1. Optimization

1-1 Stochastic Gradient Descent

1-1-1

$$L_i(X_i, w_t) = \frac{1}{2}\|X_i w_t - t_i\|_2^2$$

$$= \frac{1}{2}(X_i w_t - t_i)^T (X_i w_t - t_i)$$

$$= \frac{1}{2}(w_t^T X_i^T - t_i^T)(X_i w_t - t_i)$$

$$= \frac{1}{2}(w_t^T X_i^T X_i w_t - w_t^T X_i^T t_i - t_i^T X_i w_t + t_i^T t_i)$$

$$\Rightarrow \frac{\partial L_i}{\partial w_t} = X_i^T X_i w_t - X_i^T t_i$$

Initialization: $w_0 = 0 \Rightarrow \frac{\partial L_i}{\partial w_0} = -X_i^T t_i$

first iteration: $w_1 \leftarrow w_0 - \eta \frac{\partial L_i}{\partial w_0} = 0 + \eta X_i^T t_i = \eta X_i^T t_i = a X_i^T \quad a \in R$

$$\frac{\partial L_i}{\partial w_1} = X_i^T X_i (a X_i^T) - X_i^T t_i$$

second iteration: $w_2 \leftarrow w_1 - \eta \frac{\partial L_i}{\partial w_1} = a X_i^T - \eta (X_i^T X_i (a X_i^T) - X_i^T t_i)$

$$= a X_i^T - \eta (a X_i^T X_i X_i^T - t_i X_i^T)$$

$$= [a - \eta \underbrace{(a X_i X_i^T - t_i)}_{\text{constant}}] X_i^T$$

$$= b X_i^T \quad \text{for } b \in R$$

$\Rightarrow$ By definition of stochastic gradient descent, X_i is sampled from X, therefore, X_i is contained in row space of X;

$\Rightarrow$ As we can see from above gradient updates, each weight update always lies in the span X since it's a linear combination of row vectors within X

$\Rightarrow$ therefore, assume the final weight be $\hat{w} = X^T c$ for $c \in R^n$

$$\Rightarrow X\hat{w} - t = XX^T c - t = 0 \Rightarrow c = (XX^T)^{-1} t$$

$$\Rightarrow \hat{w} = X^T (XX^T)^{-1} t \quad (1)$$

same answer as w^* in HW1

- proof of min-norm:

⇒ still, assume the optimization problem has a solution $\hat{w} = X^T c \in \text{span } X$ and let w_t be any zero-loss solution to the model

⇒ therefore, we have: $Xw_t = t \Rightarrow w_t = X^{-1}t \Rightarrow w_t^T = t^T (X^{-1})^T$

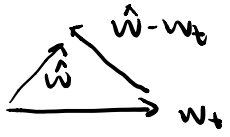
According to ①: $\hat{w}^T = t^T [(XX^T)^{-1}]^T X$

②: $\hat{w}^T \hat{w} = t^T [(XX^T)^{-1}]^T X X^T (XX^T)^{-1} t$

$= t^T (XX^T)^{-1} (XX^T)(XX^T)^{-1} t = t^T (XX^T)^{-1} t.$

③: $w_t^T \hat{w} = t^T (X^{-1})^T X^T (XX^T)^{-1} t = t^T (XX^T)^{-1} t.$

②-③: $(\hat{w} - w_t)^T \hat{w} = 0$



By Pathagorean theorem: $\|w_t\|^2 = \|\hat{w} - w_t\|^2 + \|\hat{w}\|^2 \geq \|\hat{w}\|^2$

⇒ therefore, $\hat{w}$ has the min norm solution of all solutions ⇒ $\hat{w} = w^*$

(1.1.2)

Initialization: $w_0 = 0, \delta_0 = 0 \Rightarrow \frac{\partial L_i}{\partial w_0} = -x_i^T t_i \Rightarrow \delta_1 = -\eta \frac{\partial L_i}{\partial w_0} = \eta x_i^T t_i.$

first iteration: $w_1 \leftarrow w_0 + \delta_1 = 0 + \eta x_i^T t_i = \eta x_i^T t_i = a x_i^T \quad a \in R$

$\frac{\partial L_i}{\partial w_1} = a x_i^T x_i x_i^T - x_i^T t_i, \quad \delta_2 = a\eta x_i^T x_i x_i^T - (\eta - a\eta) x_i^T t_i.$

second iteration: $w_2 \leftarrow w_1 - \eta \frac{\partial L_i}{\partial w_1} = a x_i^T - a\eta x_i^T x_i x_i^T + (\eta - a\eta) x_i^T t_i.$

$= a x_i^T - [a\eta x_i x_i^T - (\eta - a\eta) t_i] x_i^T$

$= [a - \underbrace{a\eta x_i x_i^T + (\eta - a\eta) t_i}_{\text{constant}}] x_i^T$

$= b x_i^T$

⇒ the weight updates still follow the span of X which means momentum does not influence the linearity of SGD update steps, therefore, SGD with momentum will still converge to the min norm solution.

1.2 Adaptive Method

1.2.1 take a counter-example: $X_1 = [2, 1]$, $W_0 = [0, 0]$, $t = 2$, $G_{-1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

$$\frac{\partial L_i}{\partial W_t} = X_i^T X_i W_t - X_i^T t_i$$

Initialization: $\frac{\partial L_i}{\partial W_0} = -X_i^T t_i = \begin{bmatrix} -4 \\ -2 \end{bmatrix}$, $G_0 = \begin{bmatrix} 16 \\ 4 \end{bmatrix}$

first iteration: $W_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} \frac{-4\eta}{\sqrt{16+\epsilon}} \\ \frac{-2\eta}{\sqrt{4+\epsilon}} \end{bmatrix} \approx \begin{bmatrix} \eta \\ \eta \end{bmatrix}$ not a span of X_1

$$\frac{\partial L_i}{\partial W_1} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} [2 \ 1] \begin{bmatrix} \eta \\ \eta \end{bmatrix} - \begin{bmatrix} 2 \\ 1 \end{bmatrix} \times 2 = \begin{bmatrix} 6\eta - 4 \\ 3\eta - 2 \end{bmatrix}$$

$$G_1 = \begin{bmatrix} 16 \\ 4 \end{bmatrix} + \begin{bmatrix} (6\eta-4)^2 \\ (3\eta-2)^2 \end{bmatrix} = \begin{bmatrix} (6\eta-4)^2+16 \\ (3\eta-2)^2+4 \end{bmatrix}$$

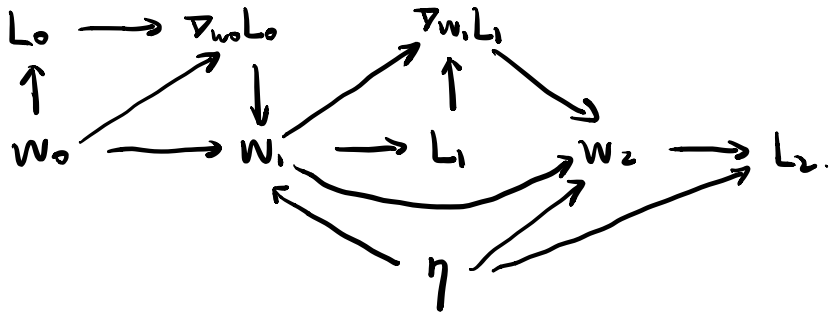
Second iteration: $W_2 = \begin{bmatrix} \eta \\ \eta \end{bmatrix} - \begin{bmatrix} \frac{(6\eta-4)\eta}{\sqrt{(6\eta-4)^2+16+\epsilon}} \\ \frac{(3\eta-2)\eta}{\sqrt{(3\eta-2)^2+4+\epsilon}} \end{bmatrix}$ also not a span of X_1

$\Rightarrow$ therefore, the weight update no longer follows the span of X_1 which may due to the adaptive gradient influences the linearity, and won't follow the minimum norm in the end.

2. Gradient-based hyperparameter optimization

2-1 Computation Graph

2-1-1



2-1-2

Forward: $O(1)$

Backward: $O(t)$ since η goes into every weight update and it needs to store all variables for t iterations.

2-2 Optimal Learning Rates

2-2-1

$$L_0 = \frac{1}{2}\|Xw_0 - t\|_2^2$$

$$= \frac{1}{2}(Xw_0 - t)^T(Xw_0 - t)$$

$$= \frac{1}{2}(w_0^T X^T - t^T)(Xw_0 - t)$$

$$= \frac{1}{2}(w_0^T X^T X w_0 - w_0^T X^T t - t^T X w_0 + t^T t)$$

$$\Rightarrow \quad \frac{\partial L_0}{\partial w_0} = X^T X w_0 - X^T t$$

$$\Rightarrow \quad w_1 \leftarrow w_0 - \eta \frac{\partial L_0}{\partial w_0} = w_0 - \eta X^T(Xw_0 - t)$$

$$= w_0 - \eta X^T a \qquad a \in \mathbb{R}^n$$

$$\Rightarrow \quad L_1 = \frac{1}{2}\|Xw_1 - t\|_2^2$$

$$= \frac{1}{2}\|Xw_0 - \eta X X^T a - t\|_2^2$$

$$= \frac{1}{2}\|a - \eta X X^T a\|^2$$

$$= \frac{1}{2}(a - \eta X X^T a)^T (a - \eta X X^T a)$$

$$= \frac{1}{2}(a^T - \eta a^T X X^T)(a - \eta X X^T a)$$

$$= \frac{1}{2}(a^T a - \eta a^T X X^T a - \eta a^T X X^T a + \eta^2 a^T X X^T X X^T a)$$

$$= \frac{1}{2}(a^T a - 2\eta a^T X X^T a + \eta^2 a^T X X^T X X^T a)$$

$$= \frac{1}{2} a^T (-\eta X X^T + I)^2 a \qquad I \in \mathbb{R}^{n \times n}$$

(2.2.2) $$\frac{\partial L_1}{\partial \eta} = a^T(-\eta X X^T + I)(-X X^T) a$$

$$\frac{\partial L_1^2}{\partial^2 \eta} = a^T(-X X^T)^2 a \Rightarrow \text{convex. since second derivative is non-negative.}$$

(2.2.3) $$\frac{\partial L_1}{\partial \eta} = a^T(-\eta X X^T + I)(-X X^T) a$$

$$\Rightarrow a^T(-\eta X X^T + I)(-X X^T) a = 0.$$

$$(-\eta a^T X X^T + a^T)(-X X^T) a = 0.$$

$$(\eta a^T X X^T X X^T - a^T X X^T) a = 0.$$

$$\eta a^T X X^T X X^T a = a^T X X^T a$$

$$\eta (X X^T a)^2 = (X^T a)^2$$

$$\Rightarrow \eta = \frac{(X^T a)^2}{(X X^T a)^2}$$

2.3 Multiple Inner-loop Iterations

(2.3.1) $$w_1 = w_0 - \eta X^T a \qquad L_1 = \frac{1}{2} a^T (-\eta X X^T + I)^2 a$$

$$\Rightarrow \frac{\partial L}{\partial w_1} = X^T X w_1 - X^T t = X^T X (w_0 - \eta X^T a) - X^T t$$

$$w_2 \leftarrow w_1 - \eta \frac{\partial L}{\partial w_1} = w_1 - \eta X^T X (w_0 - \eta X^T a) + \eta X^T t.$$

finding pattern:

$$= w_0 - \eta X^T a - \eta X^T X (w_0 - \eta X^T a) + \eta X^T t.$$

$$= w_0 - \eta X^T X w_0 - \eta X^T a + \eta^2 X^T X X^T a + \eta X^T t.$$

$$L_2 = \frac{1}{2}\|Xw_2 - t\|_2^2$$

$$= \frac{1}{2}\|Xw_0 - \eta XX^TXw_0 - \eta XX^Ta + \eta^2 XX^TXX^Ta + \eta XX^Tt - t\|_2^2$$

$$= \frac{1}{2}\|a - \eta XX^Ta - \eta XX^Ta + \eta^2 XX^TXX^Ta\|_2^2$$

$$= \frac{1}{2}\|a - 2\eta XX^Ta + \eta^2 XX^TXX^Ta\|_2^2.$$

$$= \frac{1}{2}(a - 2\eta XX^Ta + \eta^2 XX^TXX^Ta)^T(a - 2\eta XX^Ta + \eta^2 XX^TXX^Ta)$$

$$= \frac{1}{2}(a^T - 2\eta a^TXX^T + \eta^2 a^TXX^TXX^T)(a - 2\eta XX^Ta + \eta^2 XX^TXX^Ta)$$

$$= \frac{1}{2}(a^Ta - 2\eta a^TXX^Ta + \eta^2 a^TXX^TXX^Ta - 2\eta a^TXX^Ta + 4\eta^2 a^TXX^TXX^Ta$$

$$-2\eta^3 a^TXX^TXX^TXX^Ta + \eta^2 a^TXX^TXX^Ta - 2\eta^3 a^TXX^TXX^TXX^Ta + \eta^4 a^TXX^TXX^TXX^TXX^Ta)$$

$$= \frac{1}{2}(a^Ta - 4\eta a^TXX^Ta + 6\eta^2 a^TXX^TXX^Ta - 4\eta^3 a^TXX^TXX^TXX^Ta +$$

$$\eta^4 a^TXX^TXX^TXX^TXX^Ta)$$

$$= \frac{1}{2}a^T(I - 4\eta XX^T + 6\eta^2 XX^TXX^T - 4\eta^3 XX^TXX^TXX^T + \eta^4 XX^TXX^TXX^TXX^T)a.$$

$$\Rightarrow \quad L_2 = \frac{1}{2}a^T(-\eta XX^T + I)^4 a.$$

By observation, the constants for each term in the brackets are binomial coefficients a common pattern can be found $\Rightarrow$ $L_t = \frac{1}{2}a^T(-\eta XX^T + I)^{2t}a$.

proof by induction:

- Base: $t=0 \Rightarrow L_0 = \frac{1}{2}a^Ta = \frac{1}{2}\|Xw_0 - t\|^2$
- Assume $L_t = \frac{1}{2}a^T(-\eta XX^T + I)^{2t}a$
- Induction: need to prove $L_{t+1} = \frac{1}{2}a^T(-\eta XX^T + I)^{2(t+1)}a$.

from the assumption: $L_t = \frac{1}{2}\|Xw_t - t\|^2 = \frac{1}{2}a^T(-\eta XX^T + I)^{2t}a$

$$\Rightarrow Xw_t - t = (-\eta XX^T + I)^t a$$

$$\frac{\partial L_t}{\partial w_t} = X^TXw_t - X^Tt = X^T(Xw_t - t) = X^T(-\eta XX^T + I)^t a$$

$$\Rightarrow w_{t+1} \leftarrow w_t - \eta\frac{\partial L_t}{\partial w_t} = w_t - \eta X^T(-\eta XX^T + I)^t a.$$

$$\Rightarrow L_{t+1} = \frac{1}{2}\|Xw_{t+1} - t\|^2$$

$$= \frac{1}{2}\|X(w_t - \eta X^T(-\eta XX^T + I)^t a) - t\|^2$$

$$= \frac{1}{2}\|Xw_t - \eta XX^T(-\eta XX^T + I)^t a - t\|^2$$

$$= \frac{1}{2}\|(-\eta XX^T + I)^t a - \eta XX^T(-\eta XX^T + I)^t a\|^2$$

$$= \frac{1}{2}\|(-\eta XX^T + I)^{t+1} a\|^2 = \frac{1}{2} a^T(-\eta XX^T + I)^{2(t+1)} a$$

QED!

(2.3.2) By spectral decomposition $\Rightarrow XX^T = Q\Lambda Q^T$

$$\Rightarrow L_t = \frac{1}{2} a^T(-\eta Q\Lambda Q^T + I)^{2t} a = \frac{1}{2} a^T Q(-\eta\Lambda + I)^{2t} Q^T a$$

since Λ is a diagonal matrix and L_t will result in a scalar value:

$$\Rightarrow L_t = \frac{1}{2}\sum_{i=1}^{n} c_i^2(-\eta\lambda_i + 1)^{2t}$$ where c_i is the i^{th} element of a^TQ, λ_i is the i^{th} element is the diagonal matrix.

$$\Rightarrow \frac{\partial L_t}{\partial \eta} = \sum_{i=1}^{n} t c_i^2(-\eta\lambda_i + 1)^{2t-1}$$

$$\Rightarrow \frac{\partial^2 L_t}{\partial \eta^2} = \sum_{i=1}^{n} t(2t-1) c_i^2(-\eta\lambda_i + 1)^{2t-1} \geq 0$$ since $\begin{cases} t(2t-1) \geq 0 \\ c_i^2 \geq 0 \\ (-\eta\lambda_i + 1)^{2t-1} \geq 0 \text{ because of the square.} \end{cases}$

3. Convolutional Neural Networks

3.1 Convolutional filters

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

3.2 Size of conv nets

output size:

conv3-64 : $(112-3+2\times1)+1 = 112$

maxpool. : $(112-2)/2+1 = 56$.

conv3-128 : $(56-3+2\times1)+1 = 56$.

maxpool : $(56-2)/2+1 = 28$

conv3-256 : $(28-3+2\times1)+1 = 28$

conv3-256 : $(28-3+2\times1)+1 = 28$.

maxpool : $(28-2)/2+1 = \underline{14}$.

1) Number of parameters.

Conv3-64 : $3\times3\times3\times64+64 = 1792$

conv3-128 : $3\times3\times64\times128+128 = 73856$.

conv3-256 : $3\times3\times128\times256+256 = 296148$

conv3-256 : $3\times3\times256\times256+256 = 590080$.

FC-1024 : $14\times14\times256\times1024+1024 = 51381248$

FC-100 : $1024\times100+100 = 102500$

$\Rightarrow$ Sum = 52446644 parameters

2) Number of neurons.

conv3-64 : $112 \times 112 \times 64 = 802816$

maxpool. : $56 \times 56 \times 64 = 200704$

conv3-128 : $56 \times 56 \times 128 = 401408$

maxpool : $28 \times 28 \times 128 = 100352$

conv3-256 : $28 \times 28 \times 256 = 200704$

conv3-256 : $28 \times 28 \times 256 = 200704$

maxpool : $14 \times 14 \times 256 = 50176$

FC-1024 : 1024

FC-100 : 100

$\Rightarrow$ 1957988 neurons.

3) Number of connections.

conv3-64 : $112 \times 112 \times 3 \times 3 \times 64 \times 3 = 21676032$

maxpool. : $56 \times 56 \times 2 \times 2 \times 64 = 802816$

conv3-128 : $56 \times 56 \times 3 \times 3 \times 128 \times 64 = 231211008$

maxpool : $28 \times 28 \times 2 \times 2 \times 128 = 401408$

conv3-256 : $28 \times 28 \times 3 \times 3 \times 256 \times 128 = 231211008$

conv3-256 : $28 \times 28 \times 3 \times 3 \times 256 \times 256 = 462422016$

maxpool : $14 \times 14 \times 2 \times 2 \times 256 = 200704$

FC-1024 : $14 \times 14 \times 256 \times 1024 = 51380224$

FC-100 : $1024 \times 100 = 102400$

$\Rightarrow$ 999407616 connections.

3.3 Receptive Field

1. Depth of convolutional architecture. If a CNN has deeper or more layers, the RF will be larger since the input are feature maps from previous layers and it gets downsampled every layer;
2. kernel size. A larger kernel size will "extract" more information from the input and make the RF become bigger in size.
3. Stride. Higher stride will increase the size of RF since it indicates the "jump" between each region and bigger jumps will make the entire RF increase its size by certain proportions.