1.1.1

$\nabla_{wt} = 2(X_i\hat{w} - t)X_i$

$w_{t+1} \leftarrow w_t - \eta \nabla_{wt} L_i(X_i, w_t)$

$w_{t+1} \leftarrow w_t - 2\eta(X_i\hat{w} - t)X_i$

$\Leftrightarrow w_{t+1} \leftarrow w_t + aX_i \qquad a \in \mathbb{R}$

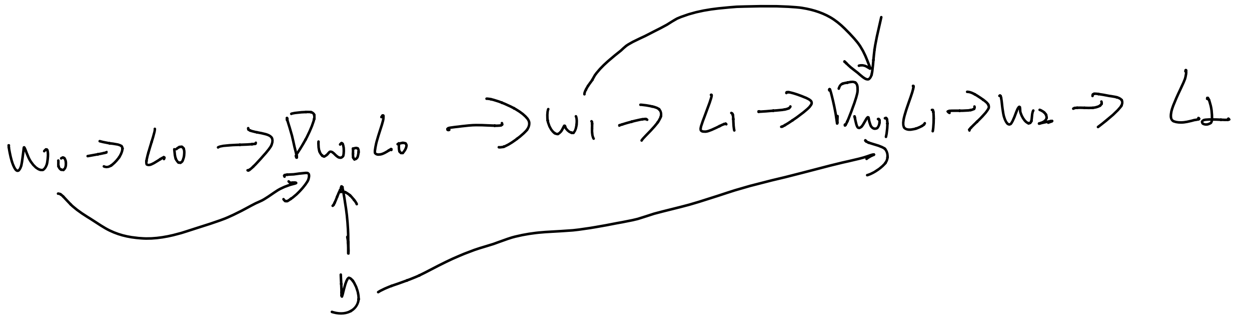
if $w_0 = 0$, then SGD $\hat{w} \in \text{span}\{x_1, x_2 \dots x_n\}$

From HW1 Q3-4, we know GD $w^* \in \text{span}\{x_1, x_2, \dots x_n\}$

If both w^*, $\hat{w}$ converge and share same objective function, then $\hat{w} = w^*$.

2.1 Computation Graph of Learning Rates.

2.1.1



2.1.2

forward: $O(1)$

backward: $O(t)$

2.1.3

The memory cost linearly increases when taking many iterations for using back propagation.

2.2 Learning Learning Rates

2.2.1

$$\frac{\partial L}{\partial w_0} = \frac{2}{n} X^T (X w_0 - t)$$

$$w_1 = w_0 - \frac{2\eta}{n} X^T (X w_0 - t)$$

$$= w_0 - \frac{2\eta}{n} X^T a$$

$$L_1 = \frac{1}{n} \| X w_1 - t \|_2^2$$

$$= \frac{1}{n} \| X (w_0 - \frac{2\eta}{n} X^T a) - t \|_2^2$$

2.2.2

$$\frac{dL_1}{dw_1} = \frac{2}{n} X^T (X w_1 - t) \qquad \frac{dw_1}{d\eta} = -\frac{2}{n} x^T a$$

$$\frac{dL_1}{d\eta} = \frac{dL_1}{dw_1} \cdot \frac{dw_1}{d\eta}$$

$$= \frac{2}{n} \left(-\frac{2}{n} x^T a \right)^T x^T (x w_1 - t)$$

$$= -\frac{4}{n} (x^T a)^T \; x^T (x w_1 - t)$$

$$\frac{d^2 L_1}{dy^2} = \frac{d}{dy}\left(\frac{dL_1}{dy}\right)$$

$$= -\frac{4}{n} (x^T a)^T x^T x \left(-\frac{2}{n} x^T a\right)$$

$$= \frac{8}{n^2} (x^T a)^T x^T x \; x^T a$$

$$= \frac{8}{n^2} \; a^T x x^T \; x x^T a$$

$$= \frac{8}{n^2} \|x x^T a\|_2^2 \; > 0$$

$\Longleftrightarrow$ L_1 is convex

2.2.3

$$\frac{dL_1}{dy} = -\frac{4}{n}\left(X^T a\right)^T X^T (X w_1 - t) = 0$$

$$-\frac{4}{n}\left(X^T a\right)^T X^T X w_1 = -\frac{4}{n}\left(X^T a\right)^T X^T t$$

$$a^T X X^T X\left(w_0 - \frac{2y}{n} X^T a\right) = a^T X X^T t$$

$$a^T X X^T X w_0 - \frac{2y}{n} a^T X X^T X X^T a = a^T X X^T t$$

$$\Rightarrow y = -\frac{n}{2} \cdot \frac{a^T X X^T X w_0 - a^T X X^T t}{\|X X^T a\|_2^2}$$

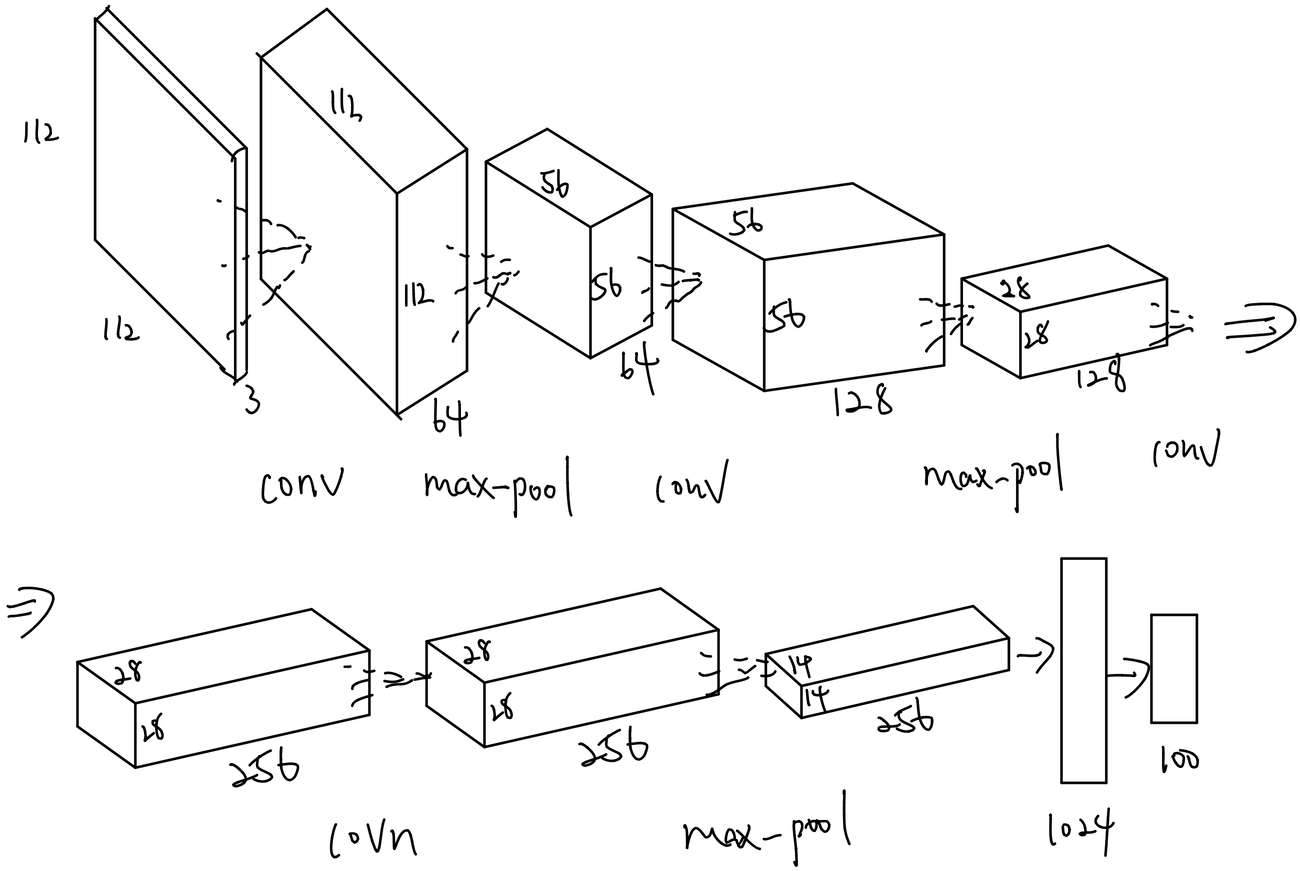
3. CNN

3.1 $$I = \begin{array}{c|ccccc|c} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \qquad J = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}$$

$$I * J = \begin{bmatrix} -1 & 2 & 2 & -2 & 0 \\ -2 & 1 & 0 & 2 & -1 \\ 3 & 0 & 0 & 1 & -1 \\ -2 & 2 & 0 & 2 & -1 \\ 0 & -2 & 3 & -2 & 0 \end{bmatrix}$$

feature: edge detect

3-2 k=3



	# of parameter
C1	3×3×3 ×64+64= 1728 +64 = 1792
C2	3×3×64×128+128 = 73728 + 128 = 73856
C3	3×3×128×256 +256 = 294912+256 = 295168
C4	3×3×256×256 +256 = 589824+256 = 590080
F1	14×14×256×1024 +1024 = 51381248
F2	1024×100 + 100 = 102500

|

total = 52444644