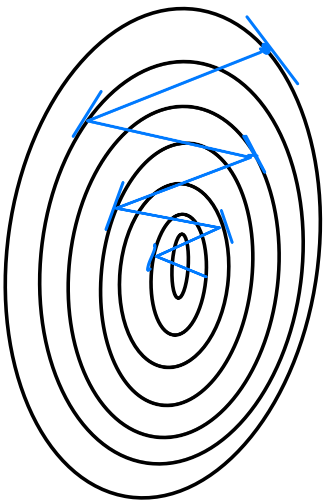
#Optimizer.

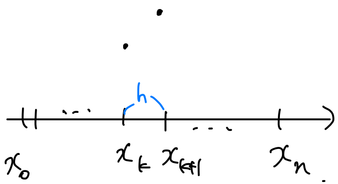
1) Momentum optimizer.



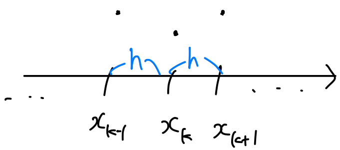
avoid zigzag.

$$\begin{cases} X_{k+1} = X_k - s \cdot \underbrace{Z_k}_{\text{new direction.}} \\ Z_k = \nabla f + \underbrace{\beta \cdot Z_{k-1}}_{\text{memory of privious step (momentum).}} \end{cases}$$

(if $Z_k = \nabla f$: simple gradient descent) check.



$$f'(x_k)? \Rightarrow f'(x_k) = \frac{f(x_{k+1}) - f(x_k)}{h}$$



$$f''(x_k)? \Rightarrow f''(x_k) = \frac{f'(x_k) - f'(x_{k-1})}{h}$$

$$= \frac{1}{h}\left\{ \frac{f(x_{k+1}) - f(x_k)}{h} - \frac{f(x_k) - f(x_{k-1})}{h} \right\}.$$

$$= \frac{f(x_{k+1}) - 2f(x_k) + f(x_{k-1})}{h^2}.$$

$$\left[\begin{array}{l} X_{k+1} = X_k - s \cdot \nabla f \longrightarrow \text{2 step} \approx \text{first order differential equation.} \\ X_{k+1} = X_k - s(\nabla f + \beta \cdot Z_{k-1}) \longrightarrow \text{3 step} \approx \text{Second order differential equation.} \end{array}\right.$$

($k \neq$ time variable but analogy)

$$\left.\begin{aligned} X_{k+1} &= X_k - s\cdot Z_k \\ Z_k &= \nabla f_k + \beta\cdot Z_{k-1} \end{aligned}\right) \longrightarrow \left.\begin{aligned} X_{k+1} &= X_k - s\cdot Z_k \\ Z_{k+1} - \nabla f_k &= \beta\cdot Z_k. \end{aligned}\right) \longrightarrow \begin{aligned} X_{k+1} &= X_k - s\cdot Z_k \\ Z_{k+1} - S X_{k+1} &= \beta\cdot Z_k. \end{aligned}$$

$k \to k+1.$

$H_f.$

$$\boxed{f = \frac{1}{2}\cdot x^T S x.}$$

$$\nabla f = Sx.$$

$$\Rightarrow \begin{bmatrix} 1 & 0 \\ -S & 1 \end{bmatrix} \begin{bmatrix} X \\ Z \end{bmatrix}_{k+1} = \begin{bmatrix} 1 & -s \\ 0 & \beta \end{bmatrix} \begin{bmatrix} X \\ Z \end{bmatrix}_k.$$

step size

$$\left(\begin{aligned} & Sg = \lambda g \longrightarrow X_k = c_k\cdot g. \\ & \qquad\qquad\qquad\quad Z_k = d_k\cdot g. \\ & S X_k = c_k\cdot\lambda\cdot g. \end{aligned} \right).$$

$$\Rightarrow \begin{bmatrix} 1 & 0 \\ -s & 1 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix}_{k+1} = \begin{bmatrix} 1 & -s \\ 0 & \beta \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix}_{k}$$

λ : S의 e.value.

$$\boxed{\frac{\lambda_{min}}{\lambda_{max}}} = \text{condition number}$$

$$\Rightarrow \begin{bmatrix} 1 & 0 \\ -\lambda & 1 \end{bmatrix} \begin{bmatrix} c_{k+1} \\ d_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & -s \\ 0 & \beta \end{bmatrix} \begin{bmatrix} c_{k} \\ d_{k} \end{bmatrix}.$$

Choose s, β.

$$\boxed{X_{k+1} = B \cdot X_k}$$

$$\begin{bmatrix} c_{k+1} \\ d_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \lambda & 1 \end{bmatrix} \begin{bmatrix} 1 & -s \\ 0 & \beta \end{bmatrix} \begin{bmatrix} c_{k} \\ d_{k} \end{bmatrix}.$$

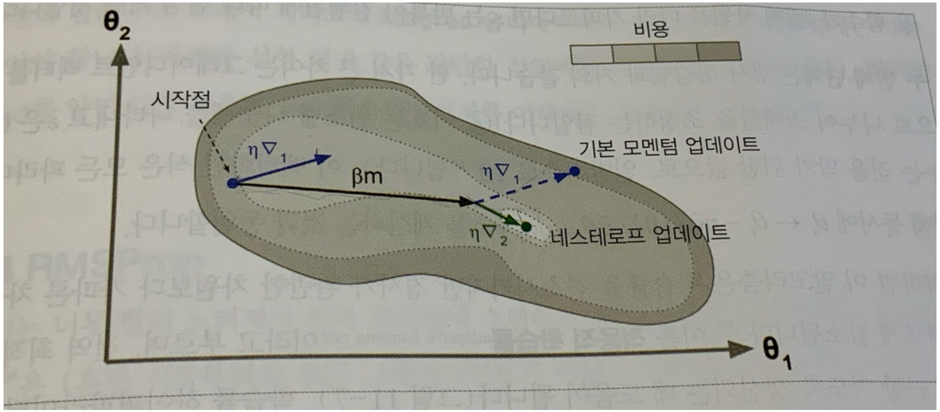
$$= \begin{bmatrix} 1 & -s \\ \lambda & \beta - \lambda s \end{bmatrix} \begin{bmatrix} c_{k} \\ d_{k} \end{bmatrix}$$

$$= B \cdot \begin{bmatrix} c_{k} \\ d_{k} \end{bmatrix}$$

we want to choose s, β

to make e.values of B as small as possible.

2) Nesterov momentum optimization.



$$X_{k+1} = X_k + \beta(x_k - x_{k-1}) - s \cdot \nabla f(x_k + \beta(x_k - x_{k-1}))$$

$$\Rightarrow \begin{bmatrix} c_{k+1} \\ d_{k+1} \end{bmatrix} = \begin{bmatrix} 0 & 1 - s\cdot\lambda \\ -\beta & (1+\beta)(1 - s\cdot\lambda) \end{bmatrix} \begin{bmatrix} c_k \\ d_k \end{bmatrix}$$

$$= R \cdot \begin{bmatrix} c_k \\ d_k \end{bmatrix}$$

3) Adaptive Method.

$$x_{k+1} = x_k - \underline{s} \nabla f.$$

$$x_{k+1} = x_k - s_k \cdot D_k.$$

$$\begin{cases} D_k = D(\nabla L_k, \nabla L_{k-1}, \cdots, \nabla L_0). \\ s_k = s(\nabla L_k, \nabla L_{k-1}, \cdots, \nabla L_0). \end{cases} \Longrightarrow D_k = \nabla L_k.$$

i) AdaGrad.

$$\underline{s \leftarrow s + \nabla_\theta J(\theta) \otimes \nabla_\theta J(\theta)} \longrightarrow \sum_{i=1}^{k} \left(\nabla_\theta J(\theta) \otimes \nabla J_\theta(\theta) \right)$$

$$\theta \leftarrow \theta - \underset{①}{\eta}\, \boxed{\nabla_\theta J(\theta)} \oslash \underset{②}{\sqrt{s+\varepsilon}}.$$

$\boxed{①} \longrightarrow$ 학습률 스케일링

③.

ii) RMSProp.

exponential decay.

$$s \leftarrow \beta s + (1-\beta) \nabla_\theta J(\theta) \otimes \nabla_\theta J(\theta). \longrightarrow (1-\beta) \sum_{i=1}^{k} \underline{\beta^{k-i}} \left(\nabla_\theta J(\theta) \otimes \nabla_\theta J(\theta) \right).$$

$$\theta \leftarrow \theta - \eta \nabla_\theta J(\theta) \oslash \sqrt{s+\varepsilon}.$$

Adam. (Adaptive + Momentum).

$$m \leftarrow \underline{\beta_1}\, m - (\underbrace{1-\beta_1}_{0.1})\, \nabla_\theta J(\theta).$$

$$s \leftarrow \beta_2\, s + (1-\beta_2)\, \nabla_\theta J(\theta) \otimes \nabla_\theta J(\theta).$$

$$\boxed{\hat{m} \leftarrow \frac{m}{1-\beta_1^t}}$$

$$\boxed{\hat{s} \leftarrow \frac{s}{1-\beta_2^t}}$$

0

$$\theta \leftarrow \theta + \eta\, \hat{m} \oslash \sqrt{\hat{s}+\varepsilon}.$$

$\beta_1 = 0.9$. $m_0 = 0$.

Regularization

- l_1 , l_2

- drop out → 규제 + 노드 개수 최적화?

- Max_norm