2018 대한의료정보학회 추계학술대회 튜토리얼

Biomedical Text Mining using Neural Networks

[Appendix] Techniques for Deep Learning – Part 1

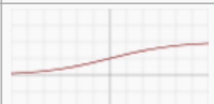
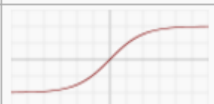
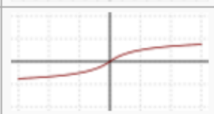
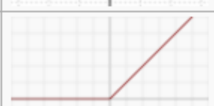
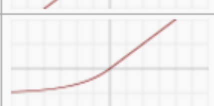
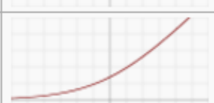
서울대학교병원 정보화실

고태훈 (taehoonko@snuh.org)

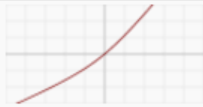
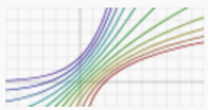
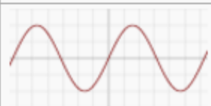
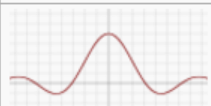
최세원 (swc@snuh.org)

Other activation functions

Other activation functions

Name	Plot	Equation	Derivative	Range
Identity		$f(x) = x$	$f'(x) = 1$	$(-\infty, \infty)$
Binary step		$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ 1 & \text{for} \quad x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for} \quad x \neq 0 \\ ? & \text{for} \quad x = 0 \end{cases}$	$\{0, 1\}$
Logistic (a.k.a Soft step)		$f(x) = \frac{1}{1 + e^{-x}}$	$f'(x) = f(x)(1 - f(x))$	$(0, 1)$
TanH		$f(x) = \tanh(x) = \frac{2}{1 + e^{-2x}} - 1$	$f'(x) = 1 - f(x)^2$	$(-1, 1)$
ArcTan		$f(x) = \tan^{-1}(x)$	$f'(x) = \frac{1}{x^2 + 1}$	$(-\frac{\pi}{2}, \frac{\pi}{2})$
Softsign [7][8]		$f(x) = \frac{x}{1 + \lvert x \rvert}$	$f'(x) = \frac{1}{(1 + \lvert x \rvert)^2}$	$(-1, 1)$
Rectifier[9]		$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ 1 & \text{for} \quad x \geq 0 \end{cases}$	$[0, \infty)$
Parameteric Rectified Linear Unit (PReLU)[10]		$f(x) = \begin{cases} \alpha x & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$	$f'(x) = \begin{cases} \alpha & \text{for} \quad x < 0 \\ 1 & \text{for} \quad x \geq 0 \end{cases}$	$(-\infty, \infty)$
Exponential Linear Unit (ELU)[11]		$f(x) = \begin{cases} \alpha(e^x - 1) & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$	$f'(x) = \begin{cases} f(x) + \alpha & \text{for} \quad x < 0 \\ 1 & \text{for} \quad x \geq 0 \end{cases}$	$(-\alpha, \infty)$
SoftPlus[12]		$f(x) = \log_e(1 + e^x)$	$f'(x) = \frac{1}{1 + e^{-x}}$	$(0, \infty)$

Other activation functions

Name	Plot	Equation	Derivative	Range
Bent identity		$f(x)=\frac{\sqrt{x^2+1}-1}{2}+x$	$f'(x)=\frac{x}{2\sqrt{x^2+1}}+1$	$(-\infty,\infty)$
SoftExponential [13]		$f(\alpha,x)=\begin{cases}-\frac{\log_e(1-\alpha(x+\alpha))}{\alpha} & \text{for } \alpha<0\\ x & \text{for } \alpha=0\\ \frac{e^{\alpha x}-1}{\alpha}+\alpha & \text{for } \alpha>0\end{cases}$	$f'(\alpha,x)=\begin{cases}\frac{1}{1-\alpha(\alpha+x)} & \text{for } \alpha<0\\ e^{\alpha x} & \text{for } \alpha\geq 0\end{cases}$	$(-\infty,\infty)$
Sinusoid		$f(x)=\sin(x)$	$f'(x)=\cos(x)$	$[-1,1]$
Sinc		$f(x)=\begin{cases}1 & \text{for } x=0\\ \frac{\sin(x)}{x} & \text{for } x\neq 0\end{cases}$	$f'(x)=\begin{cases}0 & \text{for } x=0\\ \frac{\cos(x)}{x}-\frac{\sin(x)}{x^2} & \text{for } x\neq 0\end{cases}$	$[\approx -.217234, 1]$
Gaussian		$f(x)=e^{-x^2}$	$f'(x)=-2xe^{-x^2}$	$(0,1]$

https://en.wikipedia.org/wiki/Activation_function

Gradient descent,
Learning rate and Momentum

Gradient descent

- **기울기 하강: Gradient descent**
  - 함수의 테일러 전개

$$f(x+\Delta x) = f(x) + \frac{f'(x)}{1!}\Delta x + \frac{f''(x)}{2!}\Delta x^2 + \cdots$$

  - 목적함수가 최소화인 경우 함수의 1차 도함수 값이 0이 아니면 1차 도함수의 반대 방향으로 이동해야 목적함수의 값을 감소시킬 수 있음

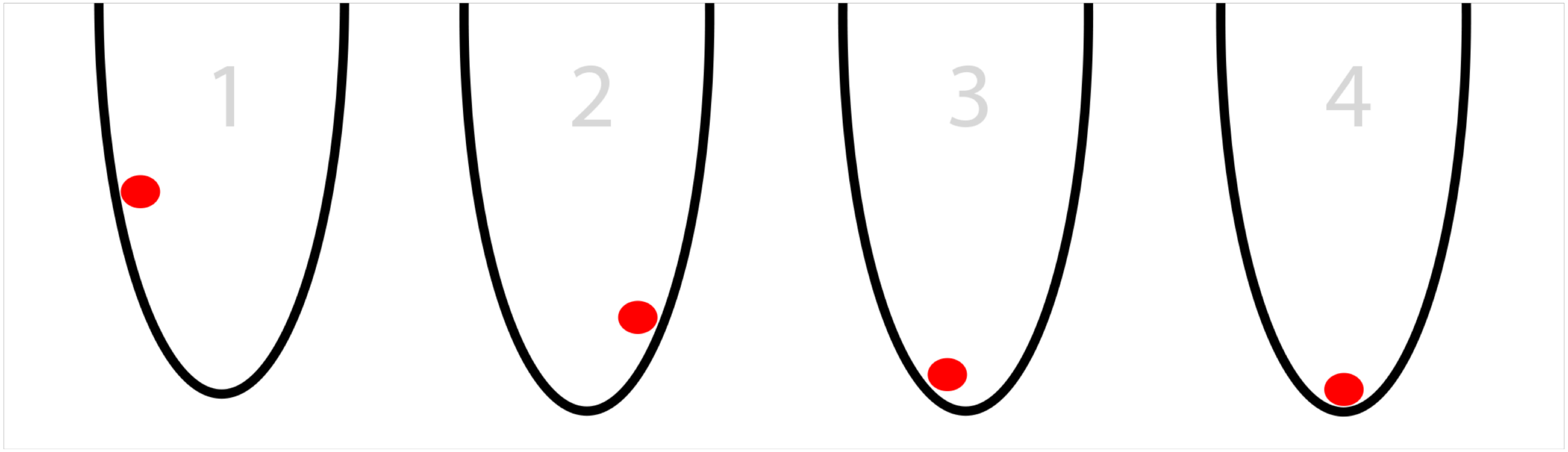
$$x_{new} = x_{old} - \eta f'(x), \quad where \quad 0 < \eta < 1$$

  - 이동 후의 새로운 함수 값은 이동 전의 함수 값보다 작음

$$f(x_{new}) = f(x_{old} - \eta f'(x)) \cong f(x_{old}) - \eta \left| f'(x) \right|^2 < f(x_{old})$$

Gradient descent: Simple example

- **Imagine that you had a red ball inside of a rounded bucket like in the picture below.**

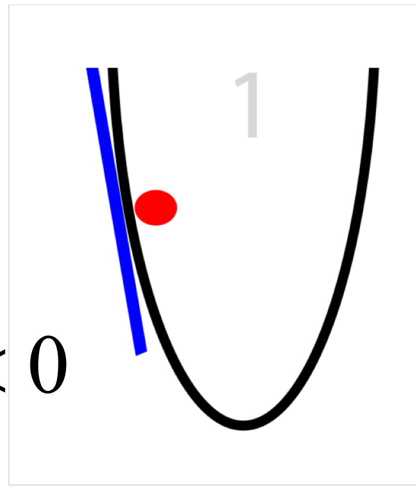


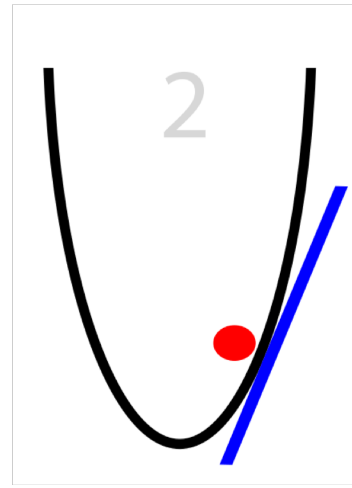
http://iamtrask.github.io/2015/07/27/python-network-part2/

Gradient descent: Simple example

- **The red ball moves in the opposite direction of the gradient.**

$$f'(x) < 0$$





$$f'(x) > 0$$

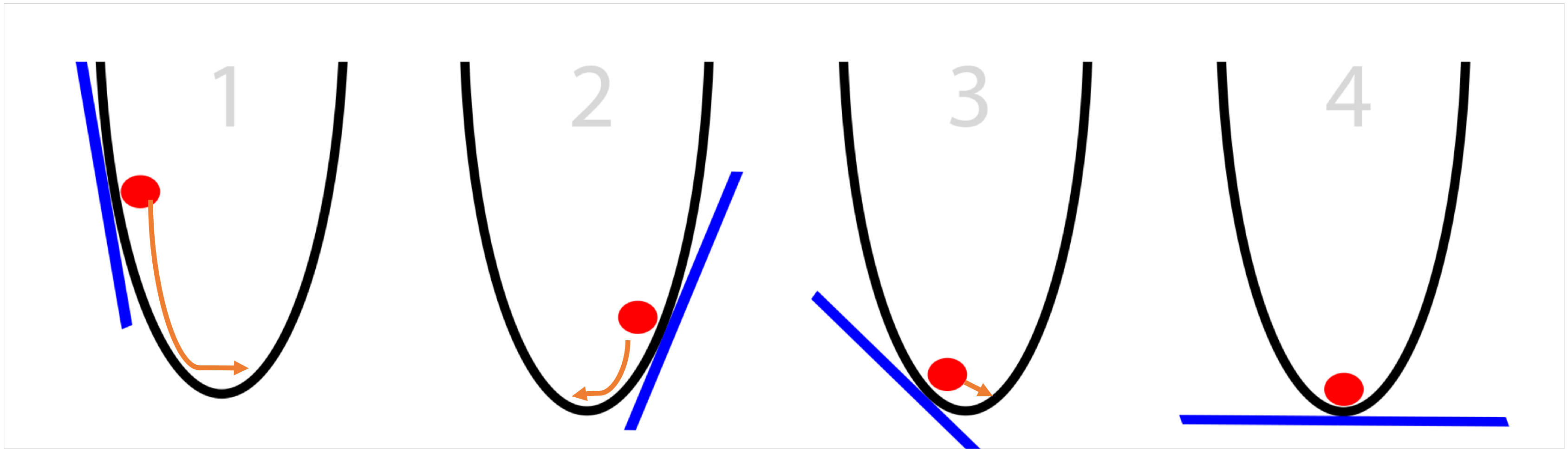
It moves in the direction x increases.

It moves in the direction x decreases.

http://iamtrask.github.io/2015/07/27/python-network-part2/

Gradient descent: Simple example

- If the **absolute value** of gradient is **large**, the red ball **moves a lot**.

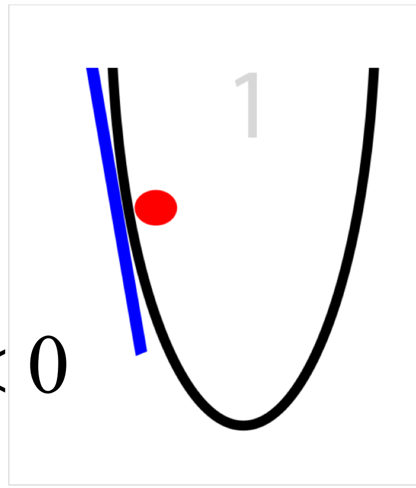


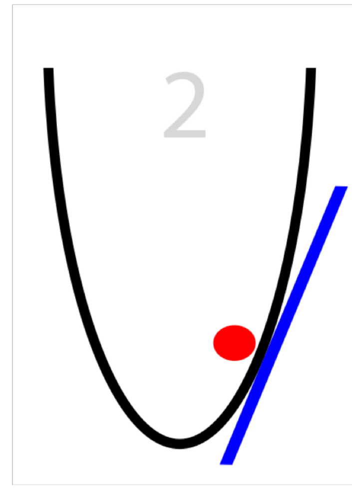
http://iamtrask.github.io/2015/07/27/python-network-part2/

Gradient descent: Simple example

- **The red ball moves in the opposite direction of the gradient.**
- **If the absolute value of gradient is large, the red ball moves a lot.**

$f'(x) < 0$





$f'(x) > 0$

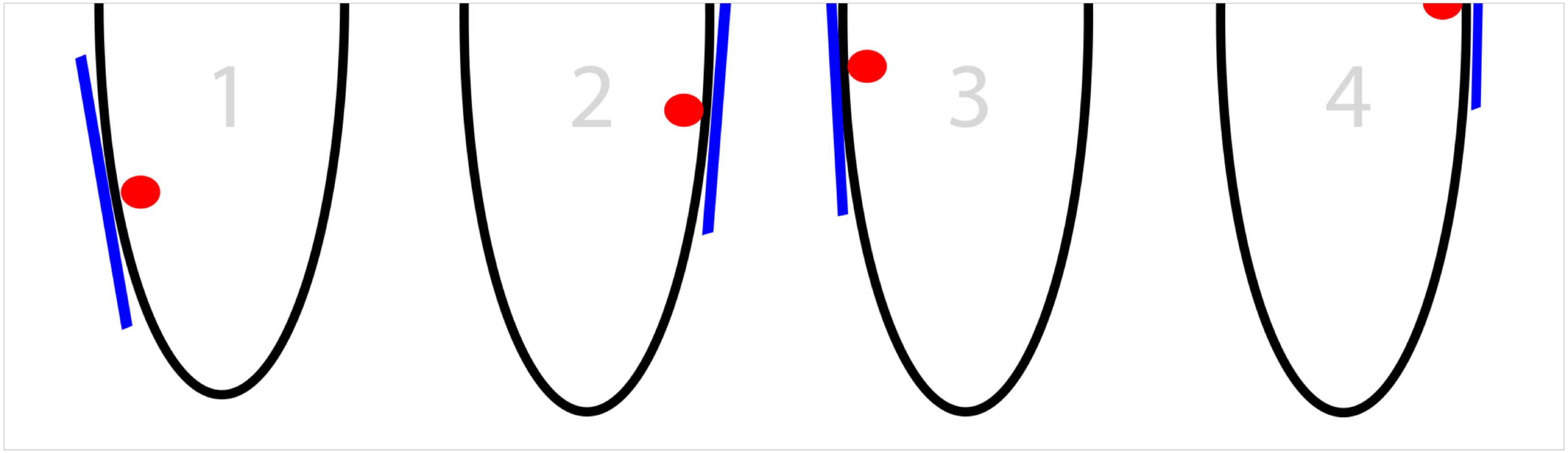
$$\boxed{x_{t+1} = x_t - \eta f'(x_t)}$$

η : Learning rate ($\eta \in [0,1]$)

http://iamtrask.github.io/2015/07/27/python-network-part2/

Learning rate (학습률)

- **Why the learning rate is used?**
  - Sometimes, calculated gradients are too big so that the ball may not arrive at the destination(global minimum).
  - The learning rate is used to reduce the gradient gradually.



Learning rate (학습률)

- **In neural networks, users set the learning rate as**
  - a constant
    - When the learning rate is tiny, the derivatives almost never changed direction and the weights ended up being reasonably small too.
    - When it is huge, the derivatives changed directions a medium amount and the weights got huge too.
  - **a value that decreases over iterations**
    - In the beginning of training process, randomly initialized weights should be updated a lot.
    - As the number of iteration increases, the learning rate decreases.

How to avoid local minimums: Momentum

- **The momentum method is a technique for accelerating gradient descent that accumulates a velocity vector in directions of persistent reduction in the objective across iterations.**
- **현재 gradient가 업데이트되고 있는 속도를 고려하겠다는 의미**
- **최근 많이 이용되는 딥러닝 학습 방법들에서 모멘텀이 자주 이용됨**

$$x_{t+1} = x_t - \eta f'(x_t)$$

$$x_{t+1} = x_t + v_{t+1}$$

$$v_{t+1} = \mu v_t - \eta f'(x_t)$$

v_{t+1} : Update value to x_t

μ : weight of the previous update ($\mu \in [0,1]$)