Differentiation

Gradient Vanishing
(기울기 소실)

소프트웨어 꼰대 강의

노기섭 교수

(kafa46@cju.ac.kr)

Course Overview

Topic	Contents
01. Orientation 오리엔테이션	Course introduction, motivations, final objectives 과정 소개, 동기부여, 최종 목표
02. Learning in deeplearning 딥러닝 학습	How does the deeplearing learns knowledge from data 어떻게 딥러닝은 데이터로부터 지식을 배우는가?
03. Principle of differentiation 미분의 원리	Basics of differentiation (concepts, notation, operations) 미분 기본지식 (개념, 표기, 연산)
04. Partial differentiation 편미분	Concept & operation of partial differenciation 편미분 개념, 연산
05. Gradient descent 경사 하강법	Concept, interpretation and learning in gradient descent 경사하강 알고리즘 개념, 해석 및 학습
06. Chain rule 연쇄법칙	Concept & operation of chain rule 연쇄법칙 개념 및 연산
07. Matrix differentiation 행렬미분	Partial differentiation in linear system 선형시스템에서의 편미분
08. Back propagation 역전파 학습	The mechanism of back propagation 역전파 학습의 작동 방법
09. Gradient vanishing 기울기 소실	Quick overview on activation function, cause root of gradient vanishing and its counter-measure 활성함수 간단 소개, 기울기 소실 근본원인과 대책

갑자기 공부를 못할 때...

평범한 부모님의 걱정 우리 애는 공부는 열심히 하는데 성적은 늘 그대로 ㅠㅠ

이미지 출처: https://youtu.be/hY2HU2lji8A?si=zKoqgDcaqS0Vjjo-

딥러닝에도 이런 일이 생길까요?

종종 발생합니다 ^^

딥러닝에서는 왜 이런 일이?

딥러닝 엔지니어 선택의 문제

1. 잘못된 학습 방법 (model selection)

딥러닝 학습(미분)의 태생적 문제

이번 강의에서 공부하고자 하는 이슈

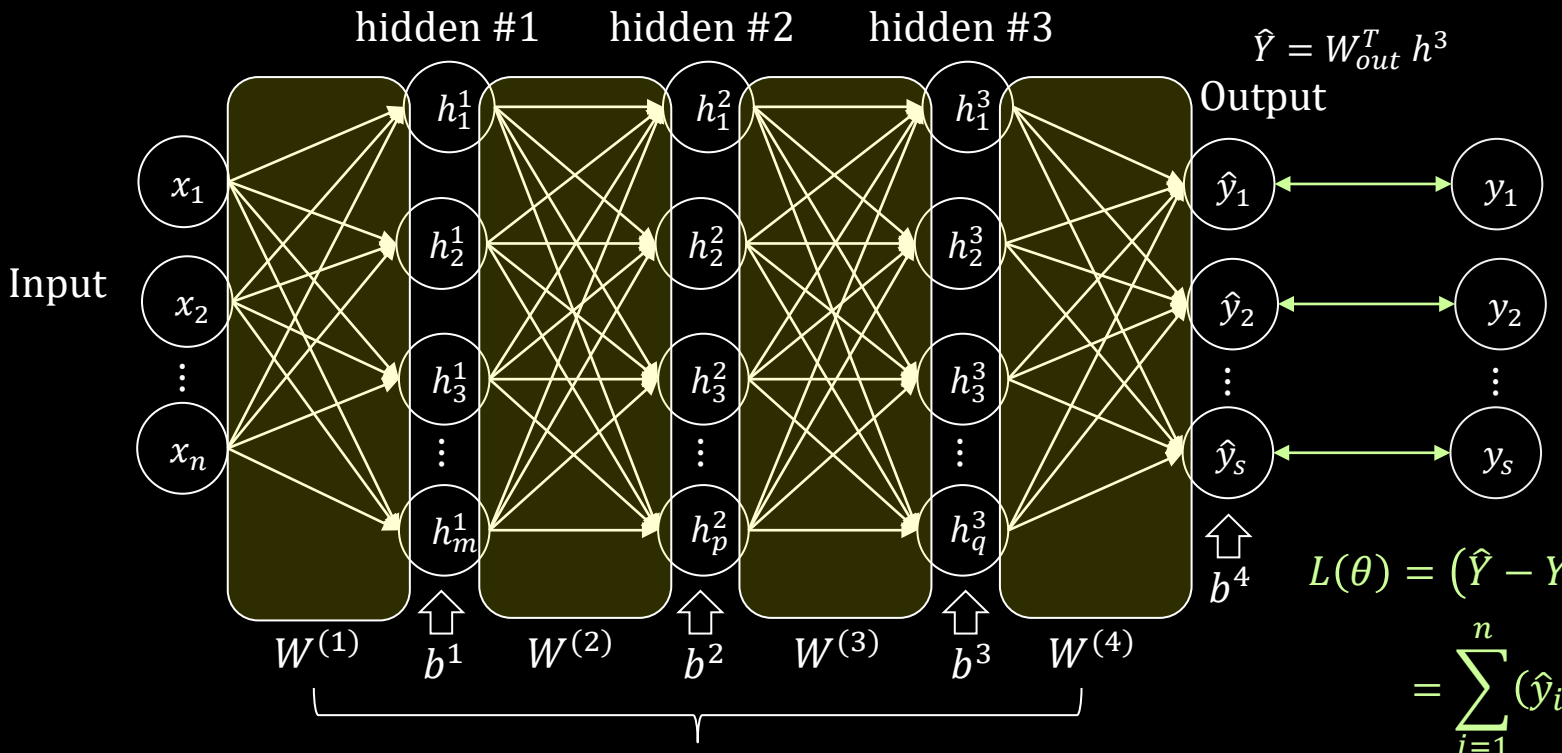
2. 책만 펴놓고 실제는 딴생각 (Gradient vanishing)

Recap: Learning in Deeplearning

Prediction: Forward Propagation

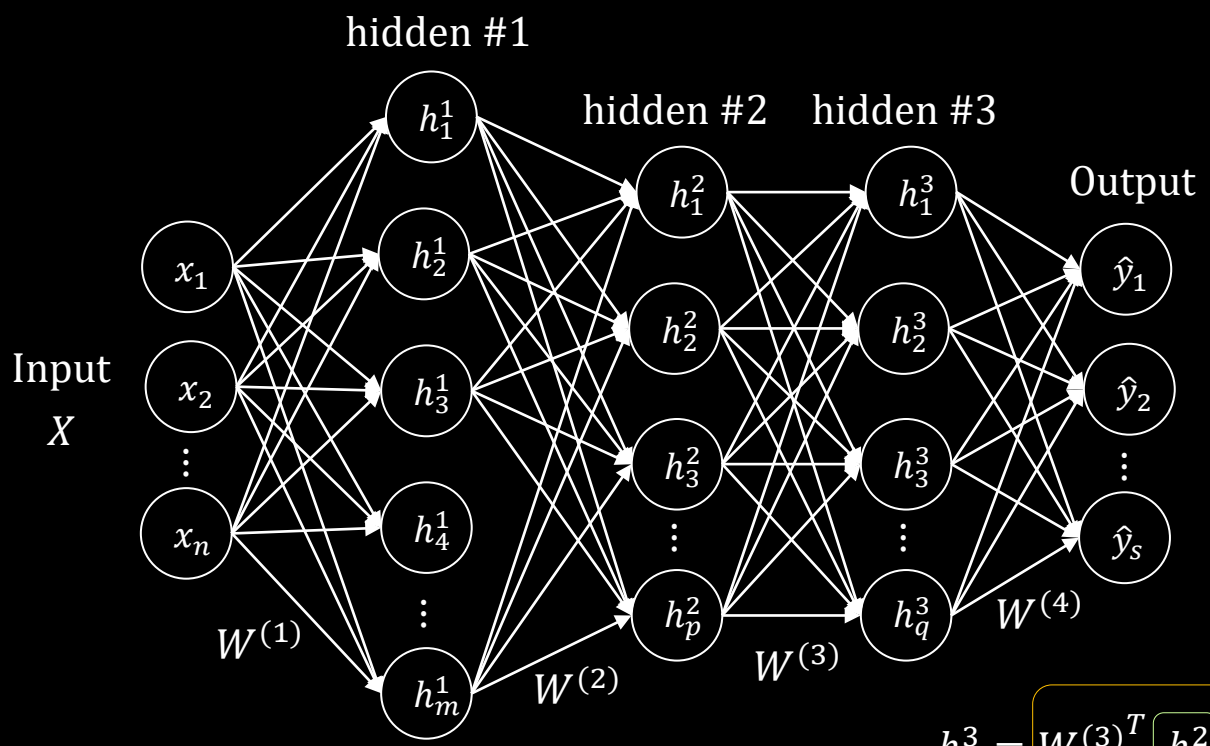
Goal: Find the set of θ that minize total loss $L(\theta)$

where $\theta = \{W^i, b^i \mid i = the\ index\ of\ layer\ (1, 2, \cdots, l)\}$



Update (optimize) all θ values using backprop

Introducing Activation Function (a.k.a. σ)



Bias 표시는 생략할게요 ^^

선형변환의 의미는 이전 강의 참고하세요^^
[선형대수]_17. 차원 축소 및 확장 (dimension reduction & expansion)
YouTube: https://youtu.be/j1D1jY71Wjg

$$= W^{(4)^T} W^{(3)^T} W^{(2)^T} W_1^{(1)^T} X$$

$$\hat{Y} = W^{(4)^T} h^3$$

$$h^3 = W^{(3)^T} W^{(2)^T} W^{(1)^T} X$$

$$h^3 = W^{(3)^T} h^2$$

$$h^2 = W^{(2)^T} h^1$$

$$h^1 = W^{(1)^T} X$$

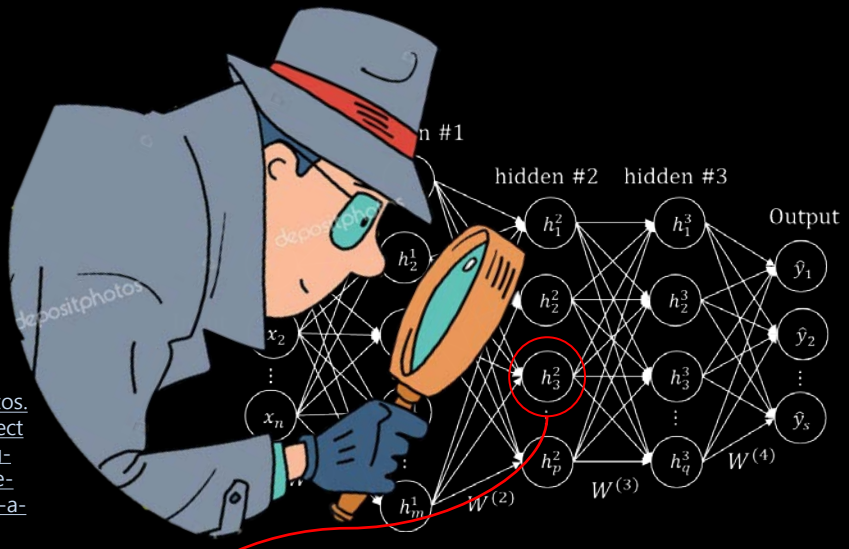
$$h^2 = W^{(2)^T} W^{(1)^T} X$$

$$h^1 = W^{(1)^T} X$$

여전히 선형....
선형공간에서
공간 확장/수축만 한다!
(차원 확장/축소)

Back-prop 과정에서
비선형 분포를 배울 수 없다.
ㅠㅠ

Solution: Hidden Layer에 비선형 함수를 추가하자!



이미지 출처:
https://depositphotos.com/ko/vector/detective-and-magnifying-glass-icon-a-private-detective-a-man-in-a-coat-hat-and-547568454.html

$$\begin{pmatrix} w_{11} & \cdots & w_{13} & \cdots & w_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{i1} & \cdots & w_{i3} & \cdots & w_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{m3} & \cdots & w_{mp} \end{pmatrix} \qquad \begin{pmatrix} h_1^1 \\ h_2^1 \\ \vdots \\ h_m^1 \end{pmatrix}$$

$$h_3^2 = w_{13}^2 h_1^1 + w_{23}^2 h_2^1 + w_{33}^2 h_3^1 + \cdots + w_{m3}^2 h_3^1$$

$$= \sum_{i=1}^{m} w_{i3}^2 \, h_i^1 \quad = W_{:3}^{(1)^T} \; h^1$$

비선형 함수 아무거나 선택한다.

시그마(sigma)로 표시하고,

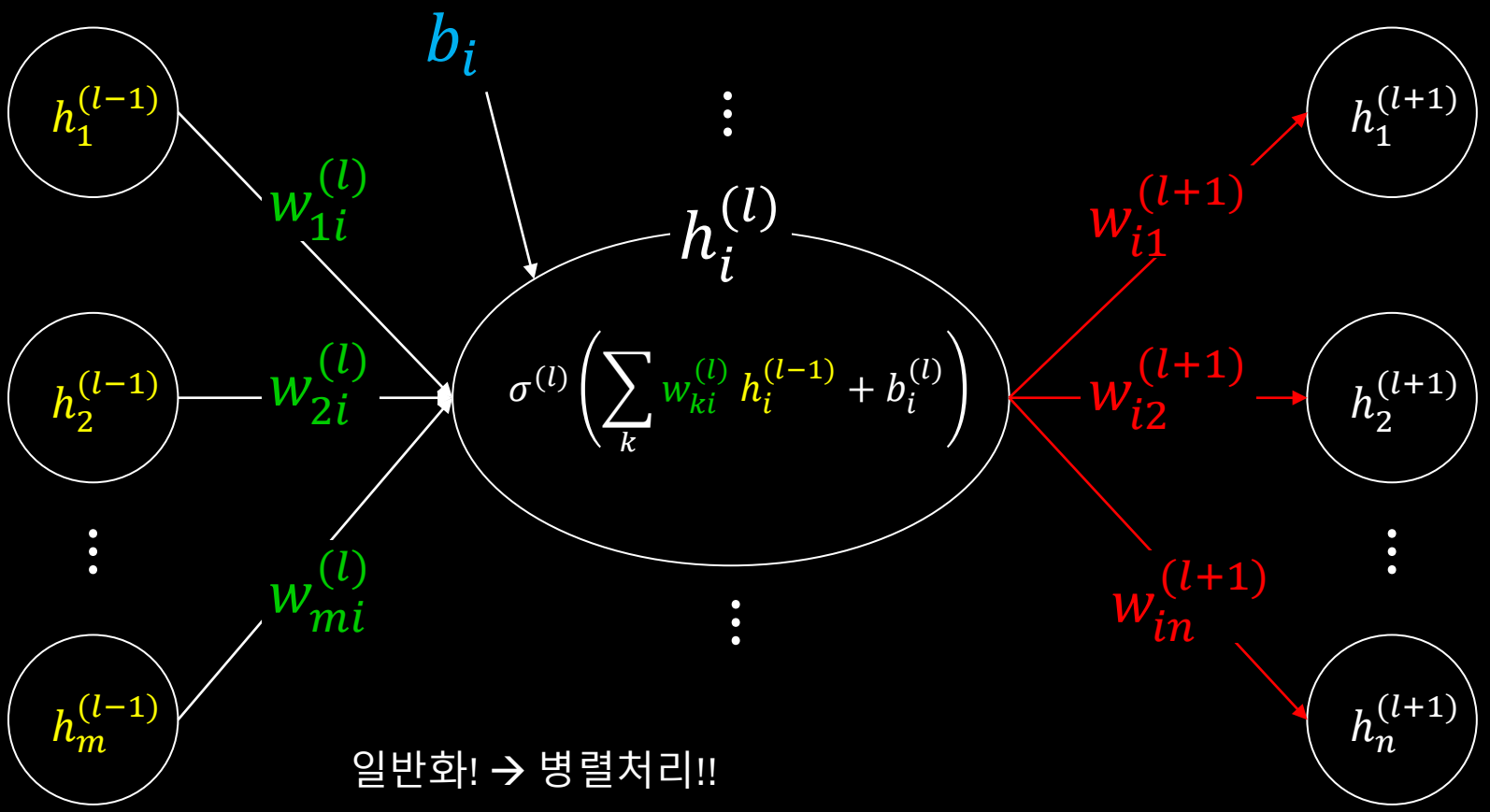
'활성함수'라고 부르기로 하자!

$\sigma(x)$

다음 노드로 출발하기 직전에
활성함수 통과한 값 z으로 보내자!

h_3^2 값을 ⇨ $\sigma(h_3^2)$ 값으로 변경

Generalization of Activation Function in DL Network



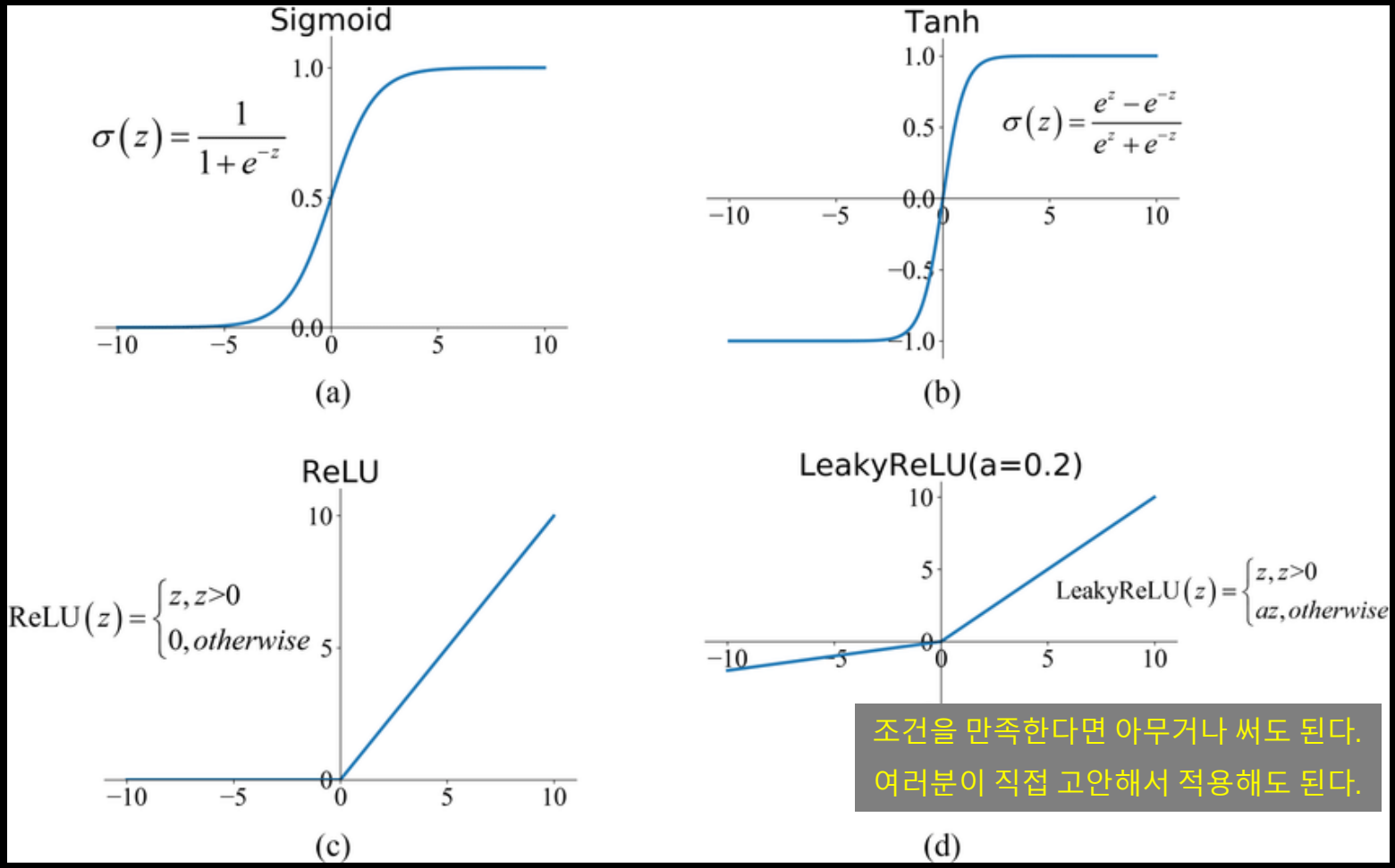
일반화! → 병렬처리!!

$$h^{(l)} = \sigma\left(W^{(l)^T} \cdot h^{(l-1)} + b\right)$$

어떤 활성 함수를 써야 할까?

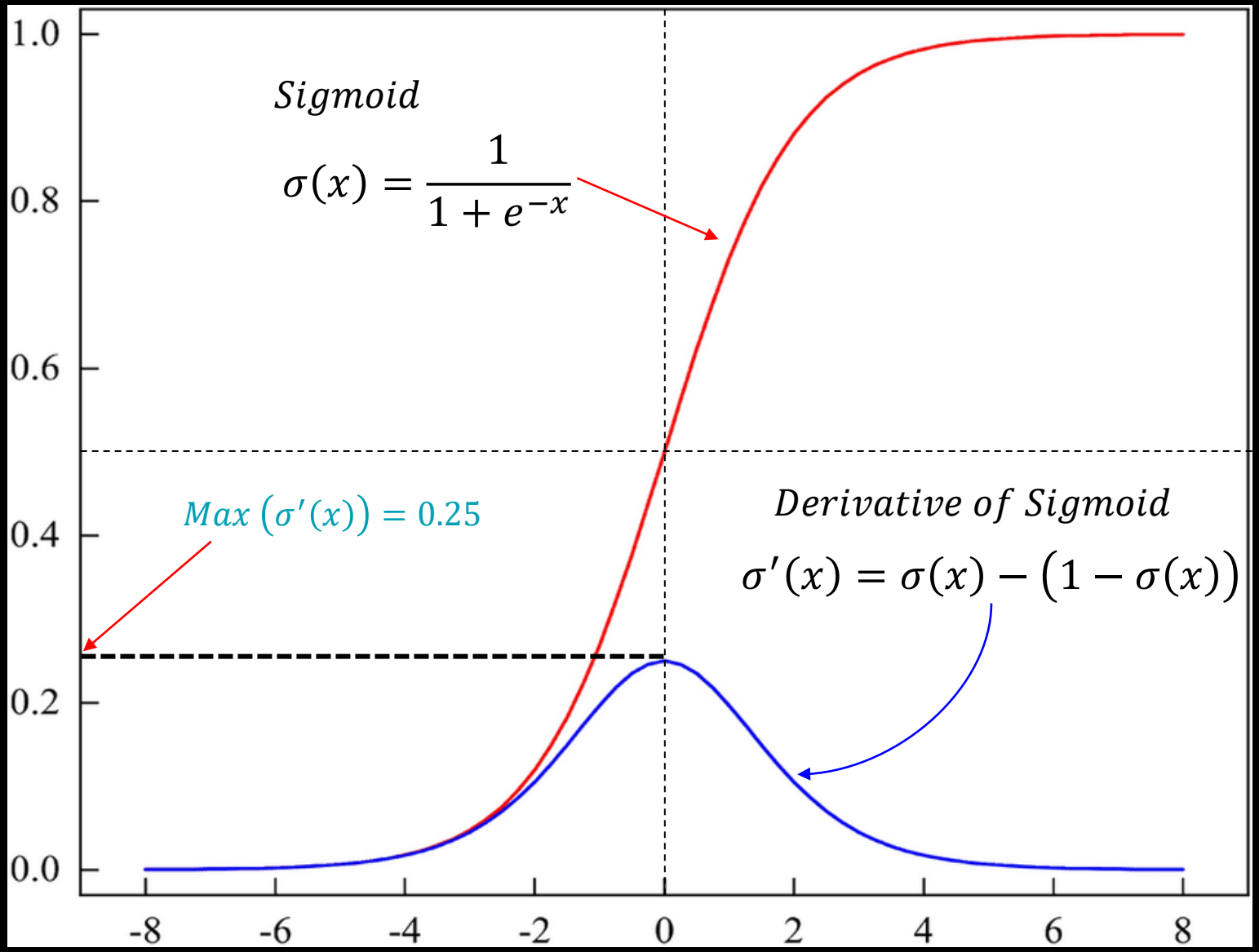
조건 1. 비선형이어야 한다.
조건 2. 미분 가능해야 한다.

Feng, Junxi et. al., (2019). Reconstruction of porous media from extremely limited information using conditional generative adversarial networks. Physical Review E. 100.
DOI: https://doi.org/10.1103/PhysRevE.100.033308

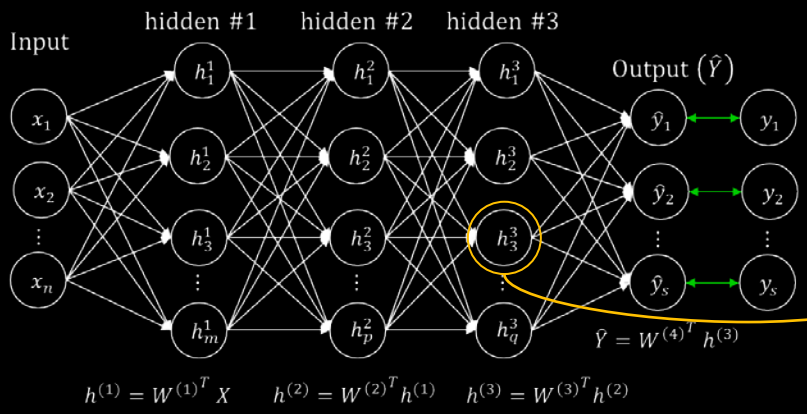


조건을 만족한다면 아무거나 써도 된다.
여러분이 직접 고안해서 적용해도 된다.

Gradient Vanishing 주범 Sigmoid



Vanishing Problem during Back-propagation



$$L(\theta) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

$$h_i^{(l)} = \sigma^{(l)}\left(\sum_k w_{ki}^{(l)}\, h_i^{(l-1)} + b_i^{(l)}\right)$$

전체 네트워크 역전파 구하기

어떤 수에 (0, 1) 사이의
값을 곱하면 0에 가까워 진다.
(양수, 음수 관계없이 공통사항)

$$\frac{\partial L(\theta)}{\partial W^{(4)^T}} = \frac{\partial L(\theta)}{\partial \hat{Y}} \times \frac{\partial \hat{Y}}{\partial \sigma^{(4)}} \times \frac{\partial \sigma^{(4)}}{\partial W^{(4)^T}}$$

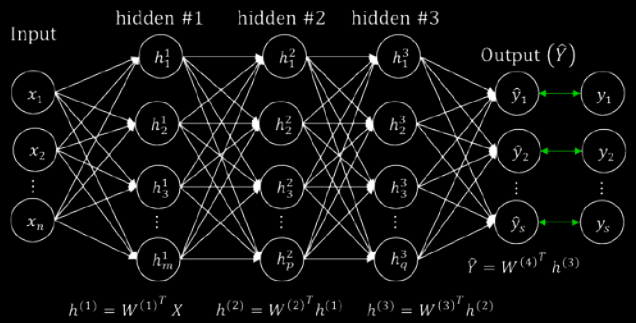
$$\frac{\partial L(\theta)}{\partial W^{(3)^T}} = \frac{\partial L(\theta)}{\partial \hat{Y}} \times \frac{\partial \hat{Y}}{\partial \sigma^{(4)}} \times \frac{\partial \sigma^{(4)}}{\partial W^{(4)^T}} \times \frac{\partial W^{(4)^T}}{\partial \sigma^{(3)}} \times \frac{\partial \sigma^{(3)}}{\partial W^{(3)^T}}$$

$$\frac{\partial L(\theta)}{\partial W^{(2)^T}} = \frac{\partial L(\theta)}{\partial \hat{Y}} \times \frac{\partial \hat{Y}}{\partial \sigma^{(4)}} \times \frac{\partial \sigma^{(4)}}{\partial W^{(4)^T}} \times \frac{\partial W^{(4)^T}}{\partial \sigma^{(3)}} \times \frac{\partial \sigma^{(3)}}{\partial W^{(3)^T}} \times \frac{\partial W^{(3)^T}}{\partial \sigma^{(2)}} \times \frac{\partial \sigma^{(2)}}{\partial W^{(2)^T}}$$

$$\frac{\partial L(\theta)}{\partial W^{(1)^T}} = \frac{\partial L(\theta)}{\partial \hat{Y}} \times \frac{\partial \hat{Y}}{\partial \sigma^{(4)}} \times \frac{\partial \sigma^{(4)}}{\partial W^{(4)^T}} \times \frac{\partial W^{(4)^T}}{\partial \sigma^{(3)}} \times \frac{\partial \sigma^{(3)}}{\partial W^{(3)^T}} \times \frac{\partial W^{(3)^T}}{\partial \sigma^{(2)}} \times \frac{\partial \sigma^{(2)}}{\partial W^{(2)^T}} \times \frac{\partial W^{(2)^T}}{\partial \sigma^{(1)}} \times \frac{\partial \sigma^{(1)}}{\partial W^{(1)^T}}$$

Reasoning for Gradient Vanishing

어떤 수에 (0, 1) 사이의
값을 곱하면 0에 가까워 진다.
(양수, 음수 관계없이 공통사항)



$$L(\theta) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

$$\frac{\partial L(\theta)}{\partial W^{(1)^T}} = \frac{\partial L(\theta)}{\partial \hat{Y}} \times \frac{\partial \hat{Y}}{\partial \sigma^{(4)}} \times \frac{\partial \sigma^{(4)}}{\partial W^{(4)^T}} \times \frac{\partial W^{(4)^T}}{\partial \sigma^{(3)}} \times \frac{\partial \sigma^{(3)}}{\partial W^{(3)^T}} \times \frac{\partial W^{(3)^T}}{\partial \sigma^{(2)}} \times \frac{\partial \sigma^{(2)}}{\partial W^{(2)^T}} \times \frac{\partial W^{(2)^T}}{\partial \sigma^{(1)}} \times \frac{\partial \sigma^{(1)}}{\partial W^{(1)^T}}$$

업데이트 규칙

$$W = W - \alpha \times \frac{\partial L(\theta)}{\partial W}$$

아주 작은 값
(learning rate)

0 에 가까운 값
(Gradient)

0 (zero) 으로
수렴

아무리 공부해도
학습 효과가 없다!!

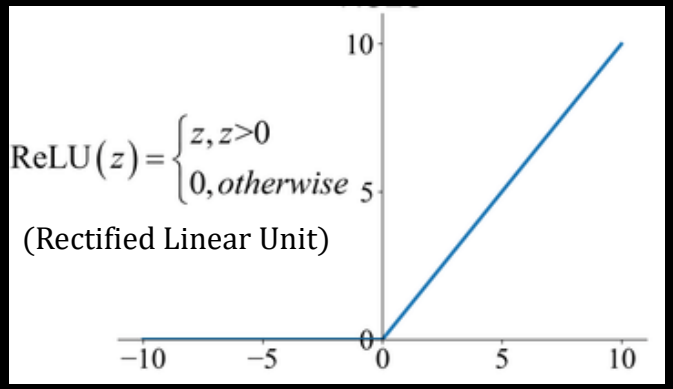
아무 일도 일어나지 않음
$W = W$ (*No update!*)

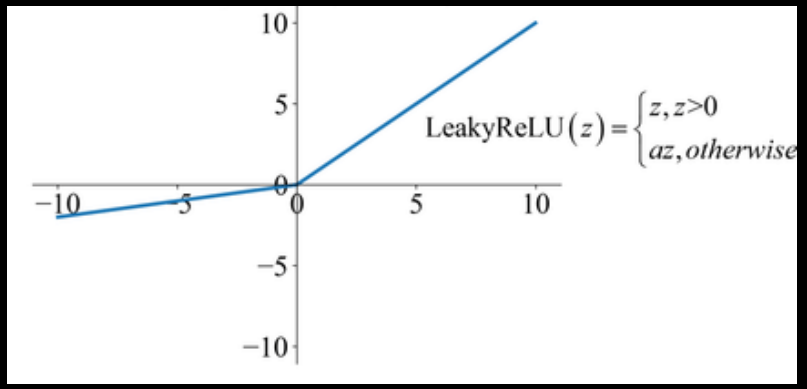
⇨ 기울기 소실
(Gradient Vanishing)

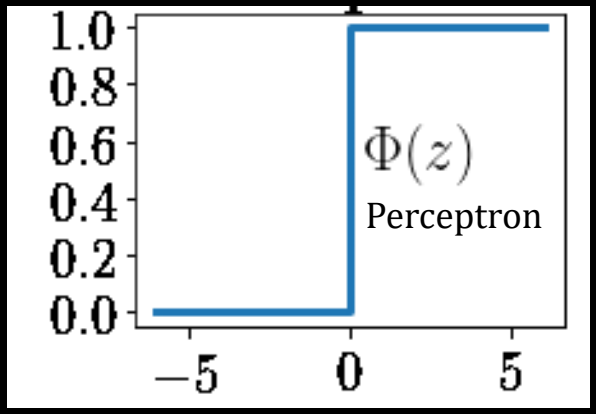
Solution against Gradient Vanishing

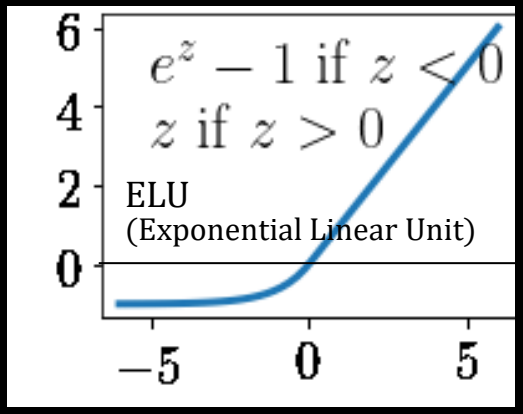
어떤 수에 (0, 1) 사이의
값을 곱하면 0에 가까워 진다.
(양수, 음수 관계없이 공통사항)

활성함수의 미분값이
(0, 1) 사이가 되지 않는
함수를 선택하면 된다.









수고하셨습니다 ..^^..