LEC2 DSA5204

$\boxed{\text{Recap}}$

1. Task, Experience, Performance
   ↳ $\boxed{\text{Learning Framework}}$

2. Linear Regression

$$\mathcal{H}_{linear} = \{ f : f(x) = w^T x,\ w \in \mathbb{R}^d \}$$

3. Linear Basis Model

$$\mathcal{H}_{LBM}^{\phi} = \{ f : f(x) = w^T \phi(x),\ w \in \mathbb{R}^D \}$$

$\boxed{\phi : x \in \mathbb{R}^d \longmapsto \phi(x) \in \mathbb{R}^D}$ is fixed feature map

4. Linear Basis Model for classification

$$\mathcal{H} = \{ f : f(x) = \text{softmax}(W \phi(x)),\ W \in \mathbb{R}^{K \times m} \}$$

This Lecture

① Idea of "Adaptive Basis Function"

$$\text{Basis Function} \updownarrow \boxed{\text{feature map}}$$

→ DSA5105 example

$x \in \mathbb{R}^{100}$, then $x = \sum_{i=1}^{100} \alpha_i e_i$

$\boxed{\text{Question}}$: we want to use 10 BASIS to recover x!

⇒ Linear Basis Model ⟺ fix $\{e_1, \dots, e_{10}\}$, learn the factors

⇒ NN ⟺ learn $\{e_{i_1}, \dots, e_{i_{10}}\}$ and corresponding factors

↳ $\boxed{\text{More accurate}}$

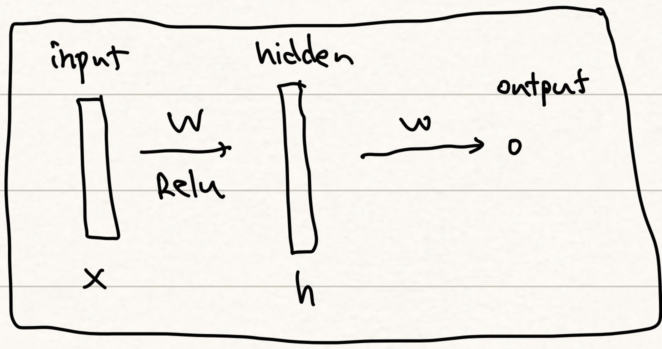
② $\boxed{\text{Neural Network}}$ → Adaptive Linear Basis Model

→ For Shallow NN, then it can be viewed as :

$$f(x) = w^T \cdot \mathrm{Relu}(Wx + b)$$

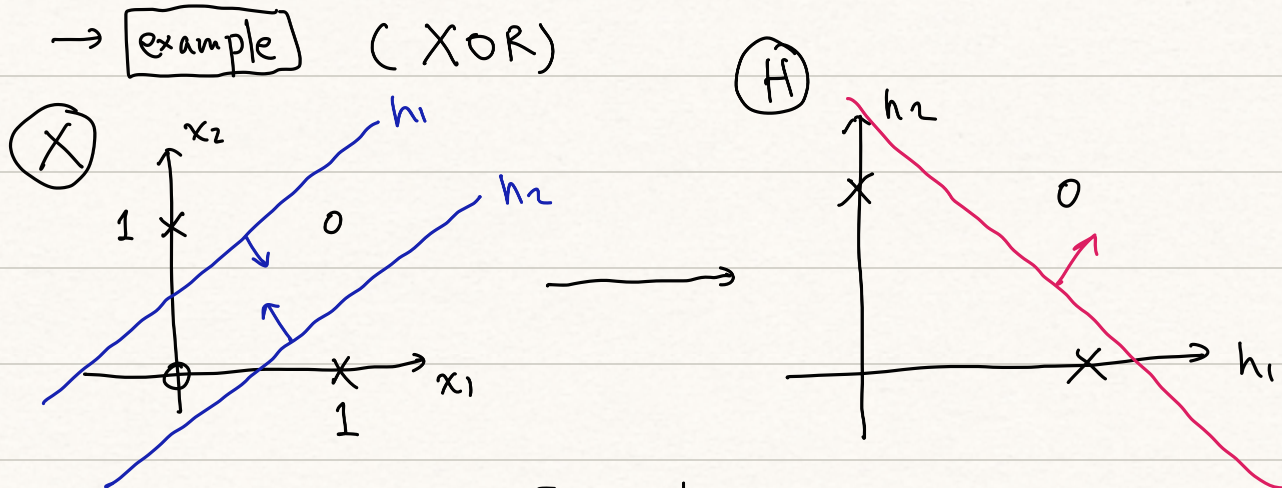
$$:= w^T \phi(x; W, b)$$

$$\phi(x; W, b) := \mathrm{Relu}(Wx + b)$$



architecture

→ $\boxed{\text{example}}$ (XOR)



$$h_1 = \mathrm{sign}\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}^T x + \frac{1}{2} \right)$$

$$h_2 = \mathrm{sign}\left(\begin{pmatrix} -1 \\ 1 \end{pmatrix}^T x + \frac{1}{2} \right)$$

③ NN will not suffer from Curse of Dimensionality

↳ $\boxed{\text{theoretical result}}$

④ <u>Gradient Descent</u> (Optimization Algorithm)

1. For <u>LR</u> case

$$obj = \frac{1}{2} \| Xw - y \|_2^2 \quad \text{(Loss)}$$

$$\rightarrow \quad \hat{w} = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \ \frac{1}{2} \| Xw - y \|_2^2$$

→ <u>GD Framework</u> :

$$\begin{aligned} w^{(k+1)} &= w^{(k)} - \varepsilon \nabla J(w^{(k)}) \\ &= w^{(k)} - \varepsilon (X^T X w^{(k)} - X^T y) \\ &= (I - \varepsilon X^T X) w^{(k)} + \varepsilon X^T y \end{aligned}$$

$$\boxed{\hat{w} := w^{(\infty)} = (X^T X)^{-1} X^T y}$$

2. Convergence Analysis for <u>General GD Framework</u>