Deep Learning Basic

Jaewon Kim, Dankook Univ.

Chapter 4-2

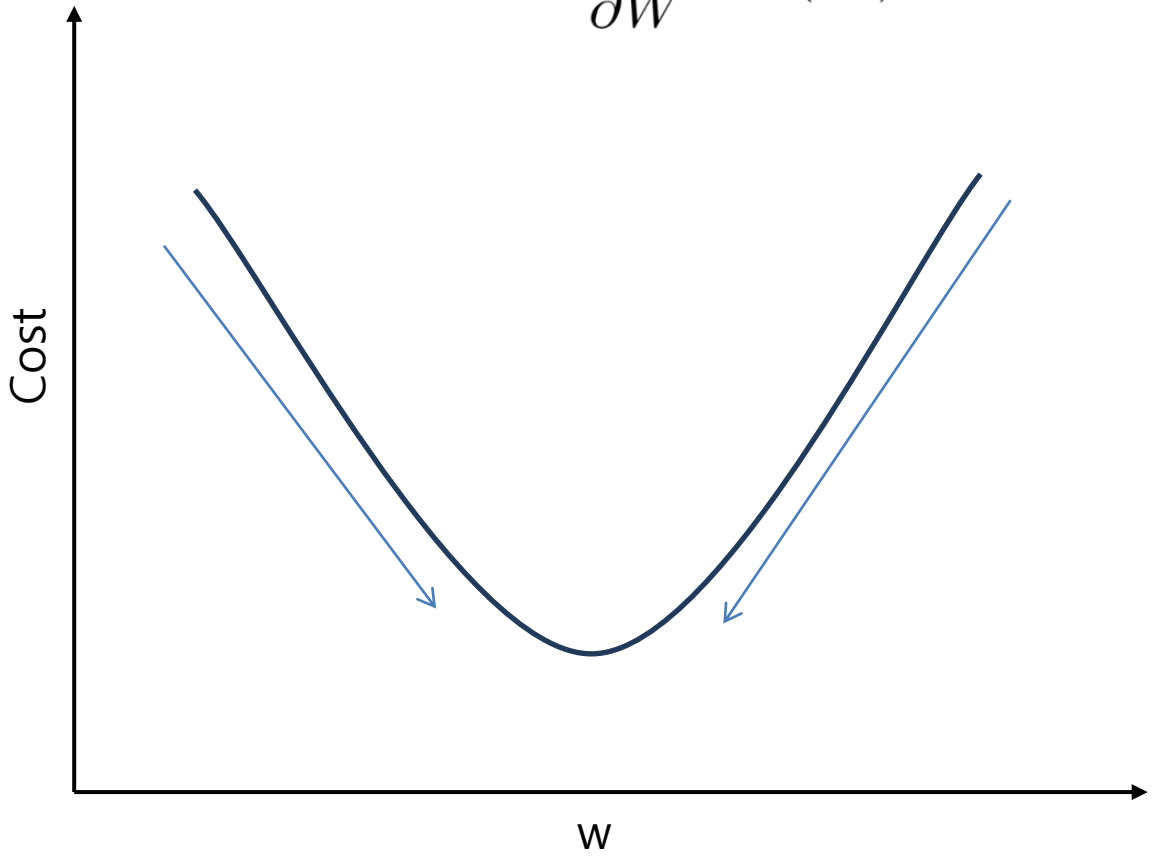
Contents

Part 1. Optimizer

- Gradient Decent Review
- Kinds of Optimizer

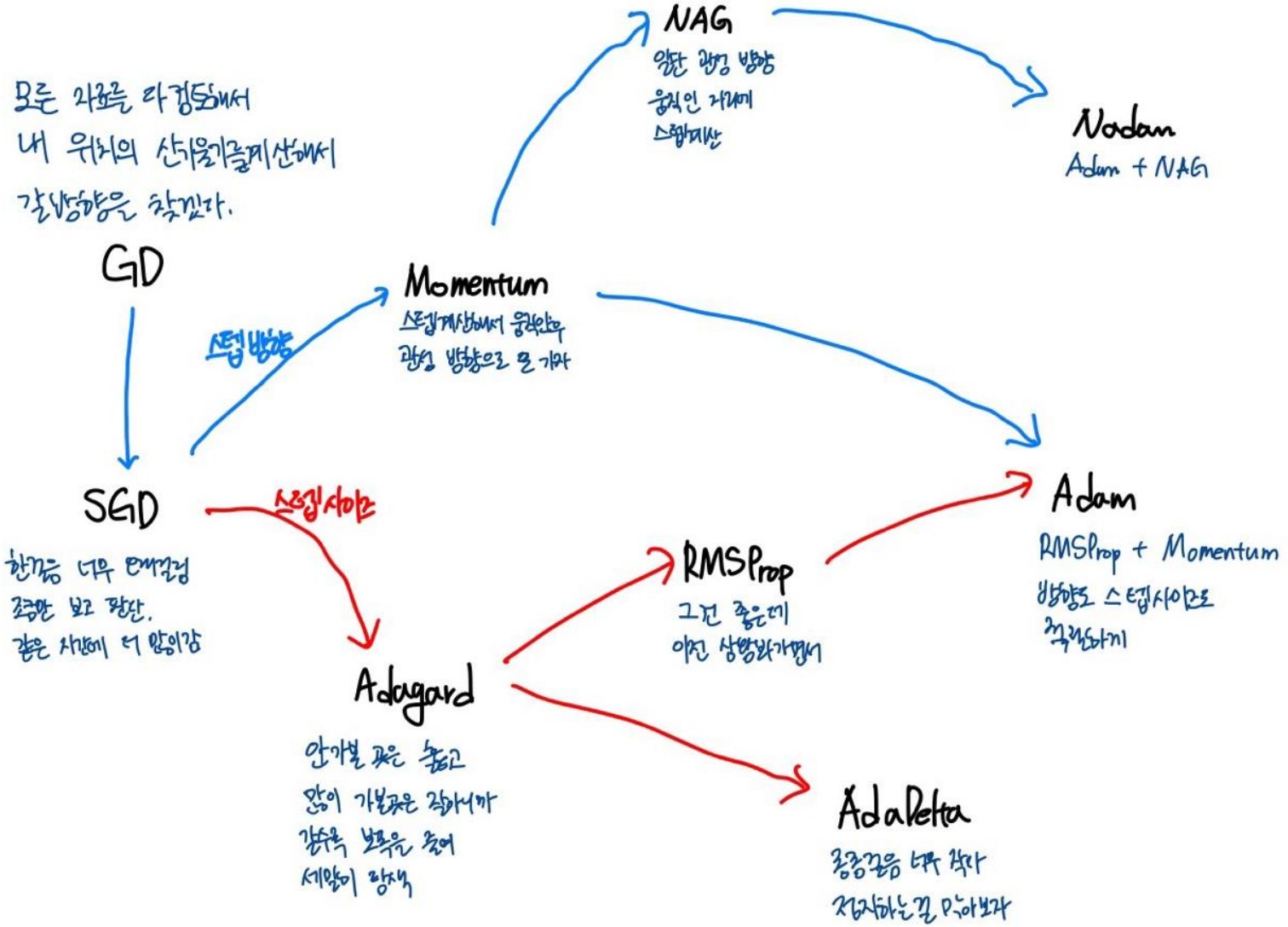
Optimization (Gradient Decent)

$$W := W - \alpha \frac{\partial}{\partial W} cost(W)$$



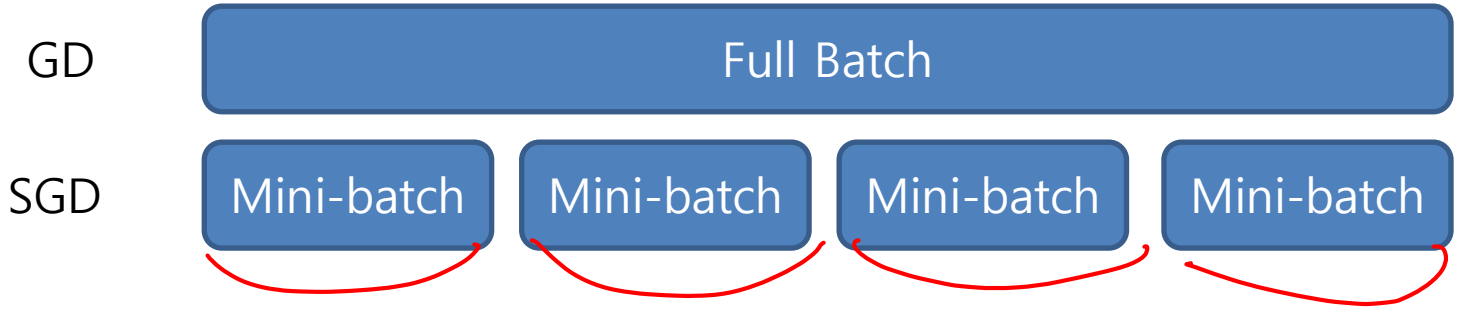
But too slow, local minimum

Kinds of Optimizer



출처 : https://www.slideshare.net/yongho/ss-79607172

Stochastic Gradient Decent (SGD)



batch size로 학습진행

↓

local minimum에 빠지기 쉽다.

Momentum

- ~~SDG~~에 Momentum 개념을 추가
  SGD

가속도계수(0.9, 1이하의값)

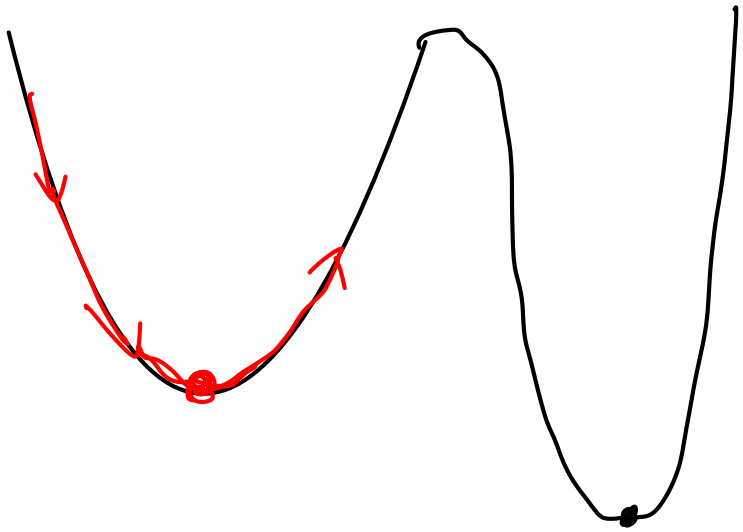
$$v \leftarrow \alpha v - \eta \frac{\partial L}{\partial W}$$

이전의 값.

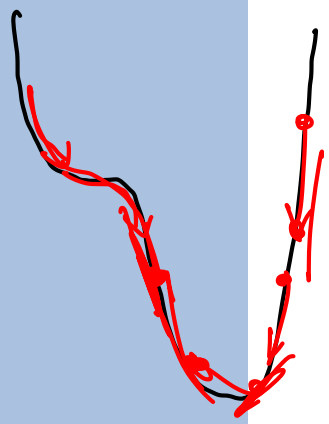
$$W \leftarrow W + v$$

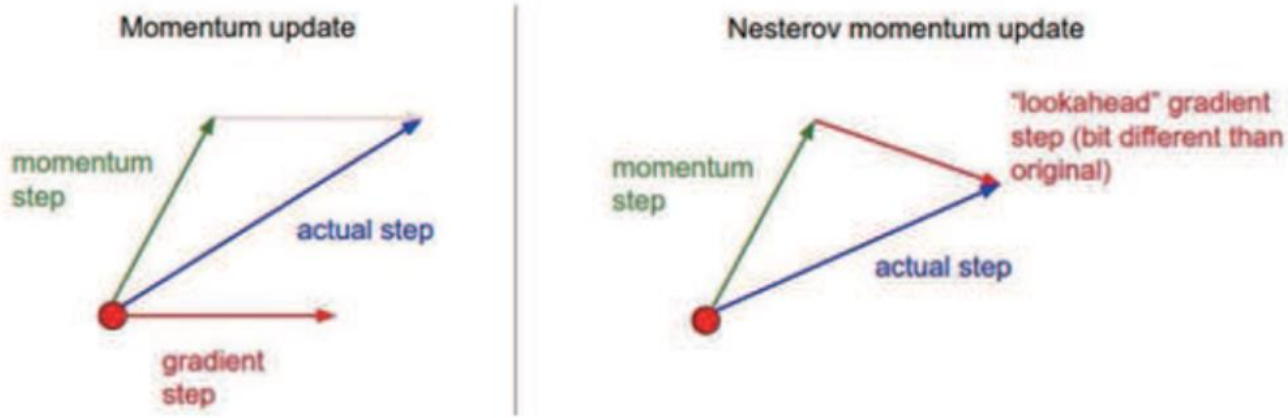
현재 batch x , 이전의 결과도 반영.

관성,,



NAG (Nesterov Accelarated Gradient)





$$\theta = \theta - v_t$$

$$v_t = \gamma v_{t-1} + \eta \nabla_\theta J(\theta - \gamma v_{t-1})$$

Adagrad (Adaptive Gradient)

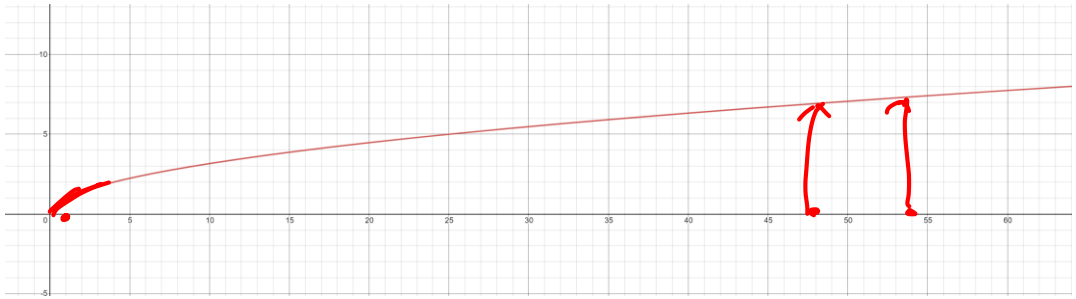
$$\theta_{t+1} = \theta - \frac{\eta}{\sqrt{G_t + \epsilon}} \cdot \nabla_\theta J(\theta_t)$$

$$G_t = G_{t-1} + (\nabla_\theta J(\theta_t))^2$$

↳ 얼마나 많이 변해왔는가.

- 크게 변했던 w 학습률↓.
- 작게 " w " ↑

$y = \sqrt{x}$



RMSProp

$$h_i \leftarrow \rho h_{i-1} + (1-\rho)\frac{\partial L_i}{\partial W} \odot \frac{\partial L_i}{\partial W}$$

decay constant (0~1사이값)

1. 최신기울기를 더 반영.

2. ρ hyperparameter 추가.

AdaDelta

$$\theta_{t+1} = \theta_t - \Delta_\theta$$

$$\Delta_\theta = \frac{\sqrt{s+\epsilon}}{\sqrt{G+\epsilon}} \cdot \nabla_\theta J(\theta_t)$$

$$s_{t+1} = \gamma s_t + (1-\gamma)\Delta_\theta$$

$$G_{t+1} = \gamma G_t + (1-\gamma)(\nabla_\theta J(\theta_t))^2$$

파라미터가
얼마나 변하는지
고려,

Adam

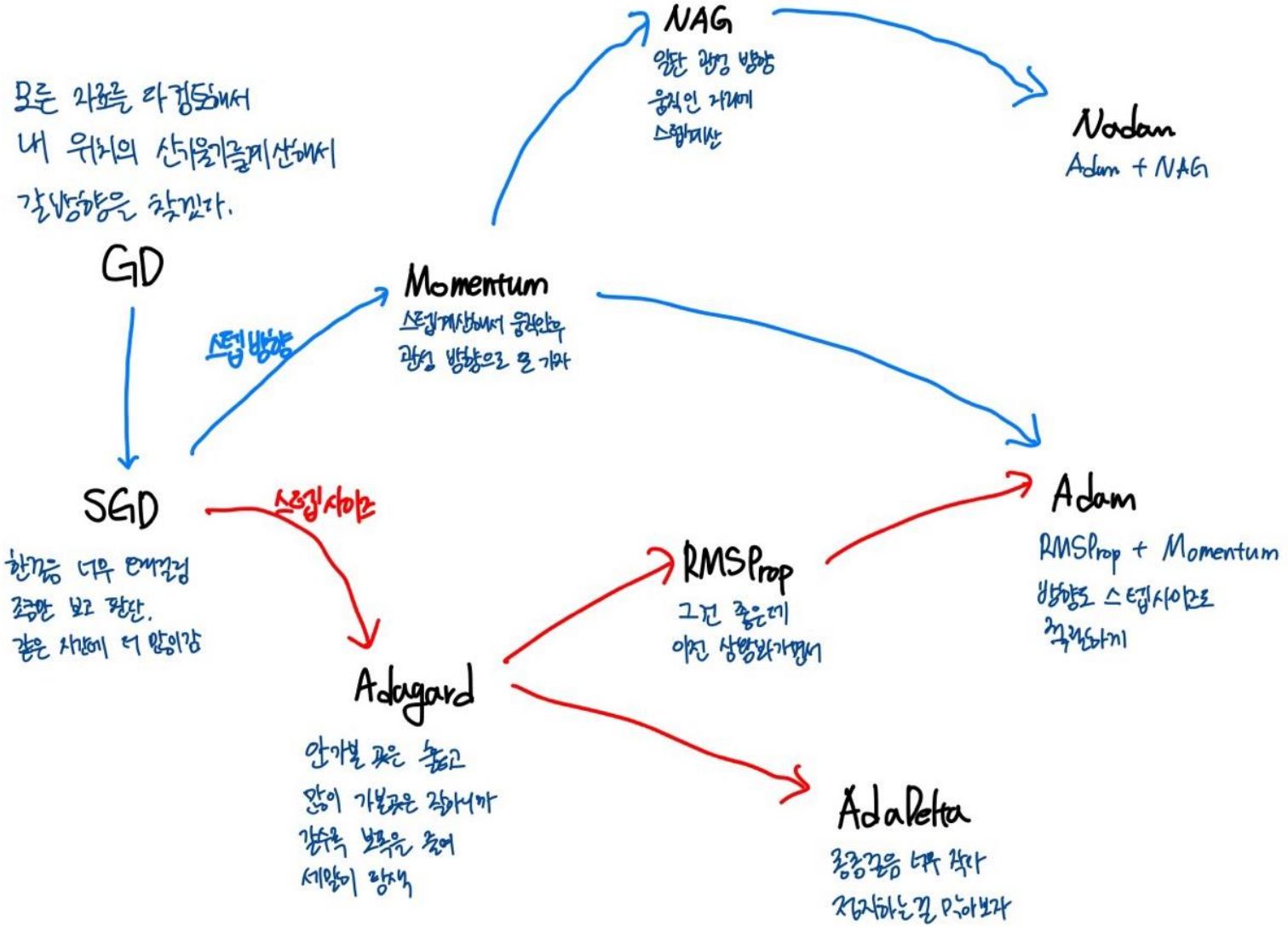
(momentum) $$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$

(adapive lr) $$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \qquad \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t$$

Kinds of Optimizer



출처 : https://www.slideshare.net/yongho/ss-79607172

Optimizer

https://gomguard.tistory.com/187

Thank you..!!!