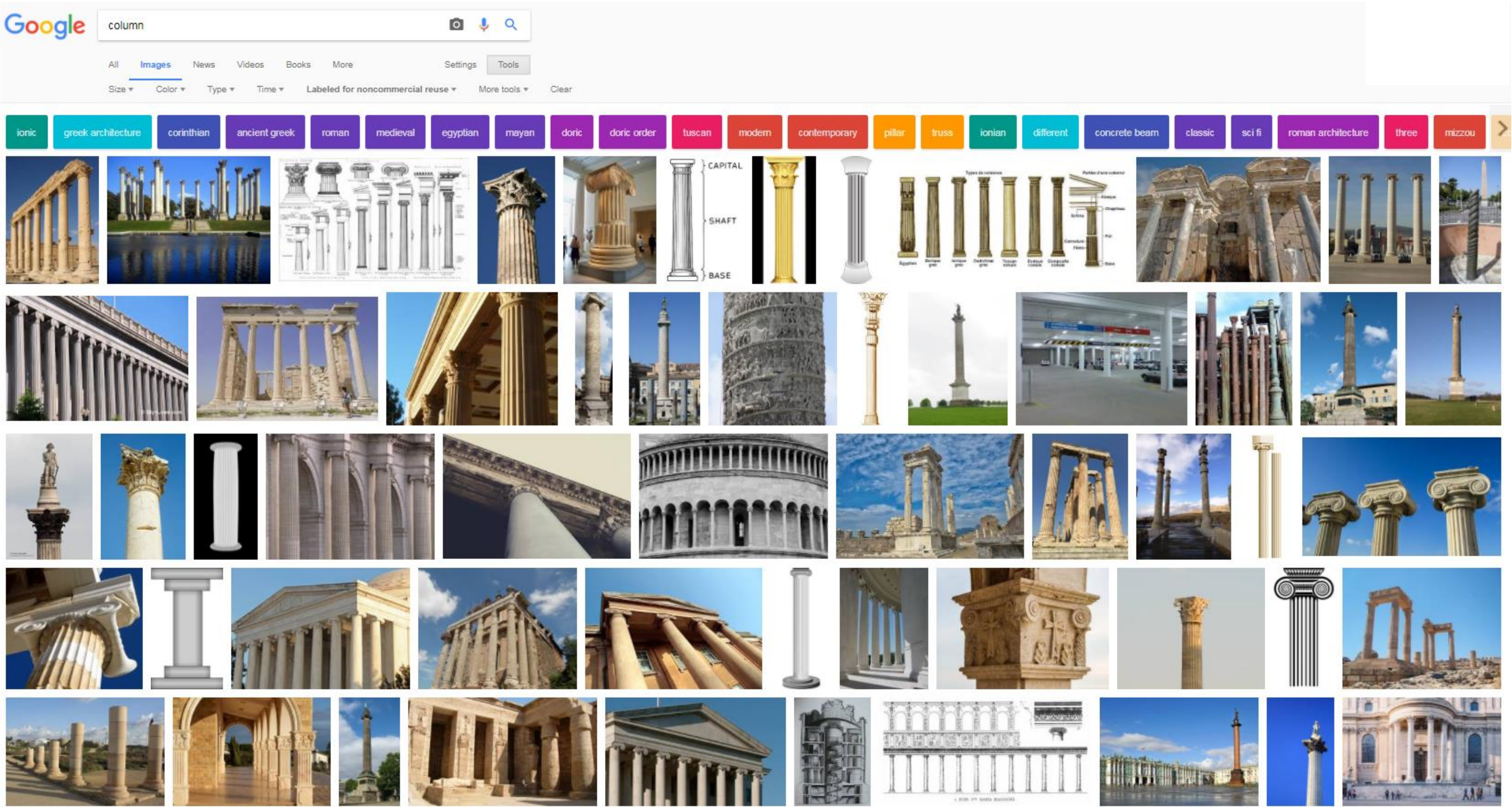
Lecture Overview

- 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

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: 3×2 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., $A_{2,:} = [3 \quad 4]$
- $A_{:,i}$: i-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
- $A(BC) = (AB)C$: Associative
- $(AB)^T = B^T A^T$: Property of transpose

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

Linear Equation

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

$$a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = b,$$

  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 $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$.

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:

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

- 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

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

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

$$\begin{aligned}\mathbf{a}_1^T\mathbf{x} &= 66\\ \mathbf{a}_2^T\mathbf{x} &= 74\\ \mathbf{a}_2^T\mathbf{x} &= 78\end{aligned}$$

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

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

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

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:**

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

- 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 $\det A$.

Does a Matrix Have an Inverse Matrix?

- $\det A$ 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/

Inverse Matrix Larger than 2×2

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

- No, but one can compute it.
- We skip details, but you can study Gaussian elimination in Lay Ch1.2 and then study Lay Ch2.2.

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

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

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

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

- Recall m = #equations and n = #variables.
- $m < n$: 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.

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

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

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

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

- 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:

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

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

- 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

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

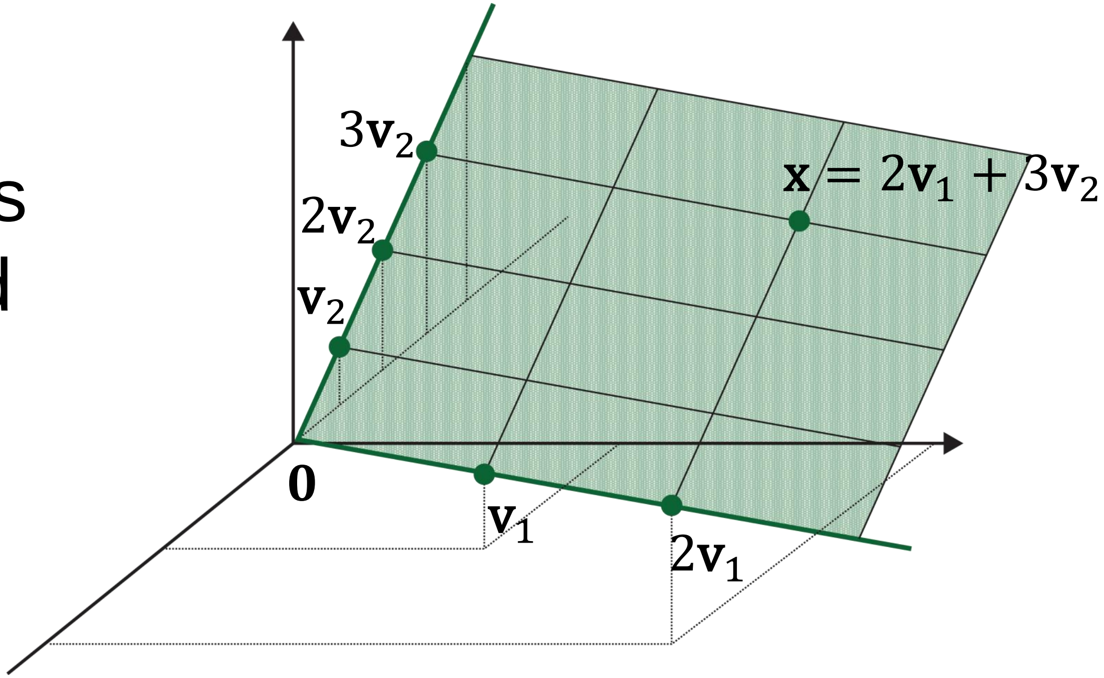
with arbitrary scalars $c_1, \cdots, c_p$.

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

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 $\mathbf{b} \in \text{Span}\ \{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$.

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:
  - $\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = A\mathbf{x} = [\mathbf{a}_1 \quad \mathbf{a}_2 \quad \mathbf{a}_3] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \mathbf{a}_3 x_3$

Matrix Multiplications as Column Combinations

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

One column on the right	*Multi-columns on the right*

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

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

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

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

Matrix Multiplications as Row Combinations

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

One row on the left

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

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 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

Lecture Overview

- Elements in linear algebra
- Linear system
- Linear combination, vector equation,
  Four views of matrix multiplication
- <u>Linear independence, span, and subspace</u>
- Linear transformation
- Least squares
- Eigendecomposition
- Singular value decomposition

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

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

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

- Or, a vector equation is written as

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

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

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

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

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

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

- 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

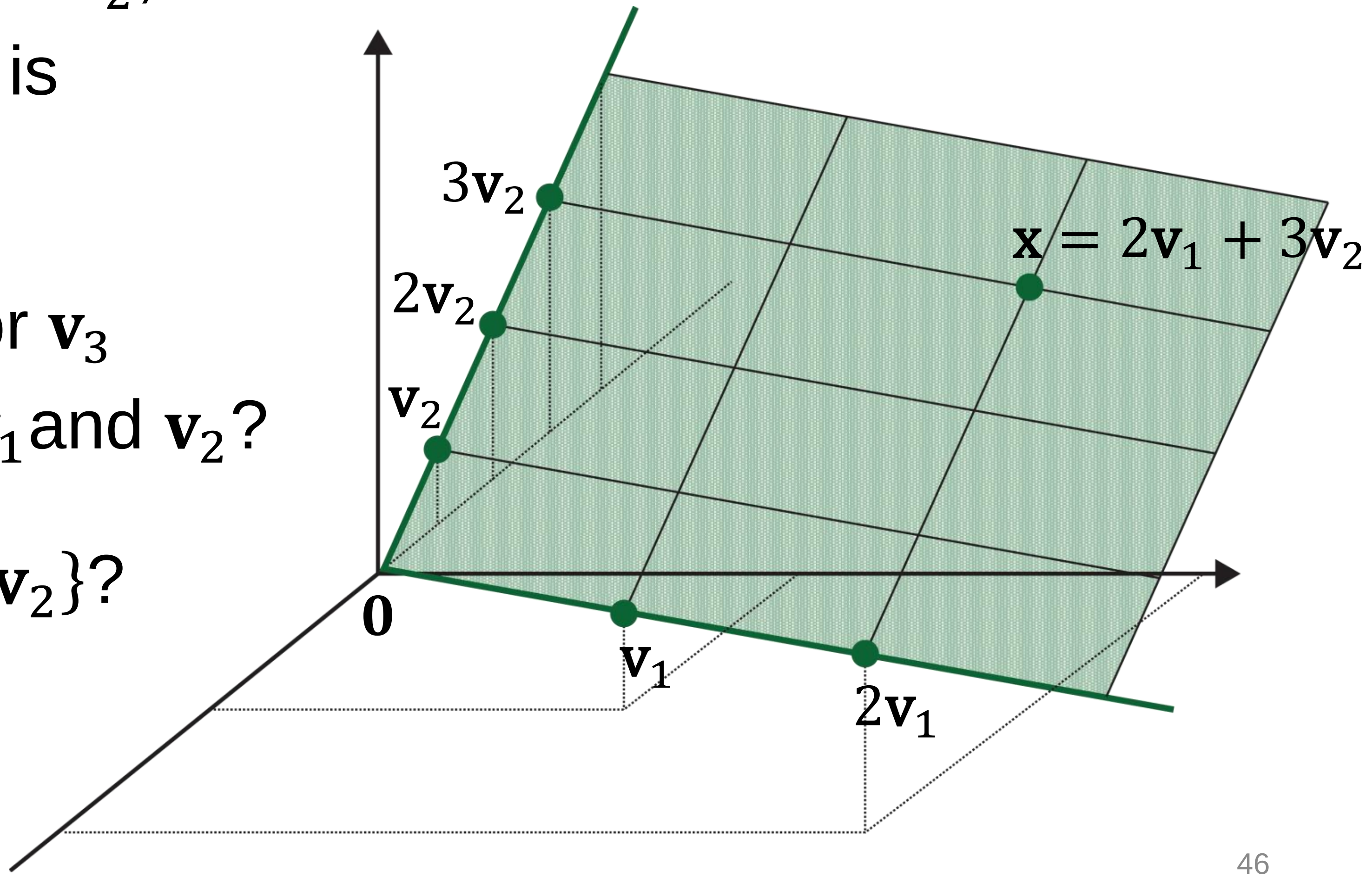
- If $\mathbf{v}_1, \cdots, \mathbf{v}_p$ are linearly dependent, consider a nontrivial solution.
- In the solution, let's denote j as the last index such that $x_j \neq 0$.
- Then, one can write $x_j \mathbf{v}_j = -x_1 \mathbf{v}_1 - \cdots - x_{j-1} \mathbf{v}_{j-1}$,
  and safely divide it by x_j, 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 $\mathbf{v}_3 \in \text{Span}\{\mathbf{v}_1, \mathbf{v}_2\}$, 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 $\mathbf{v}_1, \mathbf{v}_2$.

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 more 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 $\mathbf{b} \in \mathrm{Span}\,\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$.

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

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

- 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?
  - $\mathbf{u}_1 = a_1\mathbf{v}_1 + \cdots + a_p\mathbf{v}_p$, $\mathbf{u}_2 = b_1\mathbf{v}_1 + \cdots + b_p\mathbf{v}_p$
  - $$\begin{aligned} c\mathbf{u}_1 + d\mathbf{u}_2 &= c\left(a_1\mathbf{v}_1 + \cdots + a_p\mathbf{v}_p\right) + d\left(b_1\mathbf{v}_1 + \cdots + b_p\mathbf{v}_p\right) \\ &= (ca_1 + db_1)\mathbf{v}_1 + \cdots + \left(ca_p + db_p\right)\mathbf{v}_p \end{aligned}$$
- In fact, a subspace is always represented as Span $\{\mathbf{v}_1, \cdots, \mathbf{v}_p\}$.

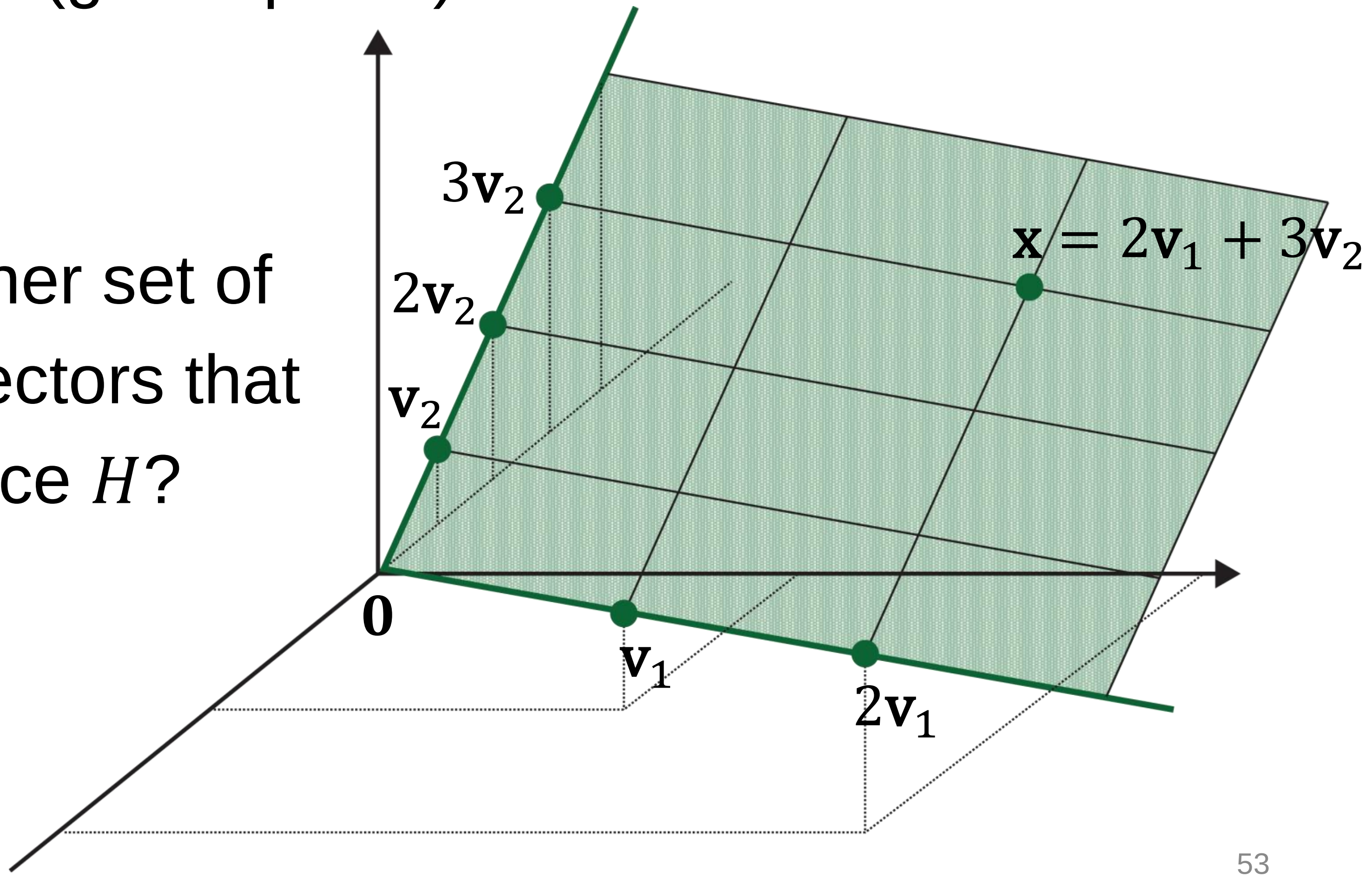
Basis of a Subspace

- **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 $H = \text{Span}\,\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, 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.

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

- 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.

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

- 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

Lecture Overview

- Elements in linear algebra
- Linear system
- Linear combination, vector equation,
  Four views of matrix multiplication
- Linear independence, span, and subspace
- <u>Linear transformation</u>
- Least squares
- Eigendecomposition
- Singular value decomposition

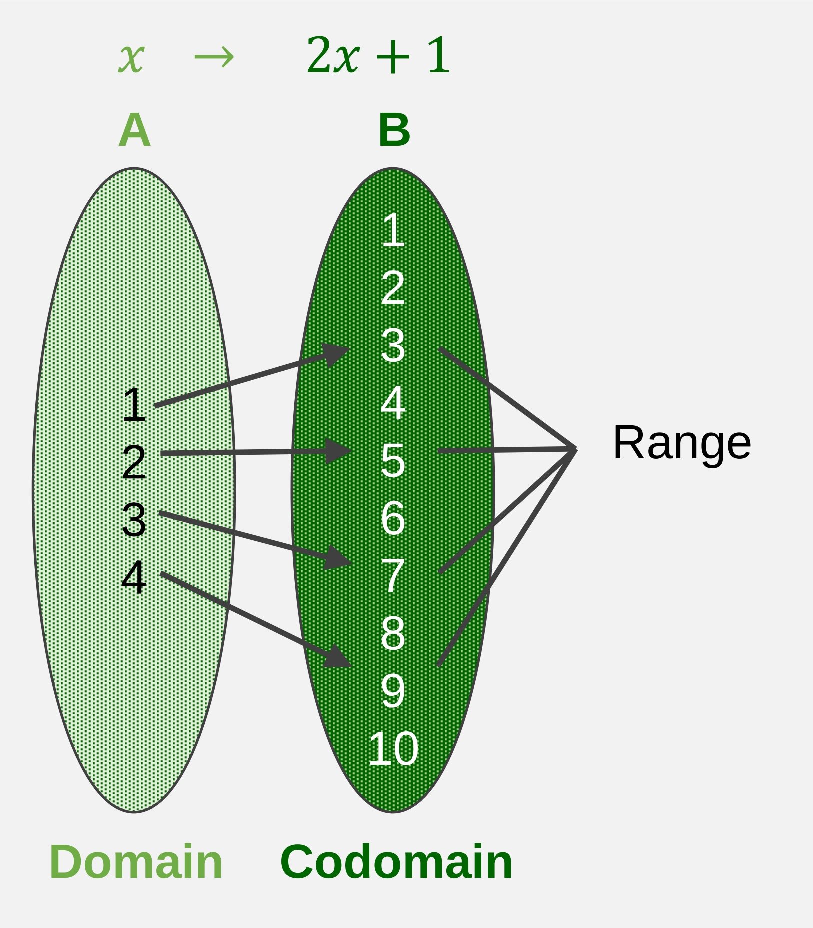
Summary So Far

- 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
- **Co-domain**: Set of all the possible values of y
- **Image**: a mapped output y, given x
- **Range**: Set of all the output values mapped by each x in the domain

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

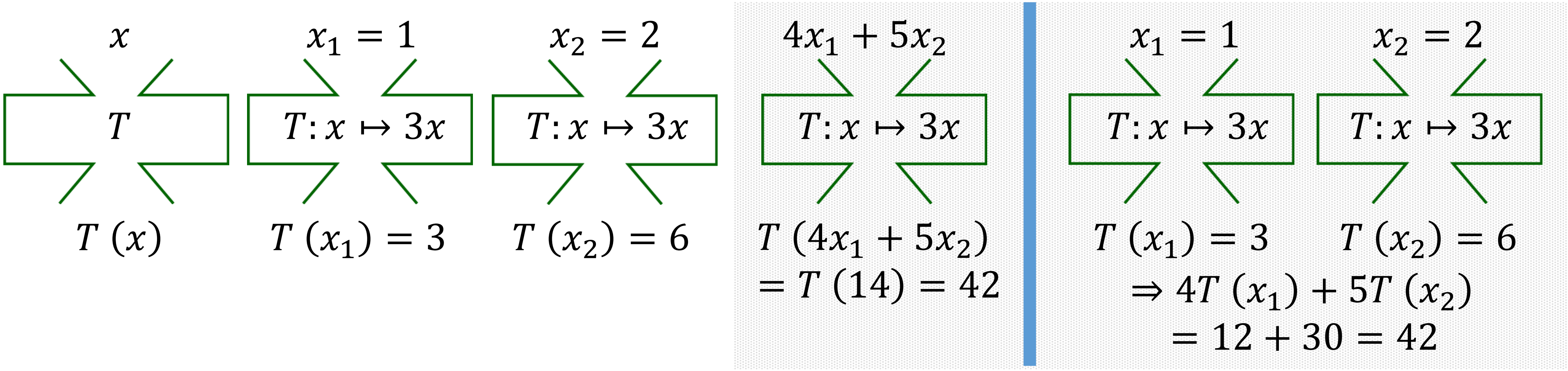


Linear Transformation

- **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: $x \mapsto y$, $T(x) = y\ = 3x$



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 $\mathbb{R}^2$

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

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

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

Matrix of Linear Transformation

- 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 $\mathbb{R}^{n\times n}$:

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

- 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

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

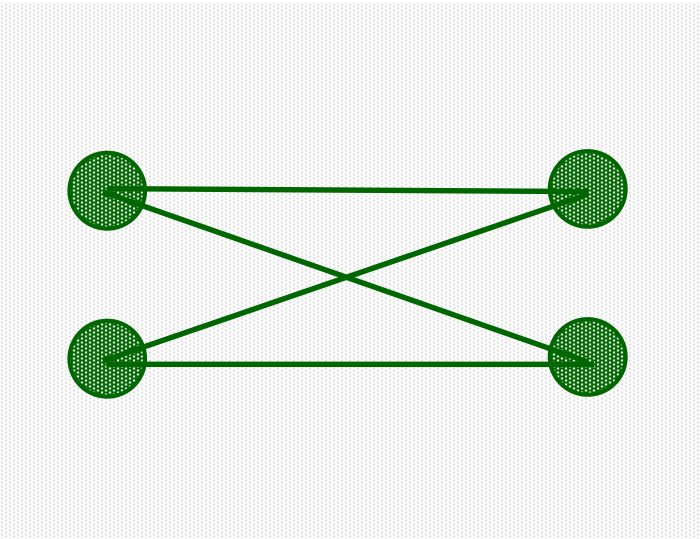
$$\mathbf{x}=\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=x_1\begin{bmatrix}1\\0\\0\end{bmatrix}+x_2\begin{bmatrix}0\\1\\0\end{bmatrix}+x_3\begin{bmatrix}0\\0\\1\end{bmatrix}$$

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

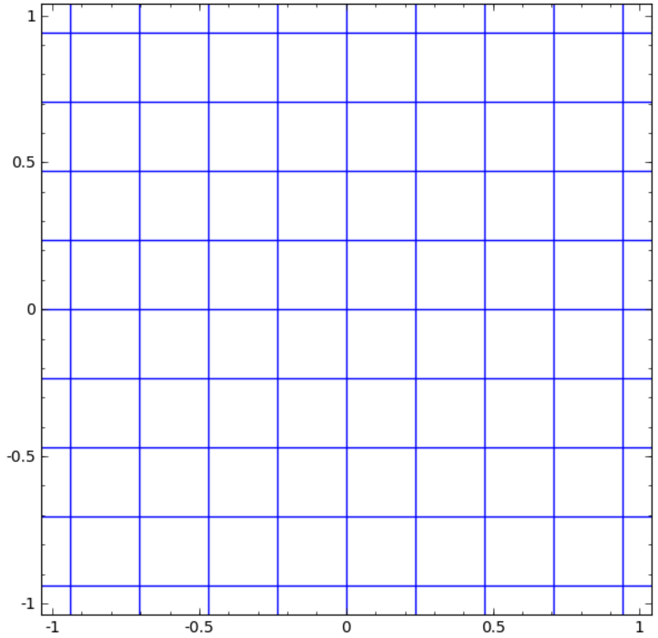
$$=x_1\begin{bmatrix}1\\2\end{bmatrix}+x_2\begin{bmatrix}4\\3\end{bmatrix}+x_2\begin{bmatrix}5\\6\end{bmatrix}=\begin{bmatrix}1&4&5\\2&3&6\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=A\mathbf{x}$$

Linear Transformation in Neural Networks

- Fully-connected layers (linear layer)



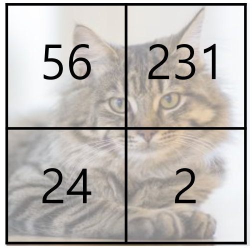
$$\mathbf{x} \xrightarrow[T_1]{} \mathbf{y}$$

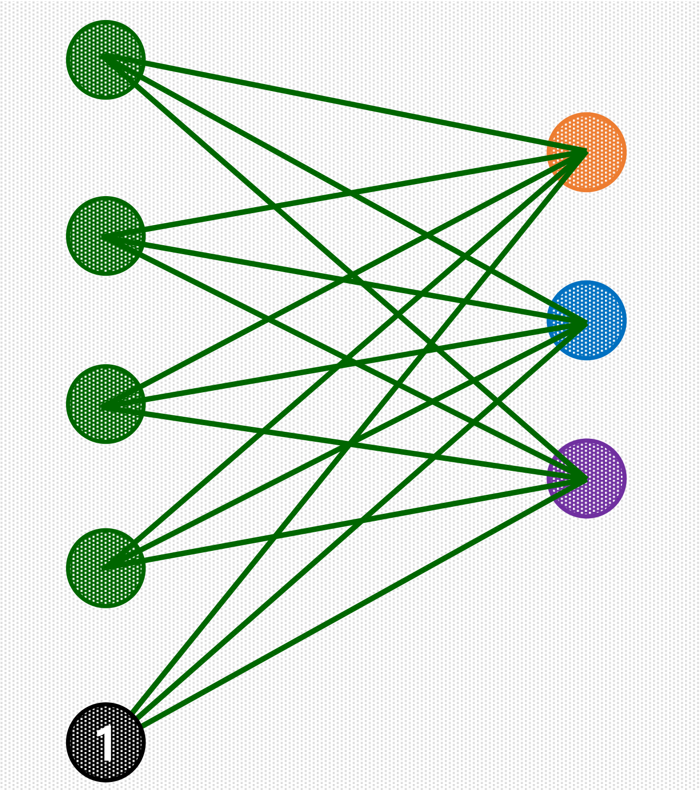


https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

Affine Layer in Neural Networks

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





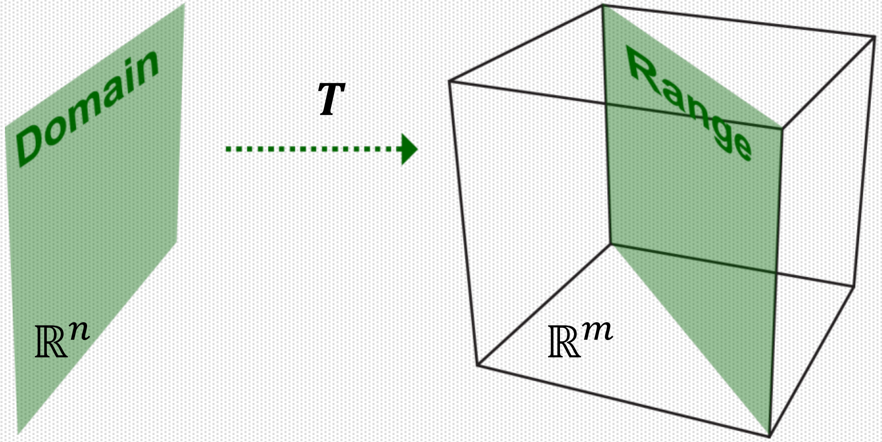
$$\begin{bmatrix}0.2&-0.5&0.1&2\\1.5&1.3&2.1&1\\-.2&0.3&0.7&-1.3\end{bmatrix}\begin{bmatrix}56\\231\\24\\2\end{bmatrix}+\begin{bmatrix}1.1\\3.2\\-1.2\end{bmatrix}=\begin{bmatrix}-96.8\\439.9\\71.1\end{bmatrix}$$

$$=56\begin{bmatrix}0.2\\1.5\\-.2\end{bmatrix}+231\begin{bmatrix}-0.5\\1.3\\0.3\end{bmatrix}+24\begin{bmatrix}0.1\\2.1\\0.7\end{bmatrix}+2\begin{bmatrix}2\\1\\-1.3\end{bmatrix}+1\begin{bmatrix}1.1\\3.2\\-1.2\end{bmatrix}$$

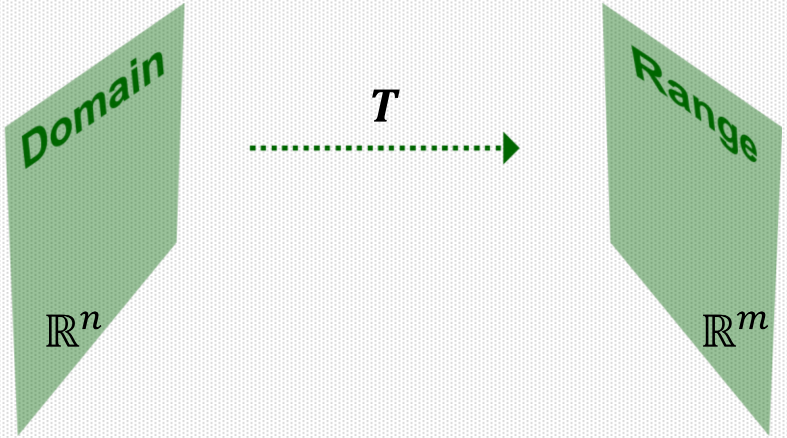
$$=\begin{bmatrix}0.2&-0.5&0.1&2&1.1\\1.5&1.3&2.1&1&3.2\\-.2&0.3&0.7&-1.3&-1.2\end{bmatrix}\begin{bmatrix}56\\231\\24\\2\\1\end{bmatrix}$$

ONTO and ONE-TO-ONE

- **Definition:** A mapping $T : \mathbb{R}^n \to \mathbb{R}^m$ 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.



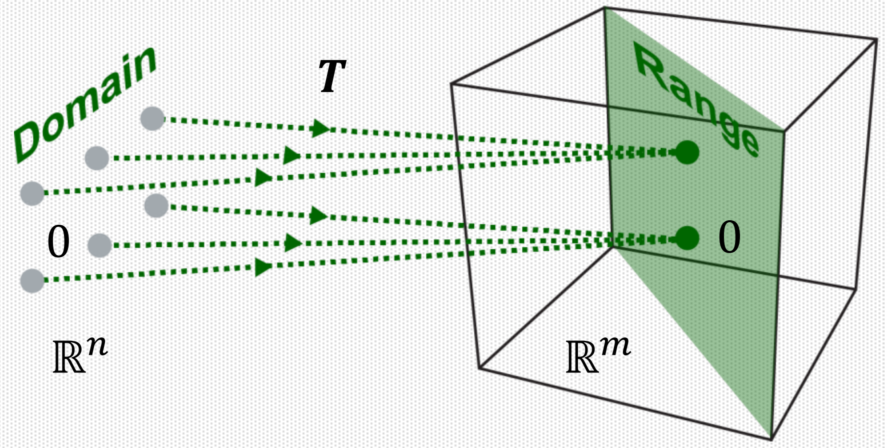
T is NOT onto $\mathbb{R}^m$



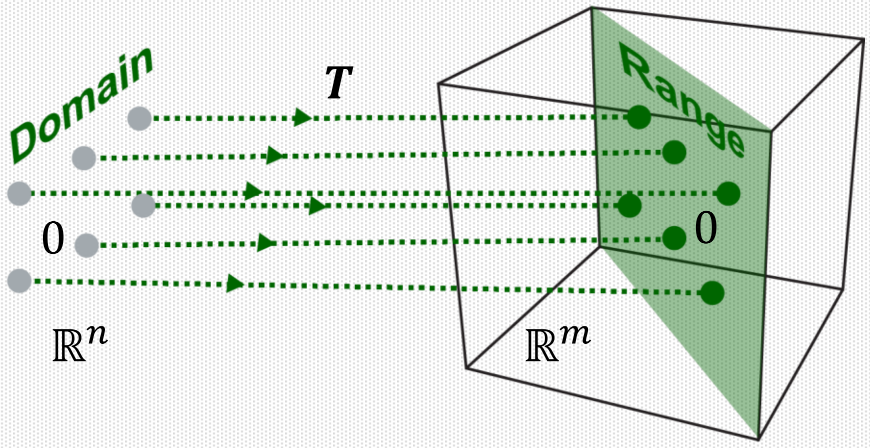
T is onto $\mathbb{R}^m$

ONTO and ONE-TO-ONE

- **Definition:** A mapping $T : \mathbb{R}^n \to \mathbb{R}^m$ 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.



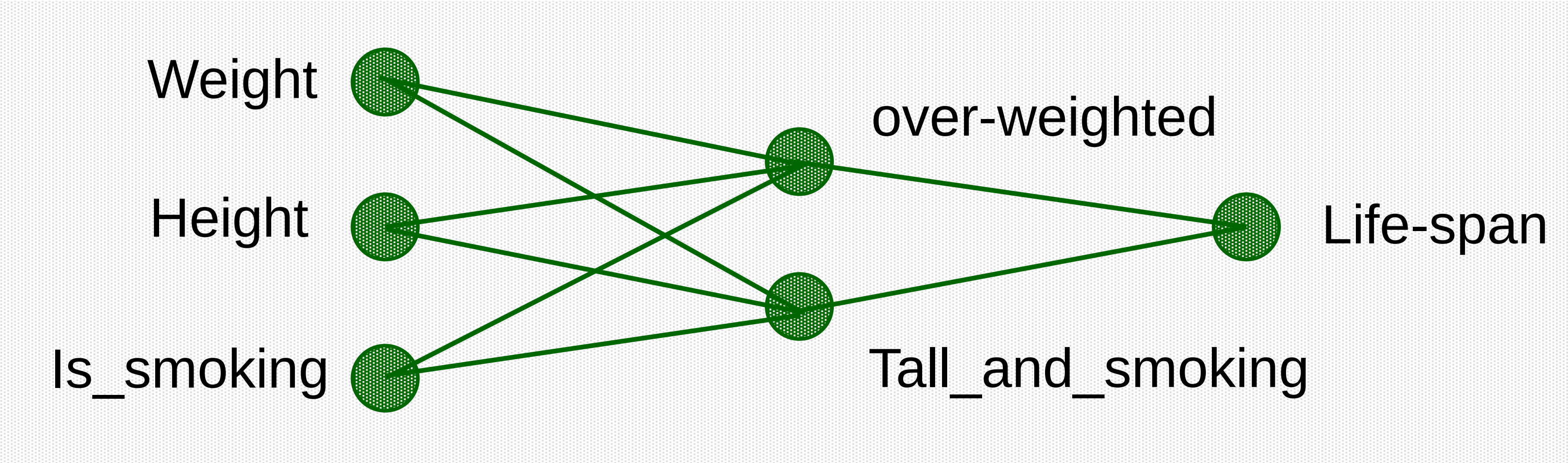
T is NOT one-to-one



T is one-to-one

Neural Network Example

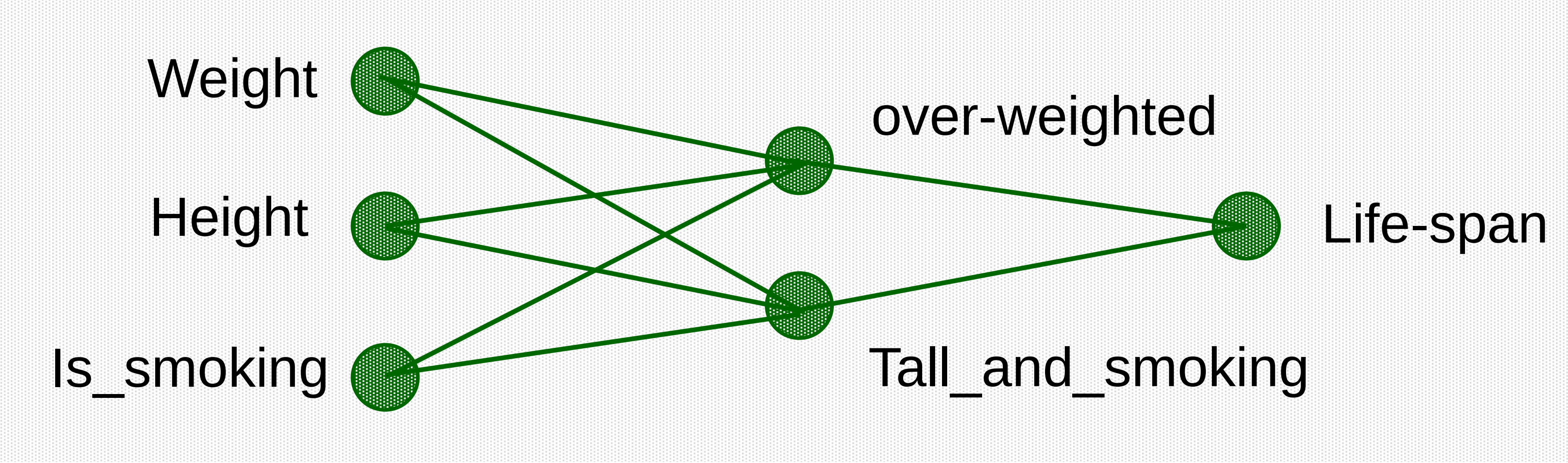
- Fully-connected layers



$$\mathbf{x} \underset{T_1}{\longrightarrow} \mathbf{y} \underset{T_2}{\longrightarrow} \mathbf{z}$$

Neural Network Example: ONE-TO-ONE

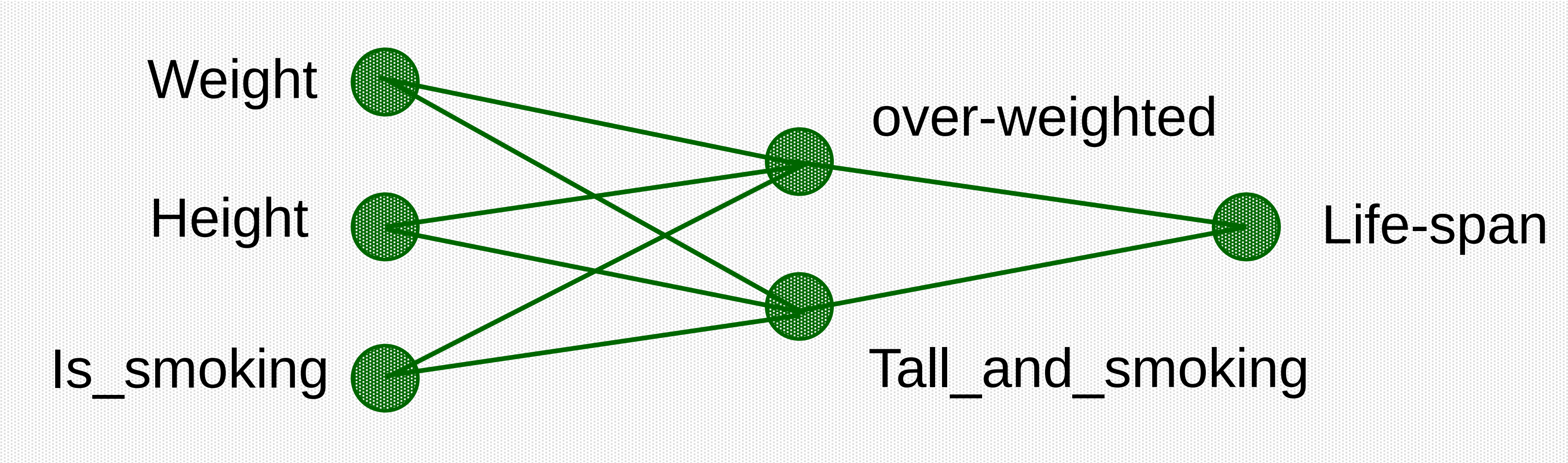
- Will there be many (or unique) people mapped to the same (over_weighted, tall_and_smoking)?



$$\mathbf{x} \underset{T_1}{\longrightarrow} \mathbf{y} \underset{T_2}{\longrightarrow} \mathbf{z}$$

Neural Network Example: ONTO

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



$$\mathbf{x} \underset{T_1}{\longrightarrow} \mathbf{y} \underset{T_2}{\longrightarrow} \mathbf{z}$$

ONTO and ONE-TO-ONE

- Let $T : \mathbb{R}^n \to \mathbb{R}^m$ 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.

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

ONTO and ONE-TO-ONE

- **Example:**

  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?
- Does T map $\mathbb{R}^2$ onto $\mathbb{R}^3$?

ONTO and ONE-TO-ONE

- **Example:**

  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?
- Does T map $\mathbb{R}^3$ onto $\mathbb{R}^2$?

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
- <u>Least squares</u>
- Eigendecomposition
- Singular value decomposition

Over-determined Linear Systems (#equations $\gg$ #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 $\gg$ #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
⋮	⋮	⋮	⋮	⋮

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

- Matrix equation:

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

$m \gg n$: more equations than variables

➡ Usually no solution exists

Vector Equation Perspective

- Vector equation form: $$\begin{bmatrix} 60 \\ 65 \\ 55 \\ \vdots \end{bmatrix} x_1 + \begin{bmatrix} 5.5 \\ 5.0 \\ 6.0 \\ \vdots \end{bmatrix} x_2 + \begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \end{bmatrix} x_3 = \begin{bmatrix} 66 \\ 74 \\ 78 \\ \vdots \end{bmatrix}$$
$$\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \mathbf{a}_3 x_3 = \mathbf{b}$$

- 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 1×1 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:

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

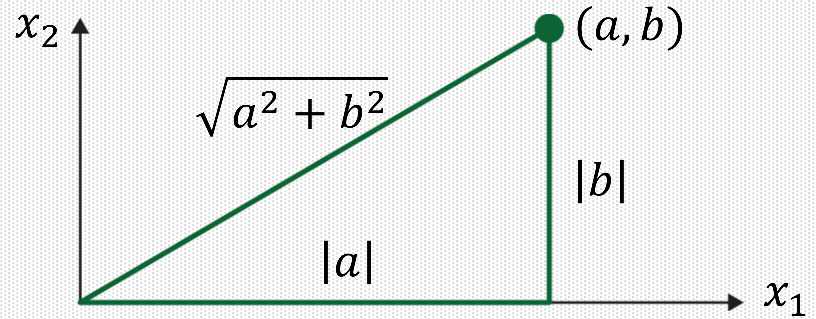
Vector Norm

- 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 $\|\mathrm{v}\|$ 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

$$\mathbf{u} = \begin{bmatrix} 6 \\ 1 \end{bmatrix} \text{ and } \mathbf{v} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}.$$

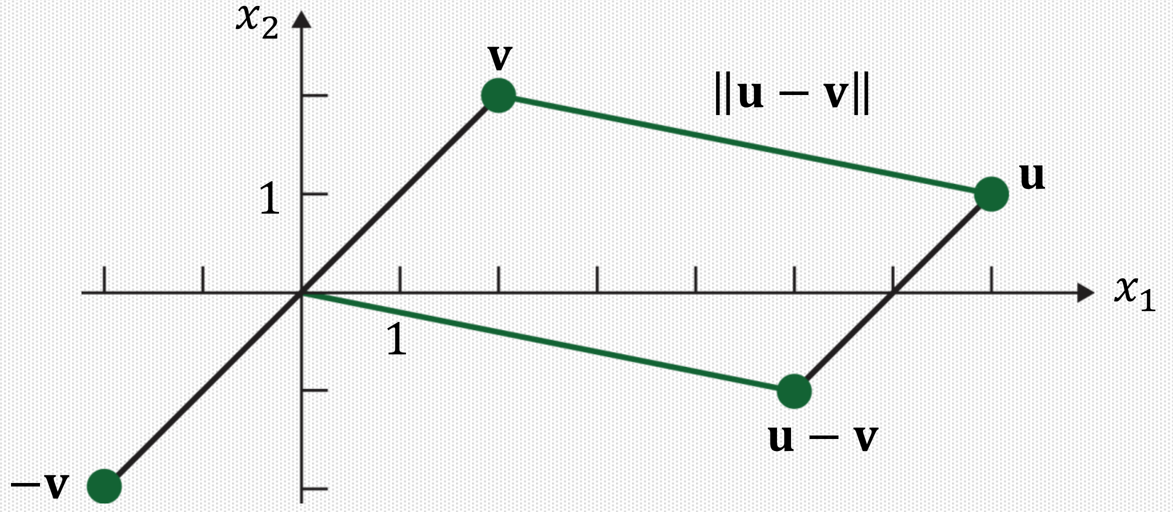
- **Solution**: Calculate

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

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

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.



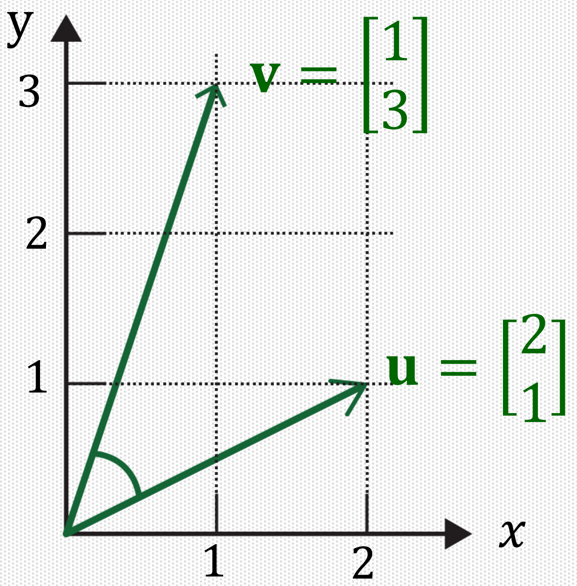
The distance between $\mathbf{u}$ and $\mathbf{v}$ is the length of $\mathbf{u}-\mathbf{v}$

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**:



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

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

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

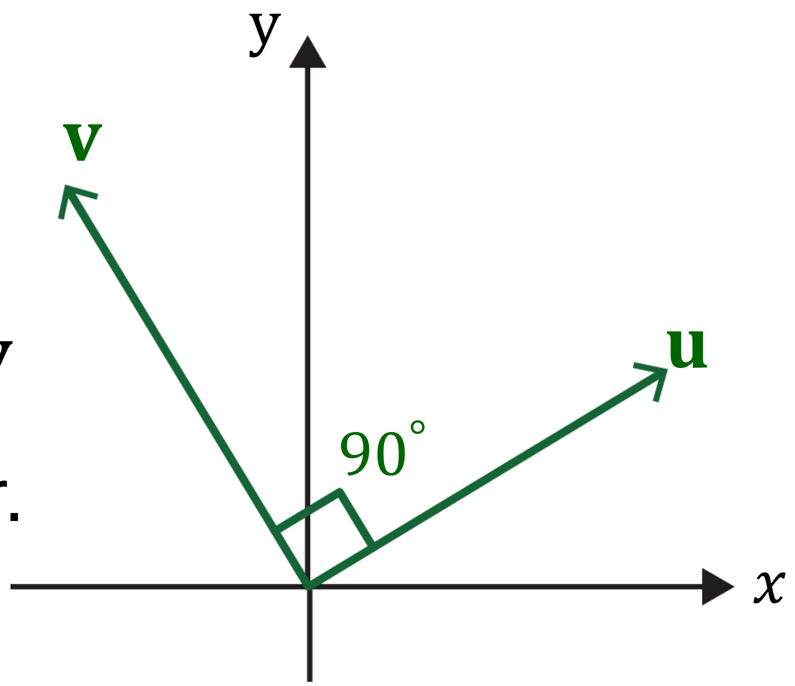
$$\Rightarrow \cos\theta = \frac{5}{\sqrt{50}} = \frac{1}{\sqrt{2}}$$

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

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,
  $\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\| \cos\theta = 0.$
  $\Rightarrow$ $\cos\theta = 0$ for nonzero vectors $\mathbf{u}$ and $\mathbf{v}$
  $\Rightarrow$ $\theta = 90°$ ($\mathbf{u} \perp \mathbf{v}$).
  $\Rightarrow$ $\mathbf{u}$ 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

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

- 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 one more 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

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

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

$$\overset{A}{\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ 50 & 5.0 & 1 \end{bmatrix}} \overset{\mathbf{x}}{\begin{bmatrix} -0.4 \\ 20 \\ -20 \end{bmatrix}} = \begin{bmatrix} 66 \\ 74 \\ 78 \\ 60 \end{bmatrix} \overset{\neq \mathbf{b}}{\neq \begin{bmatrix} 66 \\ 74 \\ 78 \\ 72 \end{bmatrix}}$$

Errors

$(\mathbf{b} - A\mathbf{x})$
0
0
0
12

Back to Over-Determined System

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

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

Errors

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

Which One is Better Solution?

Errors

$$\begin{array}{cccc} A & \mathbf{x} & \neq & \mathbf{b} & (\mathbf{b}-A\mathbf{x}) \end{array}$$

$$\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ 50 & 5.0 & 1 \end{bmatrix} \begin{bmatrix} -0.12 \\ 16 \\ -9.5 \end{bmatrix} = \begin{bmatrix} 71.3 \\ 72.2 \\ 79.9 \\ 64.5 \end{bmatrix} \neq \begin{bmatrix} 66 \\ 74 \\ 78 \\ 72 \end{bmatrix} \quad \begin{matrix} -5.3 \\ 1.8 \\ -1.9 \\ 7.5 \end{matrix}$$

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

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

Least Squares Problem

- Now, the sum of squared errors can be represented as $||\mathbf{b} - A\mathbf{x}||$.

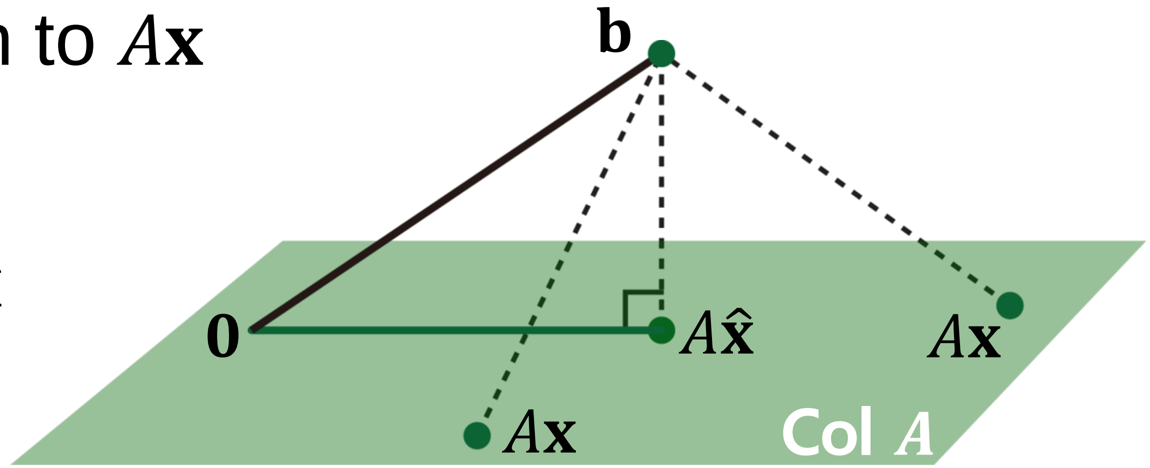
- **Definition**: Given an overdetermined system $A\mathbf{x} \simeq \mathbf{b}$ where $A \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^{n}$, and $m \gg n$, a least squares solution $\hat{\mathbf{x}}$ is defined as

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

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

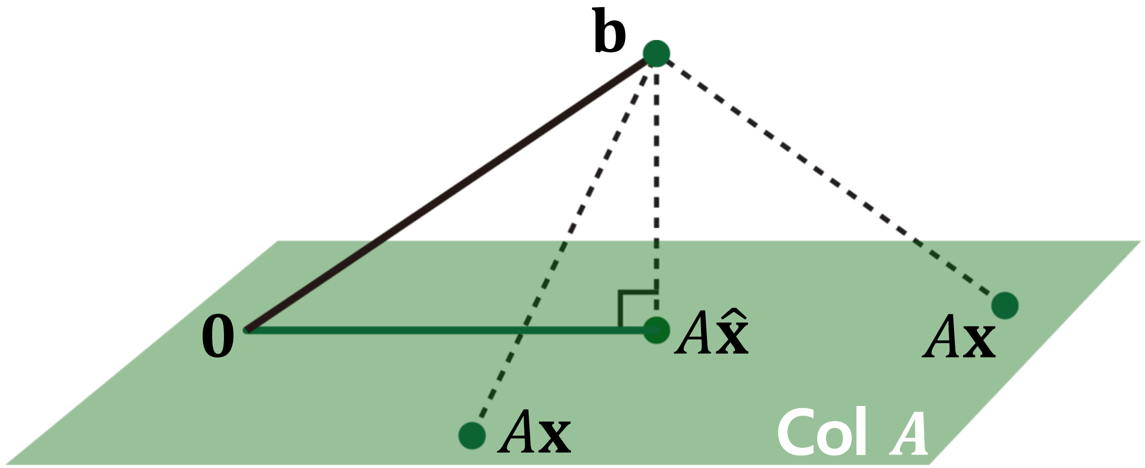


- This means $\mathbf{b} - A\hat{\mathbf{x}}$ should be orthogonal to any vector in Col A:

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

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,

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

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

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

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

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

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

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

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

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

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
- <u>Eigendecomposition</u>
- Singular value decomposition

Eigenvectors and Eigenvalues

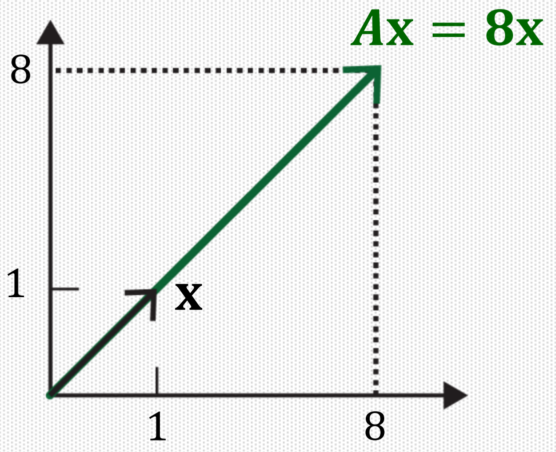
- **Definition**: An **eigenvector** of a **square** matrix $A \in \mathbb{R}^{n \times n}$ is a **nonzero** vector $\mathbf{x} \in \mathbb{R}^n$ such that $A\mathbf{x} = \lambda\mathbf{x}$ for some scalar λ
  In this case, λ is called an **eigenvalue** of A, and
  such an $\mathbf{x}$ is called an ***eigenvector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

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 λ.

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

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



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

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

- λ 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 λ**.
- The eigenspace consists of the zero vector and all the eigenvectors corresponding to λ, satisfying the above equation.

Null Space

- **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 $\mathbb{R}^n$.

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

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 λ 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

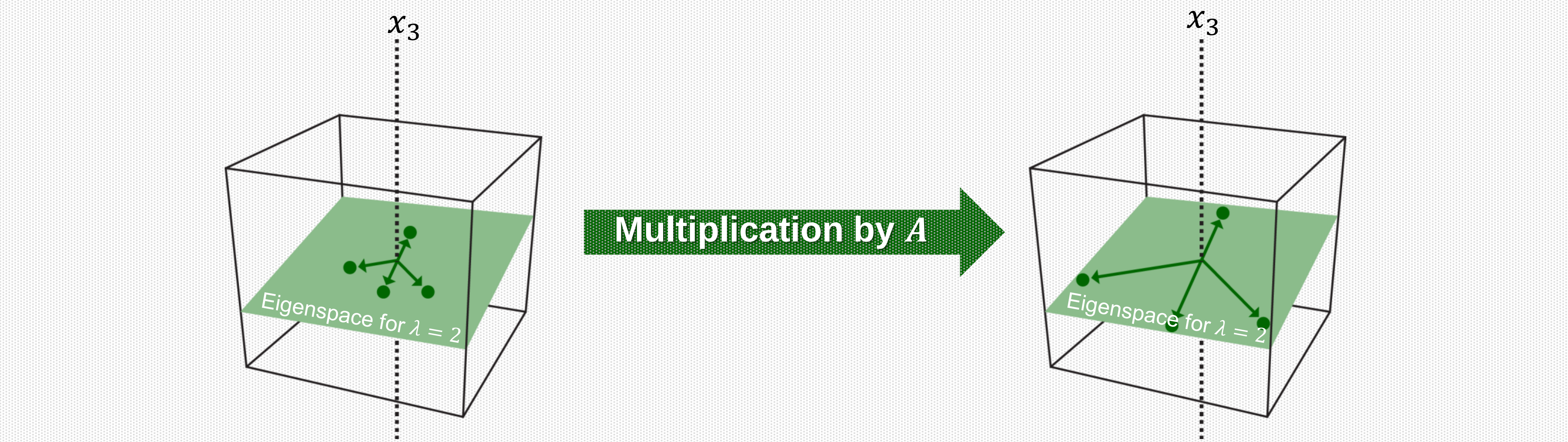
- Once obtaining eigenvalues, we compute the eigenvectors for each λ by solving

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

Eigenspace

- Note that the dimension of the eigenspace (corresponding to a particular λ) can be more 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 λ, 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, $\mathbf{v}_1$, $\mathbf{v}_2$, …, $\mathbf{v}_n$ should be eigenvectors and λ_1, λ_2, …, λ_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

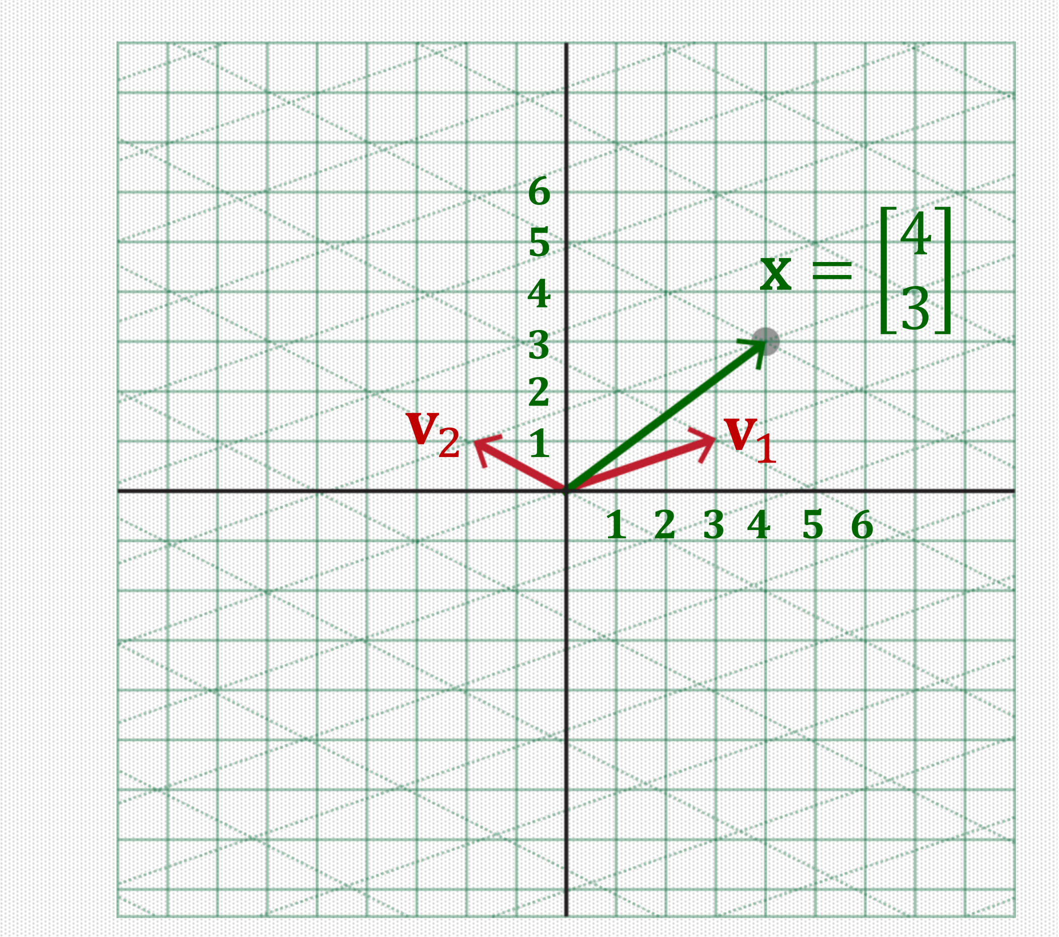
- Suppose A is diagonalizable, thus having eigendecomposition

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

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

Change of Basis

- Suppose $A\mathbf{v}_1 = -1\mathbf{v}_1$ and $A\mathbf{v}_2 = 2\mathbf{v}_2$.
- $T(\mathbf{x}) = A\mathbf{x} = PDP^{-1}\mathbf{x} = P\left(D\left(P^{-1}\mathbf{x}\right)\right)$
- 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 $\{\mathbf{v}_1, \mathbf{v}_2\}$.



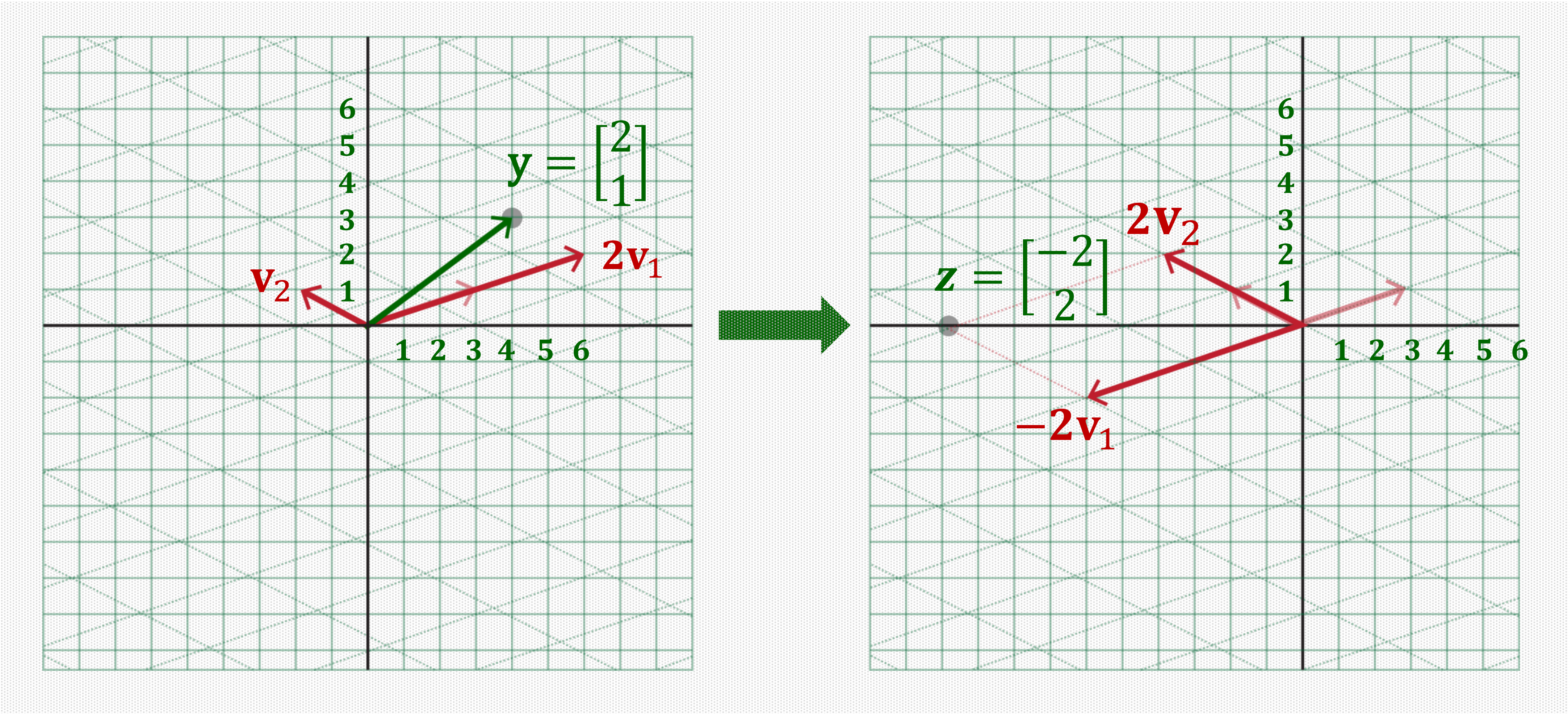
$$\mathbf{x} = \begin{bmatrix} 4 \\ 3 \end{bmatrix} = 4\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 3\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 4 \\ 3 \end{bmatrix} = P\mathbf{y} = [\mathbf{v}_1 \quad \mathbf{v}_2]\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}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

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,

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

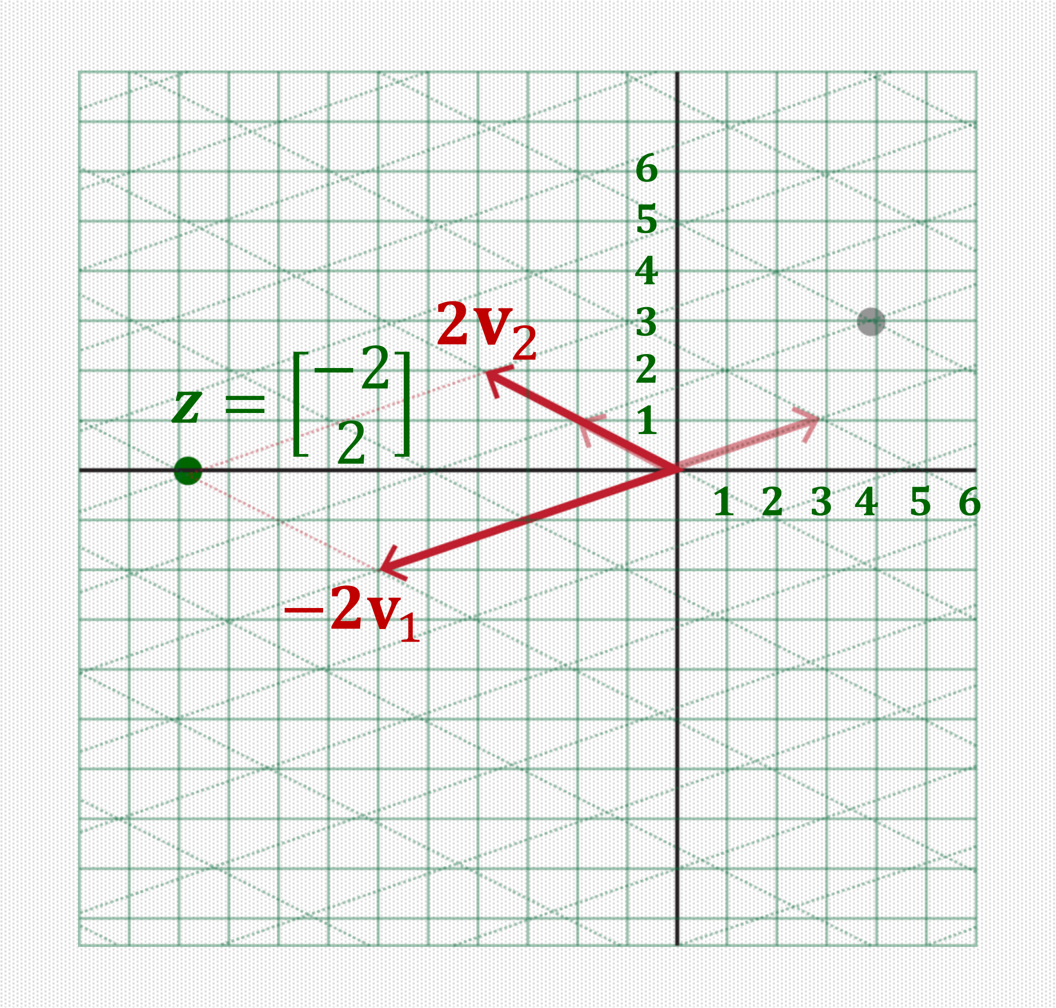
Back to Original Basis

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

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

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

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



Overview of Transformation using Eigendecomposition

$$
\begin{tikzcd}
x \arrow[r, "\boldsymbol{A}"] \arrow[d, "\boldsymbol{P^{-1}}"'] & T(x) = Ax = P\left(D(P^{-1}x)\right) \\
y \arrow[r, "\boldsymbol{D}"] & z \arrow[u, "\boldsymbol{P}"']
\end{tikzcd}
$$

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
- <u>Singular value decomposition</u>

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 (unit vectors orthogonal to each other)
- $\Sigma \in \mathbb{R}^{m \times n}$: nonzero values along diagonals

COMPUTING SVD

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

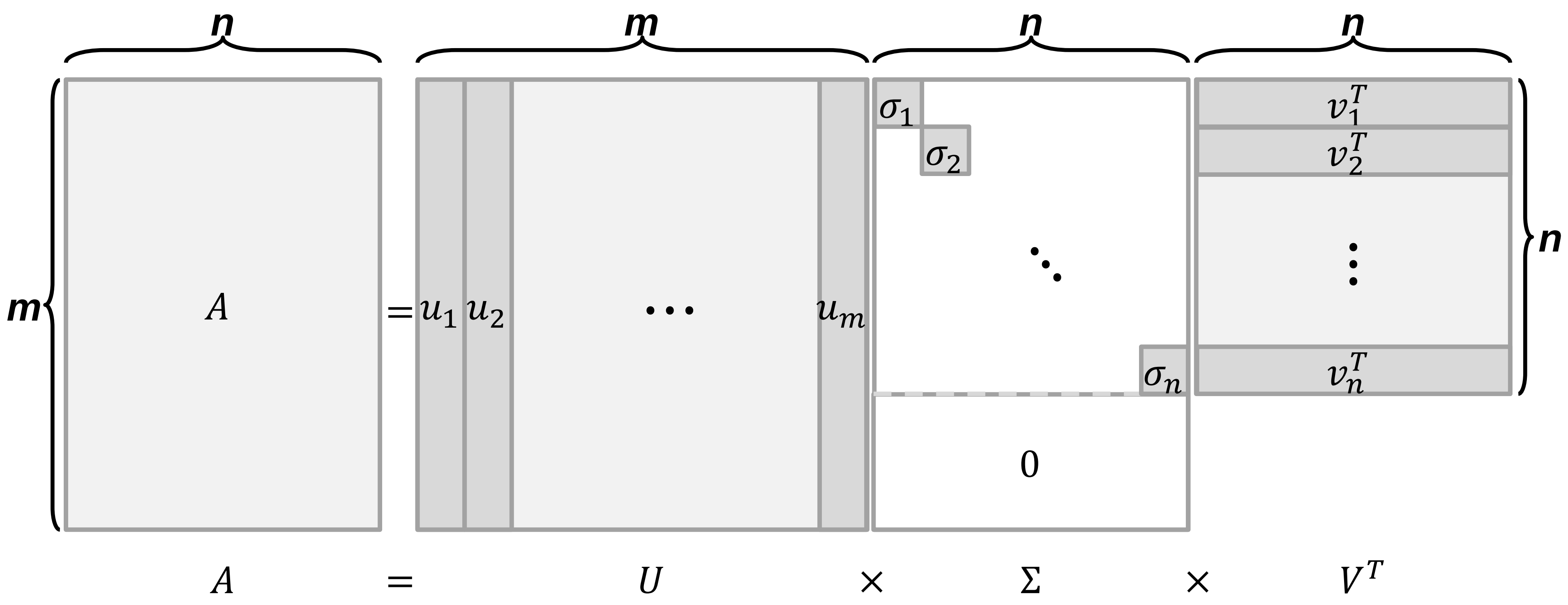
$$AA^T = U\Sigma V^T \mathrm{V}\Sigma^T U^T = U\Sigma\Sigma^T U^T = U\Sigma^2 U^T$$
$$A^T A = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T\Sigma U^T = V\Sigma^2 V^T$$

- Note that AA^T and A^TA are symmetric.

Basic Form of SVD

- Given a $m \times n$ matrix A where $m > n$, SVD gives

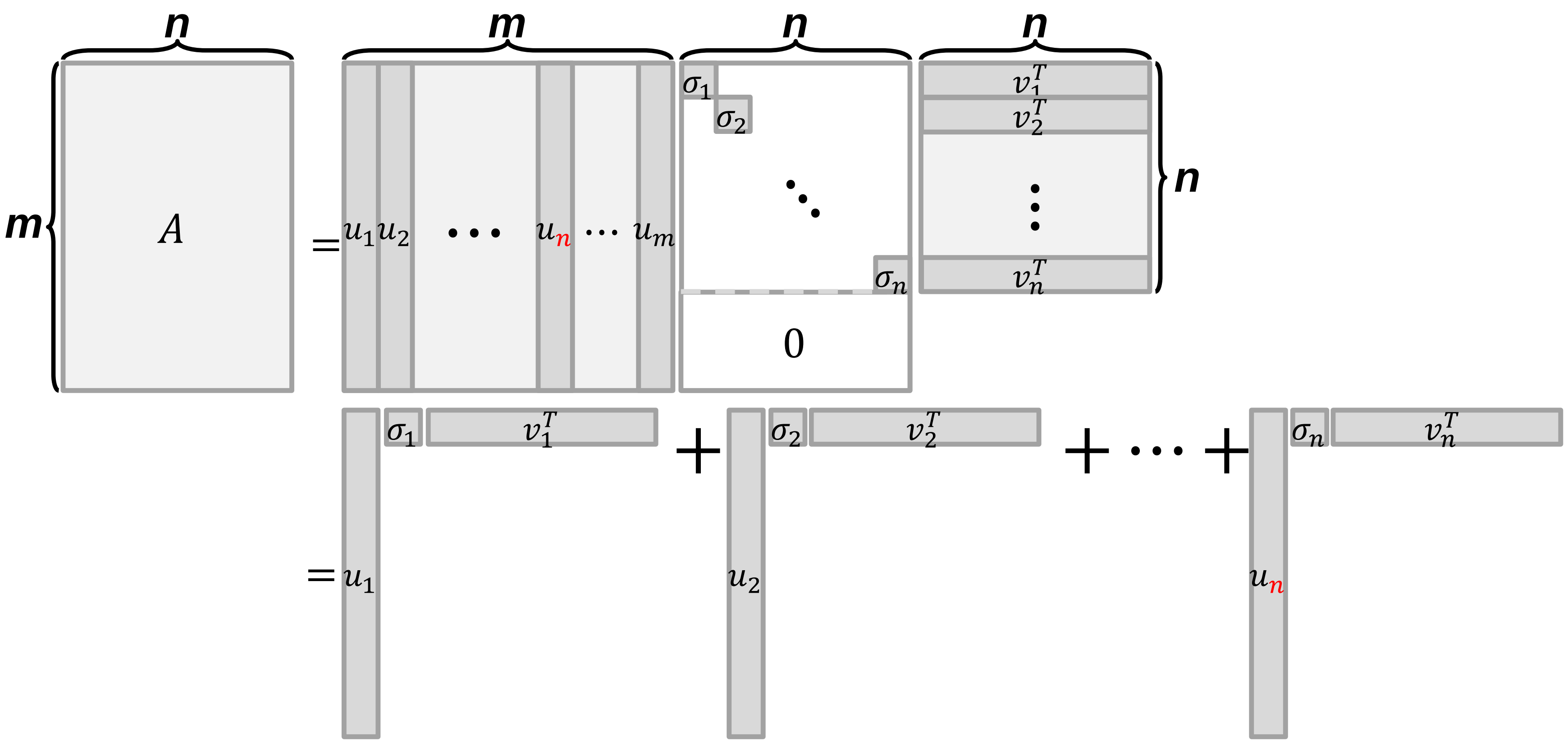
$$A = U\Sigma V^T$$



Sum of Outer Products

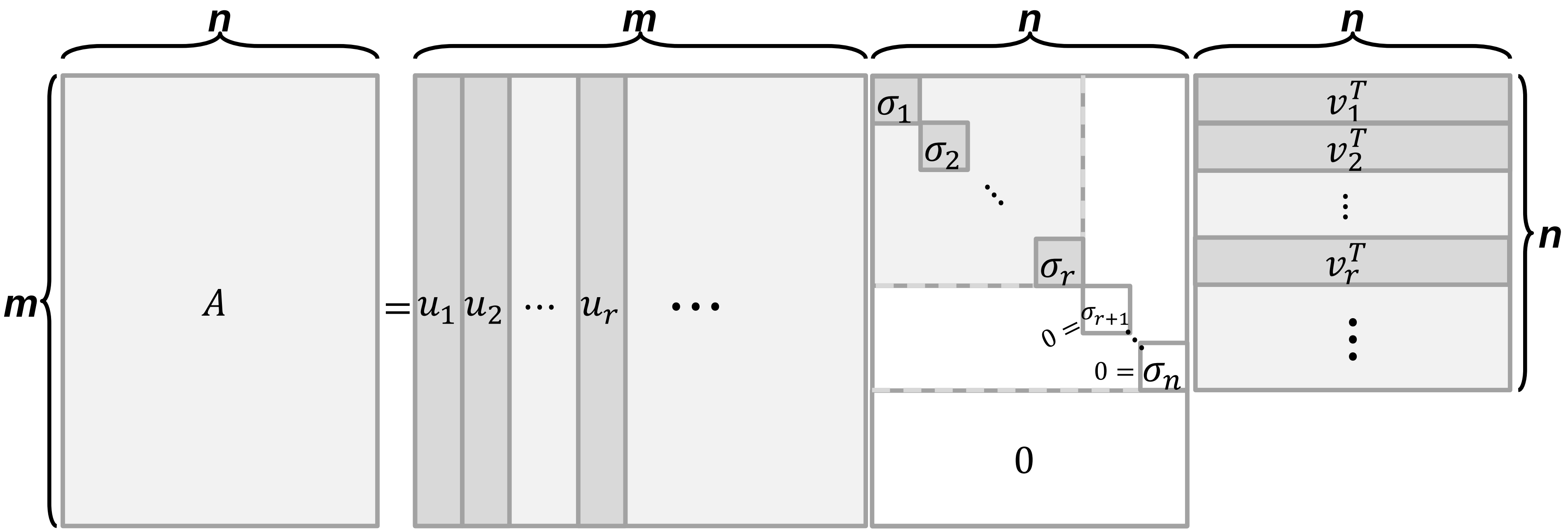
- A can also be represented as the sum of outer products

$$A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T$$



A with Rank r

- Assume the rank of A is r where $r \le n$, then $\sigma_i = 0$ for $\forall i \ge (r+1)$.



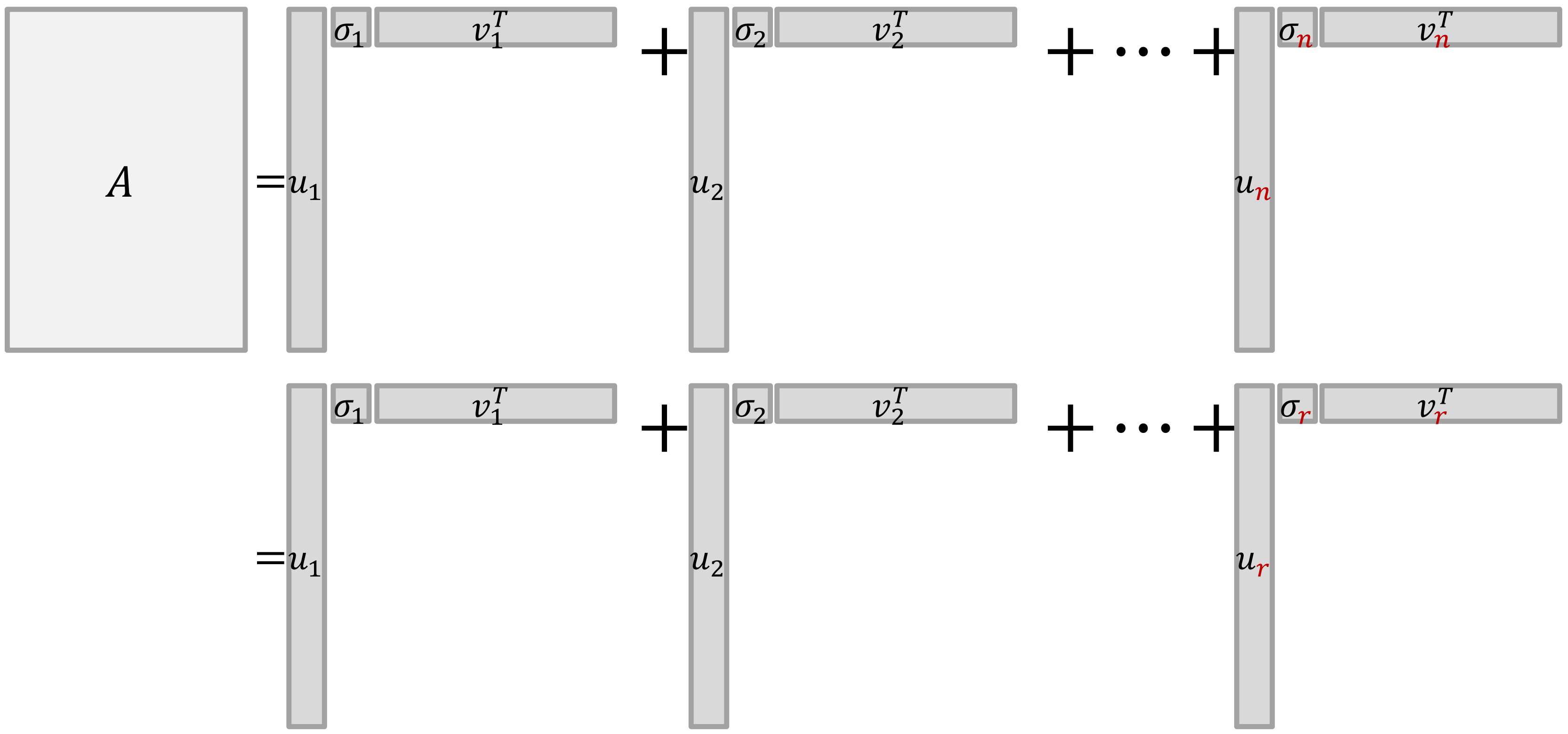
- Definition of the rank of A: the number of linearly independent columns of A.

A with Rank r: Sum of Outer Products

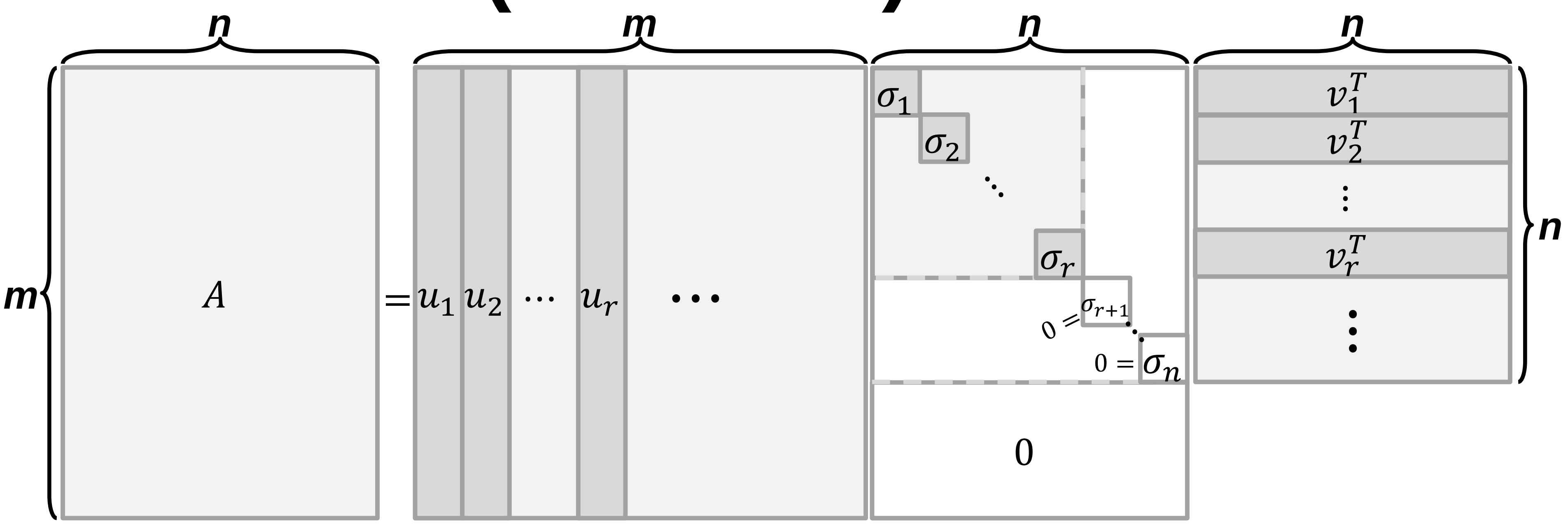
- The sum of outer products then becomes

$$A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

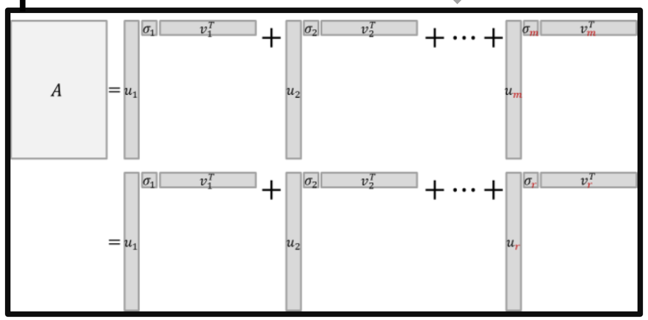
since $\sigma_i = 0$ for $\forall i \geq (r+1)$.

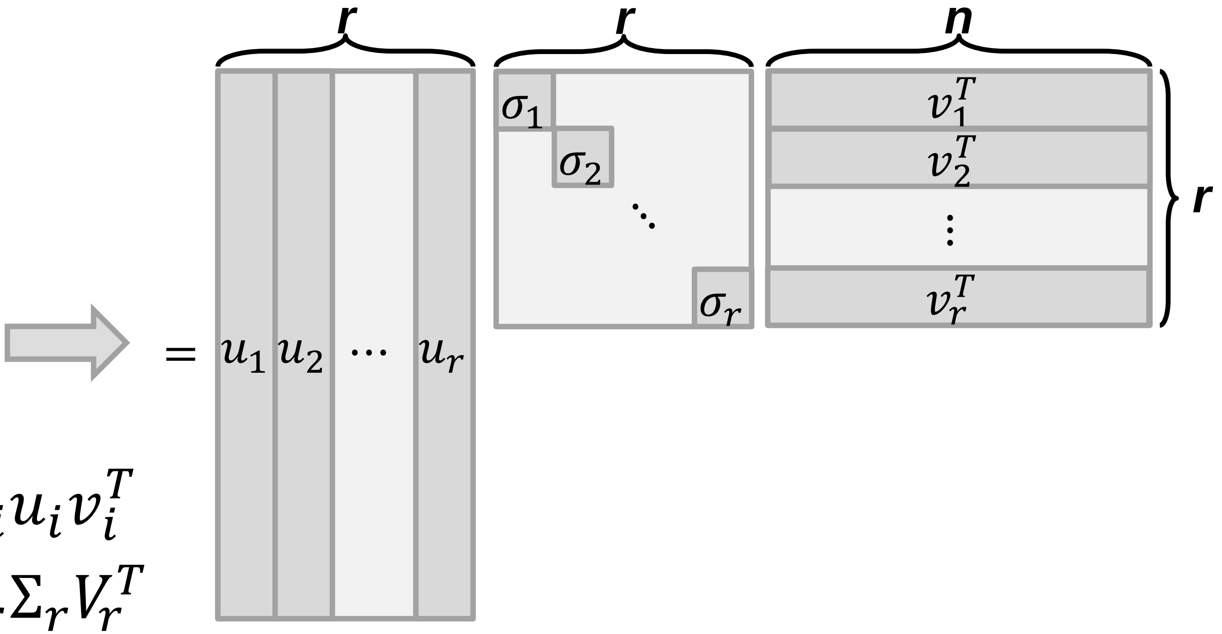


Back to (Reduced) Matrix Form



From the previous slide



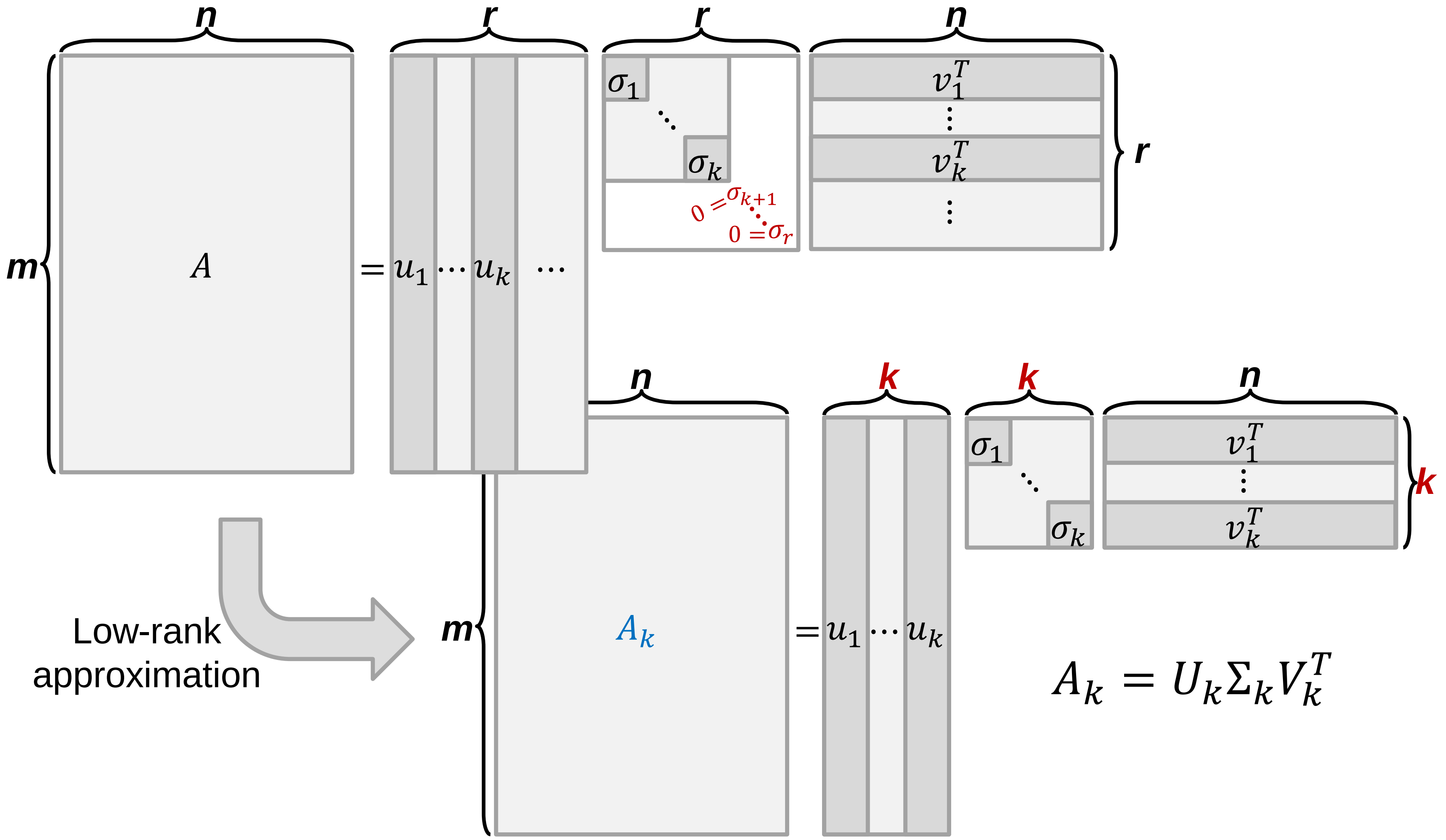


- $A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T$
  $= \sum_{i=1}^{r} \sigma_i u_i v_i^T = U_r \Sigma_r V_r^T$

Reducing Dimension (from m) to k

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

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T \simeq A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T = U_k \Sigma_k V_k^T$$



THANK YOU