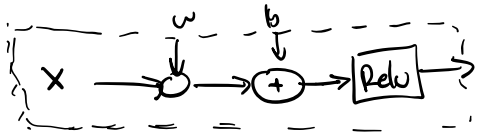
# Basic principles review part 2




$f_{w,b}(x) = \max(0, wx+b)$

Relu fn:  $-\frac{b}{w}$

## Initialization:

- Current folk wisdom:
  - (1) use what worked on a related problem
    - ↳ pretrained network as an initializer
  - (2) Random initializations using gaussians
    - ↳ how to set variance?
      - ↳ Xavier initialization $\sim N(0, ?)$ $? = \frac{1}{d}$ ← fan in of unit; total variance 1
      - ↳ for ReLU: He initialization $? = \frac{2}{d}$ ← our assumption is half of Relu outputs have 0 variance, so we double to get 1
        - b) some Relu elbows "dead" because of county-dist having for out vals ⇒ sets many to 0 ⇒ no $\nabla$

## Regularization

- focus on least squares linear regression

$$\text{Cost} = \|\vec{y} - X\vec{w}\|^2 + \lambda\|\vec{w}\|^2 \qquad \vec{w}^* = (X^TX + \lambda I)^{-1}X^T\vec{y}$$

## Data Augmentation

$$\begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix} \vec{w} = \begin{bmatrix} \vec{y} \\ 0 \end{bmatrix}$$

ols ⇒ gets same as above

*Shown in hw.

## Feature Augmentation

$$\begin{bmatrix} X & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} \vec{\omega} \\ \text{tf} \end{bmatrix} = \begin{bmatrix} \vec{y} \\ \vec{y} \end{bmatrix}$$

OLS: min norm solution w/ moore-penrose pseudo inverse

$$\begin{bmatrix} \vec{y} \\ \vec{y} \end{bmatrix} \Rightarrow \underbrace{\hat{\vec{\omega}} = X^T (XX^T + \lambda I)^{-1} \vec{y}}$$

kernalized ridge / dual view

$$\begin{bmatrix} X^T \\ \sqrt{\lambda} I \end{bmatrix} \left( \begin{bmatrix} X & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} X^T \\ \sqrt{\lambda} I \end{bmatrix} \right)^{-1} \vec{y}$$

Pseudo-inverse