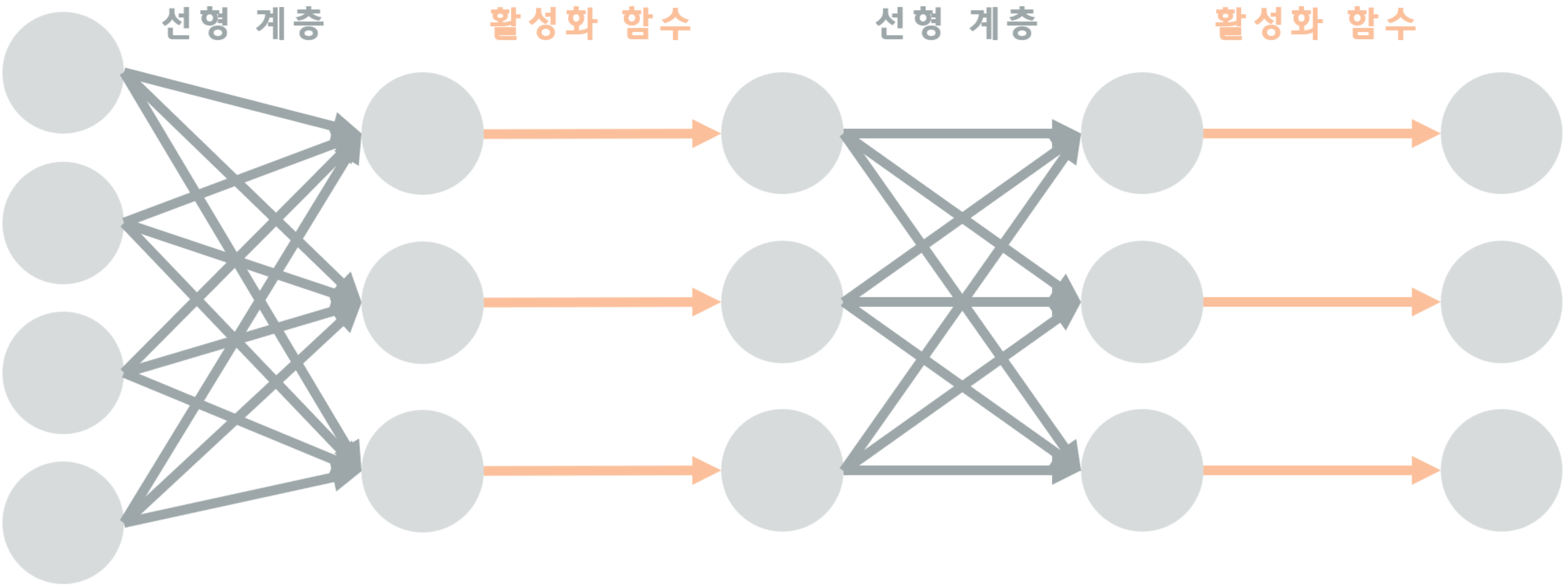
심층 신경망 1

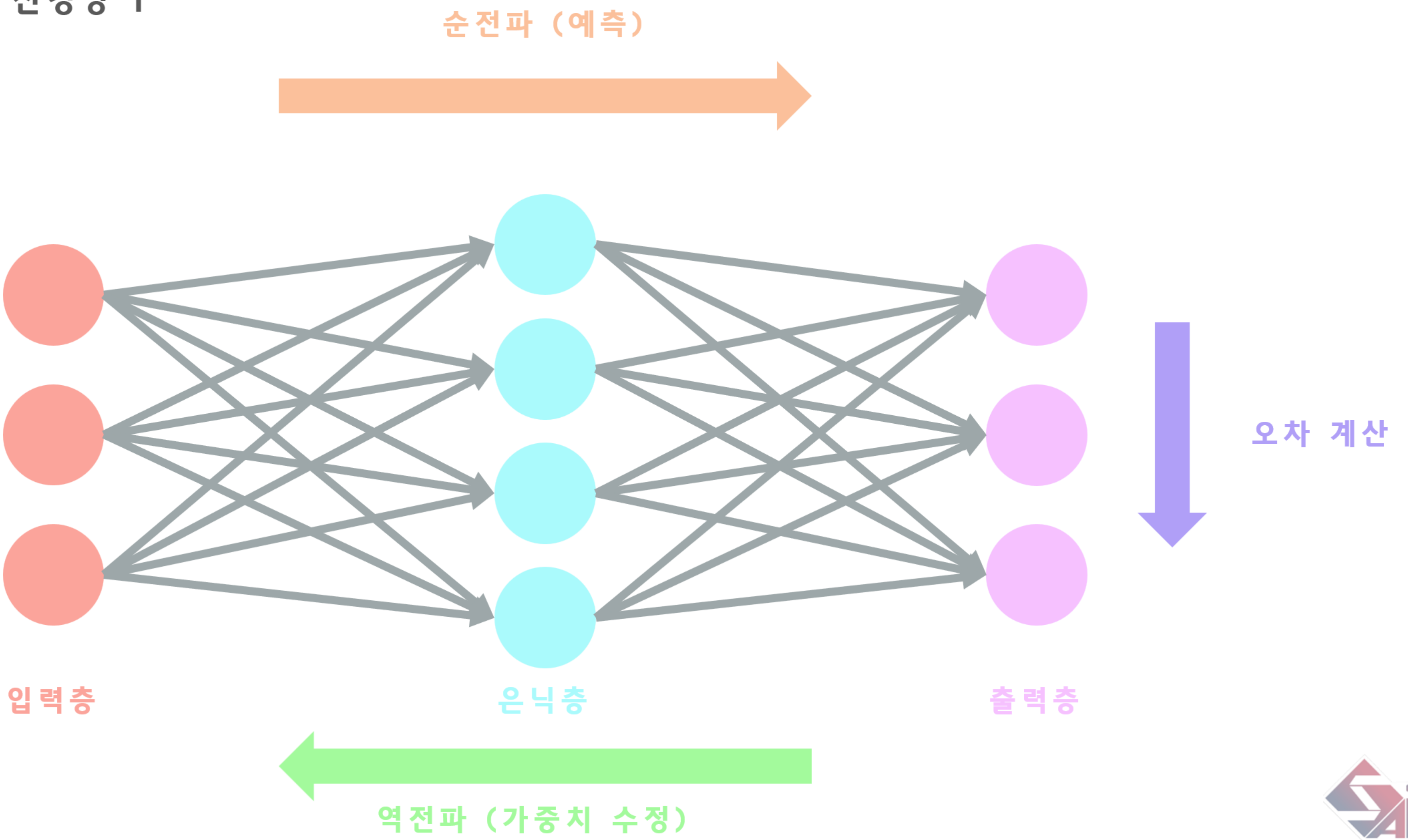
$$h = W_1 \cdot x + b_1$$

$$y = W_2 \cdot h + b_2$$

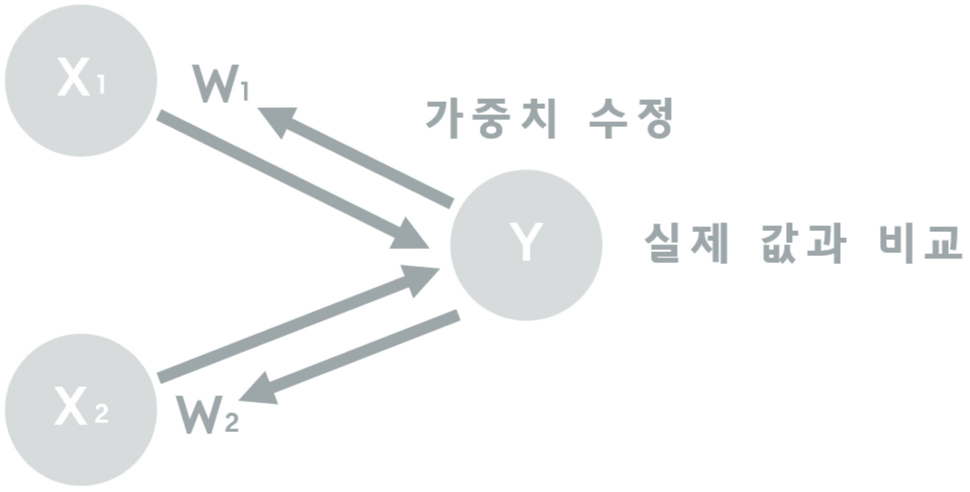
$$\begin{aligned} y &= W_2 \cdot h + b2 \\ &= W_2 \cdot (W_1 \cdot x + b_1) + b2 \\ &= W_1 \cdot W_2 \cdot x + {b_1}^\top \cdot W_2 + b_2 \\ &= W' \cdot x + b' \end{aligned}$$

→ 다시 하나의 선형 계층이 됨

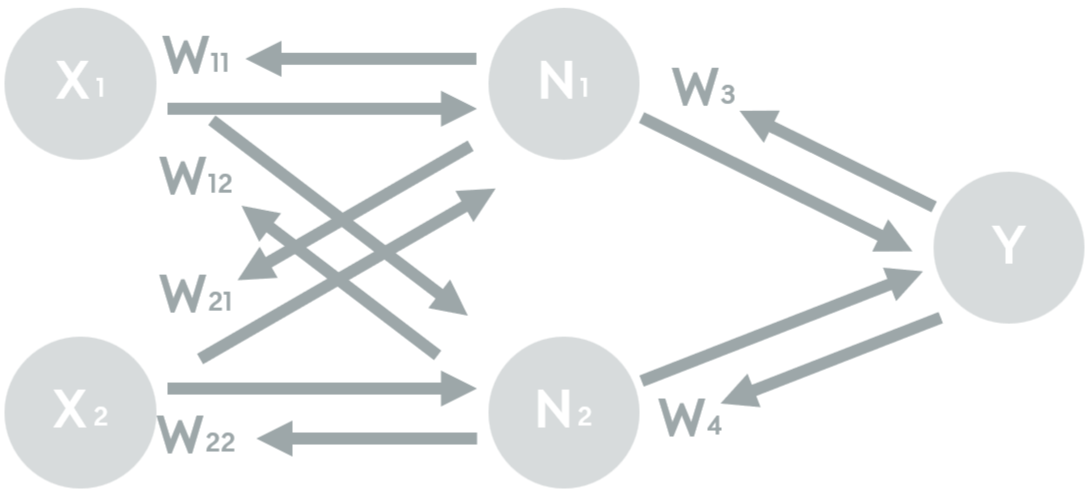


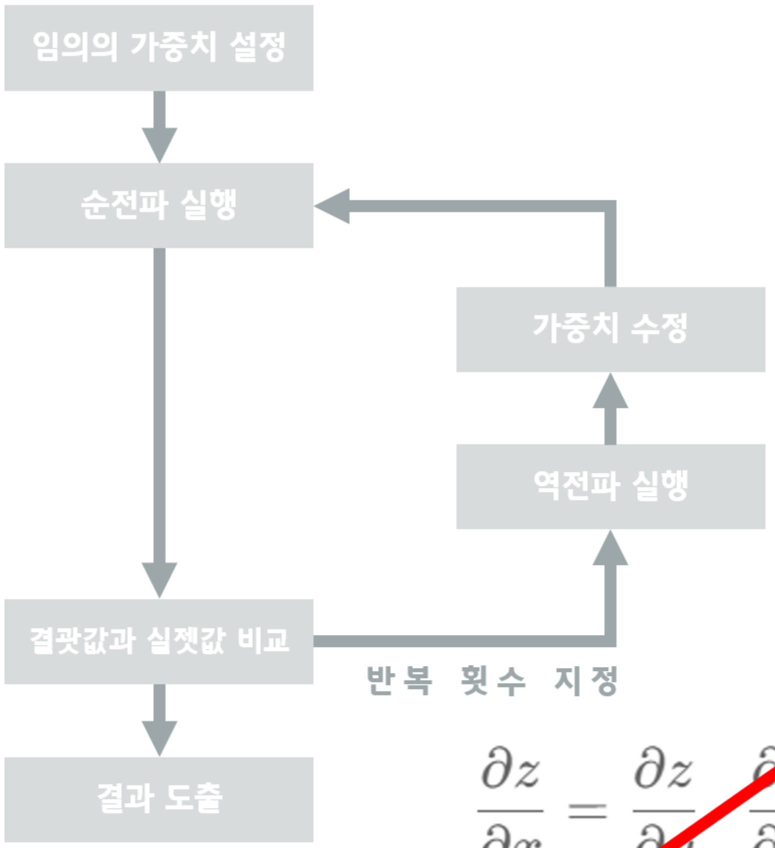

순전파 (예측)
오차 계산
입력층
은닉층
출력층
역전파 (가중치 수정)

단일 퍼셉트론



다층 퍼셉트론



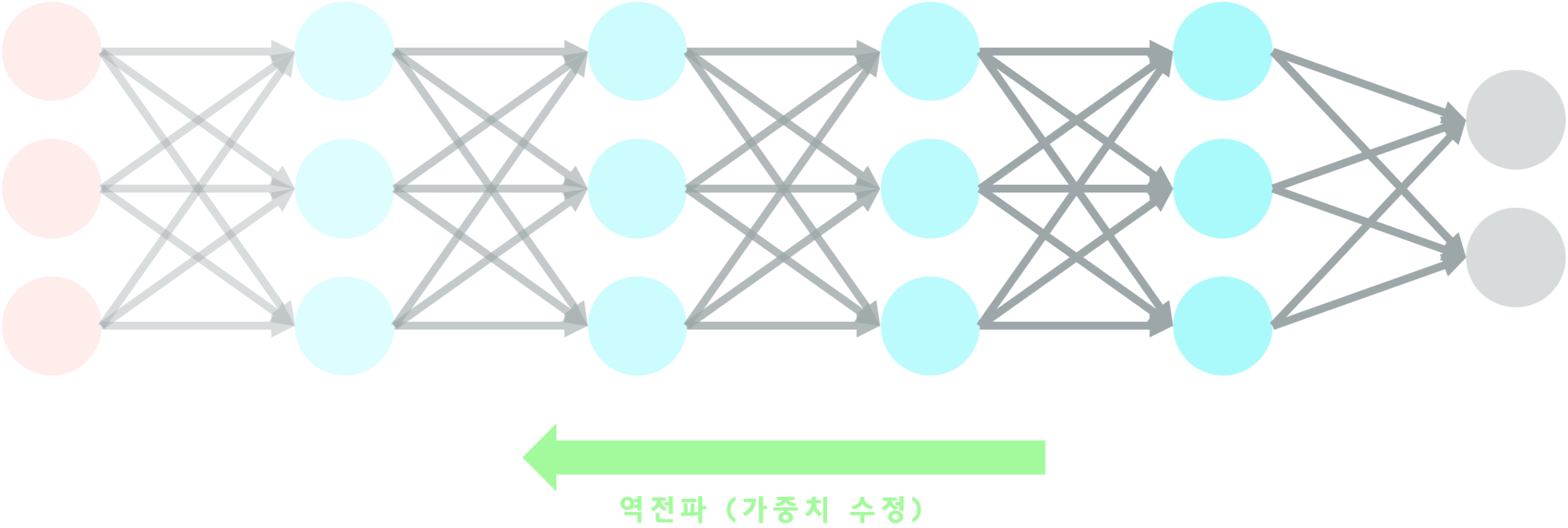


$$\frac{\partial z}{\partial x} = \frac{\partial z}{\cancel{\partial y}} \frac{\cancel{\partial y}}{\partial x}$$

연쇄 법칙

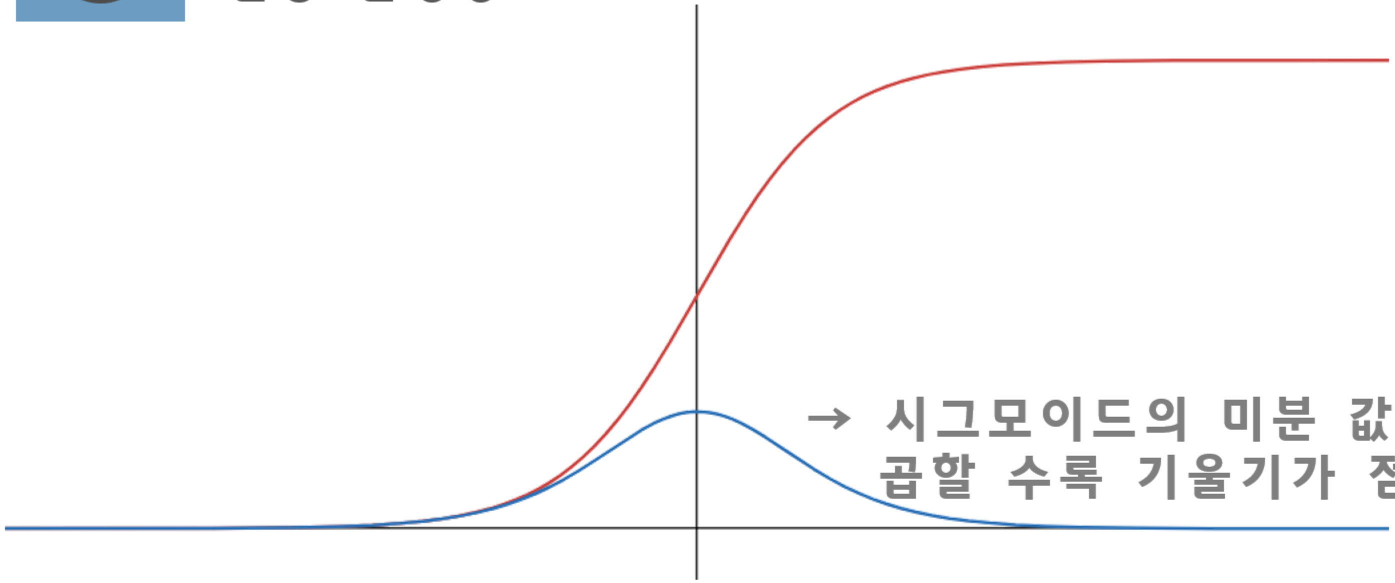
$$a = \sigma(z), \quad z = W \cdot x + b$$

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\cancel{\partial a}} \frac{\cancel{\partial a}}{\cancel{\partial z}} \frac{\cancel{\partial z}}{\partial W}$$

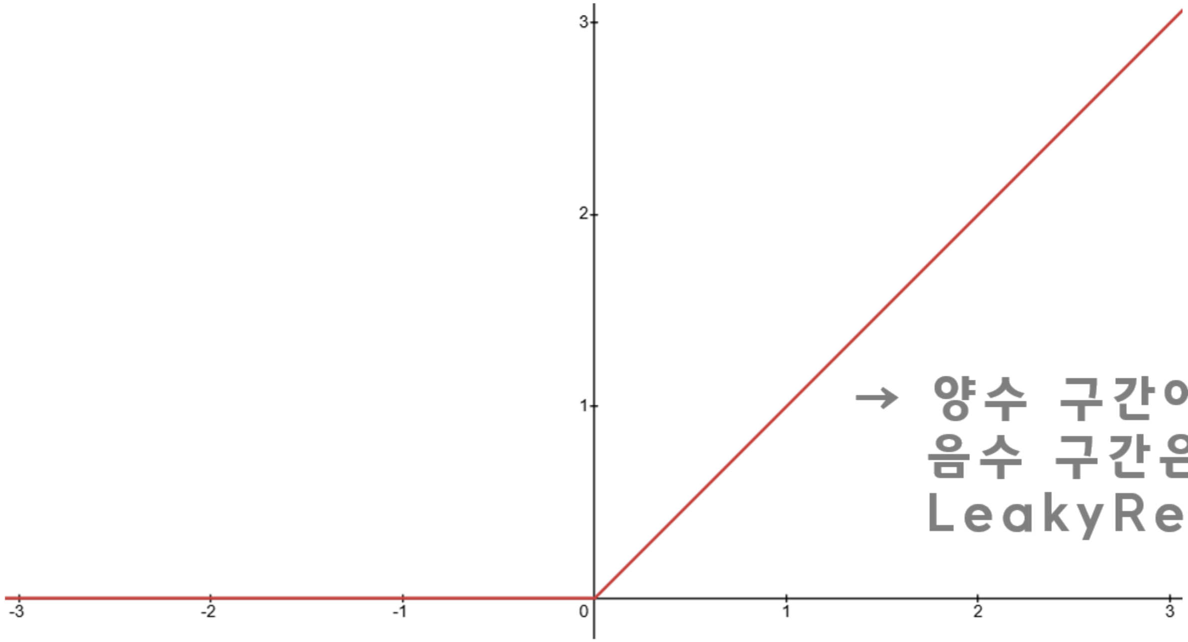

역전파 (가중치 수정)

9장

심층 신경망 1



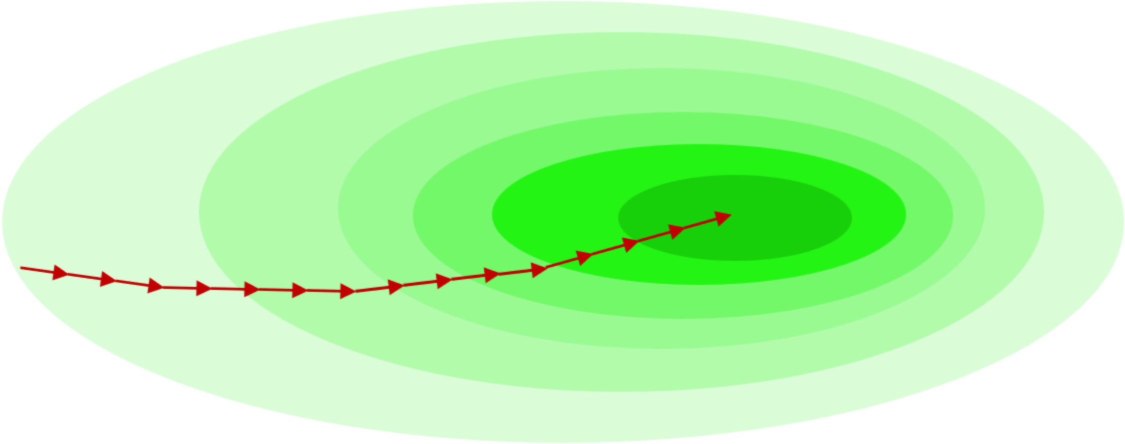
→ 시그모이드의 미분 값의 최댓값이 약 0.3...
곱할 수록 기울기가 점점 작아짐



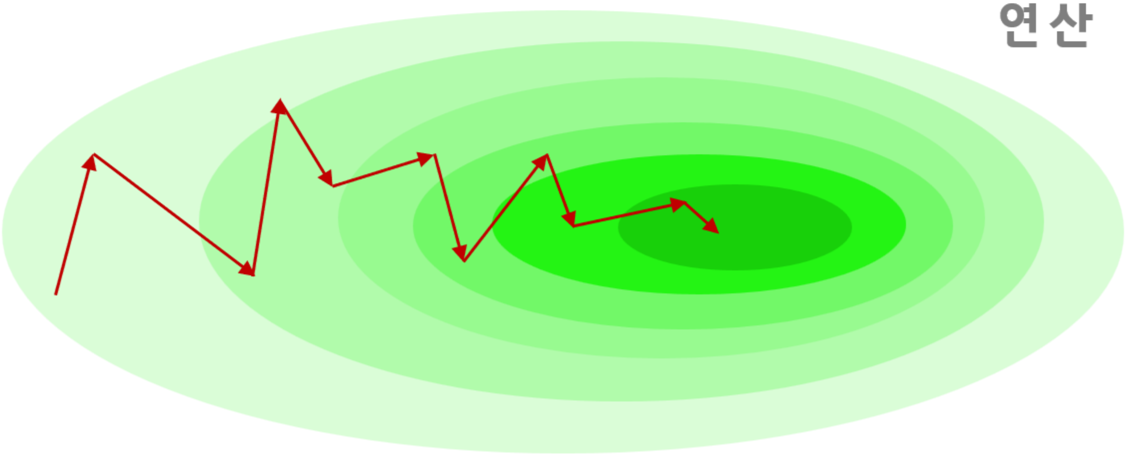
→ 양수 구간에서 기울기가 1, 빠른 최적화 가능
음수 구간은 기울기가 0이라 학습 진행 X,
LeakyReLU로 보완

확률적 경사 하강법

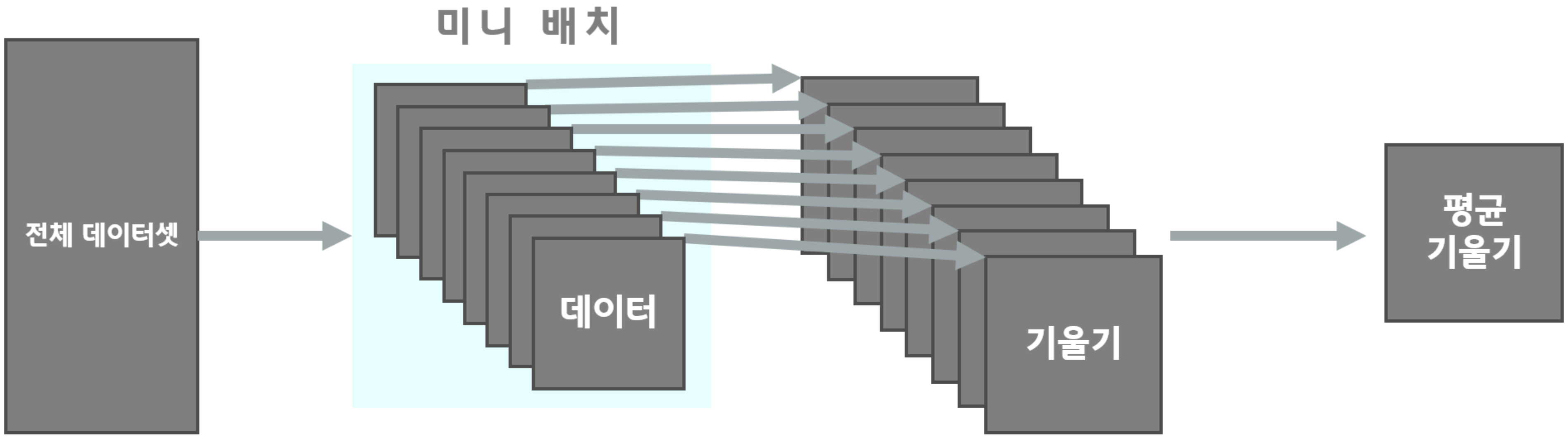
경사 하강법: 모든 데이터를 사용해 기울기 계산
안정적으로 수렴하지만 연산 횟수가 많음

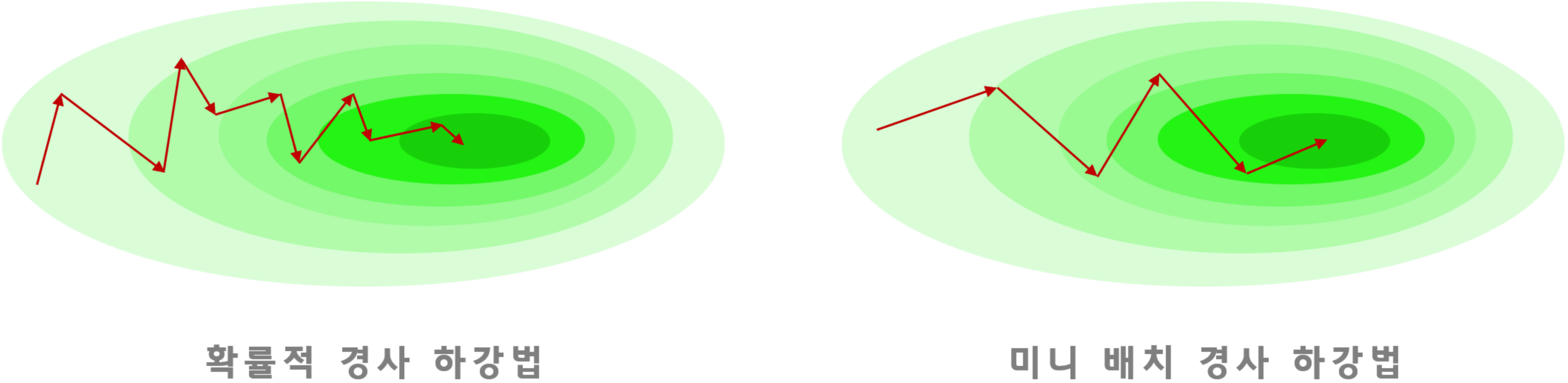


확률적 경사 하강법: 데이터를 랜덤으로 뽑아 기울기 계산
연산 횟수가 적지만 수렴이 불안정함



확률적 경사 하강법





학습률

배치 사이즈

에포크 수

최적화 함수

가중치 초기화 방법

정규화 강도

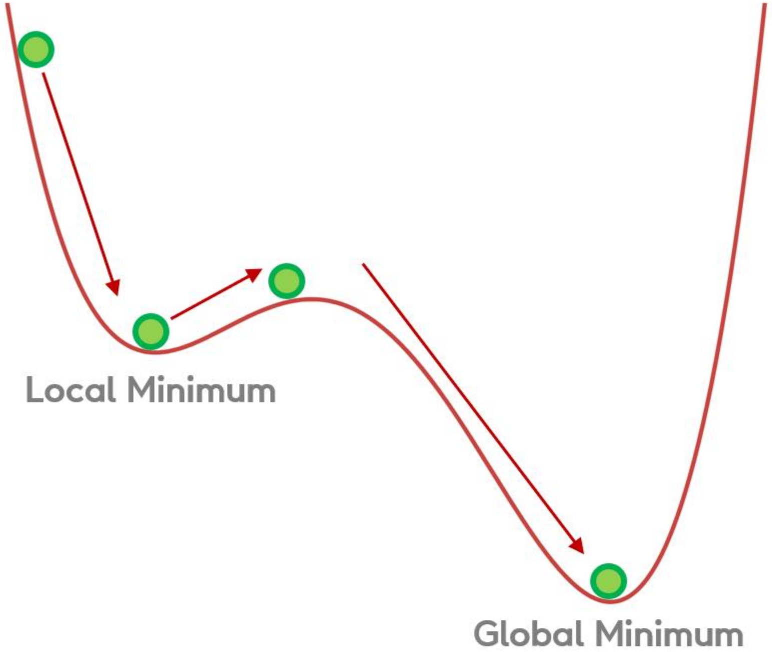
...

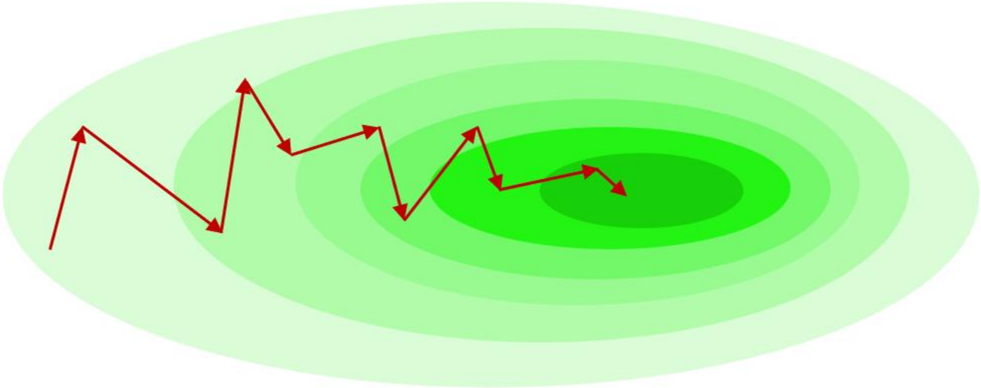
Embed Dim	Heads	Clip Len	Batch Size (Train/Val)	Epoch	Top-1 (%)

최적화

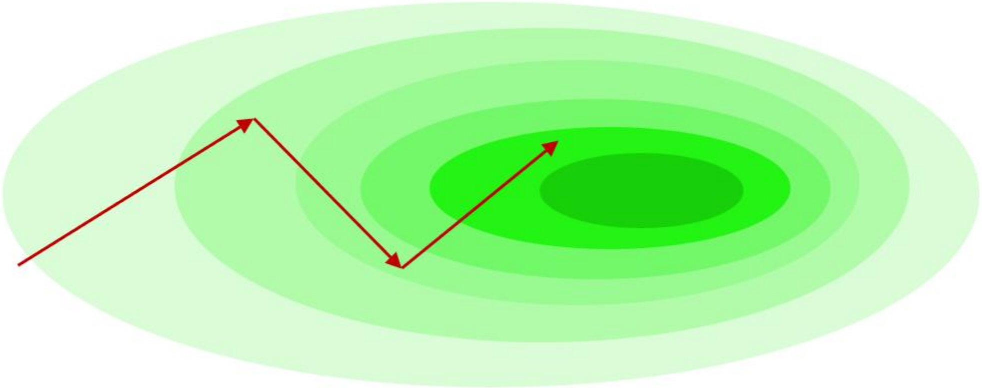
$$\mathbf{v} \leftarrow \alpha \mathbf{v} - \eta \frac{\partial L}{\partial \mathbf{W}}$$

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

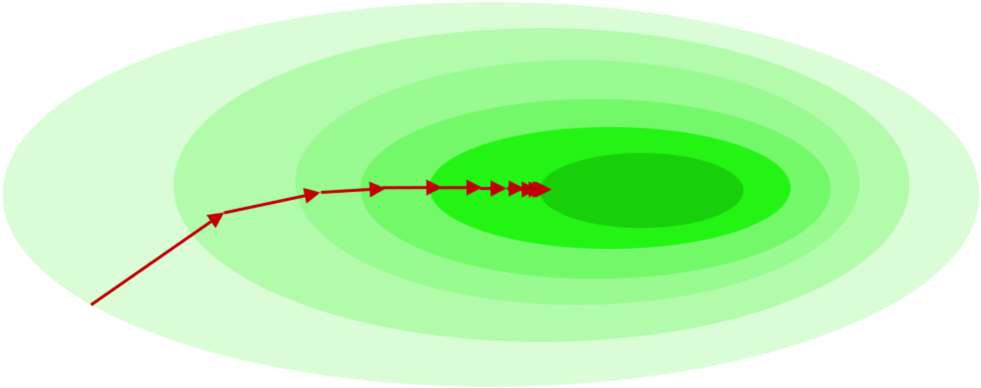




확률적 경사 하강법

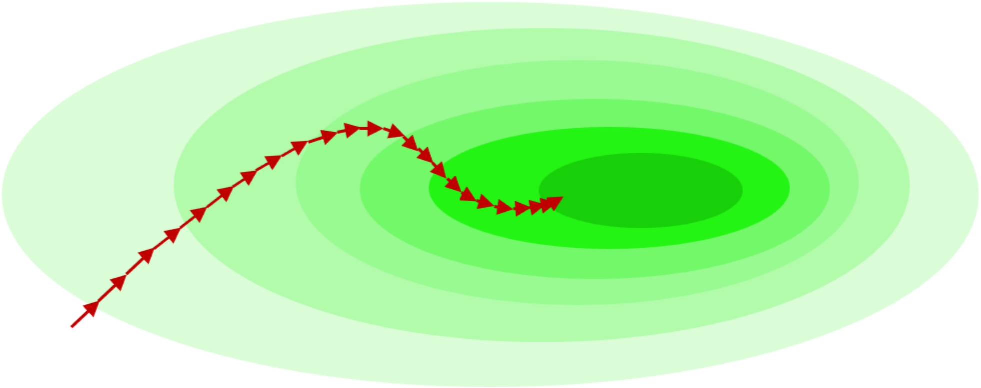


모멘텀을 사용한 경사 하강법



Adagrad

안 가본 곳은 빠르게 훑고,
많이 가본 곳은 갈수록 보폭을 줄여
세밀히 탐색하는 기법



Adam

Adagrad + 모멘텀,
이전의 변화 방향까지
고려해서 탐색하는 기법