Announcement

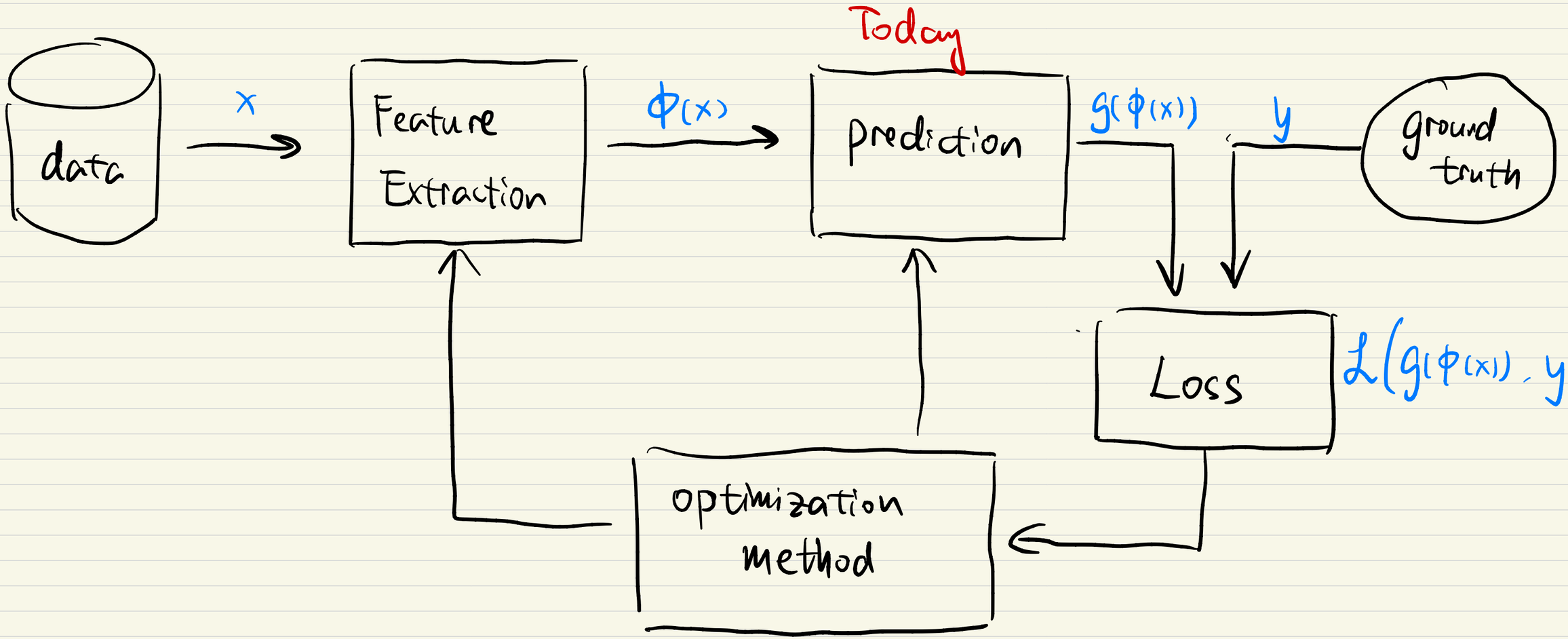
- HW 0 solution on BrightSpace (Optional)

- Quiz 0: 1/14 (Fri) 8:00AM - 1/16 (Sun) 8:00AM, 30mins, GradeScope (Optional)

- Project: survey tonight on project logistics

- Real-time sharing of lecture notes, check BrightSpace (Beta)

Outline of Supervised Learning



Today's Lecture:

Linear Regression

- Recap: Formulation, Solution

- Examples

Geometric Interpretation of Linear Regression

- Projection

- Minimum-Norm Solution

- Pseudo-Inverse

Regression

- measurements (ground truth): y^n
- inputs (data): $\vec{x}^n$
- model: $g_{\vec{\theta}}(\vec{x}^n)$. $\vec{\theta}$ coefficients (parameters)

Determine $\vec{\theta}$ such that $y^n \approx g_{\vec{\theta}}(\vec{x}^n)$

Linear Regression (LR)

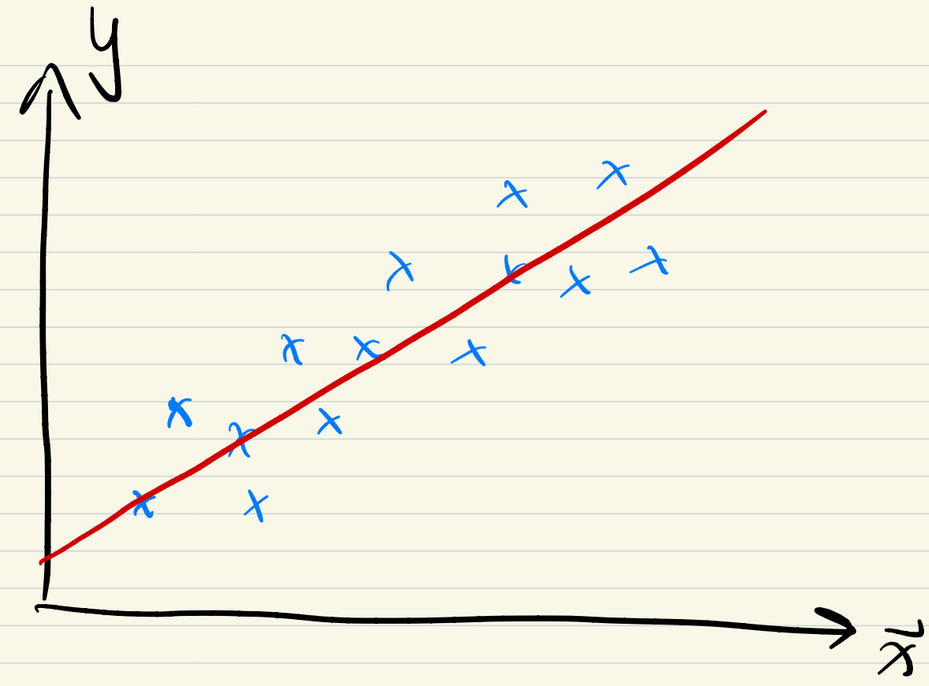
- $g_{\vec{\theta}}(\cdot)$ is a line: $g_{\vec{\theta}}(\vec{x}) = \vec{x}^T\vec{\theta} = \vec{\theta}^T\vec{x}$

Solution of LR.

Loss function $J(\vec{\theta})$

Square-error loss

$$J(\vec{\theta}) = \sum_{i=1}^{N} \left((\vec{x}^i)^T \vec{\theta} - y^i \right)^2$$



$$A = \begin{bmatrix} — (\vec{x}^1)^T — \\ \vdots \\ — (\vec{x}^N)^T — \end{bmatrix} \quad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

data (matrix)

measurement

$$J(\vec{\theta}) = \|A\vec{\theta} - \vec{y}\|^2$$

least square loss

Solution of LR.

Theorem:

For a linear regression problem

$$\hat{\vec{\theta}} = \underset{\vec{\theta}}{\operatorname{argmin}}\ J(\vec{\theta}) \overset{\text{def}}{=} \|A\vec{\theta} - \vec{y}\|^2$$

the minimizer

$$\hat{\vec{\theta}} = (A^T A)^{-1} A^T \vec{y}.$$

$$\nabla_{\vec{\theta}} J(\vec{\theta}) = 2A^T(A\vec{\theta} - \vec{y}) = \vec{0}$$

Read Tutorial on "Linear Algebra"

Example 1. Second-order Polynomial Fitting

$$y_n = ax_n^2 + bx_n + c$$

$$\vec{\theta} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \quad A = \begin{bmatrix} x_1^2 & x_1 & 1 \\ \vdots & \vdots & \vdots \\ x_N^2 & x_N & 1 \end{bmatrix} \quad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$\vec{y} = A\vec{\theta}$$

Generalizable to higher order of polynomial functions

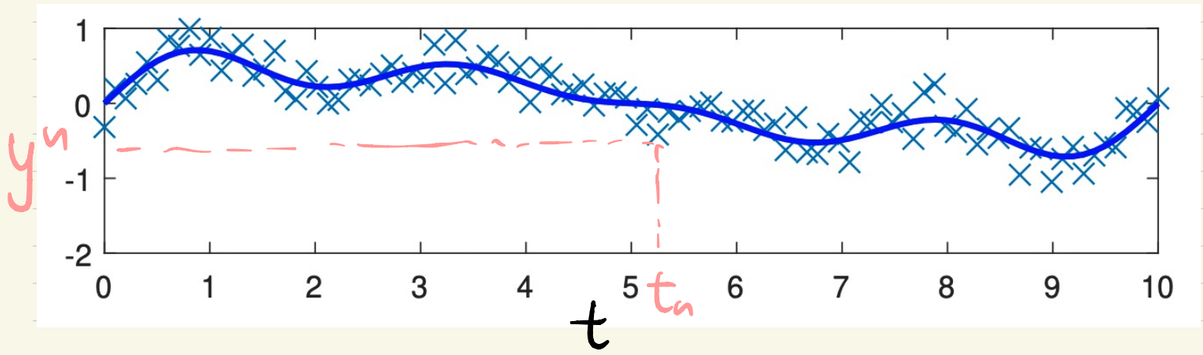
Example 2. Auto Regression

$$y_n = a y_{n-1} + b y_{n-2}$$

$$\vec{\theta} = \begin{bmatrix} a \\ b \end{bmatrix} \qquad A = \begin{bmatrix} y_2 & y_1 \\ y_3 & y_2 \\ \vdots & \vdots \\ y_{N-1} & y_{N-2} \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_3 \\ \vdots \\ y_N \end{bmatrix}$$

$$\vec{y} = A\vec{\theta}$$

Generalized Linear Regression

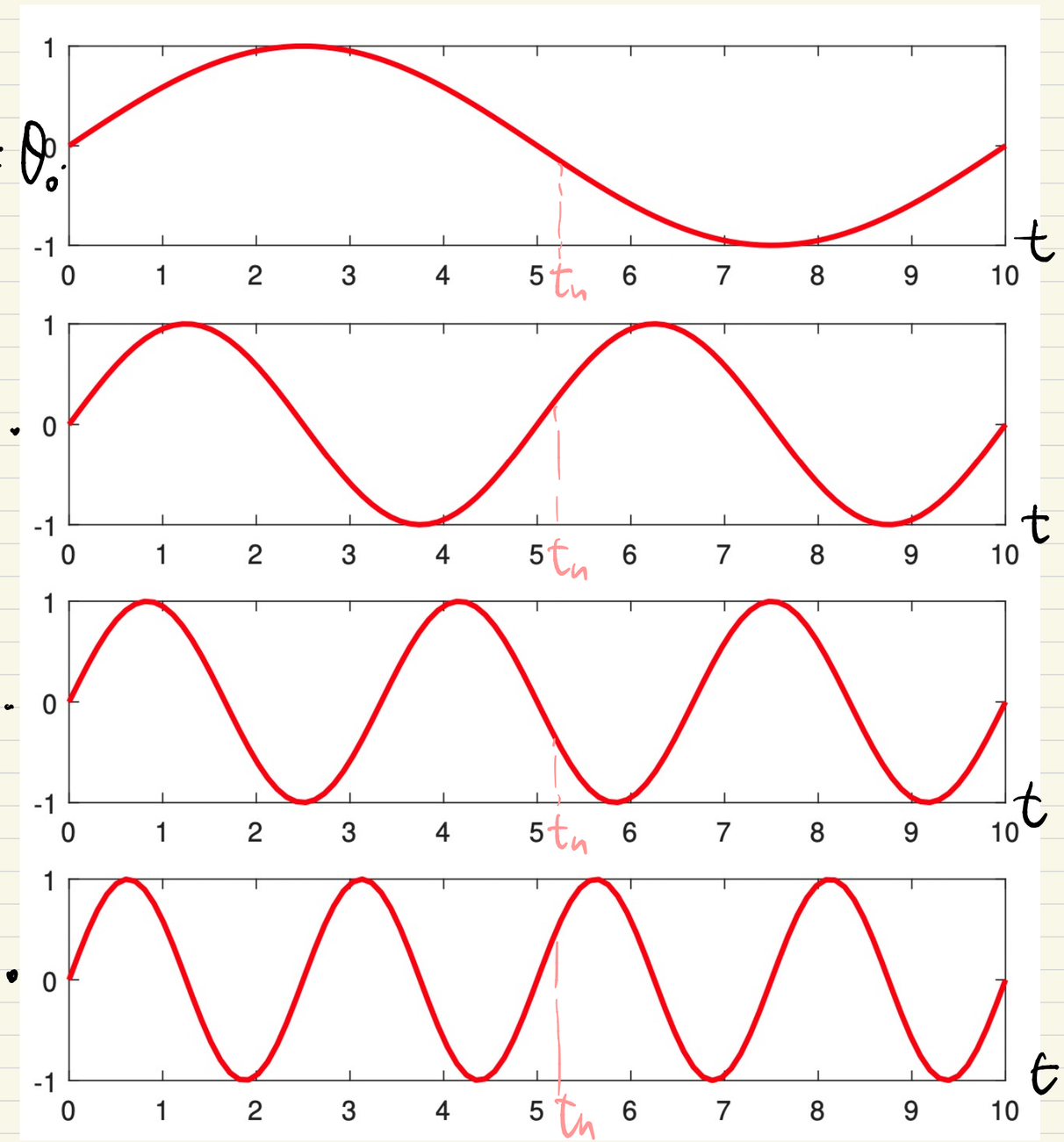


y^n

$$\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{pmatrix} \qquad y^n$$

$$x^n = \begin{pmatrix} \sin(\omega_0 t_n) \\ \sin(2\omega_0 t_n) \\ \vdots \\ \sin(d\omega_0 t_n) \end{pmatrix}$$

$$= \theta_0 \cdot$$



$$+\theta_1 \cdot$$

$$+\theta_2 \cdot$$

$$+\theta_3 \cdot$$

$$\vdots$$

Today's Lecture:

Linear Regression

- Recap: Formulation, Solution

- Examples

Geometric Interpretation of Linear Regression

- Projection

- Minimum-Norm Solution

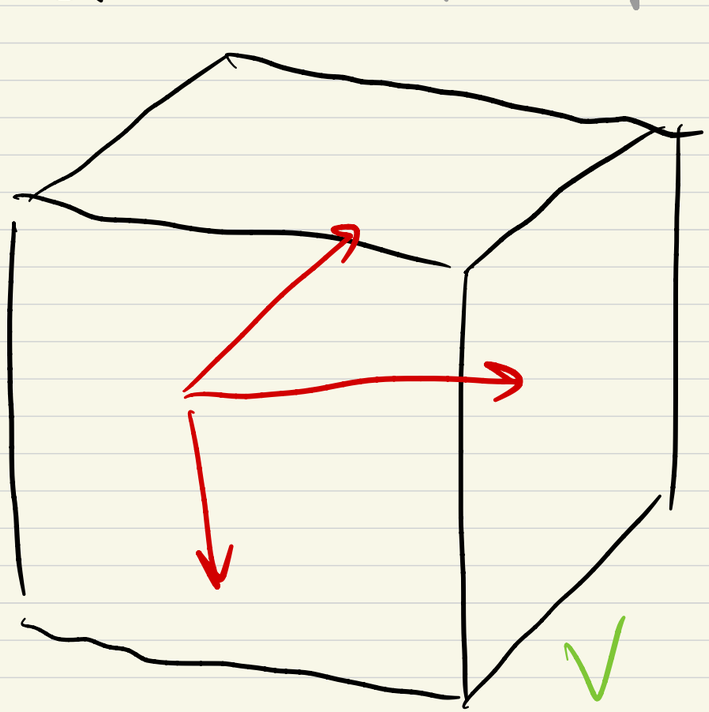
- Pseudo-Inverse

Linear Span (from Linear Algebra)

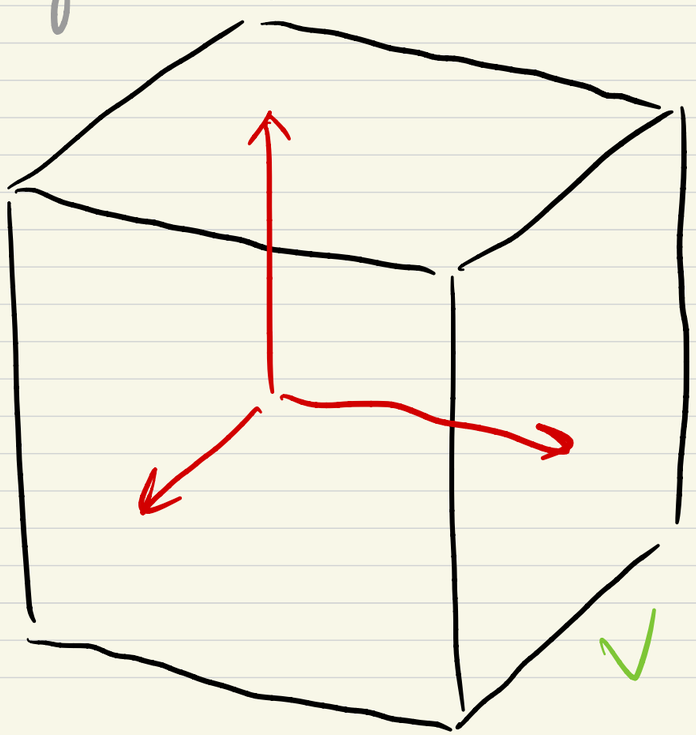
Given a set of vectors $\{\vec{a}_1, \cdots, \vec{a}_d\}$, the **span** is the set of all possible linear combinations of these vectors.

$$\text{span}\{\vec{a}_1, \cdots, \vec{a}_d\} = \{\vec{z} \mid \vec{z} = \sum_{j=1}^{d} \alpha_j \vec{a}_j\}.$$

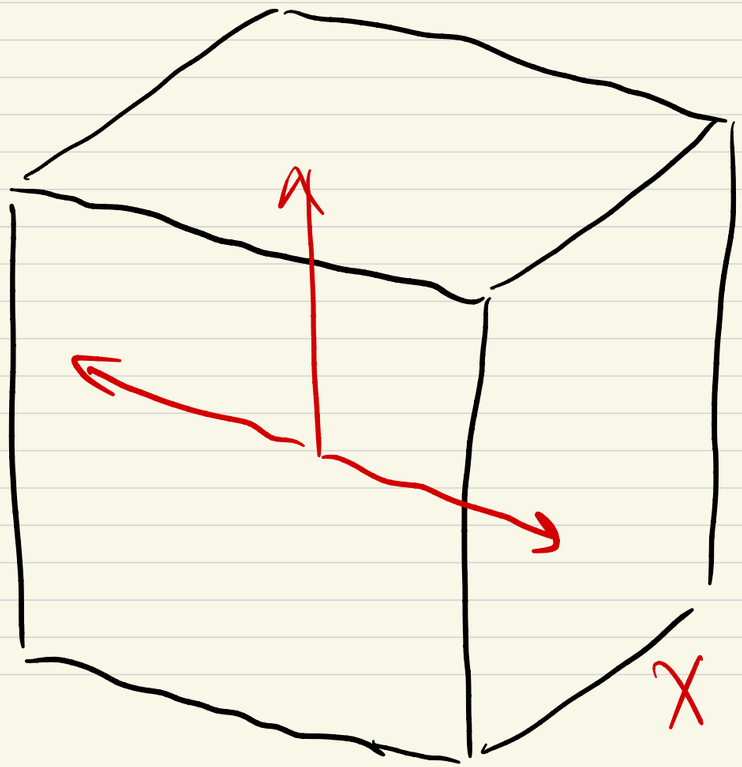
Q: which of the following sets of vectors can span $\mathbb{R}^3$?



✓



✓



✗

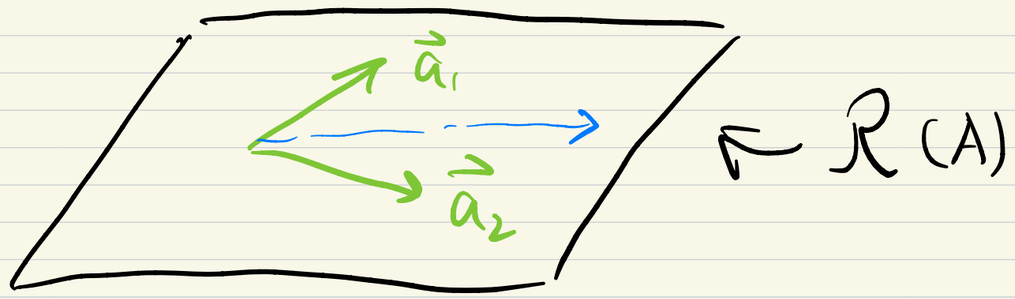
Geometry of Linear Regression

Given $\vec{\theta}$, the product $A\vec{\theta}$ can be viewed as

$$A\vec{\theta} = \begin{bmatrix} | & | & & | \\ \vec{a}_1 & \vec{a}_2 & \cdots & \vec{a}_d \\ | & | & & | \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix} = \sum_{j=1}^{d} \theta_j \vec{a}_j$$

So the set of all possible $A\vec{\theta}$'s is equivalent to span $\{\vec{a}_1, \cdots, \vec{a}_d\}$.

Define the **range** of A as $R(A) = \{\hat{\vec{y}} \mid \hat{\vec{y}} = A\vec{\theta}\}$

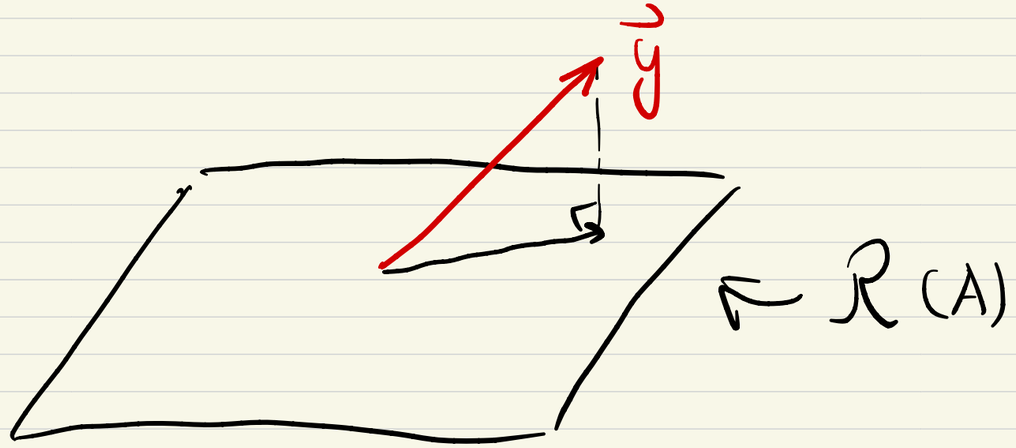


Geometry of Linear Regression

For linear regression problem, $\operatorname{argmin} \|A\vec{\theta} - \vec{y}\|^2$.

Note $\vec{y} \notin R(A)$. as $\vec{y}$ is noisy

Find a vector $\hat{\vec{y}} \in R(A)$. shortest distance to $\vec{y}$



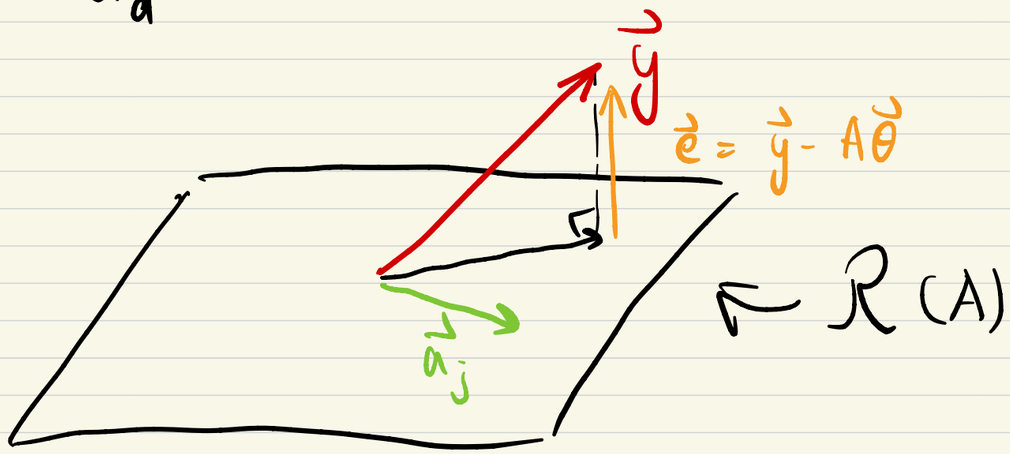
Orthogonality Principle

Error $\vec{e} = \vec{y} - A\vec{\theta}$

For the error to minimize, it must be orthogonal to $R(A)$.

This means $\vec{a}_j^T \vec{e} = 0$ for all $j = 1, \dots, d$, which implies $A^T \vec{e} = 0$.

$$\begin{pmatrix} — \ \vec{a}_1^T \ — \\ \vdots \\ — \ \vec{a}_d^T \ — \end{pmatrix} \vec{e} = 0$$



Normal Equation

- $A^T e = 0 \rightarrow A^T(\vec{y} - A\vec{\theta}) = 0 \rightarrow \boxed{A^T\vec{y} = A^T A\vec{\theta}} \rightarrow \boxed{\vec{\theta} = (A^T A)^{-1} A^T \vec{y}}$

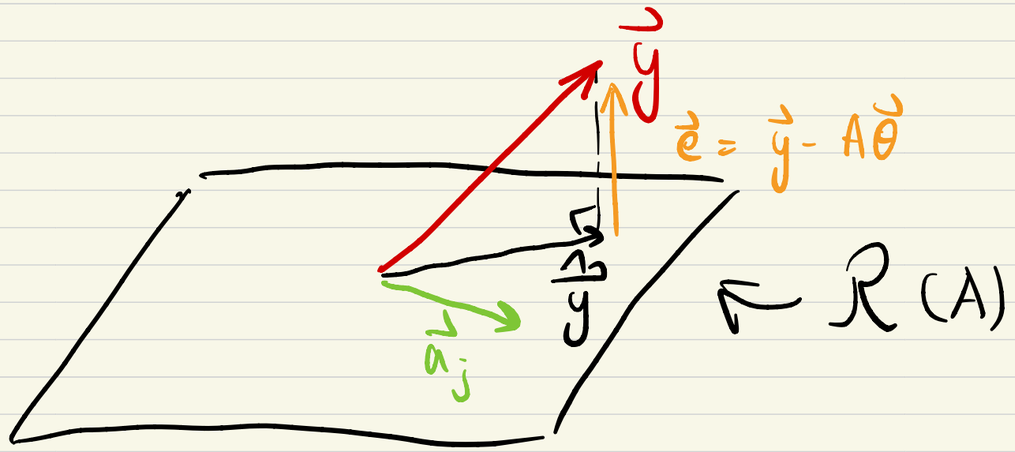
  normal equation — **LS Solution of LR**

- The closest vector $\hat{\vec{y}} = A\vec{\theta} = A(A^T A)^{-1} A\vec{y}$

- **Def:** $P = A(A^T A)^{-1} A$. $\vec{y} \xrightarrow{P} \hat{\vec{y}}$. P is called **projection matrix**.

- Q: $PP = ?\ P$

- $\vec{e} = (I - P)\vec{y}$

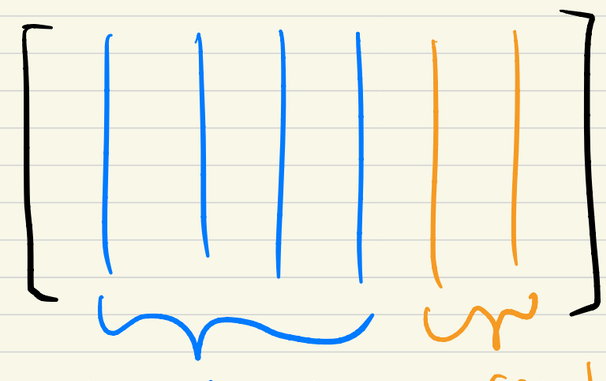


Over-determined and Under-determined Systems

Data matrix

$$A = \underset{N}{\begin{bmatrix} | & | & & | \\ \vec{a}_1 & \vec{a}_2 & \cdots & \vec{a}_d \\ | & | & & | \end{bmatrix}}_{N \times d}$$

d



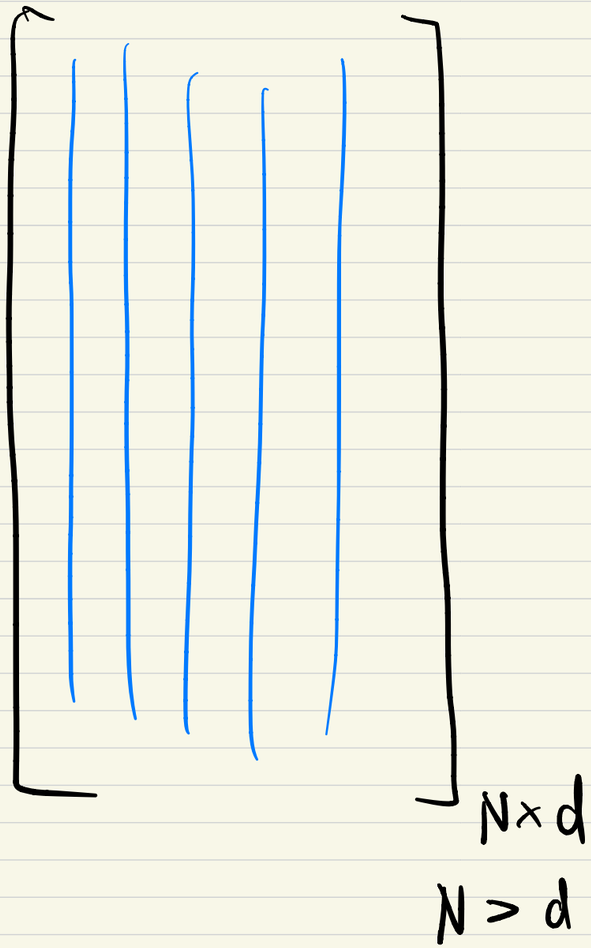
linearly independent

can be linearly represented by $|||$

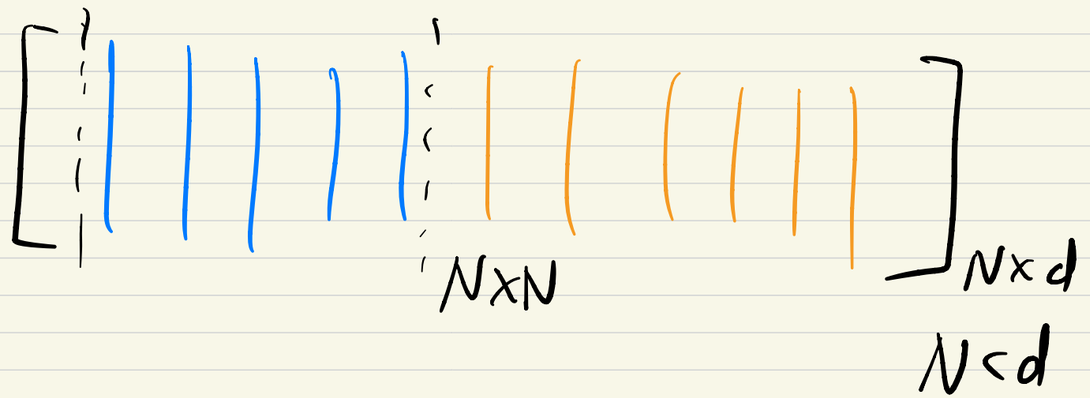
Over-determined and Under-determined Systems

Data matrix A

A full column matrix, $\text{rank}(A) = \min(N, d)$



Overdetermined

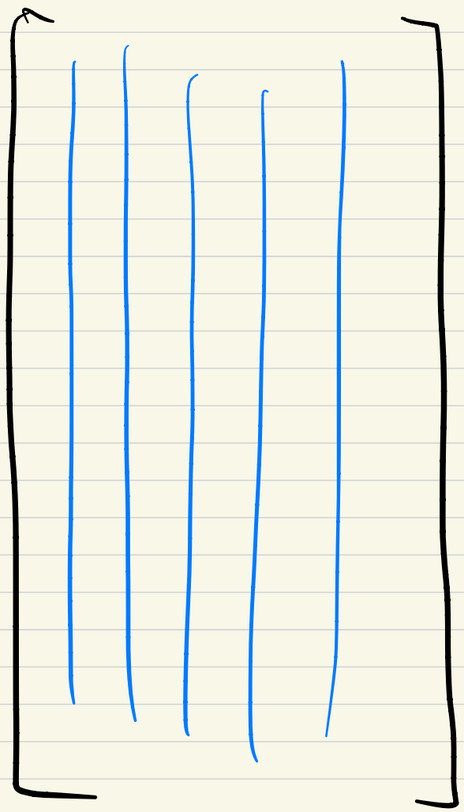


Underdetermined

Over-determined and Under-determined Systems

Overdetermined A

$$\hat{\vec{\theta}} = (A^T A)^{-1} A^T \vec{y}$$



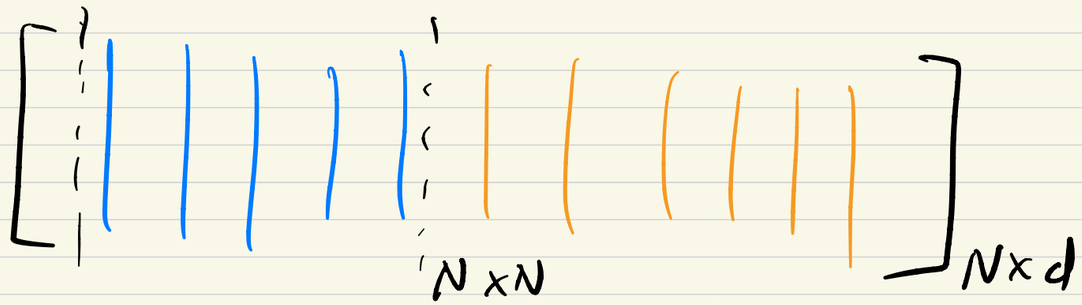
N × d

N > d

Overdetermined

Over-determined and Under-determined Systems

Underdetermined A



$N < d$

- There exists a non-trivial **Null Space**

$$\mathcal{N}(A) = \{ \vec{\theta} \mid A\vec{\theta} = 0 \}$$

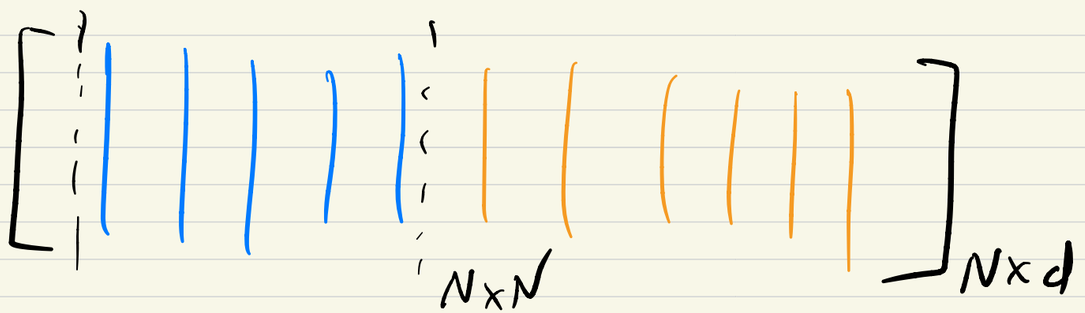
$$\Downarrow$$

if $\hat{\vec{\theta}}$ is a solution to LR,
then $\hat{\vec{\theta}} + \vec{\theta}_0$ is also a solution
as long as $\vec{\theta}_0 \in \mathcal{N}(A)$

why?

Over-determined and Under-determined Systems

Underdetermined A



$N < d$

- Since A is full column rank. $\Rightarrow$
  $R(A) \in \mathbb{R}^d$, $\forall \vec{y}$. $\vec{y} \in R(A)$
- There exists infinitely many $\hat{\vec{\theta}}$.
- s.t $A\hat{\vec{\theta}} = y$.
- Pick one !
- Pick $\hat{\vec{\theta}}$ with **minimum-norm**

$$\hat{\vec{\theta}} = \underset{\vec{\theta}}{\operatorname{argmin}} \|\vec{\theta}\|^2, \text{ s.t. } A\vec{\theta} = \vec{y}$$

Solution: $\hat{\vec{\theta}} = A^T(AA^T)^{-1}\vec{y}$

(solution in Appendix)

Solving the Minimum Norm Problem (Optional)

$$\hat{\vec{\theta}} = \underset{\vec{\theta}}{\operatorname{argmin}} \|\vec{\theta}\|^2 \quad \text{s.t.} \quad A\vec{\theta} = \vec{y}$$

Using Lagrange Multiplier $\vec{\lambda}$

$$\mathcal{L}(\vec{\theta}, \vec{\lambda}) = \|\vec{\theta}\|^2 + \vec{\lambda}^T (A\vec{\theta} - \vec{y})$$

Taking derivative

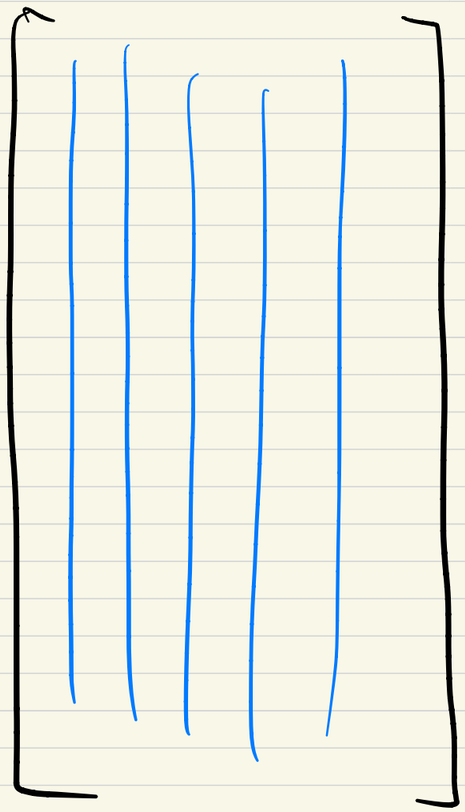
$$\nabla_{\vec{\theta}} \mathcal{L} = 2\vec{\theta} + A^T\vec{\lambda} = 0 \quad \Rightarrow \quad \vec{\theta} = -\frac{1}{2} A^T \vec{\lambda} \ \cdots ①$$

$$\nabla_{\vec{\lambda}} \mathcal{L} = A\vec{\theta} - \vec{y} = 0 \quad \longrightarrow \quad \vec{\lambda} = -2(AA^T)^{-1}\vec{y} \ \cdots ②$$

$$① \cdot ② \Rightarrow \quad \vec{\theta} = A^T(AA^T)^{-1}\vec{y}.$$

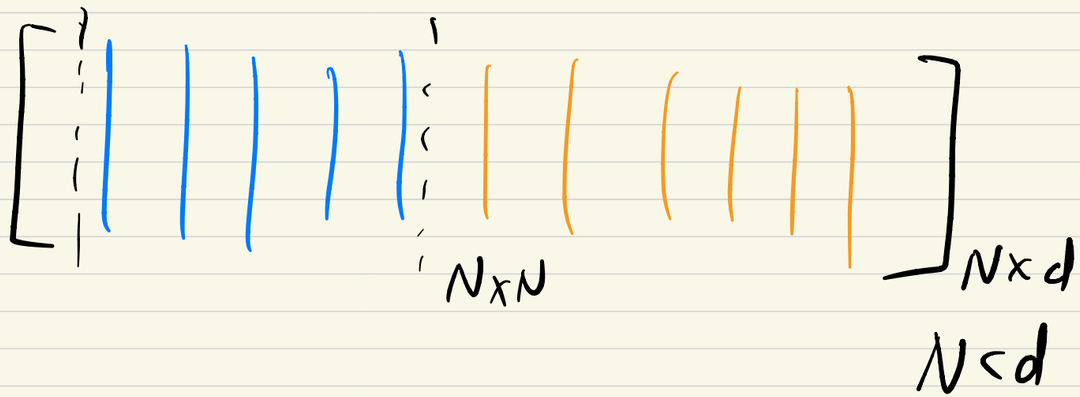
Over-determined and Under-determined Systems

Overdetermined A



$$\hat{\theta} = (A^T A)^{-1} A^T y$$

Underdetermined A



$$\hat{\theta} = A^T (AA^T)^{-1} y$$

Rank Deficiency for Over-determined System

Overdetermined A .

$N \times d$

$N > d$

$\hat{\theta} = (A^T A)^{-1} A^T y$. $A^T A$ not invertible anymore!

Approach ① : regularization . (Next lec)

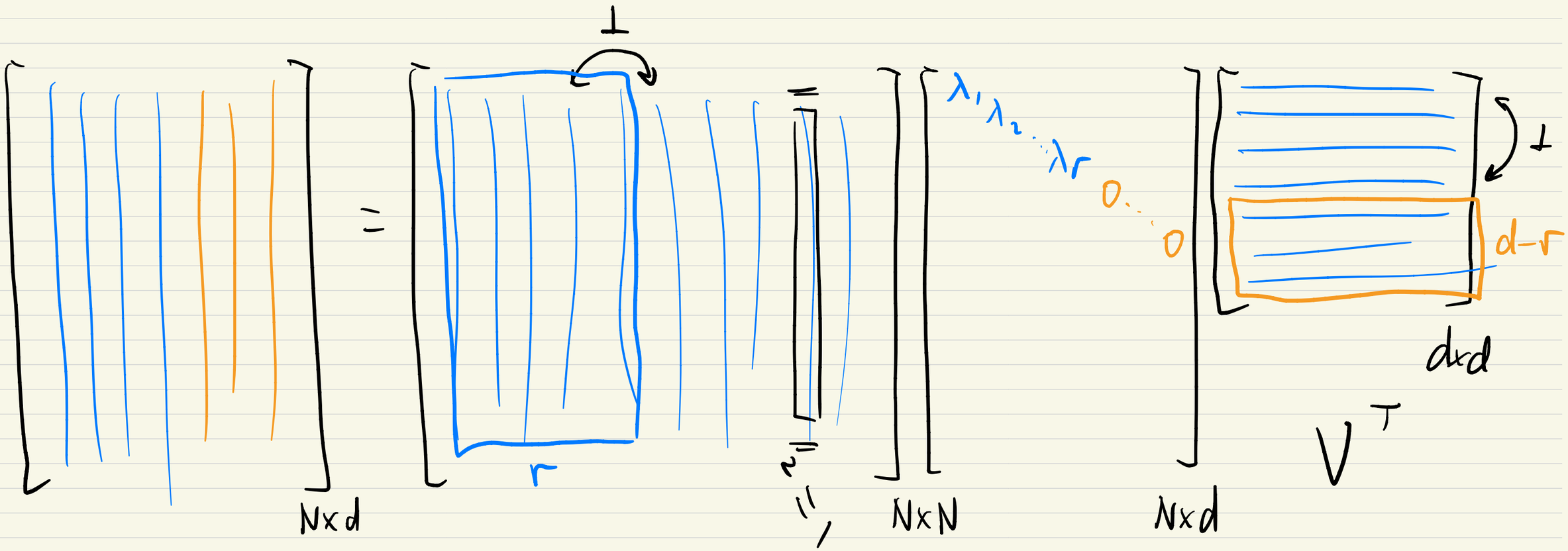
Approach ② : pseudo-inverse

Singular value decomposition

$\hat{\theta} = A^{\dagger} y$

Singular Value Decomposition

https://towardsdatascience.com/understanding-singular-value-decomposition-and-its-application-in-data-science-388a54be95d



A U S V^T

$U^T U = I$

$V^T V = I$

Pseudo-inverse $A^+ = V S^+ U^T$ $S^+ = diag(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \dots, \frac{1}{\lambda_r}, 0, \dots, 0)$

Q: $A^T A =$ $\begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \ddots 0 \end{bmatrix}$ (r)

Reading List

Linear Algebra

- Gilbert Strang, Linear Algebra and Its Applications, 5th Edition.
- Carl Meyer, Matrix Analysis and Applied Linear Algebra, SIAM, 2000.
- Univ. Waterloo Matrix Cookbook.
  https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Linear Regression

- Stanford CS 229 (Note on Linear Algebra)
  http://cs229.stanford.edu/section/cs229-linalg.pdf
- Elements of Statistical Learning (Chapter 3.2)
  https://web.stanford.edu/~hastie/ElemStatLearn/
- Learning from Data (Chapter 3.2)
  https://work.caltech.edu/telecourse