Optimizer

Explorer

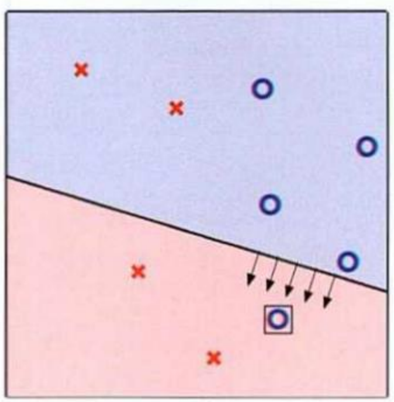
- Fining deepest valley in the world
  - Without map
  - With blindfold

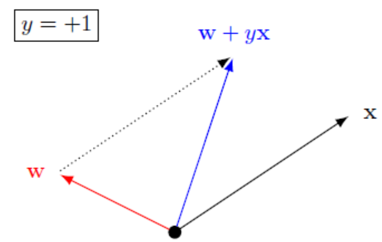
Prof. Hyerim Bae (hrbae@pusan.ac.kr)

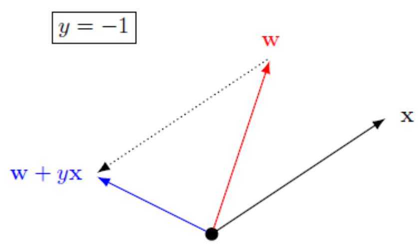
How to make ‘P’ learn

- PLA (Perceptron Learning Algorithm)

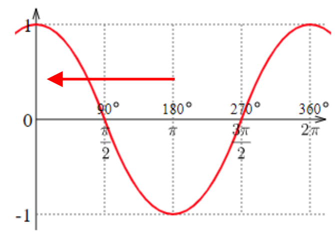
$$\mathbf{w}(t+1) = \mathbf{w}(t) + y(t)\mathbf{x}(t).$$







$w^T x$



The weight update rule in (1.3) has the nice interpretation that it moves in the direction of classifying $\mathbf{x}(t)$ correctly.

(a) Show that $y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t) < 0$. *[Hint: $\mathbf{x}(t)$ is misclassified by $\mathbf{w}(t)$.]*

(b) Show that $y(t)\mathbf{w}^{\mathrm{T}}(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^{\mathrm{T}}(t)\mathbf{x}(t)$. *[Hint: Use (1.3).]*

(c) As far as classifying $\mathbf{x}(t)$ is concerned, argue that the move from $\mathbf{w}(t)$ to $\mathbf{w}(t+1)$ is a move ‘in the right direction’.

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

What is Learning?

- Finding f such that
  - $\hat{y} = y \ (f(x) = y)$
    - Minimize prediction error (Loss Function L)
- Finding f means
  - In regression

- Elements of learning
  - Algorithm
    - Define the process that is used for learning
    - Transform input data into a particular form of useful output
  - Target function
    - The product of learning
  - Training

$$\mathbf{W} \leftarrow \mathbf{W} + c(d - f)\mathbf{X}$$

Weigh update	=	Direction reducing err. **(decent)**	x	Size of one step **(learning rate)**	x	slope **(gradient)**
$-\eta\nabla_{\theta}J(\theta)$		$-$		η		$\nabla_{\theta}J(\theta)$

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

The Widrow-Hoff Procedure

- Weight update procedure:
  - Using $f = s = \mathbf{W} \cdot \mathbf{X}$
  - Data labeled $1 \to 1$, Data labeled $0 \to -1$
- Gradient: if f =s,

$$\frac{\partial \varepsilon}{\partial \mathbf{W}} = -2(d - f)\frac{\partial f}{\partial s}\mathbf{X} = -2(d - f)\mathbf{X}$$

- New weight vector

$$\mathbf{W} \leftarrow \mathbf{W} + c(d - f)\mathbf{X}$$

- Widrow-Hoff (delta) rule
  - $(d - f) > 0 \to$ increasing $s \to$ decreasing $(d - f)$
  - $(d - f) < 0 \to$ decreasing $s \to$ increasing $(d - f)$

$w \uparrow, \quad s = x \cdot w \quad f \uparrow$

$\Rightarrow$ $|d-f|$ decreases!

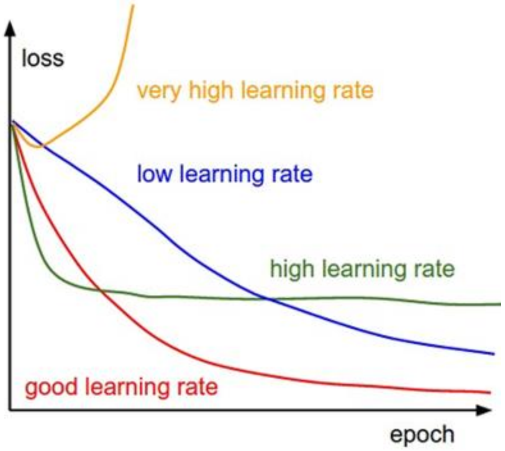
Prof. Hyerim Bae (hrbae@pusan.ac.kr)

Optimizer

- ML Optimizer
  - Minimize a loss function

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

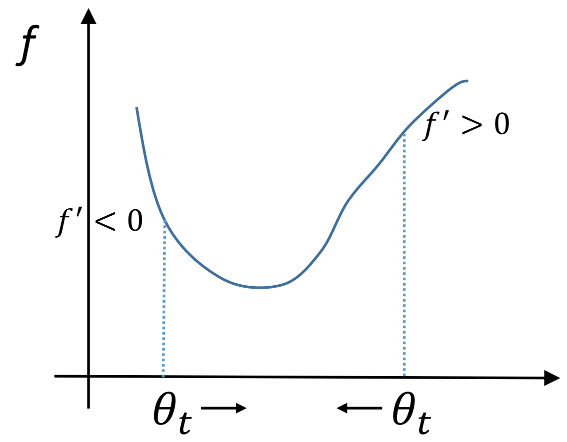
Slow learning vs. Fast learning

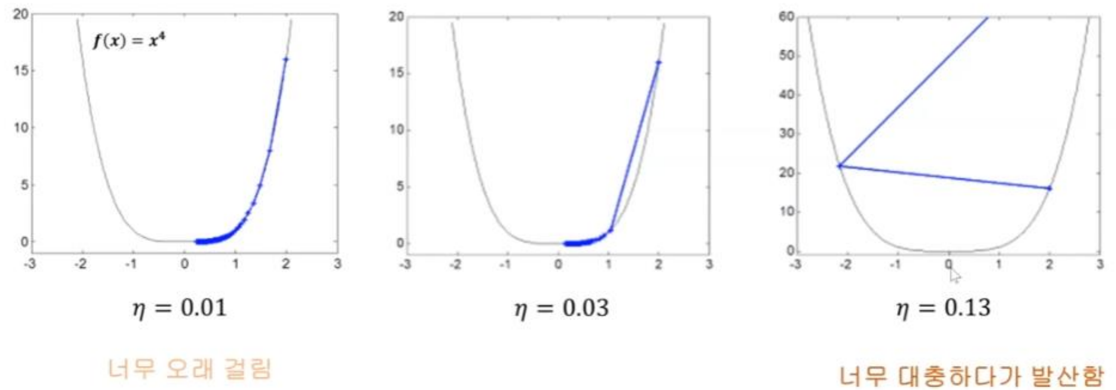


Gradient descent

- Revisit Update weights

$$\theta_{t+1} = \theta_t - \lambda f'(\theta_t)$$





Prof. Hyerim Bae (hrbae@pusan.ac.kr)

Gradient descent

- Batch gradient descent

$$\theta_t = \theta_{t-1} - \eta \cdot \nabla_\theta J(\theta)$$

- Stochastic gradient descent

$$\theta_t = \theta_{t-1} - \eta \cdot \nabla_\theta J(\theta; x^{(i)}; y^{(i)})$$

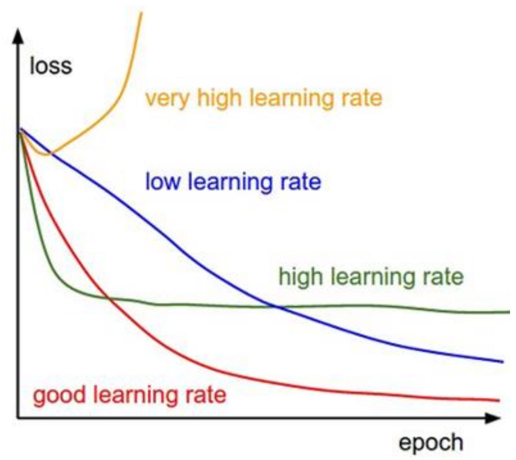
- Mini-batch gradient descent

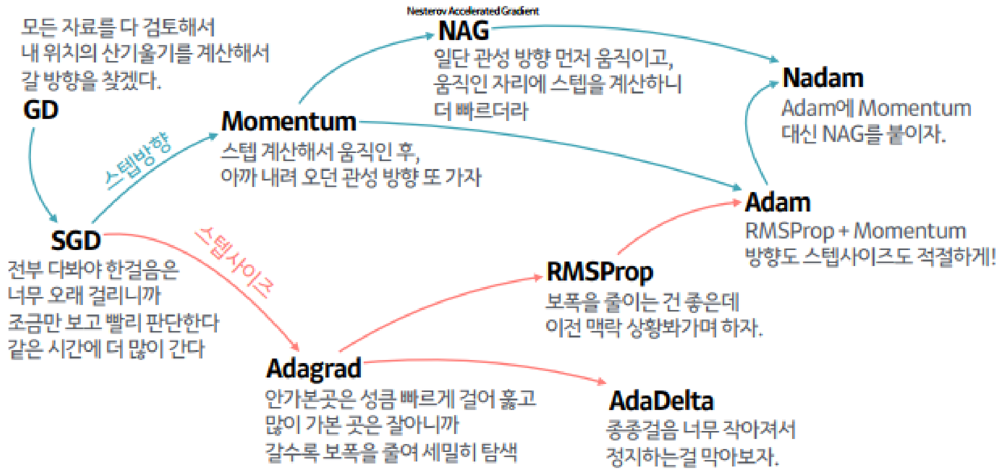
$$\theta_t = \theta_{t-1} - \eta \cdot \nabla_\theta J(\theta; x^{(i:i+n)}; y^{(i:i+n)})$$

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

2 ways of Optimizer

- Momentum
- Adaptive



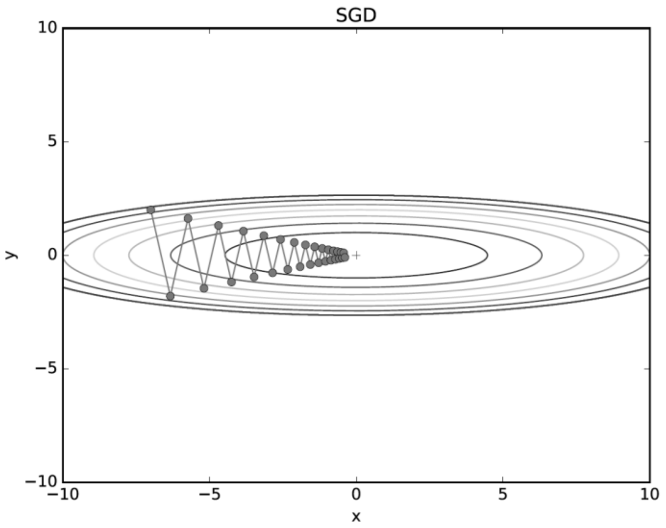


Prof. Hyerim Bae (hrbae@pusan.ac.kr)

SGD (Stochastic Gradient Descent)

- Gradient Descent by random sampling

$$\mathbf{W} \leftarrow \mathbf{W} - \eta \frac{\partial L}{\partial \mathbf{W}} \, cost(W)$$



Momentum

- 모멘텀: 과거의 경험치를 반영하자

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

$$\mathbf{W} \leftarrow \mathbf{W} + \mathbf{v}$$

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

Adagrad

- 업데이트 횟수에 따른 Learning rate의 조절

$$\mathbf{h} \leftarrow \mathbf{h} + \frac{\partial L}{\partial \mathbf{W}} \odot \frac{\partial L}{\partial \mathbf{W}}$$

$$\mathbf{W} \leftarrow \mathbf{W} - \eta \frac{1}{\sqrt{\mathbf{h}}} \frac{\partial L}{\partial \mathbf{W}}$$

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

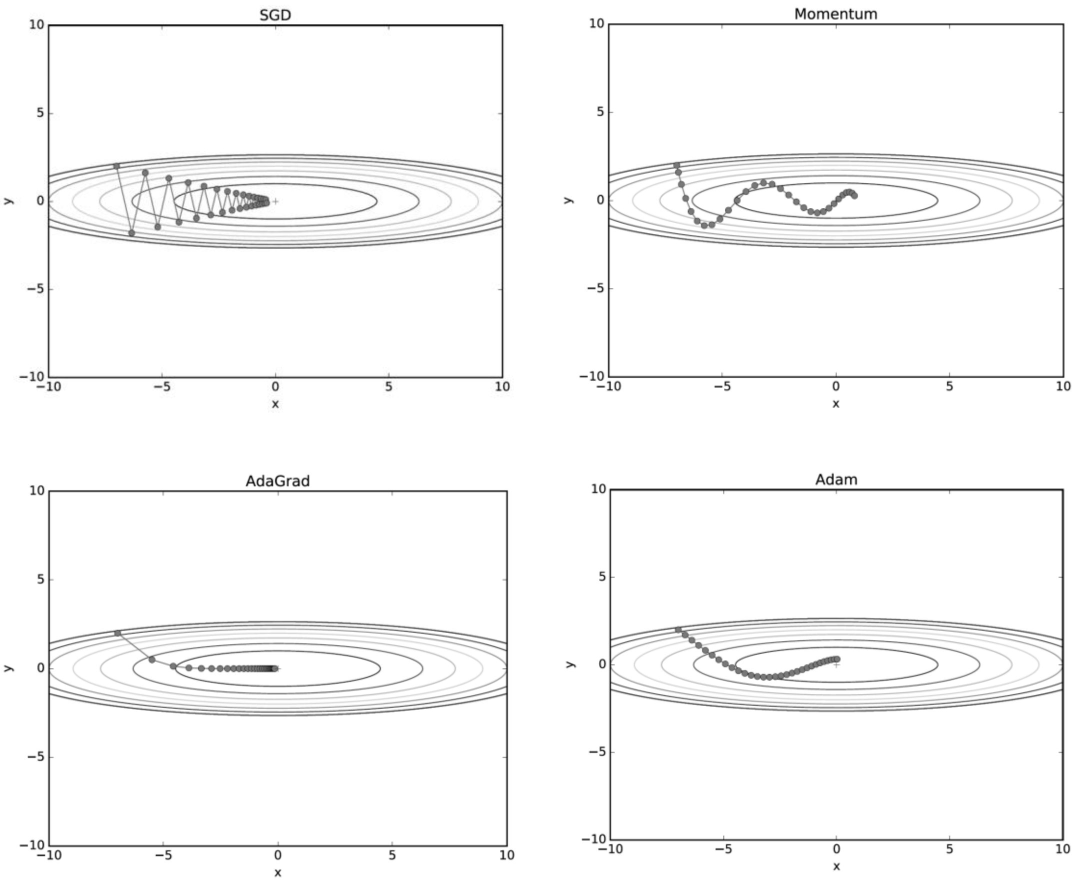
Adam

- Momentum+Adagad

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

Optimizing by Optimizers

- Which one do you prefer?



Prof. Hyerim Bae (hrbae@pusan.ac.kr)

References

- https://twinw.tistory.com/247
- https://m.blog.naver.com/PostView.naver?isHttpsRedirect=true&blogId=lego7407&logNo=221681014509
- https://www.youtube.com/watch?v=KN120w3PZIA&t=201