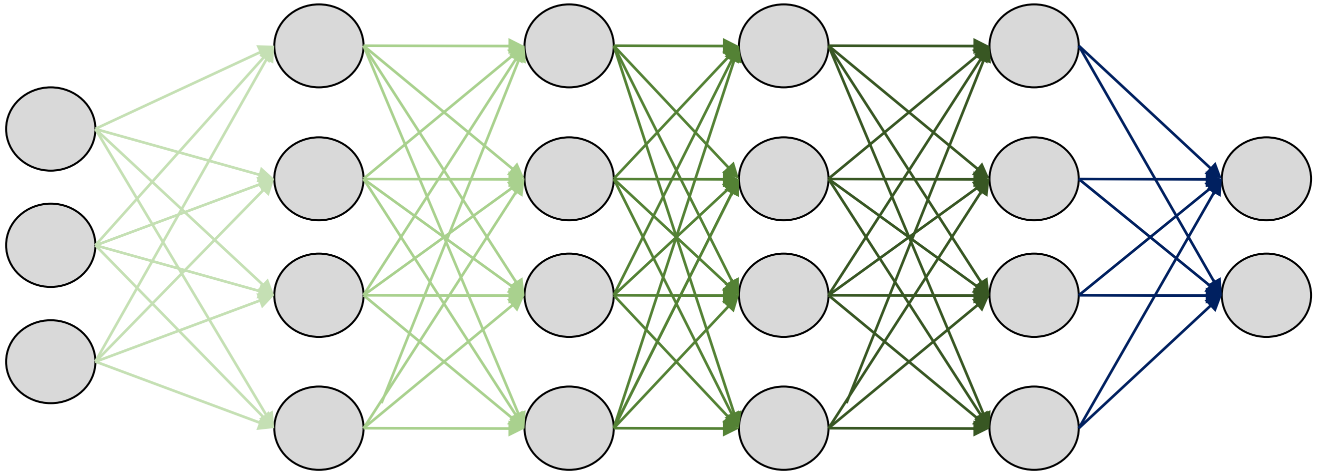
깊은 신경망을 위한 딥러닝 학습

기울기 소실(Vanishing Gradient)

- 기울기 소실 : 앞층으로 갈수록 오차가 잘 전달되지 않는 현상(학습이 이루어지지 않음)
- 계층이 깊어질수록 입력층과 가까이 있는 가중치들의 학습이 잘 일어나지 않음

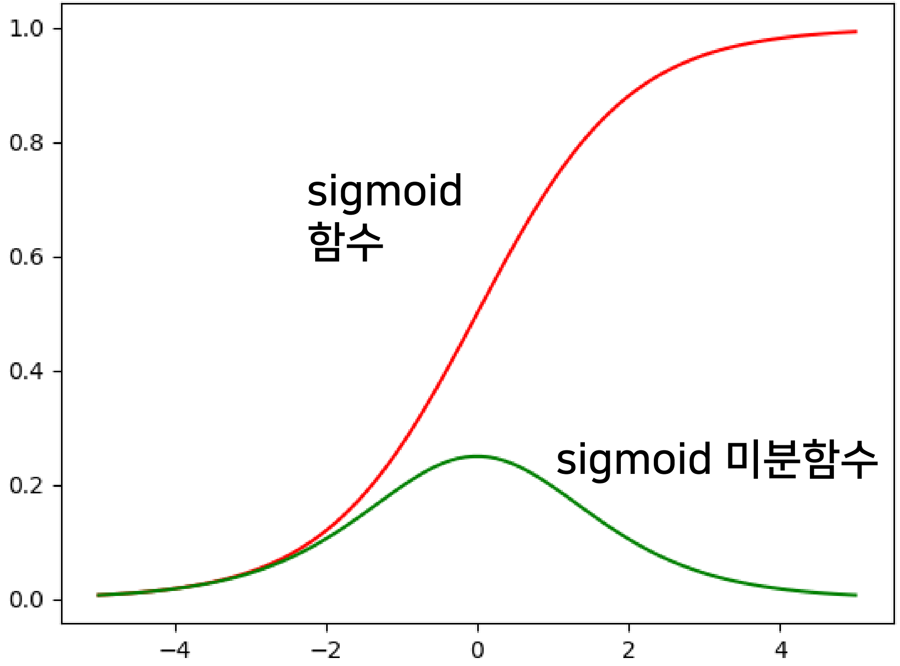


학습이 안됨 학습이 잘됨

기울기 소실(Vanishing Gradient)

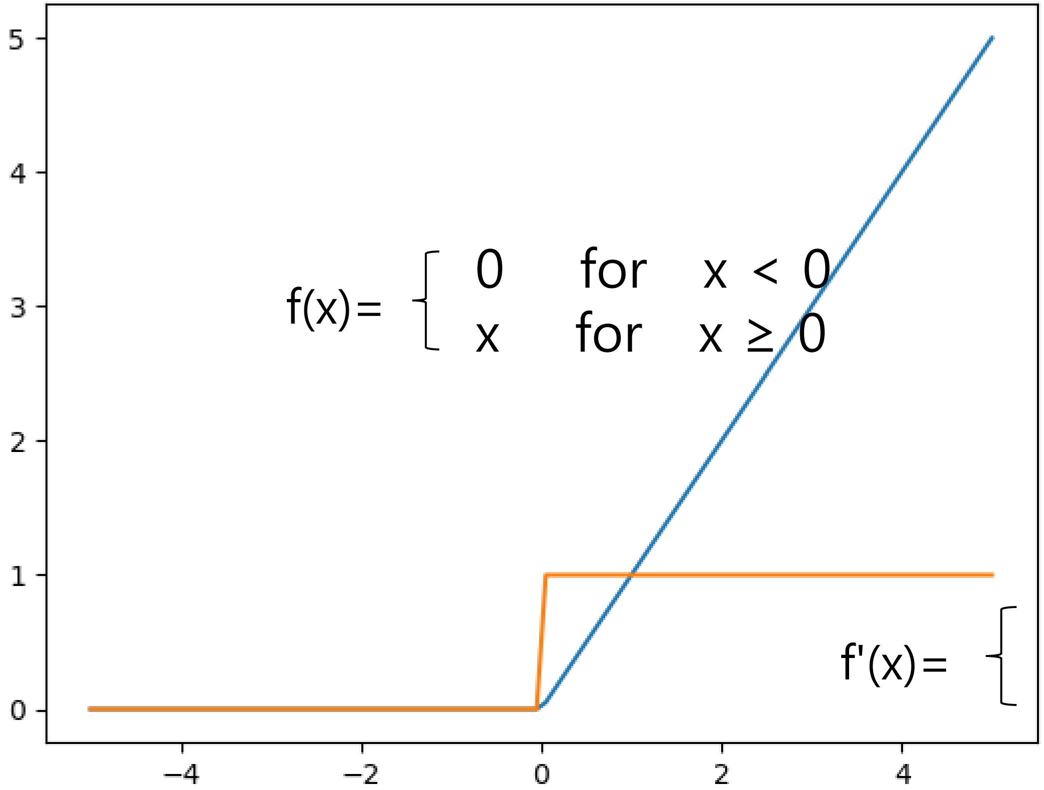
- sigmoid 함수에서 0, 1로 강제 출력하는 영역에서는 학습이 이루어지지 않음
- 도함수의 계산결과가 역방향으로 전달될 때 출력값이 현저하게 감소됨 (입력층으로 갈수록 0에 가까워짐)

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}} \qquad \frac{d}{dx}\text{sigmoid}(x) = \text{sigmoid}(x)(1\text{-sigmoid}(x))$$



ReLU(Re Rectified Linear Unit)

- ReLu 함수 : y=x라는 직선부분과 모든 부분을 0으로 출력하는 부분으로 구성
  – 음수영역에서는 미분값이 0, 양수영역에서는 전 구간 미분값이 존재하여 학습이 이루어짐

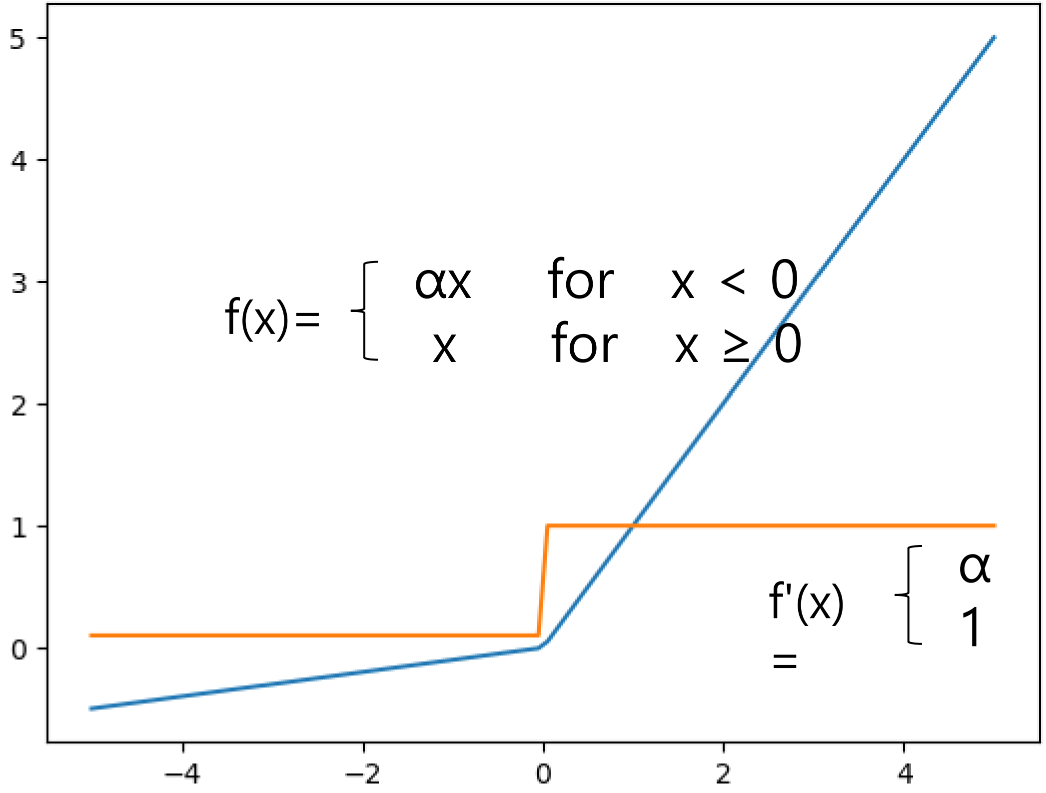


$$f(x)=\begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$$

$$f'(x)=\begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$$

Leaky ReLU

- Leaky ReLU : x<0 인 영역은 아주 작은 기울기를 가져 약간의 학습이 일어남

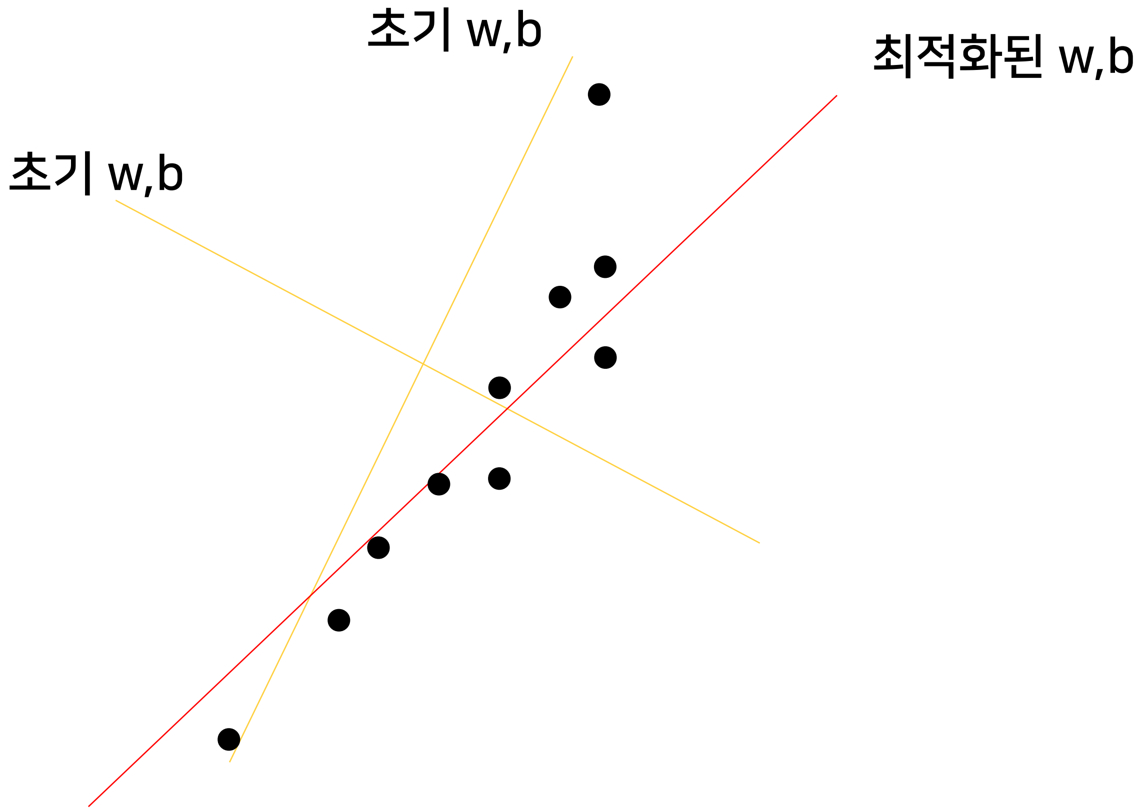


$$f(x) = \begin{cases} \alpha x & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$$

$$f'(x) = \begin{cases} \alpha & \text{for} \quad x < 0 \\ 1 & \text{for} \quad x \geq 0 \end{cases}$$

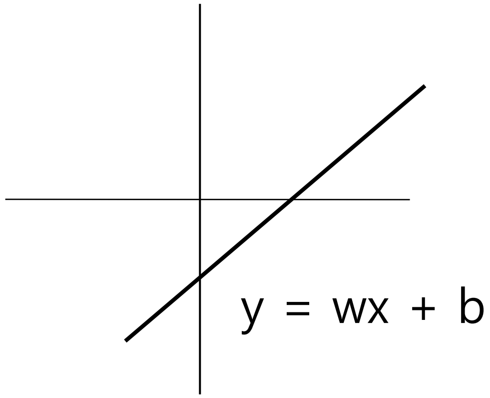
가중치 초기화

- 사전 정보가 없다면 랜덤하게 가중치를 지정해서 경사하강법으로 최적의 w와 b를 찾을 수 있음



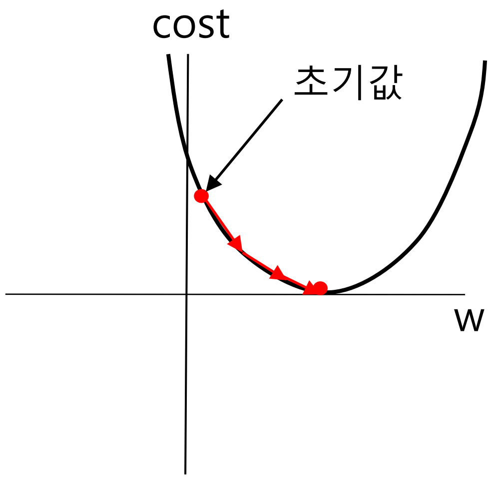
전역 최적화(Global Optimization)

- 선형함수를 사용하는 경우 어떤 w에서 시작해도 수렴이 보장됨



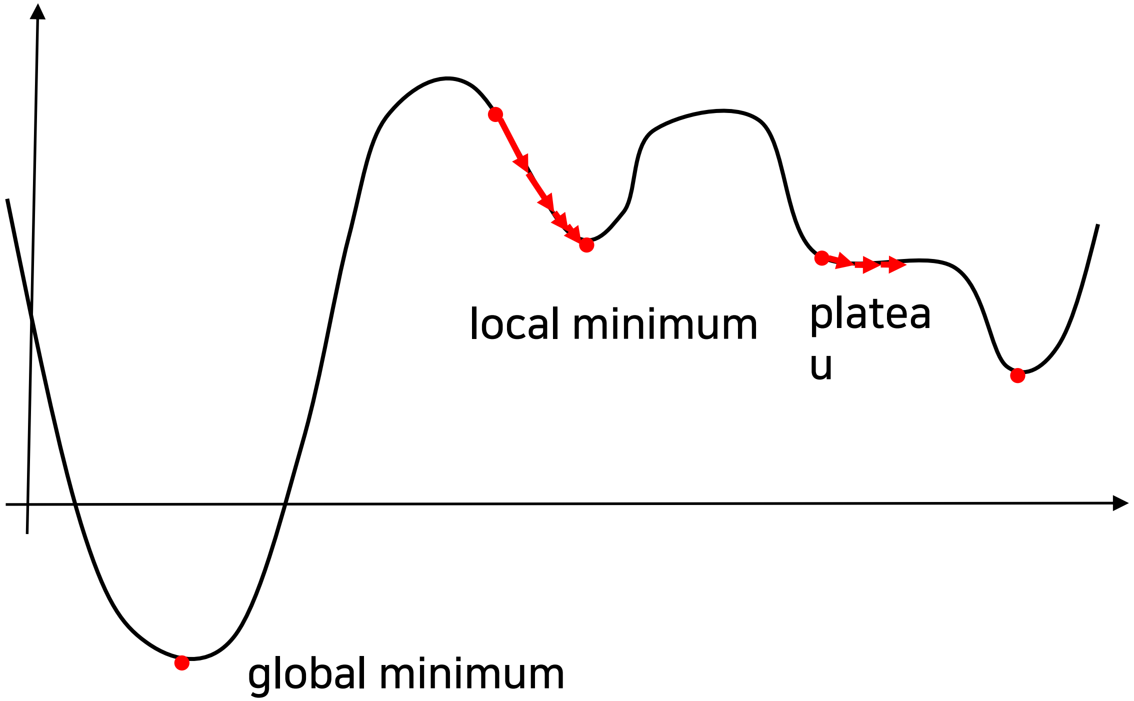
$$cost(w,b) = \frac{1}{m}\sum_{i=1}^{m}\left(H(x^{(i)}) - y^{(i)}\right)^2$$

cost 정의



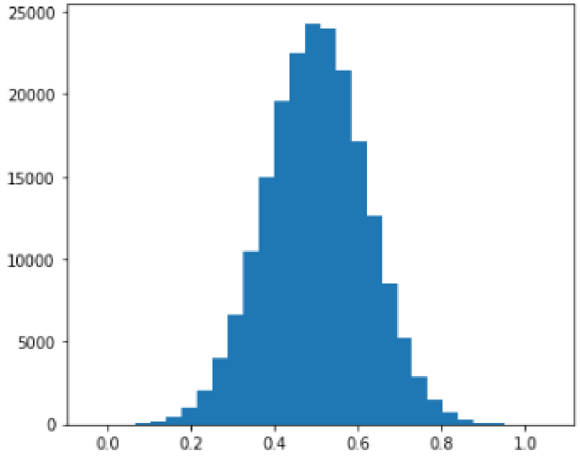
지역 최적화(Local Optimization)

- 다층 구조, 각각의 층에서 다양한 비선형 함수를 사용할 경우 w에 대한 cost함수 형태가 복잡
  - 이 때 잘못된 초기값을 선택하면 지역 최소점(local minimum)에 도달
- 원활한 학습을 위해 주어진 문제에 맞는 데이터 분포를 고려해서 적절한 초기화 수행이 필요

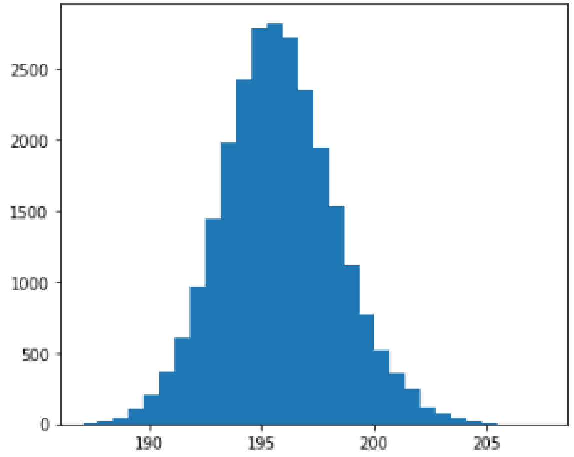


정규분포 초기화의 문제점

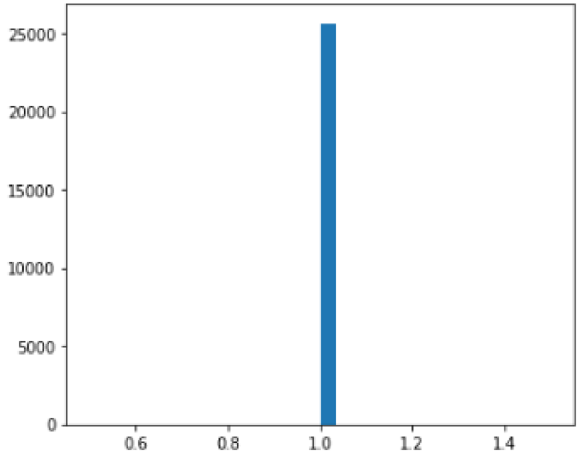
정규분포를 이용한 초기화의 경우



w 분포



wx+b



sigmoid(wx+b) 분포

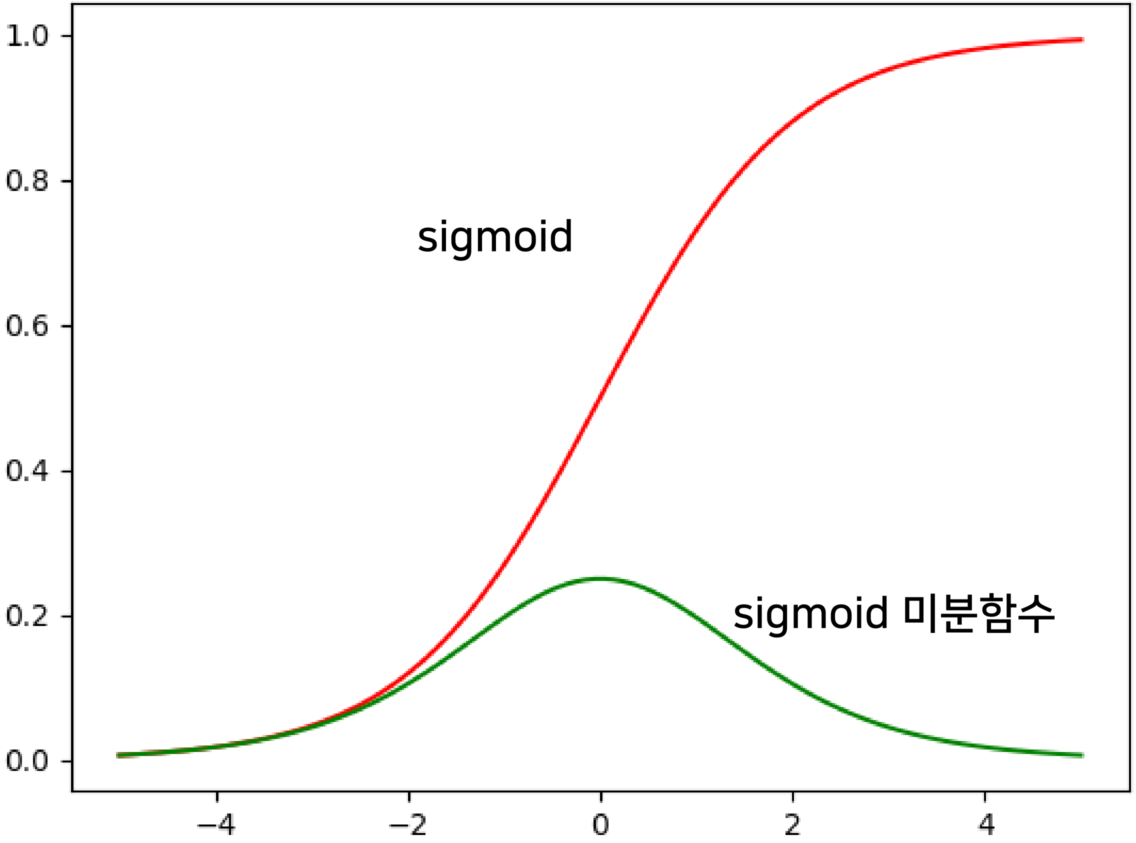
$$w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + \cdots.$$

wx+b값이 커질 경우 sigmoid 함수 적용결과는 대부분 1로 수렴

특징의 수(w의 개수)가 늘어날 경우 wx+b합이 1을 초과

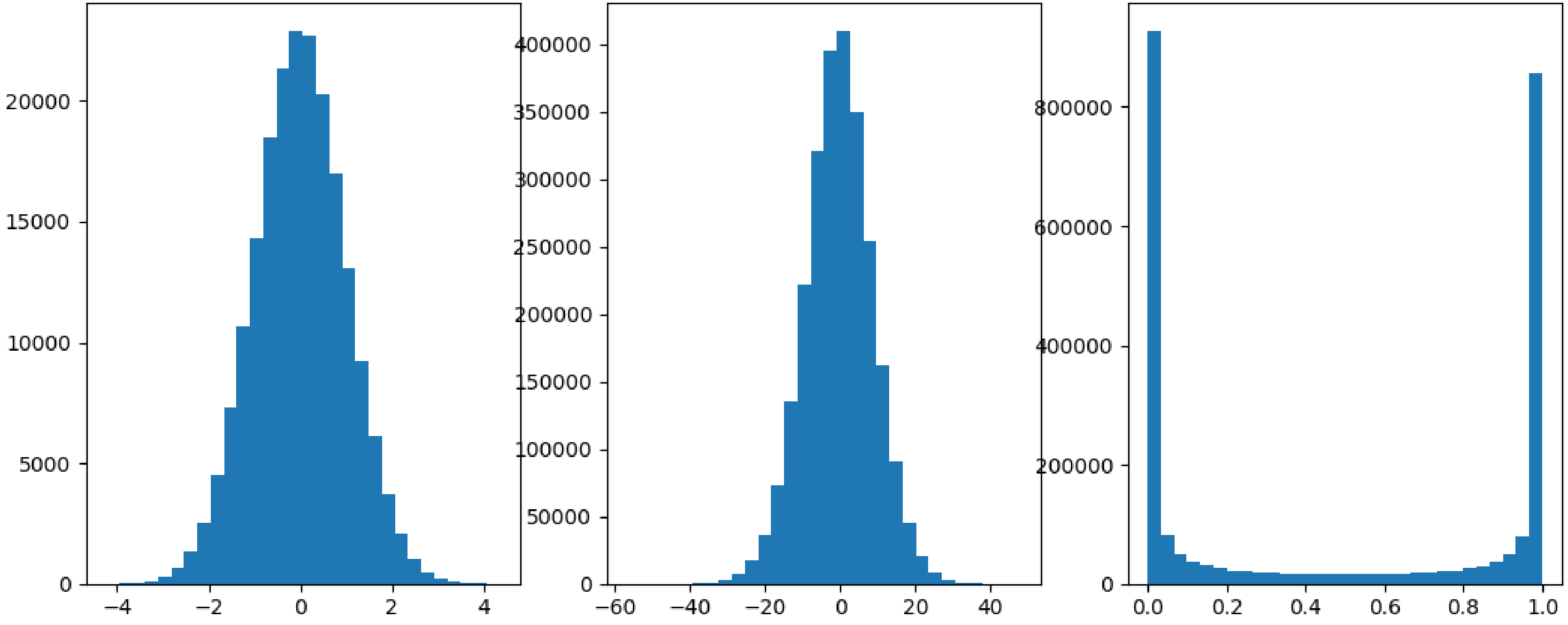
정규분포 초기화의 문제점

- 출력값이 1 근처에 모여있을 경우 가중치 값이 업데이트되지 않아 학습이 일어나지 않음



정규분포 초기화의 문제점

-4~4로 초기화할 경우



w 분포

wx+b

sigmoid(wx+b) 분포

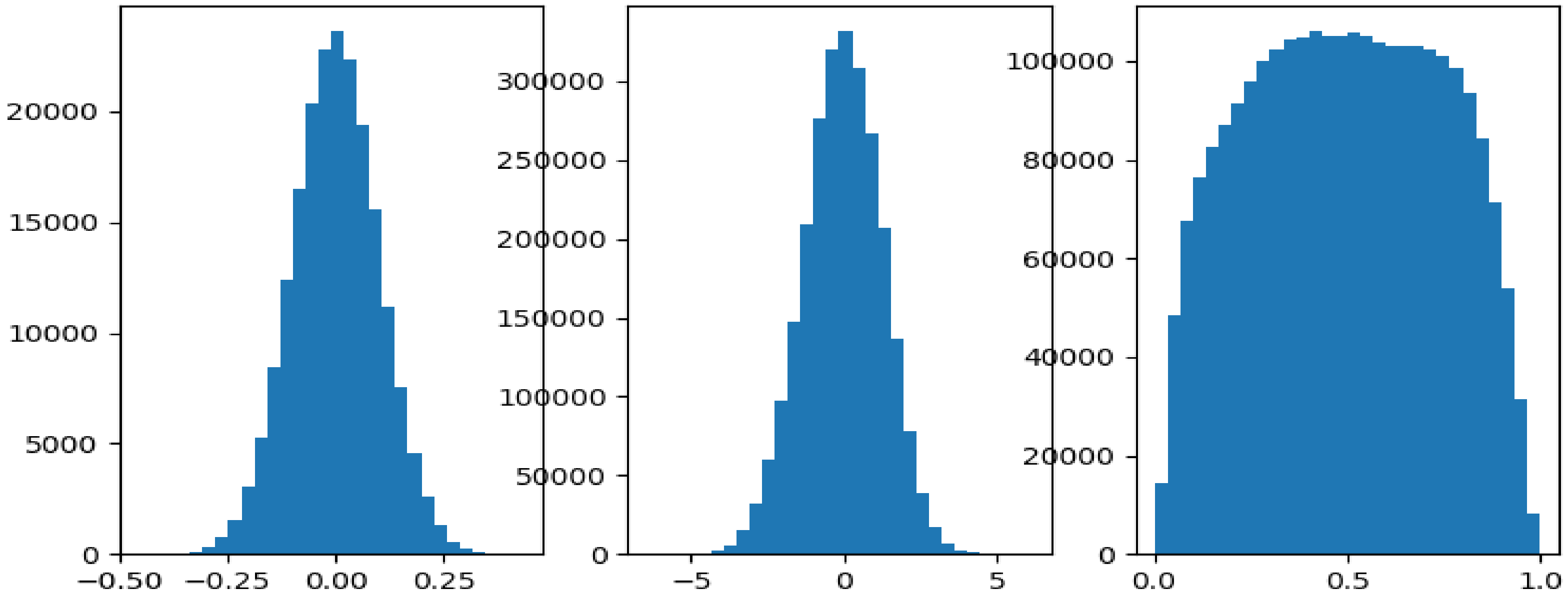
wx+b가 0을 중심으로
-40~40 사이에 분포

Sigmoid 함수 출력값이 0, 1인 경우
도함수 값이 0에 가까워 학습이 되지 않음

표준 편차를 이용한 초기화

N(0, 0.1)의 경우

– 적절한 표준편차 설정이 필요



w 분포 | wx+b | sigmoid(wx+b) 분포

wx+b가 0을 중심으로
-5~5 사이에 분포

Sigmoid 함수 출력값이 0.5에 가까워
도함수가 0이 되지 않아 학습이 이루어짐

Xavier 초기화

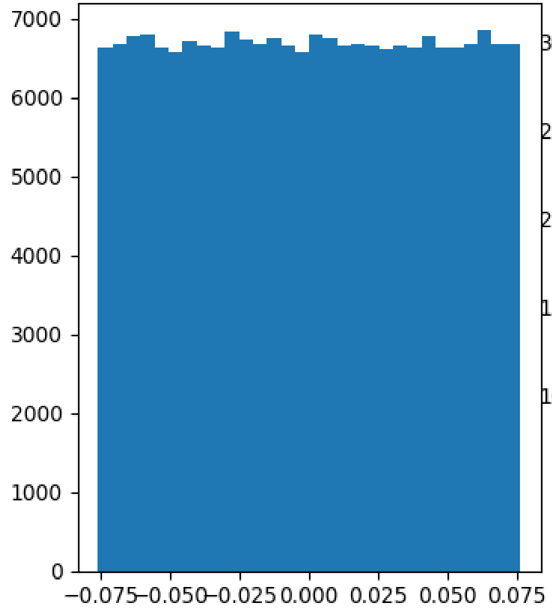
데이터의 특성(입력과 출력 노드의 수)를 이용한 초기화

1. 정규분포

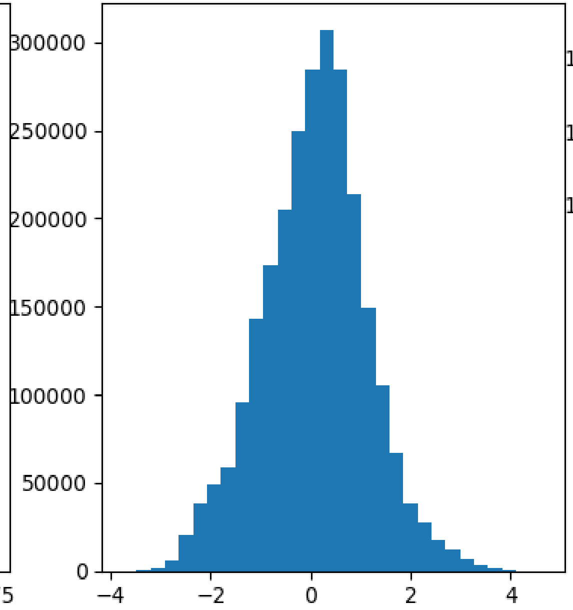
   표준편차 = $\sqrt{3.0 / (\text{input} + \text{output})}$

2. uniform 분포−range ~ +range 사이로 랜덤 초기화

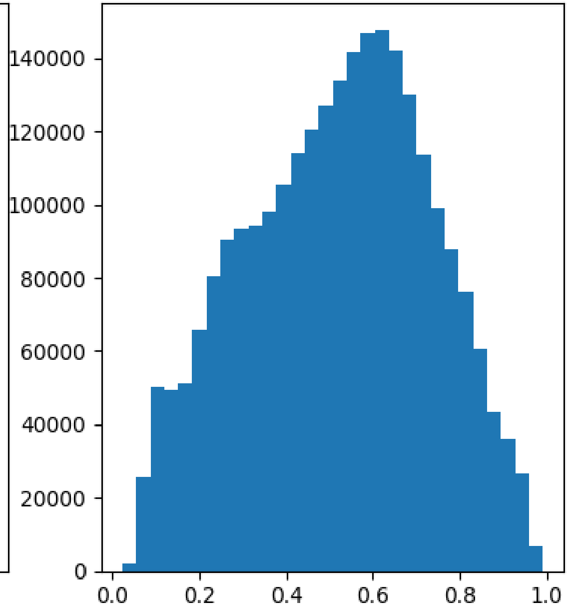
   range = $\sqrt{6.0 / (\text{input} + \text{output})}$



w 분포



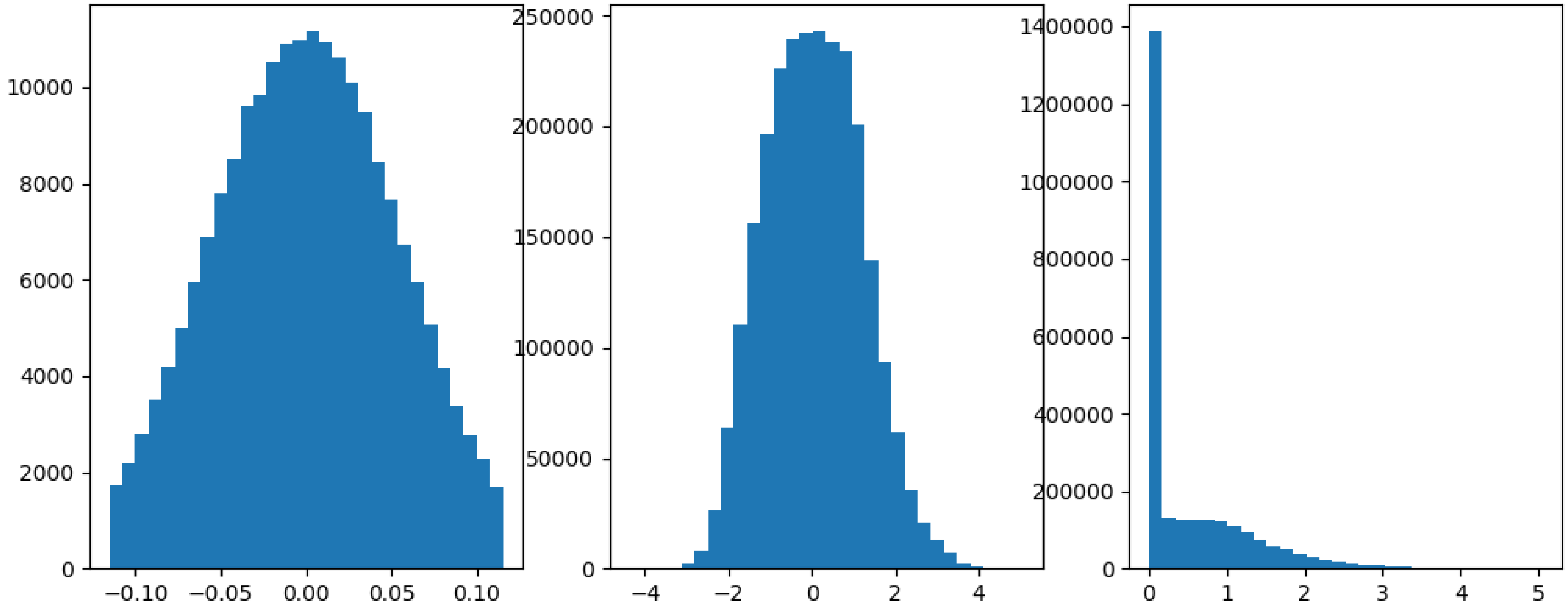
wx+b



sigmoid(wx+b) 분포

He 초기화

ReLU 활성화 함수를 사용할 때 초기화 방법
– 입력값을 반으로 나눈 제곱근을 사용

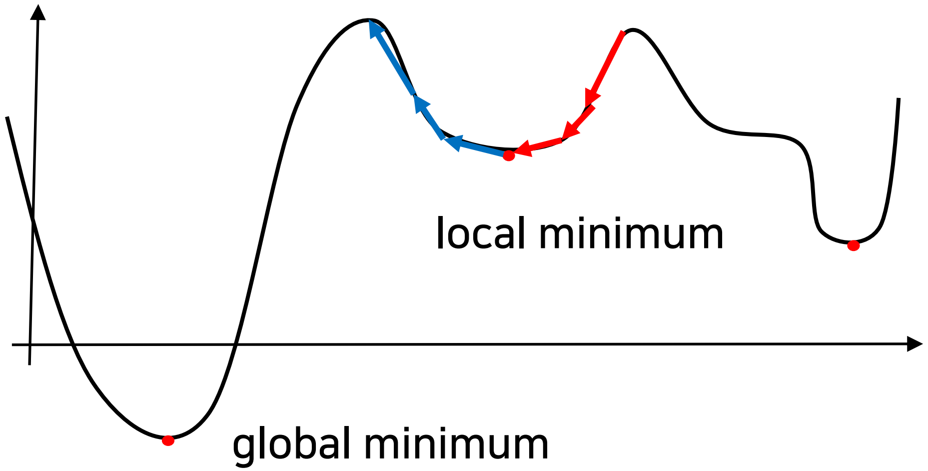


w 분포　　wx+b　　ReLU(wx+b) 분포

0보다 큰 부분은 학습이 이루어짐

모멘텀을 이용한 수렴 속도 개선

- 모멘텀(Momentum) : 진행 중인 방향으로 관성을 더해 지역 최소점에 빠지지 않도록 함
- ADAM(Adaptive Moment Estimation) : 기울기의 지수 평균과 기울기 제곱의 지수 평균
  - 속도가 클수록 기울기가 크게 업데이트 되어, 경사하강법의 단점을 보완할 수 있음



$$w_t = \mathrm{wt} - \left(\boxed{v_t} + \alpha \frac{\partial Cost}{\partial w}\right)$$

모멘텀

오래된 변화량에 작은 값, 최근 변화량에 큰 값을 주기 위해 모멘텀 상수를 지수로 곱함

$$v_t = \left(\alpha \frac{\partial Cost}{\partial w}\right)_t \boxed{r}\left(\alpha \frac{\partial Cost}{\partial w}\right)_{t-1} + \cdots + \boxed{r^n}\left(\alpha \frac{\partial Cost}{\partial w}\right)_{t-n}$$

r : 모멘텀 상수 (r<1)