9주차(1/3)

역전파 2

파이썬으로 배우는 기계학습

한동대학교
김영섭 교수

역전파 2

- 학습 목표
  - 역전파 과정에서 오차함수의 미분을 학습한다.
  - 오차 역전파로 각 층의 가중치를 조정한다.

- 학습 내용
  - 은닉층과 출력층 사이 $\Delta W^{[2]}$ 계산
  - $W^{[2]}$의 오차함수 미분
  - $W^{[1]}$의 오차함수 미분
  - 역전파의 가중치 조정

1. 지난 시간 복습

- 출력층의 오차 $E^{[2]}$
  - 레이블과 예측 값의 차이
  - 은닉층의 오차 $E^{[1]}$ 계산
- 가중치 조정 가능
  - 아달라인 이용
  - $W^{[1]}, W^{[2]}$ 조정

2. $W^{[2]}$의 오차함수 미분 : 1 단계

- 경사하강법 오차함수와 같은 형식
  - **가중치 W 조정 → 오차 E 최소화**
  - 오차함수를 **J**가 아닌 **E**로 표기
  - 오차함수 **E** 역시 가중치 **W**에 관한 함수

$$W^{[2]} := W^{[2]} - \alpha \Delta W^{[2]}$$
$$= W^{[2]} - \alpha \boxed{\frac{\partial E}{\partial W^{[2]}}}$$

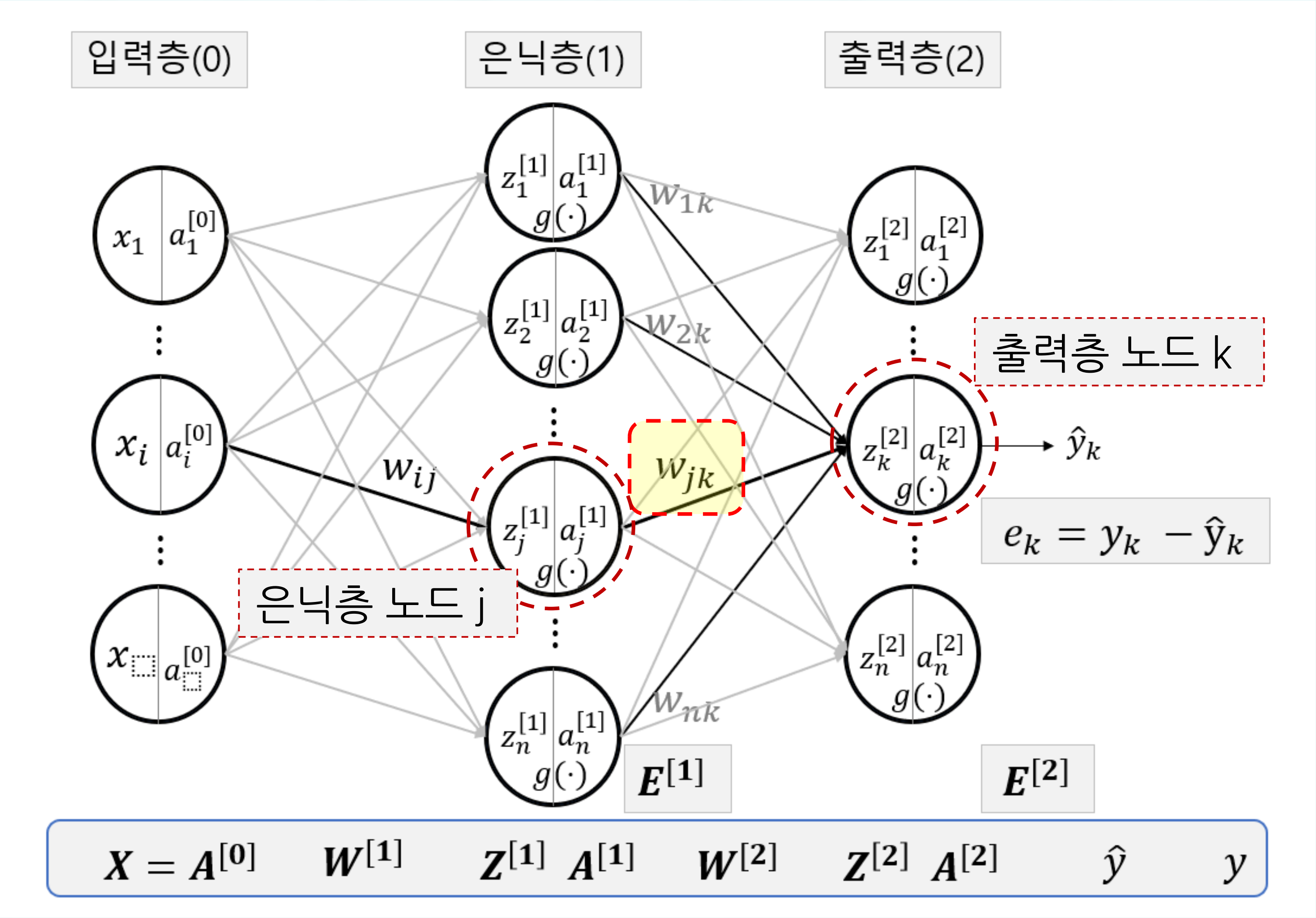
- 문제는**?**
  - 행렬 미분의 어려움
  - 해결책: $w_{jk}^{[2]}$

2. $W^{[2]}$의 오차함수 미분 : 1 단계

$$W^{[2]} := W^{[2]} - \alpha \Delta W^{[2]}$$

$$= W^{[2]} - \alpha \boxed{\frac{\partial E}{\partial W^{[2]}}}$$

- $w_{jk}^{[2]}$:
  은닉층 노드 j 와
  출력층 노드 **k** 사이 가중치
  (층번호 생략하기도 함)



2. $W^{[2]}$의 오차함수 미분 : 2 단계

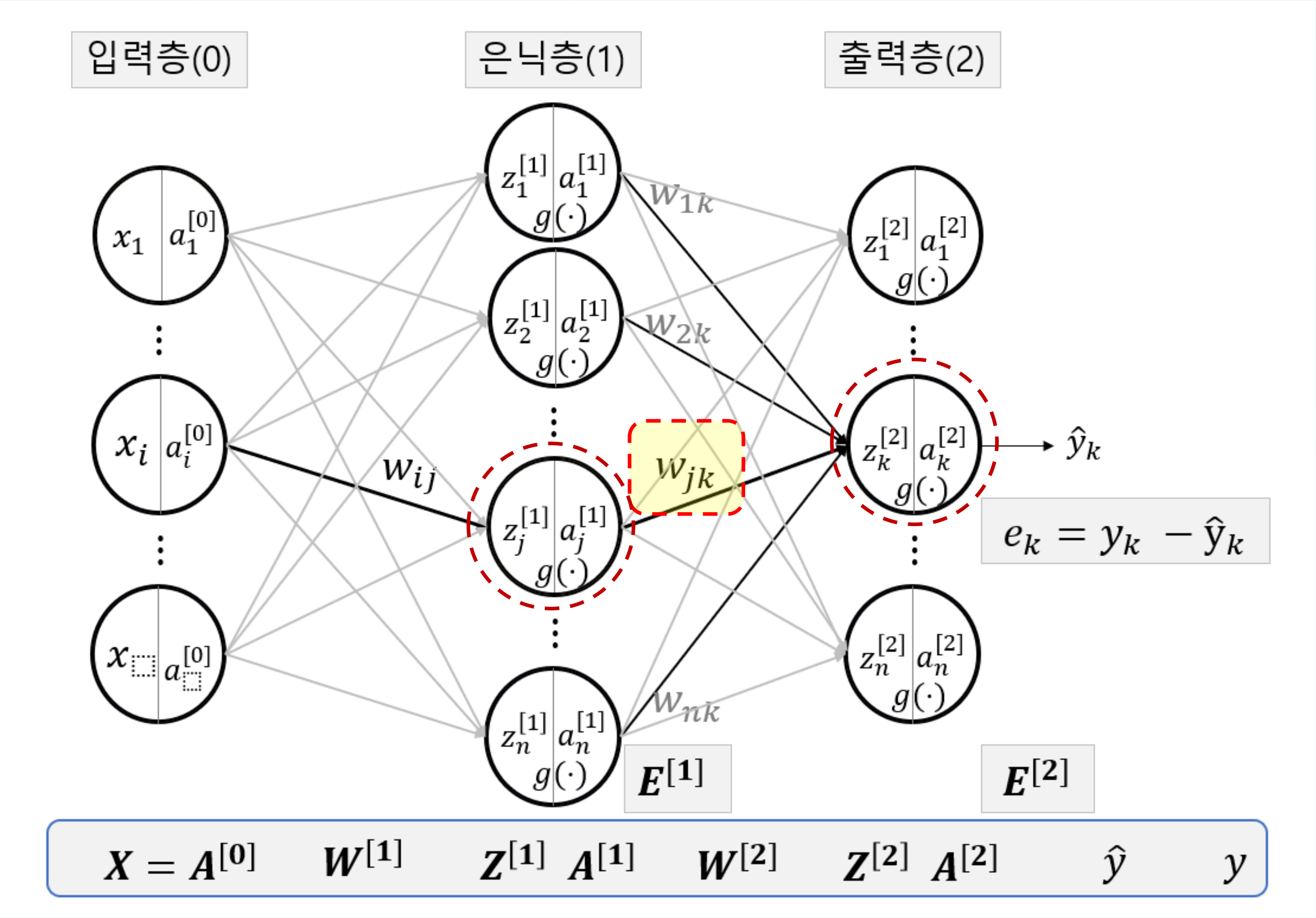
$$W^{[2]} := W^{[2]} - \alpha \Delta W^{[2]}$$

$$= W^{[2]} - \alpha \frac{\partial E}{\partial W^{[2]}}$$

$$w_{jk}^{[2]} := w_{jk}^{[2]} - \alpha \Delta w_{jk}^{[2]}$$

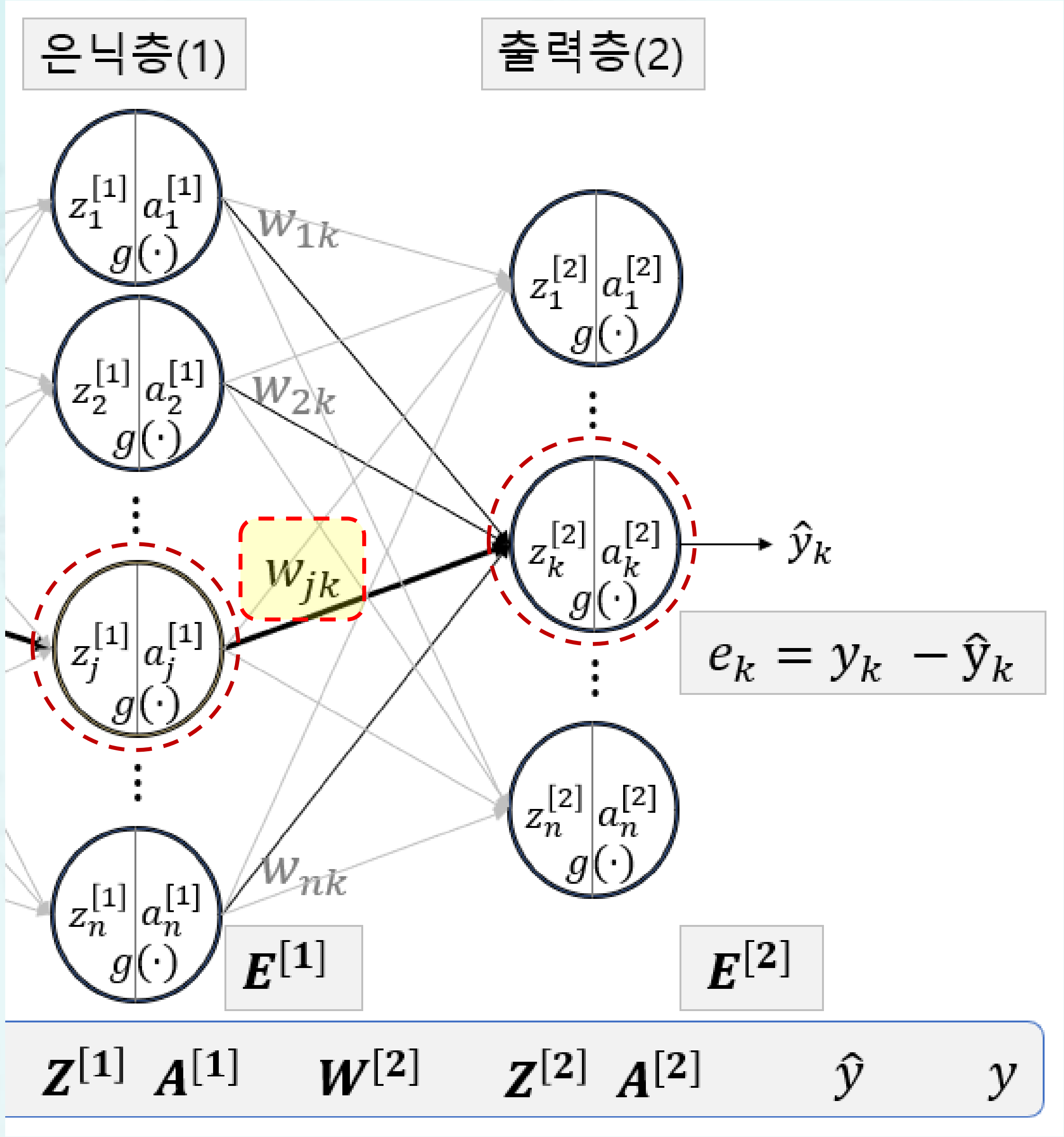
$$= w_{jk}^{[2]} - \alpha \boxed{\frac{\partial E}{\partial w_{jk}^{[2]}}}$$

2 단계



2. $W^{[2]}$의 오차함수 미분 : 2 단계

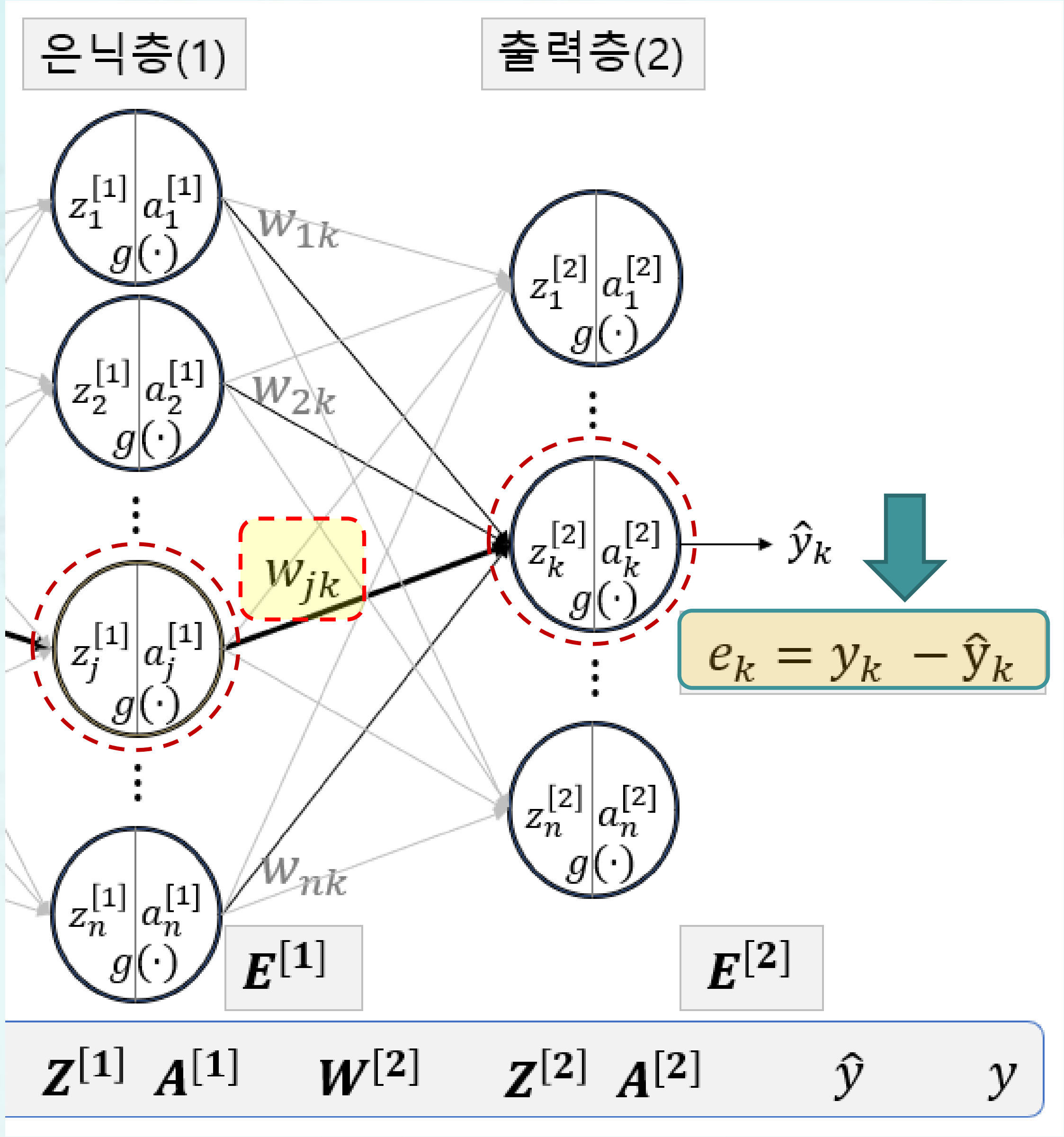
$$\frac{\partial E}{\partial w_{jk}^{[2]}} = \boxed{\frac{\partial}{\partial w_{jk}} \frac{1}{2} \sum_{m=1}^{n} (y_m - \hat{y}_m)^2}$$



2. $W^{[2]}$의 오차함수 미분 : 2 단계

$$\frac{\partial E}{\partial w_{jk}^{[2]}} = \frac{\partial}{\partial w_{jk}} \frac{1}{2} \sum_{m=1}^{n} (y_m - \hat{y}_m)^2$$

$$= \boxed{\frac{\partial}{\partial w_{jk}} \frac{1}{2} (y_k - \hat{y}_k)^2}$$



2. $W^{[2]}$의 오차함수 미분 : 2 단계

$$\frac{\partial E}{\partial w_{jk}^{[2]}} = \frac{\partial}{\partial w_{jk}} \frac{1}{2} \sum_{m=1}^{n} (y_m - \hat{y}_m)^2$$

$$= \frac{\partial}{\partial w_{jk}} \frac{1}{2} (y_k - \hat{y}_k)^2$$

$$= \boxed{\frac{1}{2} \cdot 2(y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (y_k - \hat{y}_k)}$$

합성함수 미분법

$f(g(x))' = f'(g(x))g'(x)$

2. $W^{[2]}$의 오차함수 미분 : 2 단계

$$\frac{\partial E}{\partial w_{jk}^{[2]}} = \frac{\partial}{\partial w_{jk}} \frac{1}{2} \sum_{m=1}^{n} (y_m - \hat{y}_m)^2$$

$$= \frac{\partial}{\partial w_{jk}} \frac{1}{2} (y_k - \hat{y}_k)^2$$

$$= \frac{1}{\cancel{2}} \cdot \overset{1}{\cancel{2}} (y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (\overset{0}{\cancel{y_k}} - \hat{y}_k)$$

$$= \boxed{(y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (-\hat{y}_k)}$$

합성함수 미분법

$f(g(x))' = f'(g(x))g'(x)$

2. $W^{[2]}$의 오차함수 미분 : 2 단계

$$\frac{\partial E}{\partial w_{jk}^{[2]}} = \frac{\partial}{\partial w_{jk}} \frac{1}{2} \sum_{m=1}^{n} (y_m - \hat{y}_m)^2$$

$$= \frac{\partial}{\partial w_{jk}} \frac{1}{2} (y_k - \hat{y}_k)^2$$

$$= \frac{1}{2} \cdot 2(y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (y_k - \hat{y}_k)$$

$$= (y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (-\hat{y}_k)$$

$$= \boxed{-(y_k - \hat{y}_k) \frac{\partial \hat{y}_k}{\partial w_{jk}}}$$

2. $W^{[2]}$의 오차함수 미분 : 2 단계

$$\frac{\partial E}{\partial w_{jk}^{[2]}} = \frac{\partial}{\partial w_{jk}} \frac{1}{2} \sum_{m=1}^{n} (y_m - \hat{y}_m)^2$$

$$= \frac{\partial}{\partial w_{jk}} \frac{1}{2} (y_k - \hat{y}_k)^2$$

$$= \frac{1}{2} \cdot 2(y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (y_k - \hat{y}_k)$$

$$= (y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (-\hat{y}_k)$$

$$= -(y_k - \hat{y}_k) \boxed{\frac{\partial \hat{y}_k}{\partial w_{jk}}}$$

2. $W^{[2]}$의 오차함수 미분 : 2 단계

$$\frac{\partial E}{\partial w_{jk}^{[2]}} = \frac{\partial}{\partial w_{jk}} \frac{1}{2} \sum_{m=1}^{n} (y_m - \hat{y}_m)^2$$

$$= \frac{\partial}{\partial w_{jk}} \frac{1}{2} (y_k - \hat{y}_k)^2$$

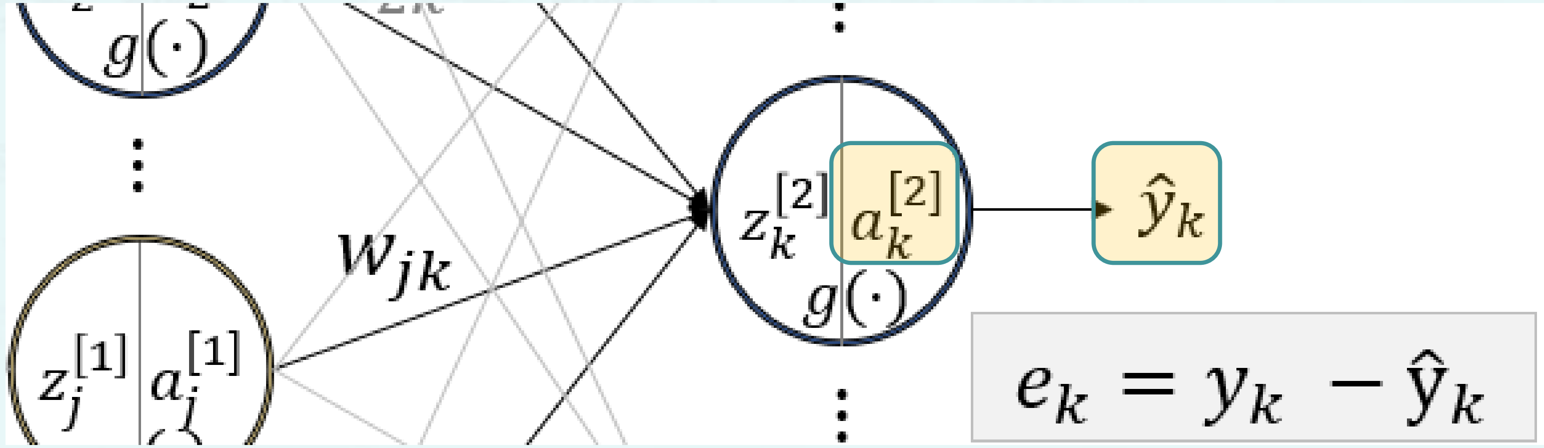
$$= \frac{1}{2} \cdot 2(y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (y_k - \hat{y}_k)$$

$$= (y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (-\hat{y}_k)$$

$$= -(y_k - \hat{y}_k) \boxed{\frac{\partial \hat{y}_k}{\partial w_{jk}}}$$

- 출력층 노드 **k**의 출력 $\hat{y}_k$의 미분

$$\frac{\partial \hat{y}_k}{\partial w_{jk}} = \boxed{}$$



$e_k = y_k - \hat{y}_k$

2. $W^{[2]}$의 오차함수 미분 : 2 단계

$$\frac{\partial E}{\partial w_{jk}^{[2]}} = \frac{\partial}{\partial w_{jk}} \frac{1}{2} \sum_{m=1}^{n} (y_m - \hat{y}_m)^2$$

$$= \frac{\partial}{\partial w_{jk}} \frac{1}{2} (y_k - \hat{y}_k)^2$$

$$= \frac{1}{2} \cdot 2(y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (y_k - \hat{y}_k)$$

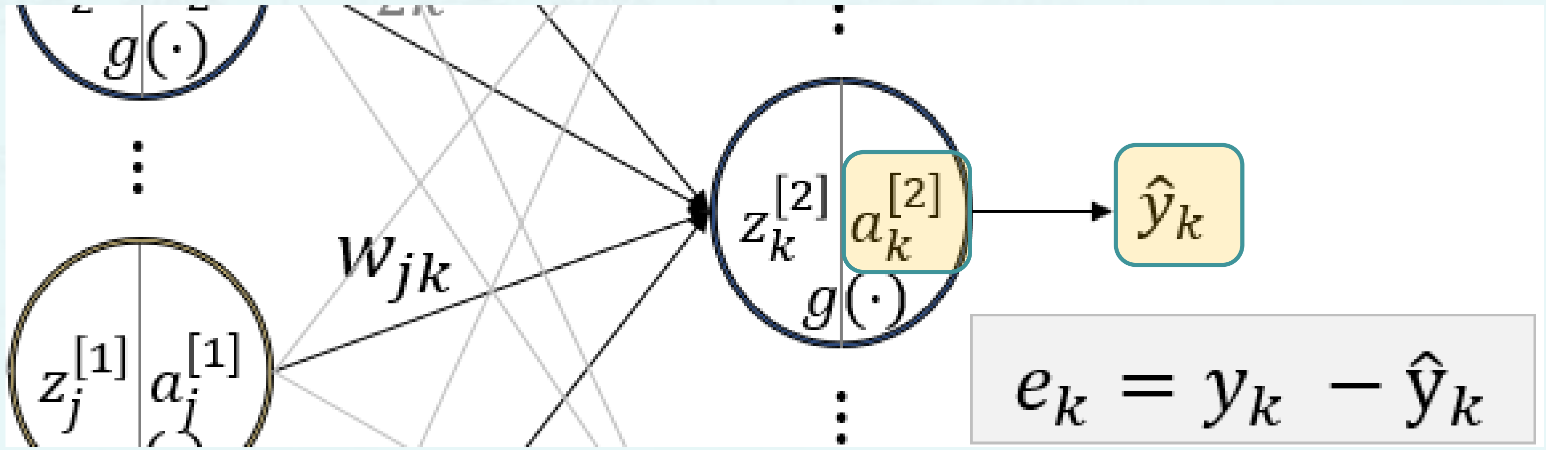
$$= (y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (-\hat{y}_k)$$

$$= -(y_k - \hat{y}_k) \boxed{\frac{\partial \hat{y}_k}{\partial w_{jk}}}$$

- 출력층 노드 **k**의 출력 $\hat{y}_k$의 미분

$$\frac{\partial \hat{y}_k}{\partial w_{jk}} = \frac{\partial}{\partial w_{jk}} a_k^{[2]}$$

$$= \boxed{}$$



2. $W^{[2]}$의 오차함수 미분 : 2 단계

$$\frac{\partial E}{\partial w_{jk}^{[2]}} = \frac{\partial}{\partial w_{jk}} \frac{1}{2} \sum_{m=1}^{n} (y_m - \hat{y}_m)^2$$

$$= \frac{\partial}{\partial w_{jk}} \frac{1}{2} (y_k - \hat{y}_k)^2$$

$$= \frac{1}{2} \cdot 2(y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (y_k - \hat{y}_k)$$

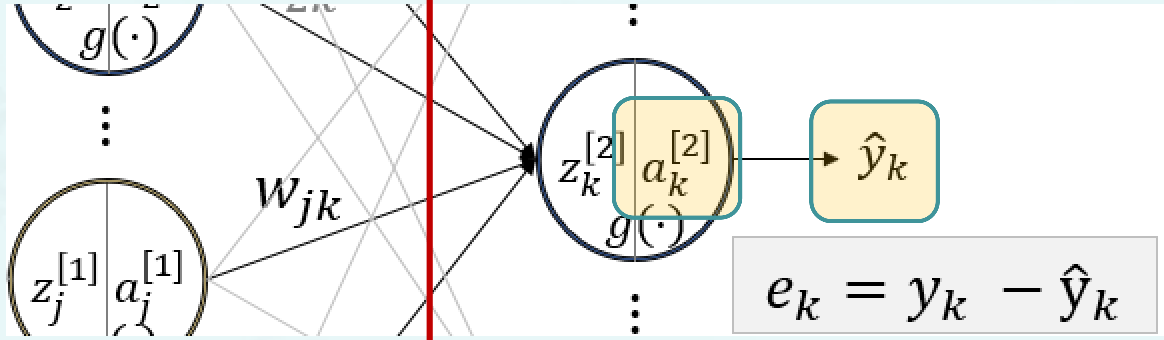
$$= (y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (-\hat{y}_k)$$

$$= -(y_k - \hat{y}_k) \boxed{\frac{\partial \hat{y}_k}{\partial w_{jk}}}$$

- 출력층 노드 **k**의 출력 $\hat{y}_k$의 미분

$$\frac{\partial \hat{y}_k}{\partial w_{jk}} = \frac{\partial}{\partial w_{jk}} a_k^{[2]}$$

$$= \boxed{\frac{\partial}{\partial w_{jk}} g(z_k^{[2]})}$$



2. $W^{[2]}$의 오차함수 미분 : 3 단계

- 3단계

$$\frac{\partial E}{\partial w_{jk}} = -(y_k - \hat{y}_k) \cdot \boxed{\frac{\partial}{\partial w_{jk}} g(z_k)}$$

- 1단계

$$w_{jk}^{[2]} := w_{jk}^{[2]} - \alpha \Delta w_{jk}^{[2]}$$

$$= w_{jk}^{[2]} - \alpha \boxed{\frac{\partial E}{\partial w_{jk}^{[2]}}}$$

- 2단계

$$\frac{\partial E}{\partial w_{jk}^{[2]}} = \boxed{\frac{\partial}{\partial w_{jk}} \frac{1}{2} \sum_{m=1}^{n} (y_m - \hat{y}_m)^2}$$

2. $W^{[2]}$의 오차함수 미분 : 3 단계

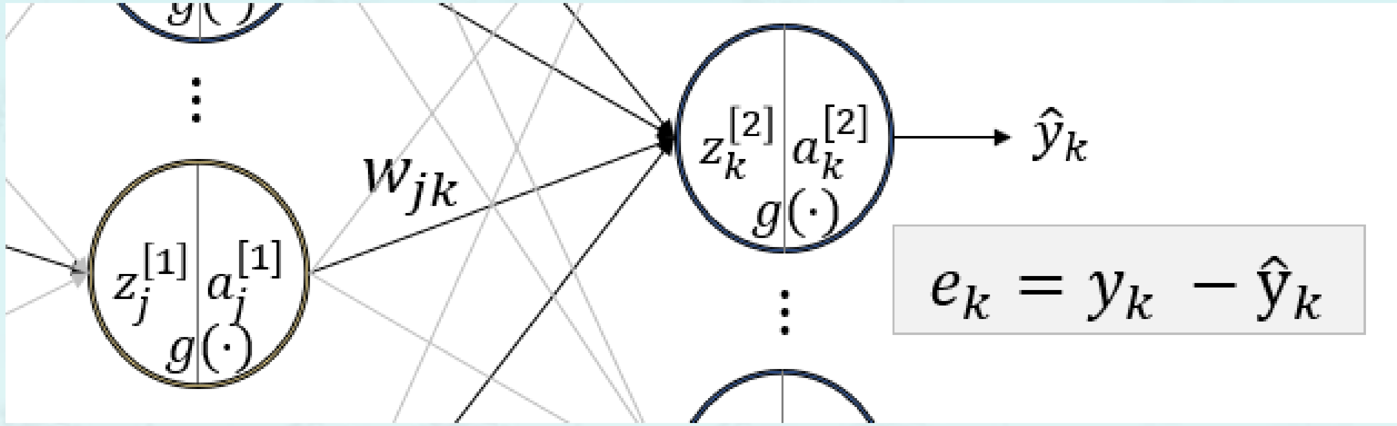
$$\frac{\partial E}{\partial w_{jk}} = -(y_k - \hat{y}_k) \cdot \frac{\partial}{\partial w_{jk}} g(z_k)$$

$$= -(y_k - \hat{y}_k) \cdot \boxed{g'(z_k) \frac{\partial z_k}{\partial w_{jk}}}$$

합성함수 미분법

$$u(v(x))' = u'(v(x))v'(x)$$

2. $W^{[2]}$의 오차함수 미분 : 3 단계



$$\frac{\partial E}{\partial w_{jk}} = -(y_k - \hat{y}_k) \cdot \frac{\partial}{\partial w_{jk}} g(z_k)$$

$$= -(y_k - \hat{y}_k) \cdot g'(z_k) \frac{\partial z_k}{\partial w_{jk}}$$

$$= -(y_k - \hat{y}_k) \cdot g'(z_k) \boxed{\frac{\partial}{\partial w_{jk}} (\sum_j w_{jk} \cdot a_j)}$$

$$\because z_k = \sum_j w_{jk}^{[2]} a_j^{[1]}$$

2. $W^{[2]}$의 오차함수 미분 : 3 단계

$$
\begin{aligned}
\frac{\partial E}{\partial w_{jk}} &= -(y_k - \hat{y}_k) \cdot \frac{\partial}{\partial w_{jk}} g(z_k) \\
&= -(y_k - \hat{y}_k) \cdot g'(z_k) \frac{\partial z_k}{\partial w_{jk}} \\
&= -(y_k - \hat{y}_k) \cdot g'(z_k) \frac{\partial}{\partial w_{jk}} (\sum_j w_{jk} \cdot a_j) \\
&= \boxed{-(y_k - \hat{y}_k) \cdot g'(z_k) \cdot a_j}
\end{aligned}
$$

2. $W^{[2]}$의 오차함수 미분 : 3 단계

$$\frac{\partial E}{\partial w_{jk}} = -(y_k - \hat{y}_k) \cdot \frac{\partial}{\partial w_{jk}} g(z_k)$$

$$= -(y_k - \hat{y}_k) \cdot g'(z_k) \frac{\partial z_k}{\partial w_{jk}}$$

$$= -(y_k - \hat{y}_k) \cdot g'(z_k) \frac{\partial}{\partial w_{jk}} (\sum_j w_{jk} \cdot a_j)$$

$$= -(y_k - \hat{y}_k) \cdot g'(z_k) \cdot a_j$$

$$= \boxed{-(y_k - \hat{y}_k) \cdot \sigma(z_k)\big(1 - \sigma(z_k)\big) \cdot a_j} \qquad \boxed{if\ g(x) = \sigma(x)}$$

2. $W^{[2]}$의 오차함수 미분 : 3 단계 각 항 설명

$$\Delta w_{jk}^{[2]} = \frac{\partial E}{\partial w_{jk}} = -(y_k - \hat{y}_k) \cdot g'(z_k) \cdot a_j$$

- 오차: 출력층 **k** 노드에서 레이블과 예측값의 차이
- 활성화 함수 미분에 z_k를 적용한 값
  - z_k: 출력층 노드 **k**의 순입력
- a_j : 은닉층 노드 **j**의 출력

2. $W^{[2]}$의 오차함수 미분 : 4 단계

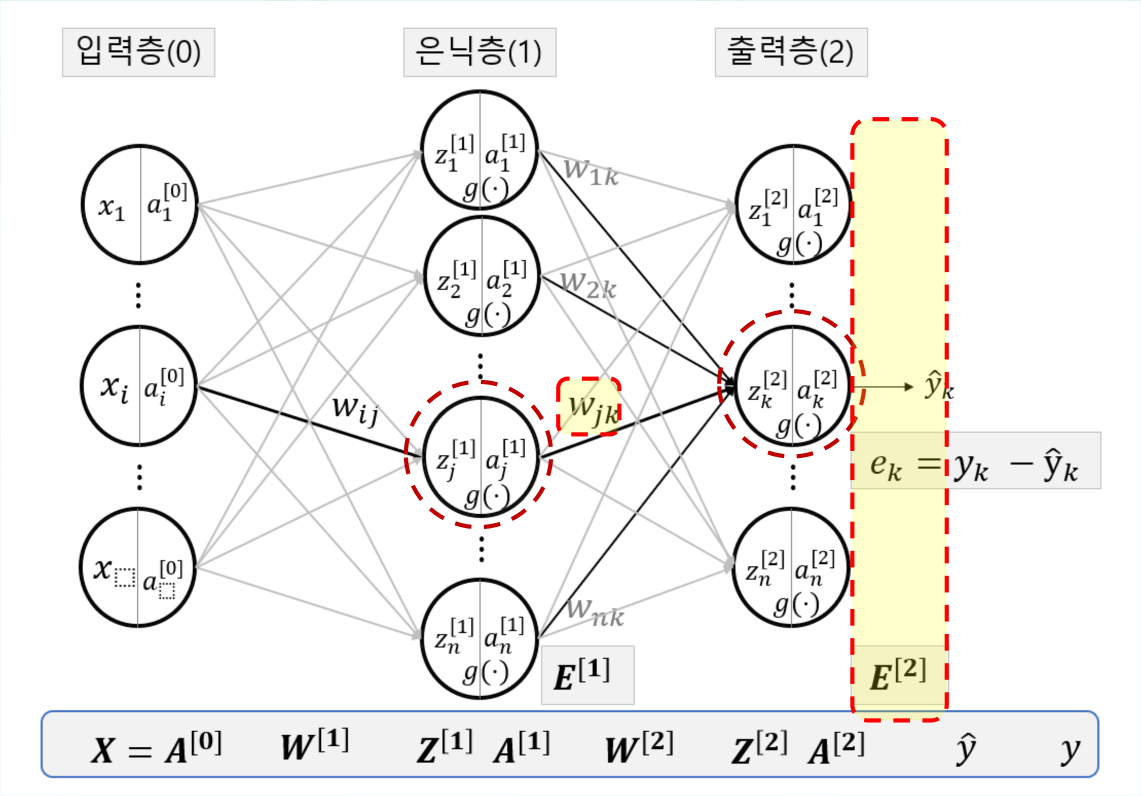
$$\Delta w_{jk}^{[2]} = \frac{\partial E}{\partial w_{jk}} = -(y_k - \hat{y}_k) \cdot g'(z_k) \cdot a_j$$

- 어떻게 **W2**로 확장할 것인가**?**

2. $W^{[2]}$의 오차함수 미분 : 4 단계

$$\Delta w_{jk}^{[2]} = \frac{\partial E}{\partial w_{jk}} = \boxed{-(y_k - \hat{y}_k)} \cdot \boxed{g'(z_k)} \cdot \boxed{a_j}$$

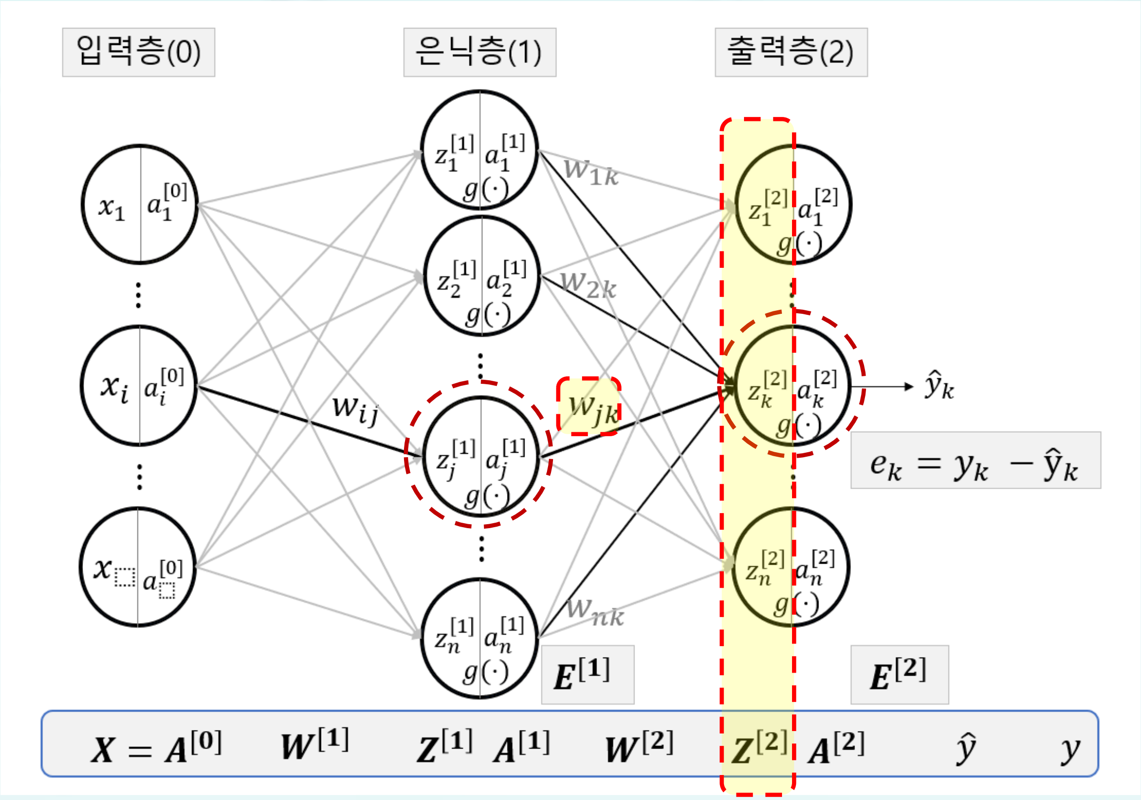
$$\Delta W^{[2]} = \frac{\partial E}{\partial W^{[2]}} = \boxed{-E^{[2]}} \cdot g'(Z^{[2]}) \cdot A^{[1]T}$$



2. $W^{[2]}$의 오차함수 미분 : 4 단계

$$\Delta w_{jk}^{[2]} = \frac{\partial E}{\partial w_{jk}} = \boxed{-(y_k - \hat{y}_k)} \cdot \boxed{g'(z_k)} \cdot \boxed{a_j}$$

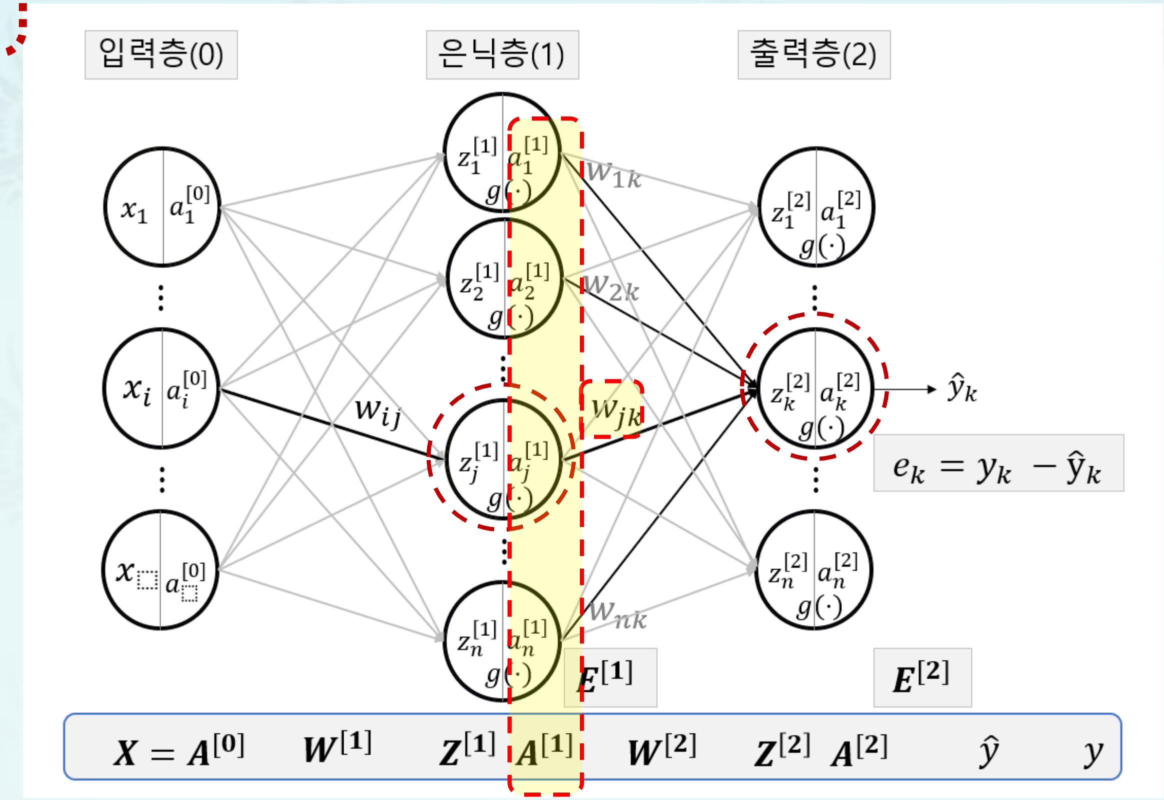
$$\Delta W^{[2]} = \frac{\partial E}{\partial W^{[2]}} = -E^{[2]} \cdot \boxed{g'(Z^{[2]})} \cdot A^{[1]T}$$



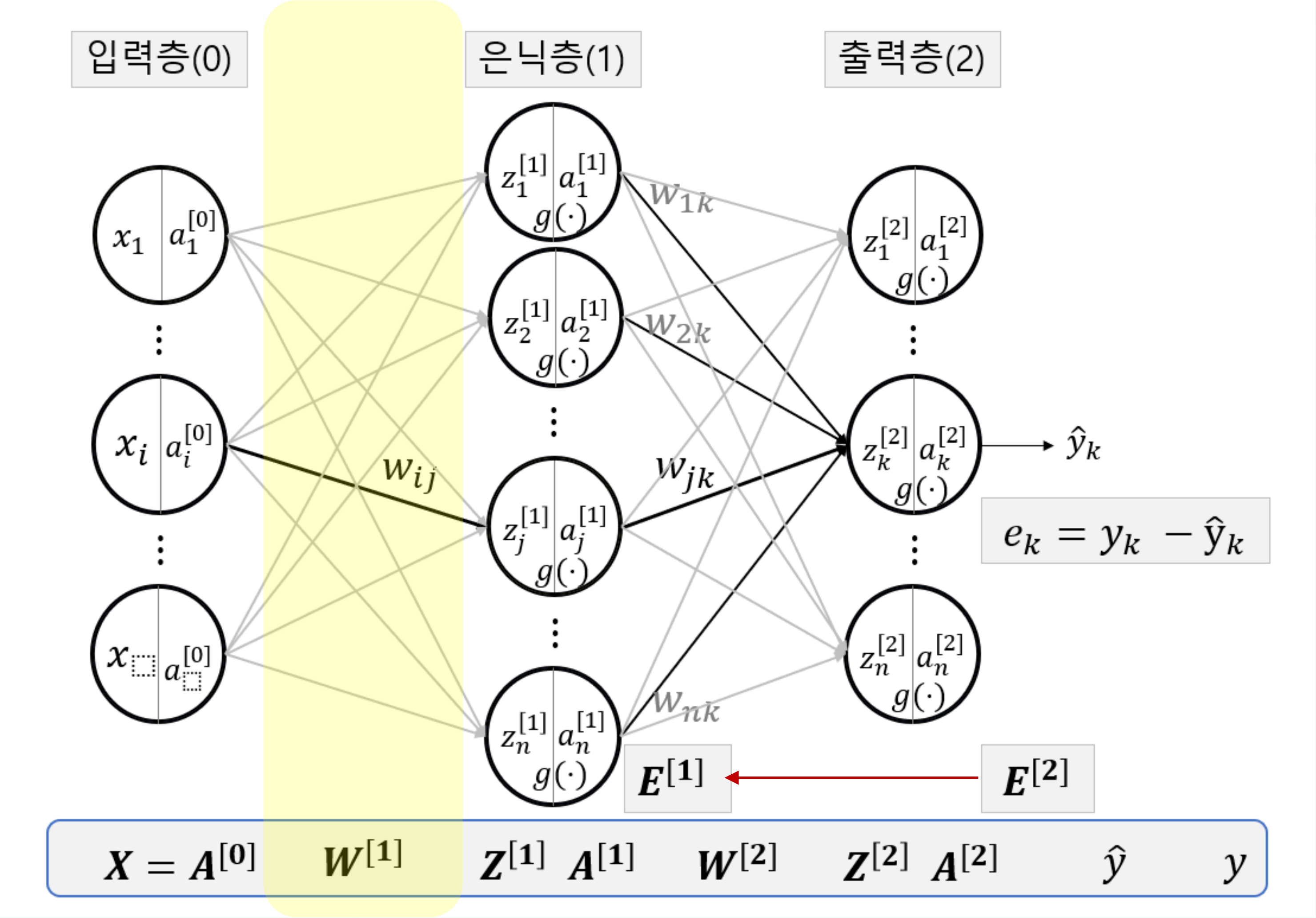
2. $W^{[2]}$의 오차함수 미분 : 4 단계

$$\Delta w_{jk}^{[2]} = \frac{\partial E}{\partial w_{jk}} = \boxed{-(y_k - \hat{y}_k)} \cdot \boxed{g'(z_k)} \cdot \boxed{a_j}$$

$$\Delta W^{[2]} = \frac{\partial E}{\partial W^{[2]}} = -E^{[2]} \cdot g'(Z^{[2]}) \cdot \boxed{A^{[1]T}}$$



3. $W^{[1]}$의 오차함수 미분



입력층(0)
은닉층(1)
출력층(2)
x_1 $a_1^{[0]}$
x_i $a_i^{[0]}$
$z_1^{[1]}$ $a_1^{[1]}$ $g(\cdot)$
$z_2^{[1]}$ $a_2^{[1]}$ $g(\cdot)$
$z_j^{[1]}$ $a_j^{[1]}$ $g(\cdot)$
$z_n^{[1]}$ $a_n^{[1]}$ $g(\cdot)$
w_{ij}
w_{1k}
w_{2k}
w_{jk}
w_{nk}
$z_1^{[2]}$ $a_1^{[2]}$ $g(\cdot)$
$z_k^{[2]}$ $a_k^{[2]}$ $g(\cdot)$
$z_n^{[2]}$ $a_n^{[2]}$ $g(\cdot)$
$\hat{y}_k$
$e_k = y_k - \hat{y}_k$
$E^{[1]}$
$E^{[2]}$
$X = A^{[0]}$ $W^{[1]}$ $Z^{[1]}$ $A^{[1]}$ $W^{[2]}$ $Z^{[2]}$ $A^{[2]}$ $\hat{y}$ y

3. $W^{[1]}$의 오차함수 미분

$$\Delta W^{[2]} = \frac{\partial E}{\partial W^{[2]}} = -E^{[2]} \cdot g'(Z^{[2]}) \cdot A^{[1]T}$$

$$\Downarrow$$

$$\boxed{\Delta W^{[1]} = \frac{\partial E}{\partial W^{[1]}} = -E^{[1]} \cdot g'(Z^{[1]}) \cdot A^{[0]T}}$$

4. 역전파의 가중치 조정 : 공식의 완성

- $W^{[2]} := W^{[2]} - \alpha \Delta W^{[2]}$

$$= W^{[2]} - \alpha \frac{\partial E}{\partial W^{[2]}}$$

$$= \boxed{\mathrm{W}^{[2]} + E^{[2]} \cdot g'(Z^{[2]}) \cdot A^{[1]T}}$$

4. 역전파의 가중치 조정 : 공식의 완성

$$W^{[2]} := W^{[2]} - \alpha \Delta W^{[2]}$$

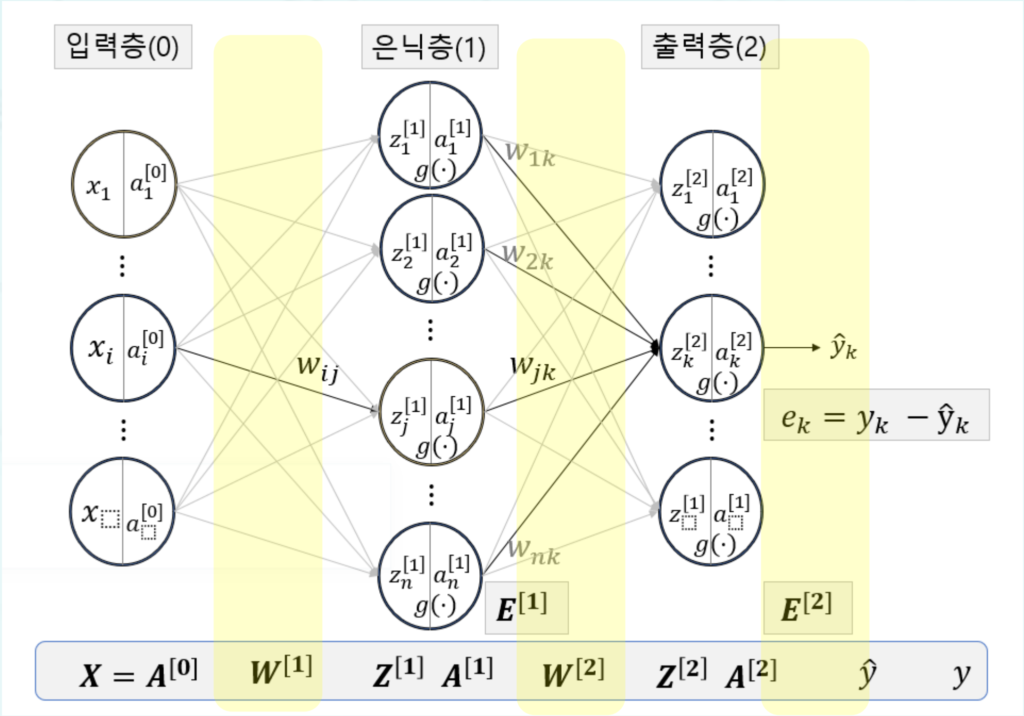
$$= W^{[2]} - \alpha \frac{\partial E}{\partial W^{[2]}}$$

$$= \boxed{W^{[2]} + E^{[2]} \cdot g'(Z^{[2]}) \cdot A^{[1]T}}$$

$$W^{[1]} := W^{[1]} - \alpha \Delta W^{[1]}$$

$$= W^{[1]} - \alpha \frac{\partial E}{\partial W^{[1]}}$$

$$= \boxed{W^{[1]} + E^{[1]} \cdot g'(Z^{[1]}) \cdot A^{[0]T}}$$



역전파 2

- 학습 정리
  - 역전파 과정에서 오차함수의 미분
  - 미분한 오차함수를 기반으로 한 신경망의 가중치 조정