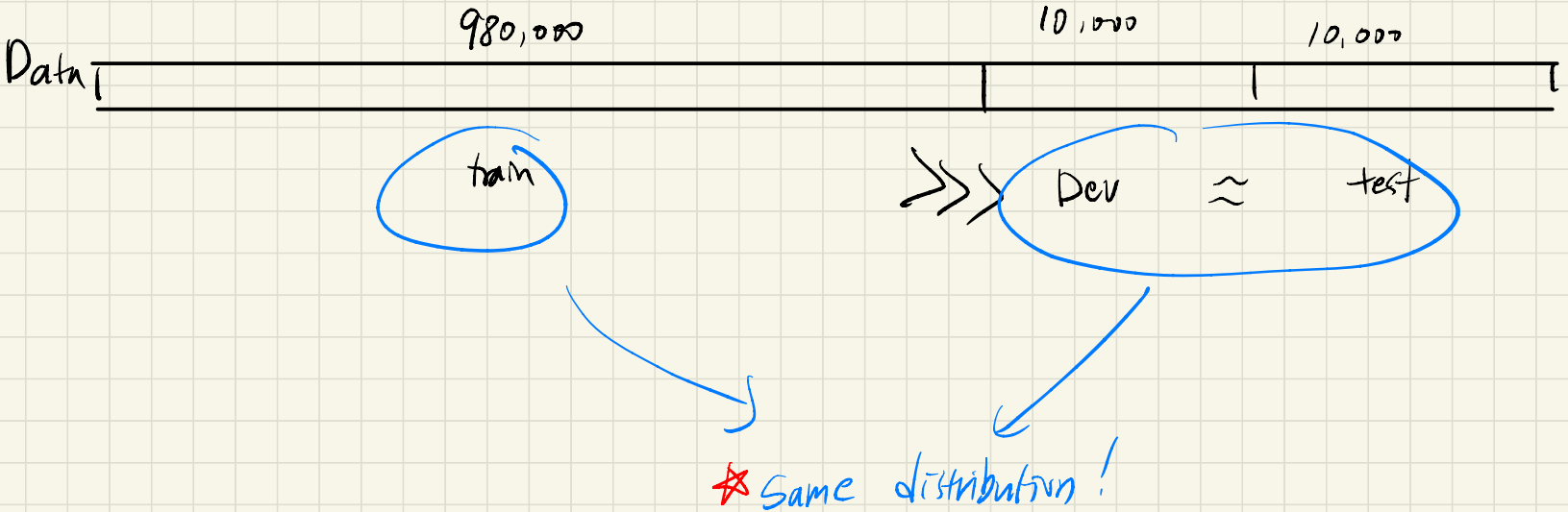
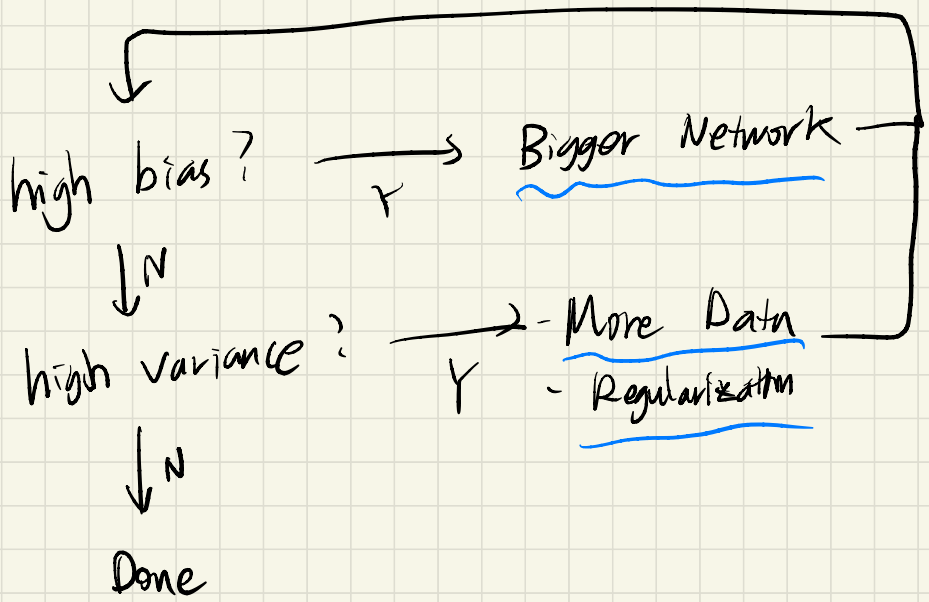
Big data era : 1,000,000 ↑ data



[Basic Recipe]



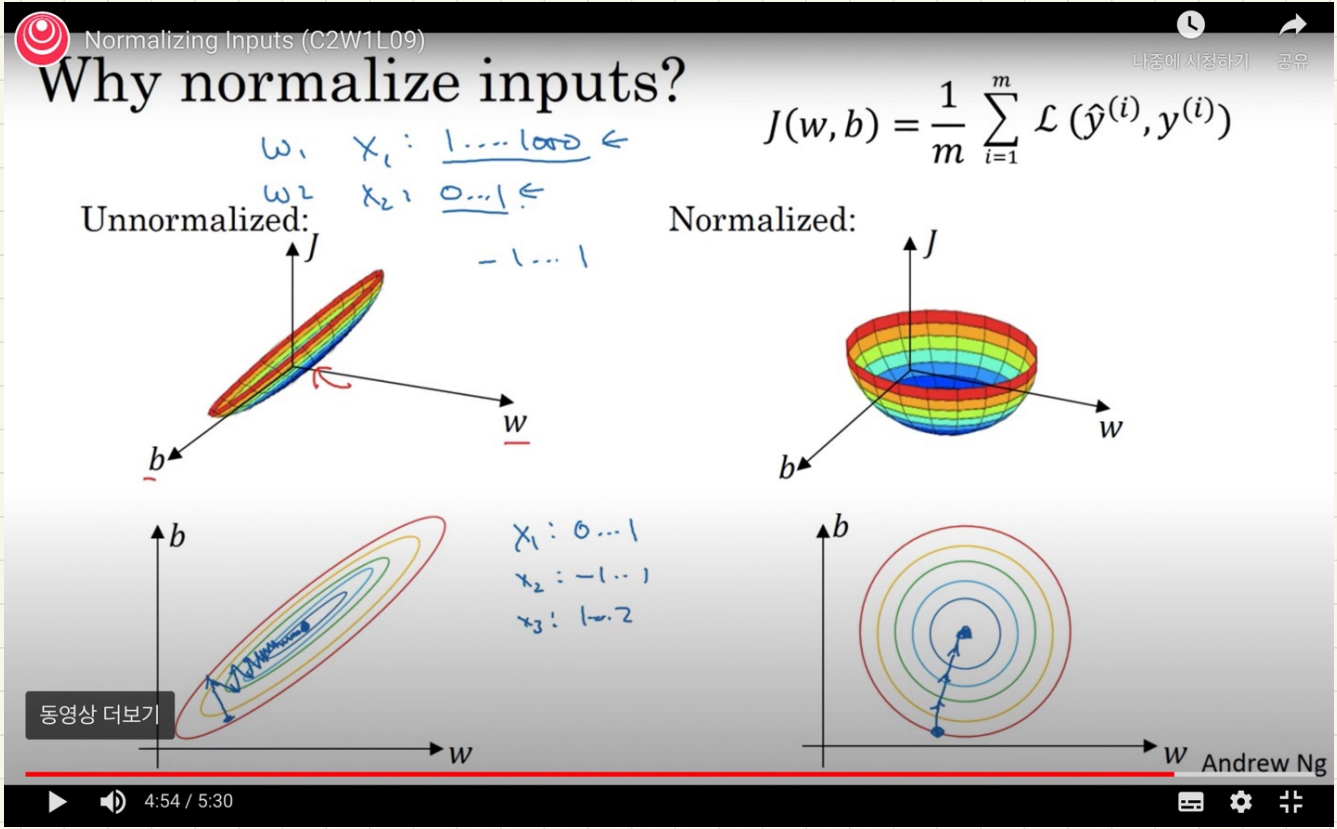
[Regularization]

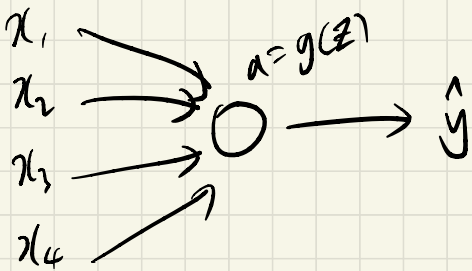
$$J(w,b) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(\hat{y}^{(i)}, y^{(i)}) + \frac{\lambda}{2m}\|w\|_2^2$$

L_2 Regularization: $\|w\|_2^2 = \sum_{j=1}^{n_x} w_j^2 = w^T w$

L_1 " : $\frac{\lambda}{2m}\sum_{i=1}^{n_0}|w| = \frac{\lambda}{2m}\|w\|_1$

[Normalize inputs]





[weights initialization]

$$z = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

initialize randomly

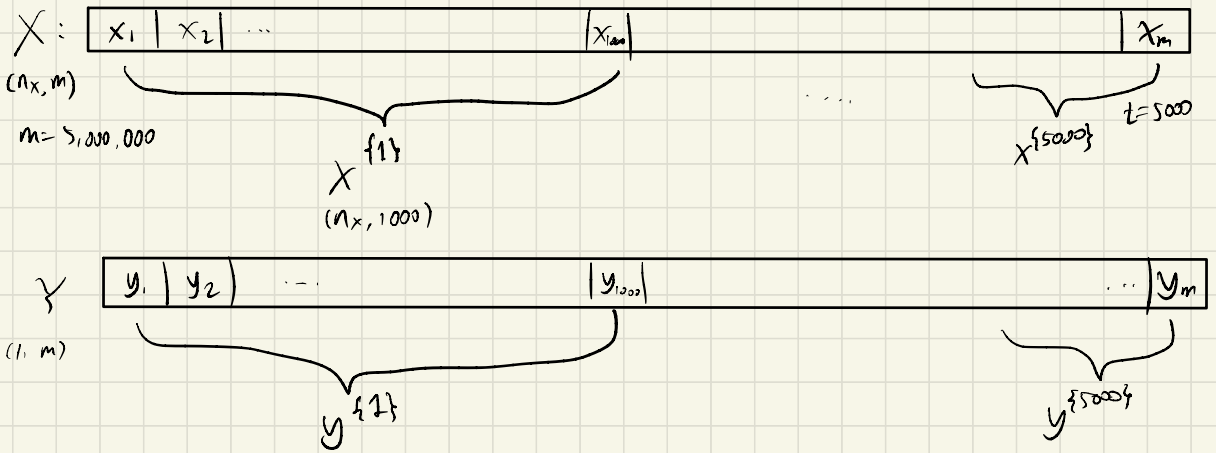
$$w^{[l]} = np.random.randn(shape) * np.sqrt\left(\frac{2}{n^{[l-1]}}\right)$$

when $g^{[l]}(z) = ReLU(z)$

tanh, $\sqrt{\frac{1}{n^{[l-1]}}}$ } Xavier initialization

or $\sqrt{\frac{2}{n^{[l-1]} + n^{[l]}}}$

[mini batch GD]



for $t = 1, \dots, 5000$

1. Forward Prop. on $X^{\{t\}}$

$$Z^{[1]} = W^{[1]} X^{\{t\}} + b^{[1]}$$
$$A^{[1]} = g^{[1]}(Z^{[1]})$$
$$\vdots$$
$$A^{[L]} = g^{[L]}(Z^{[L]})$$

} Vectorize (1,000 example)

2. Compute cost $J^{\{t\}}$

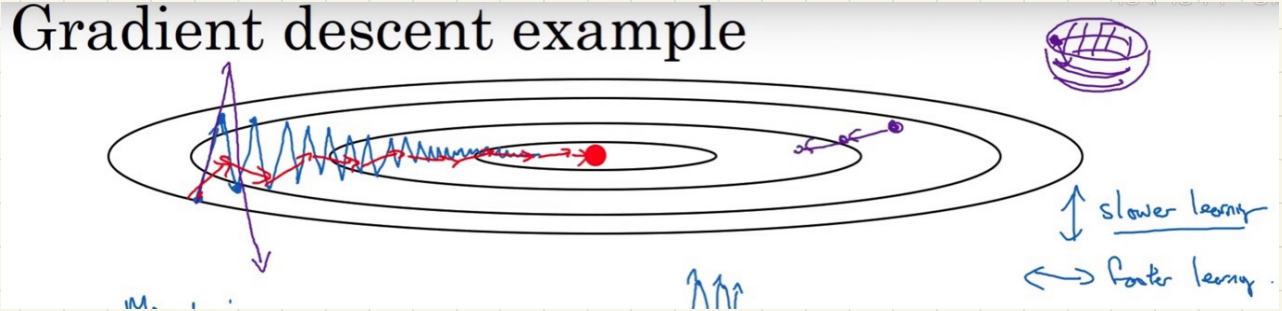
$$= \frac{1}{1000} \sum_i \mathcal{L}(\hat{y}^i, y^i) + \frac{\lambda}{2 \cdot 1000} \sum \|w^{[l]}\|_F^2$$

3. Back prop.

4. $w^{[l]} := w^{[l]} - \alpha dw^{[l]}$, $b^{[l]} := b^{[l]} - \alpha db^{[l]}$

} 1 epoch

[GD with Momentum]



On iteration t:

Compute dW, db on current mini-batch.

$$V_{dW} = \beta V_{dW} + (1-\beta) dW \quad \rightarrow \text{Moving Average of } dW$$

$$V_{db} = \beta V_{db} + (1-\beta) db$$

V_{db}: velocity; db: acceleration

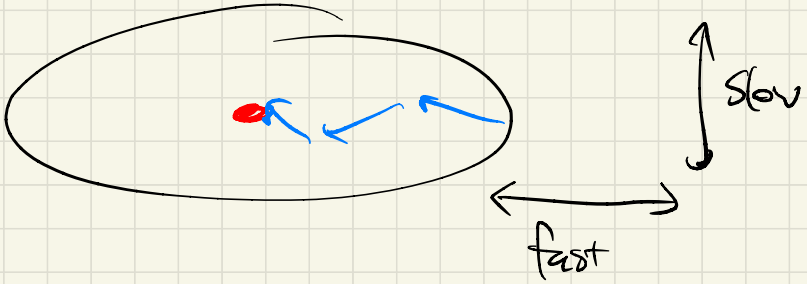
$$W := W - \alpha V_{dW} \qquad b := b - \alpha V_{db}$$

[RMS Prop]

$$S_{dw} = \beta S_{dw} + (1-\beta) dw^2$$

$$S_{db} = \beta S_{db} + (1-\beta) db^2$$

$$w := w - \alpha \frac{dw}{\sqrt{S_{dw}}} \qquad b := b - \alpha \frac{db}{\sqrt{S_{db}}}$$



[Adam]

momentum

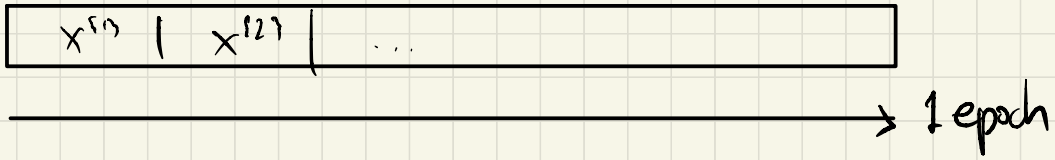
V_{dw} corrected $= V_{dw} / (1-\beta_1^t)$, $V_{db}^{corr} = V_{db} / (1-\beta_1^t)$

$S_{dw}^{corr} = S_{dw} / (1-\beta_2^t)$, $S_{db}^{corr} = S_{db} / (1-\beta_2^t)$

RMS prop

$$w := w - \alpha \frac{V_{dw}^{corr}}{\sqrt{S_{dw}^{corr}} + \varepsilon} \qquad b := b - \alpha \frac{V_{db}^{corr}}{\sqrt{S_{db}^{corr}} + \varepsilon}$$

[Learning Rate Decay]



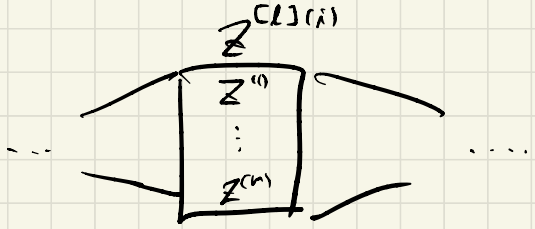
$$\alpha = \frac{1}{1 + \text{decay rate} \times \text{epoch}} \alpha_0$$

$\alpha_0 = 0.2$

decay rate $= 1$

Epoch	α
1	$\frac{1}{1+(1\times1)} \cdot 0.2 = 0.1$
2	$\frac{1}{1+(1\times2)} \cdot 0.2 = 0.067$
3	$= 0.05$
4	$= 0.04$
⋮	⋮

[Batch Normalization]



$$\mu = \frac{1}{m} \sum_i z^{(i)}$$

$$\sigma^2 = \frac{1}{m} \sum_i (z_i - \mu)^2$$

$$z_{norm}^{(i)} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2 + \varepsilon}}$$

$$\tilde{z}^{(i)} = \gamma z_{norm}^{(i)} + \beta$$

$\beta \rightarrow = \mu$

$\gamma \rightarrow = \sqrt{\sigma^2 + \varepsilon}$

then $\tilde{z}^{(i)} = z^{(i)}$

$X^{\{1\}} \xrightarrow{W^{[1]}, b^{[1]}} z^{[1]} \xrightarrow[BN]{\beta^{[1]}, \gamma^{[1]}} \tilde{z}^{[1]} \rightarrow g^{[1]}(\tilde{z}^{[1]}) = a^{[1]} \rightarrow z^{[2]} \cdots$

$X^{\{2\}} \rightarrow \cdots$

$$z^{[l]}_{(n^{[l]}, 1)} = w^{[l]} \cdot a^{[l-1]} + b^{[l]}$$

$z^{[l]}_{norm}$

$$\tilde{z}^{[l]} = \gamma^{[l]} \cdot z^{[l]}_{norm} + \beta^{[l]}$$