10주차(1/3)

역전파 2

파이썬으로 배우는 기계학습

한동대학교
김영섭 교수

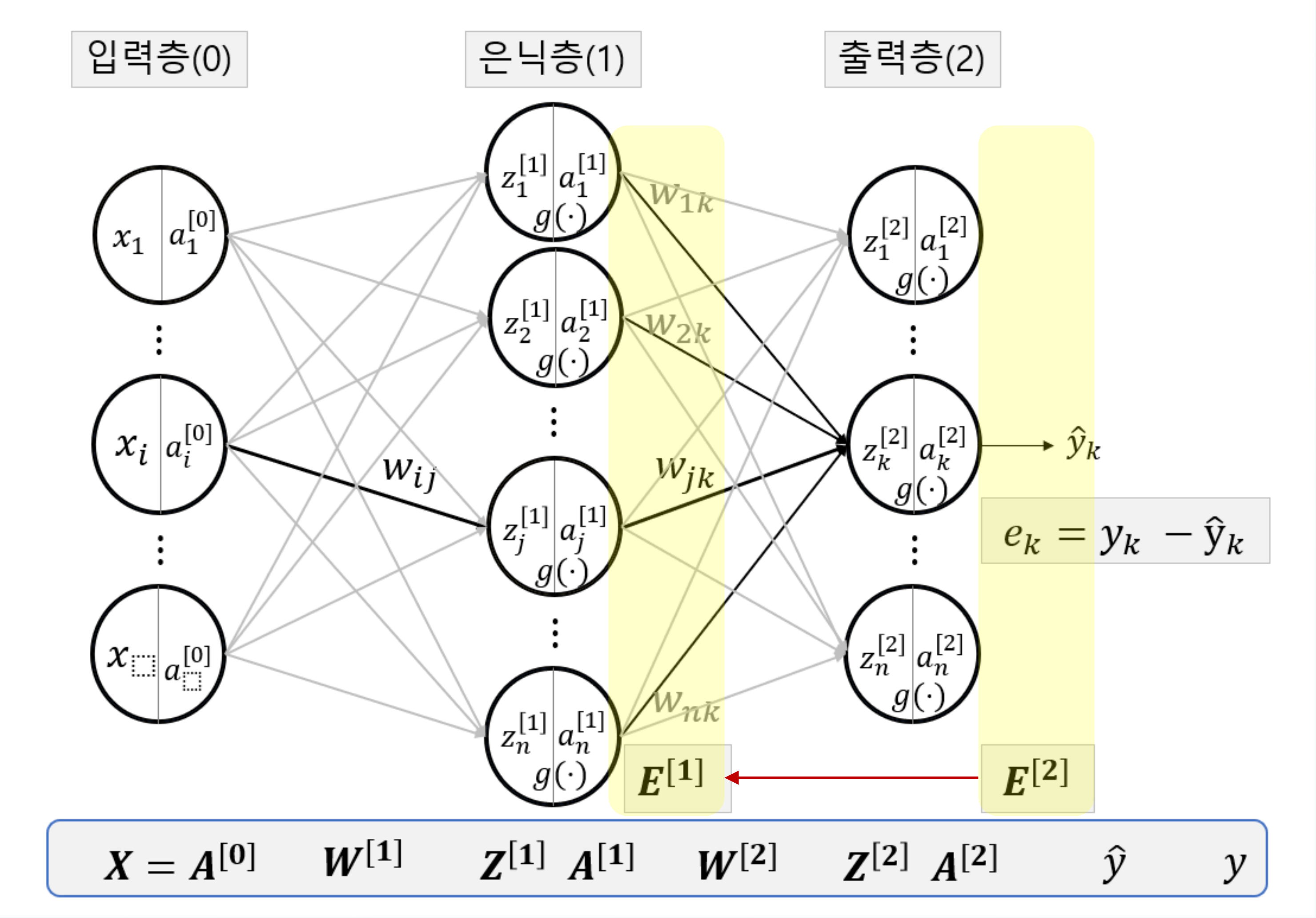
역전파 2

- 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

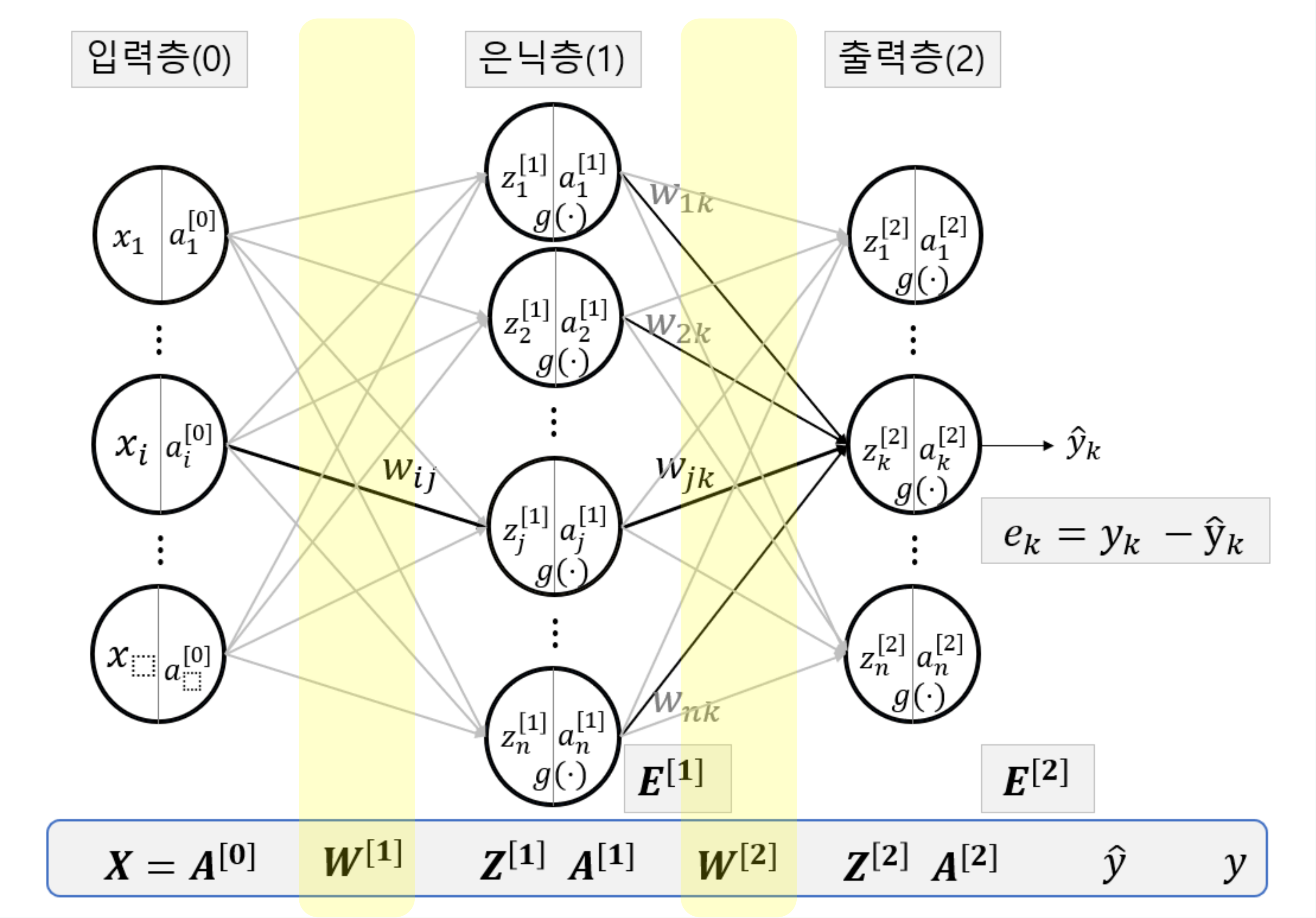
역전파 2: $W^{[2]}$의 오차함수 미분

- 출력층의 오차 $E^{[2]}$
  - 레이블과 예측값의 차이
  - 은닉층의 오차 $E^{[1]}$ 계산
- 가중치 조정 가능



역전파 2: $W^{[2]}$의 오차함수 미분

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



역전파 2: $W^{[2]}$의 오차함수 미분 – 1단계

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

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

← 1단계

- 경사하강법 오차함수와 같은 형식
  - 가중치 **W** 조정 → 오차 **E** 최소화

역전파 2: $W^{[2]}$의 오차함수 미분 – 1단계

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

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

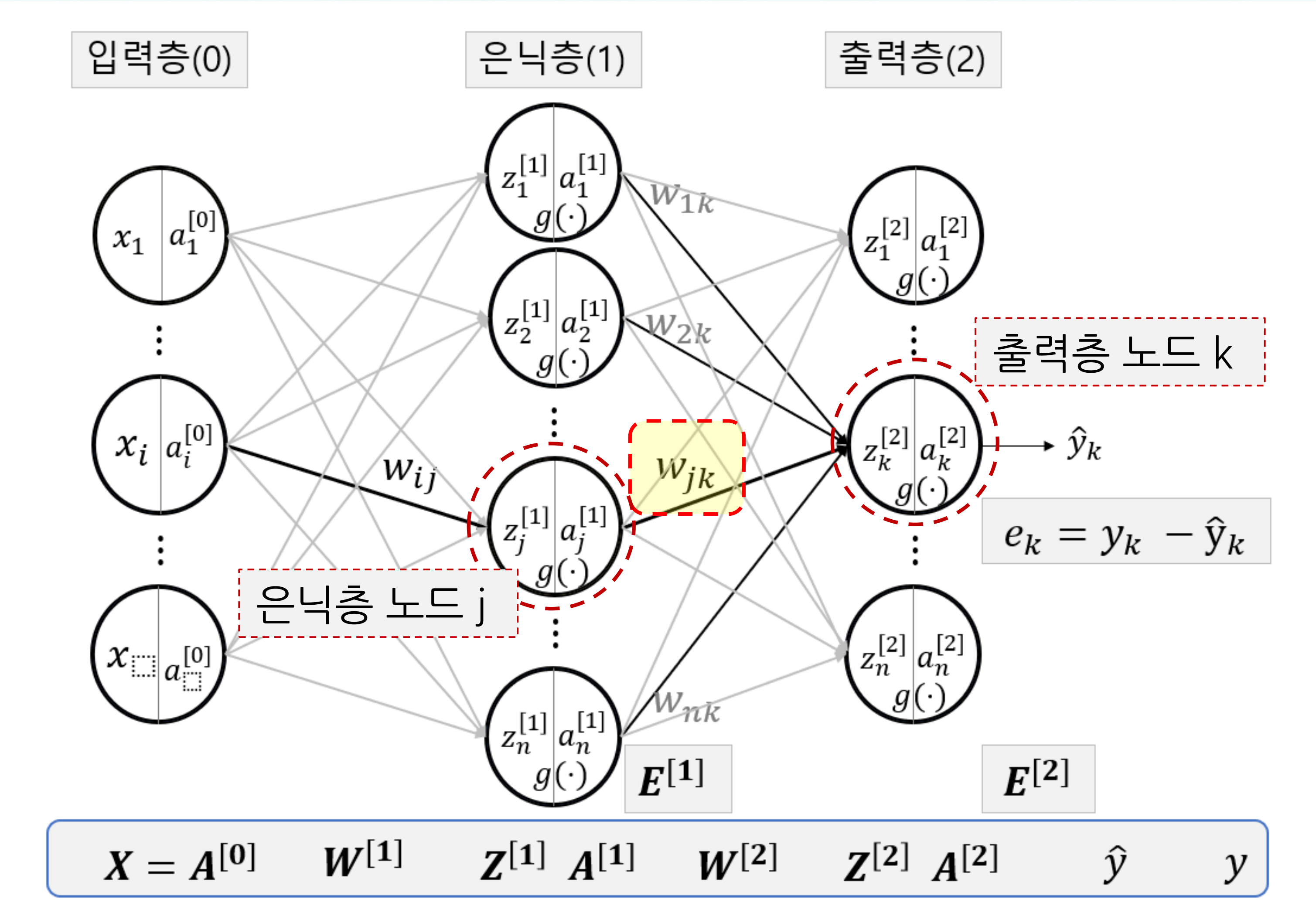
- 경사하강법 오차함수와 같은 형식
  - 가중치 **W** 조정 → 오차 **E** 최소화
- 문제는**?**
  - 행렬 미분의 어려움
  - 해결책: $w_{jk}^{[2]}$

역전파 2: $W^{[2]}$의 오차함수 미분 – 1단계

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

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

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



역전파 2: $W^{[2]}$의 오차함수 미분 – 2단계

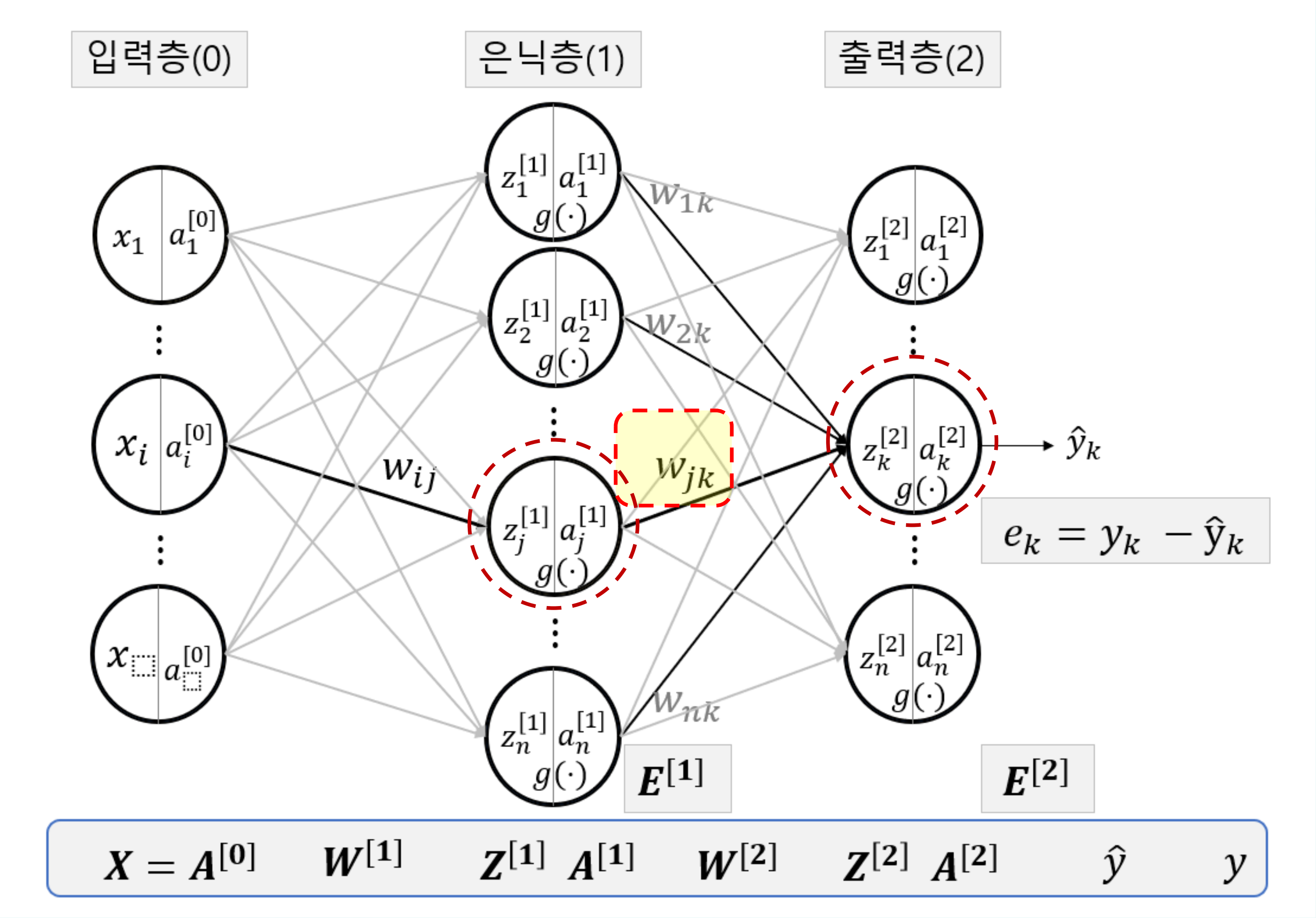
$$W^{[2]} := W^{[2]} - \alpha \Delta W^{[2]}$$

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

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

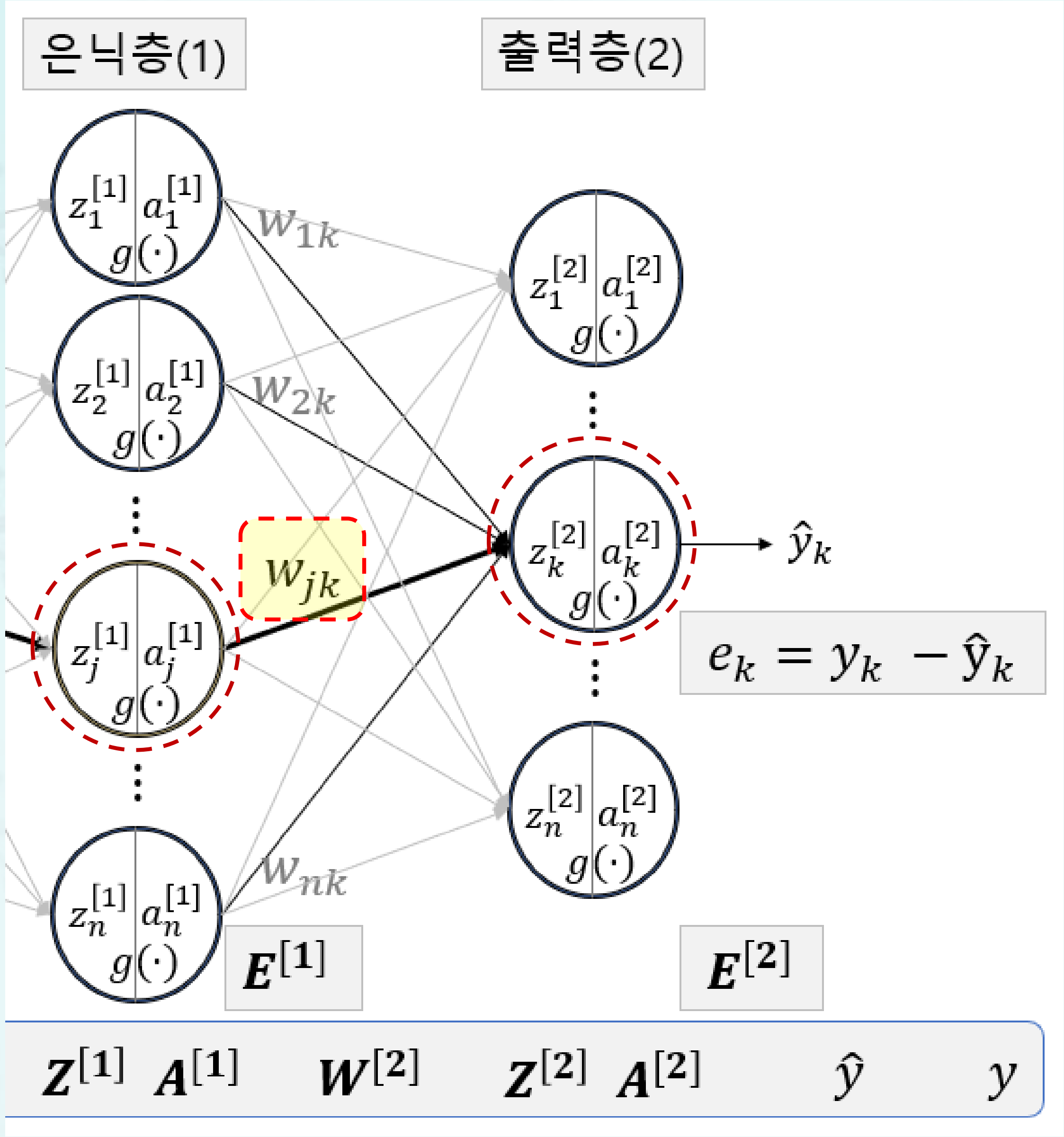
$$= w_{jk}^{[2]} - \alpha \boxed{\frac{\partial E}{\partial w_{jk}^{[2]}}}$$

2 단계



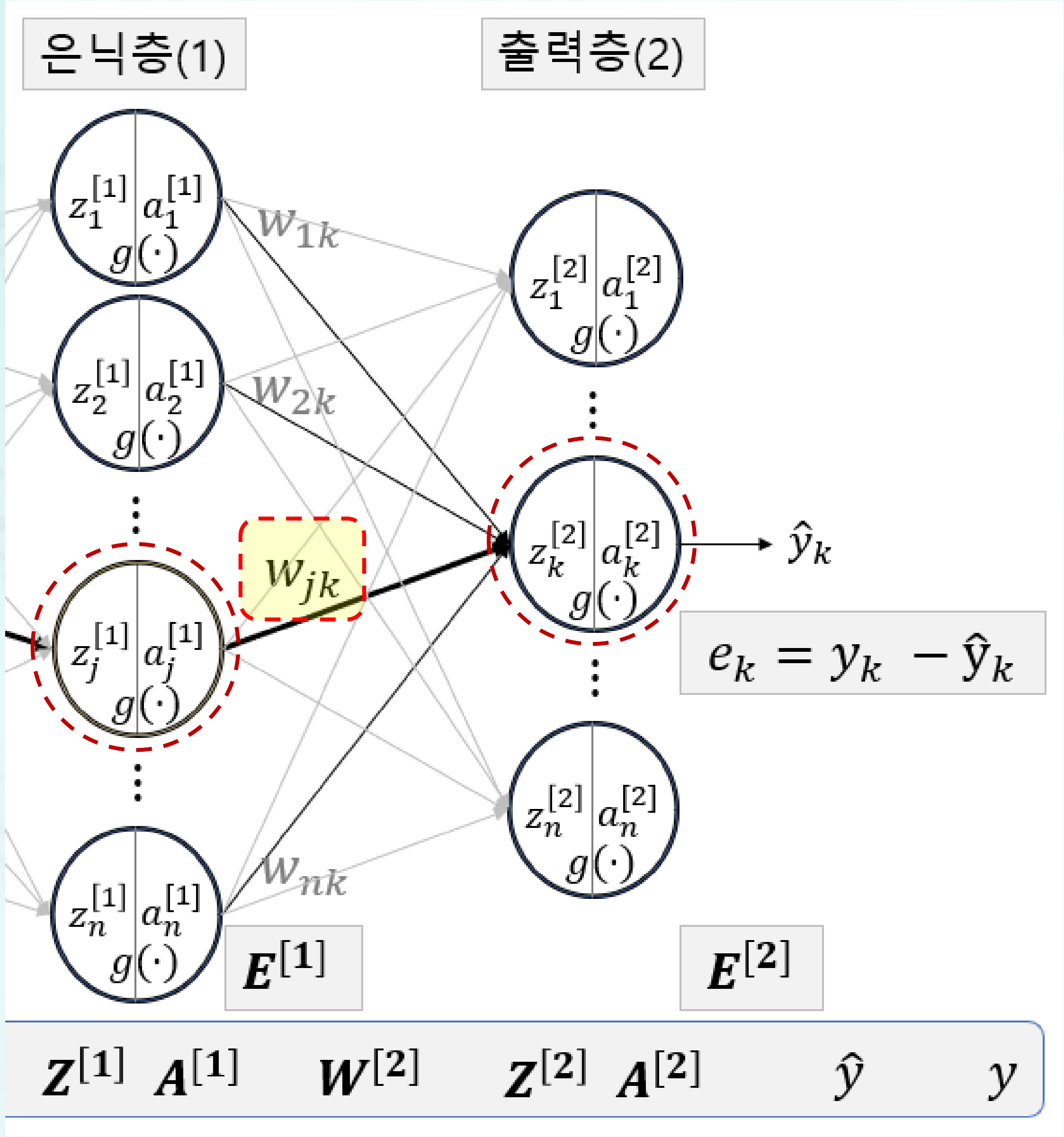
역전파 2: $W^{[2]}$의 오차함수 미분 – 2단계

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



역전파 2: $W^{[2]}$의 오차함수 미분 – 2단계

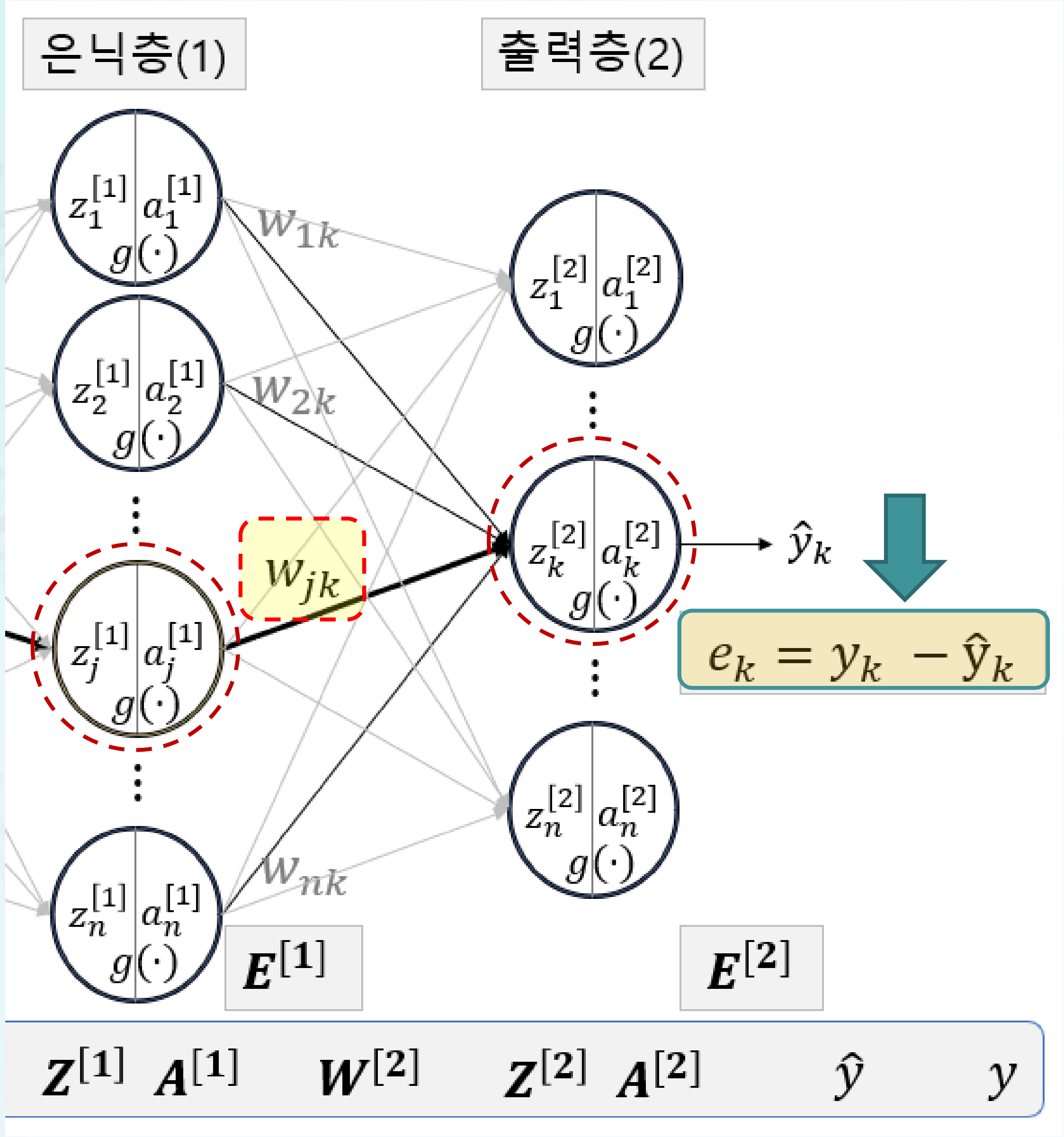
$$\frac{\partial E}{\partial w_{jk}^{[2]}} = \boxed{\frac{\partial}{\partial w_{jk}} \frac{1}{2} \sum_{m=1}^{n} (y_m - \hat{y}_m)^2}$$



역전파 2: $W^{[2]}$의 오차함수 미분 – 2단계

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

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



역전파 2: $W^{[2]}$의 오차함수 미분 – 2단계

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

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

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

합성함수 미분법

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

역전파 2: $W^{[2]}$의 오차함수 미분 – 2단계

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

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

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

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

합성함수 미분법
$f(g(x))' = f'(g(x))g'(x)$

역전파 2: $W^{[2]}$의 오차함수 미분 – 2단계

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

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

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

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

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

역전파 2: $W^{[2]}$의 오차함수 미분 – 2단계

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

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

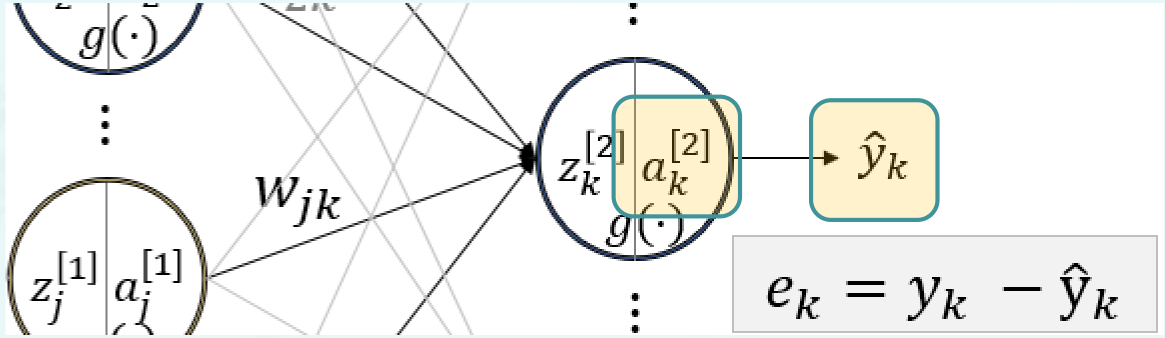
$$= \frac{1}{2} \cdot 2(y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (y_k - \hat{y}_k)$$

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

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

- 출력층 노드 **k**의 출력 $\hat{y}_k$의 미분

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



역전파 2: $W^{[2]}$의 오차함수 미분 – 2단계

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

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

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

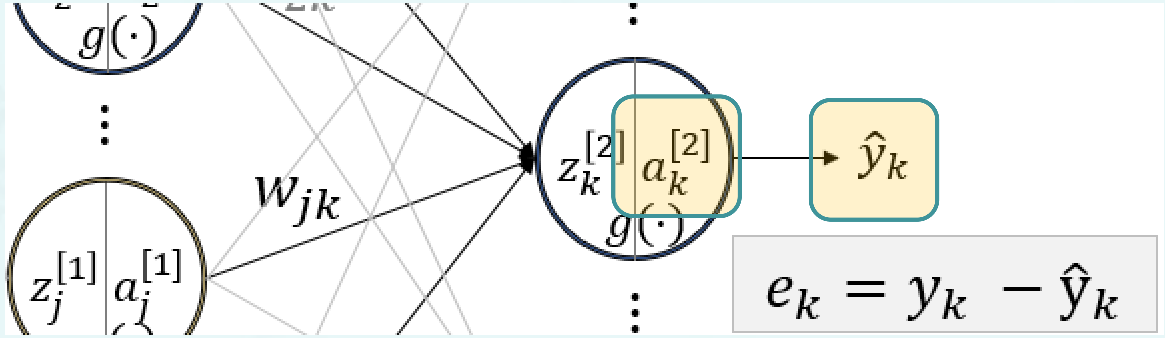
$$= (y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (-\hat{y}_k)$$

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

- 출력층 노드 **k**의 출력 $\hat{y}_k$의 미분

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

$$=$$



역전파 2: $W^{[2]}$의 오차함수 미분 – 2단계

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

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

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

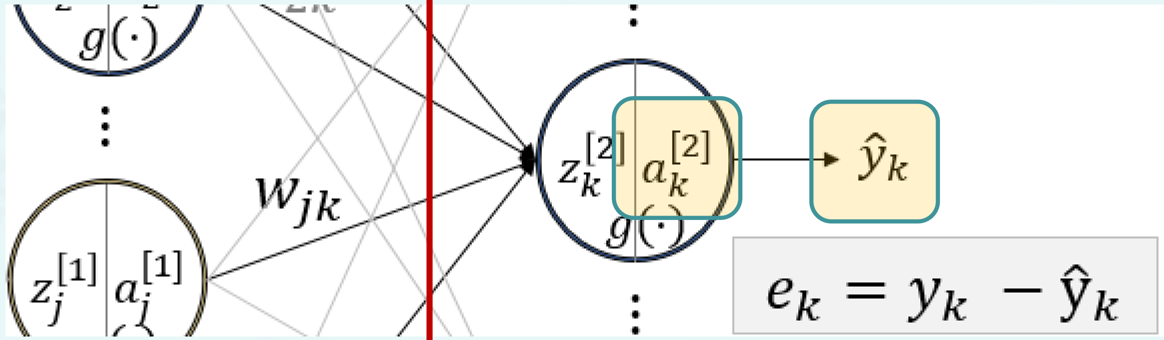
$$= (y_k - \hat{y}_k) \frac{\partial}{\partial w_{jk}} (-\hat{y}_k)$$

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

- 출력층 노드 **k**의 출력 $\hat{y}_k$의 미분

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

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



역전파 2: $W^{[2]}$의 오차함수 미분 – 3단계

- 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

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

역전파 2: $W^{[2]}$의 오차함수 미분 – 3단계

- 3단계
- $$\frac{\partial E}{\partial w_{jk}} = -(y_k - \hat{y}_k) \cdot \frac{\partial}{\partial w_{jk}} g(z_k)$$
  $$= -(y_k - \hat{y}_k) \cdot \boxed{g'(z_k) \frac{\partial z_k}{\partial w_{jk}}}$$

합성함수 미분법
$$u\big(v(x)\big)' = u'\big(v(x)\big)v'(x)$$

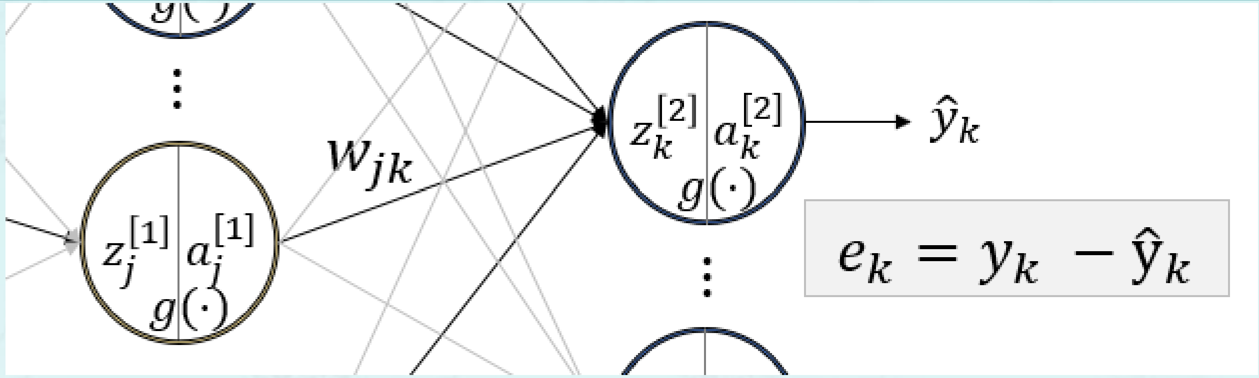
역전파 2: $W^{[2]}$의 오차함수 미분 – 3단계

- 3단계
- $$\frac{\partial E}{\partial w_{jk}} = -(y_k - \hat{y}_k) \cdot \frac{\partial}{\partial w_{jk}} g(z_k)$$
  $$= -(y_k - \hat{y}_k) \cdot g'(z_k) \boxed{\frac{\partial z_k}{\partial w_{jk}}}$$

합성함수 미분법
$$u(v(x))' = u'(v(x))v'(x)$$

역전파 2: $W^{[2]}$의 오차함수 미분 – 3단계

- 3단계



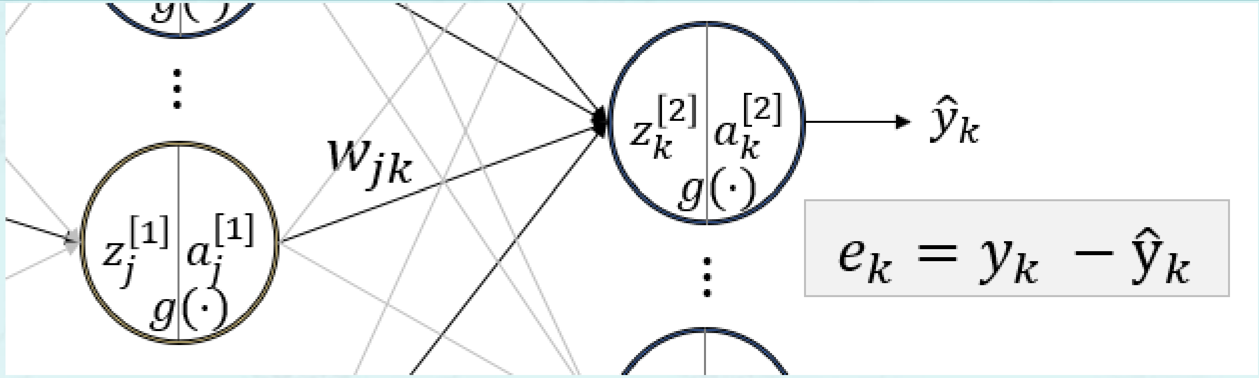
$$\frac{\partial E}{\partial w_{jk}} = -(y_k - \hat{y}_k) \cdot \frac{\partial}{\partial w_{jk}} g(z_k)$$

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

$$= -(y_k - \hat{y}_k) \cdot g'(z_k) \boxed{\frac{\partial}{\partial w_{jk}} \left(\sum_j w_{jk} \cdot a_j\right)} \qquad \because z_k = \sum_j w_{jk}^{[2]} a_j^{[1]}$$

역전파 2: $W^{[2]}$의 오차함수 미분 – 3단계

- 3단계
- $$\frac{\partial E}{\partial w_{jk}} = -(y_k - \hat{y}_k) \cdot \frac{\partial}{\partial w_{jk}} g(z_k)$$
  $$= -(y_k - \hat{y}_k) \cdot g'(z_k) \frac{\partial z_k}{\partial w_{jk}}$$
  $$= -(y_k - \hat{y}_k) \cdot g'(z_k) \frac{\partial}{\partial w_{jk}} (\sum_j w_{jk} \cdot a_j)$$
  $$= \boxed{-(y_k - \hat{y}_k) \cdot g'(z_k) \cdot a_j}$$



역전파 2: $W^{[2]}$의 오차함수 미분

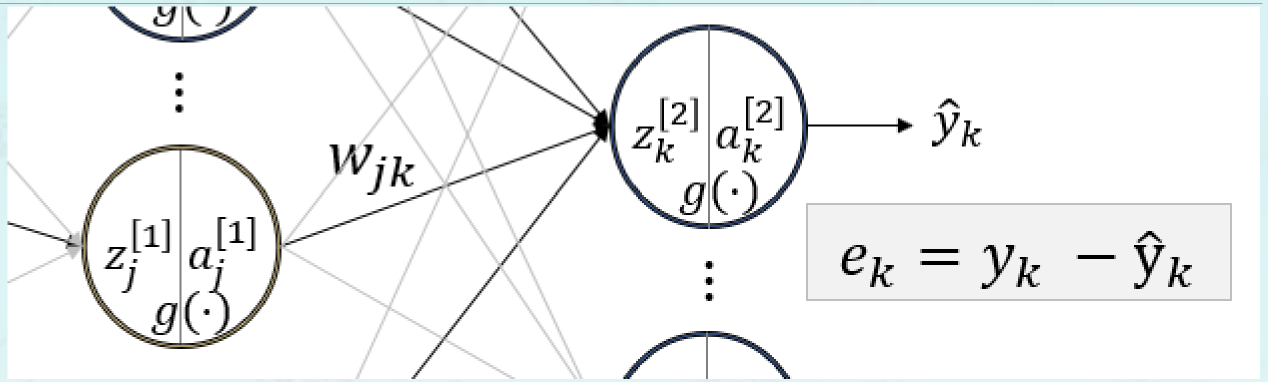
- 3단계

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

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

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

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



역전파 2: $W^{[2]}$의 오차함수 미분 – 3단계

- 3단계

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

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

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

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

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

역전파 2: $W^{[2]}$의 오차함수 미분 – 3단계

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

- 오차: 출력층 **k** 노드에서 레이블과 예측값의 차이

역전파 2: $W^{[2]}$의 오차함수 미분 – 3단계

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

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

역전파 2: $W^{[2]}$의 오차함수 미분 – 3단계

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

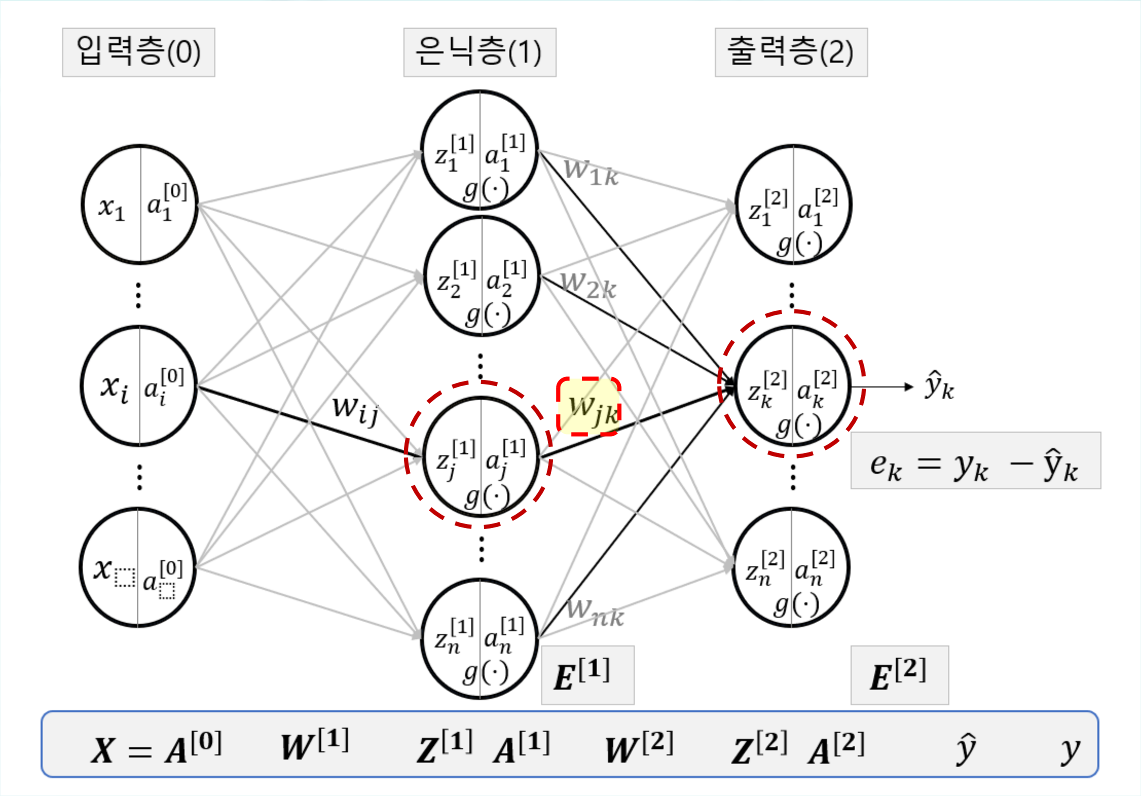
- 오차: 출력층 **k** 노드에서 레이블과 예측값의 차이
- 활성화 함수 미분에 z_k를 적용한 값
  - z_k: 출력층 노드 **k**의 순입력
- a_j : 은닉층 노드 **j**의 출력

역전파 2: $W^{[2]}$의 오차함수 미분 – 4단계

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

역전파 2: $W^{[2]}$의 오차함수 미분 – 4단계

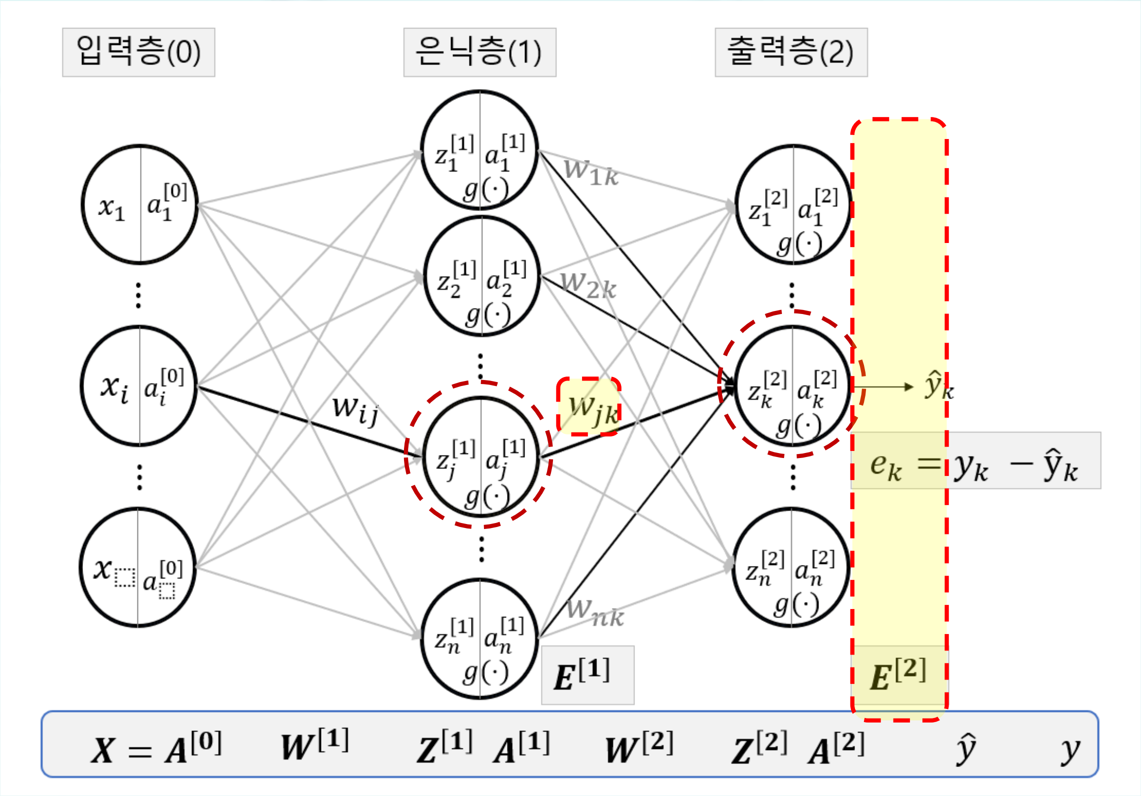
$$\Delta w_{jk}^{[2]} = \frac{\partial E}{\partial w_{jk}} = -(y_k - \hat{y}_k) \cdot g'(z_k) \cdot a_j$$



역전파 2: $W^{[2]}$의 오차함수 미분 – 4단계

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

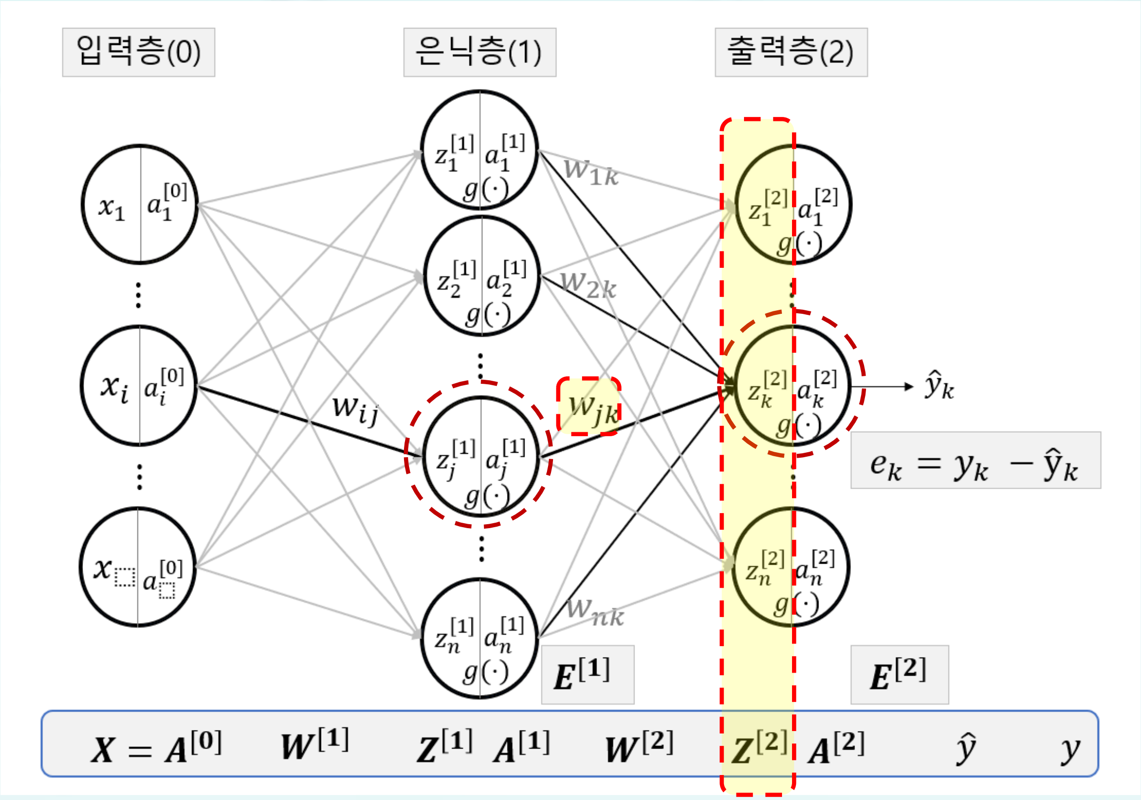
$$\Delta W^{[2]} = \frac{\partial E}{\partial W^{[2]}} = \boxed{-E^{[2]}} \cdot g'(Z^{[2]}) \cdot A^{[1]T}$$



역전파 2: $W^{[2]}$의 오차함수 미분 – 4단계

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

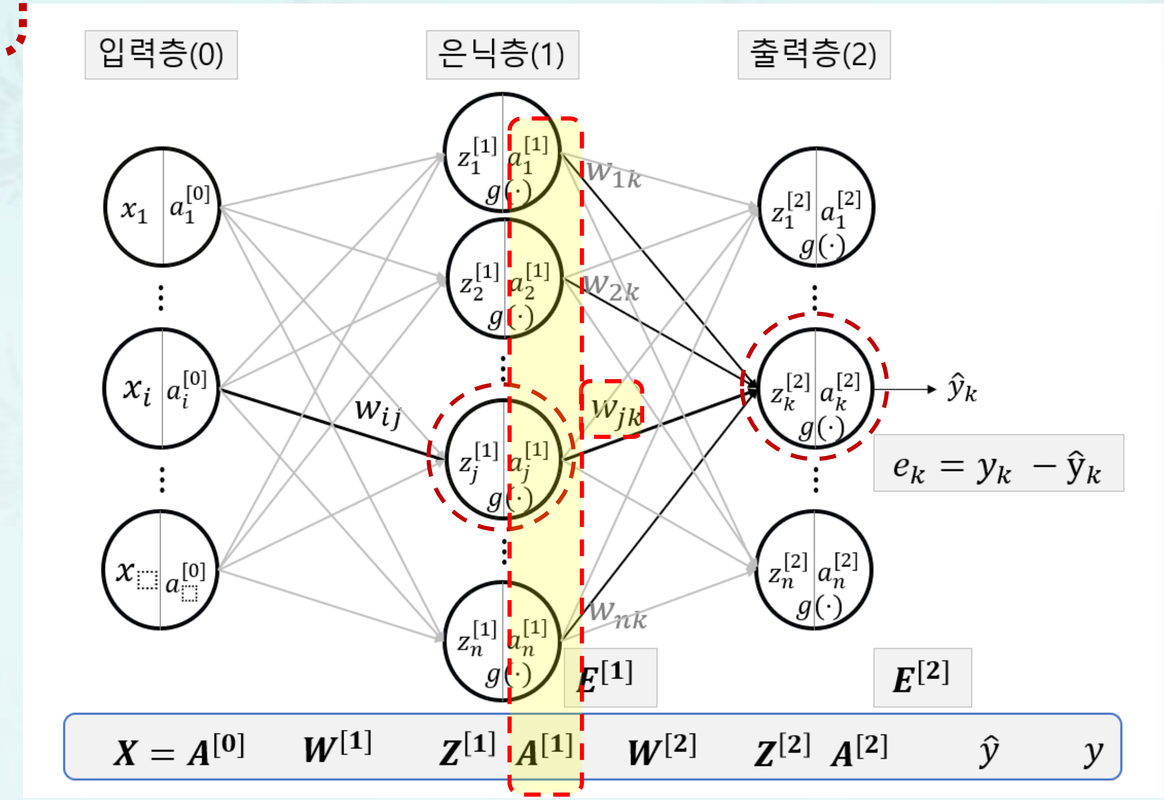
$$\Delta W^{[2]} = \frac{\partial E}{\partial W^{[2]}} = -E^{[2]} \cdot \boxed{g'(Z^{[2]})} \cdot A^{[1]T}$$



역전파 2: $W^{[2]}$의 오차함수 미분 – 4단계

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

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

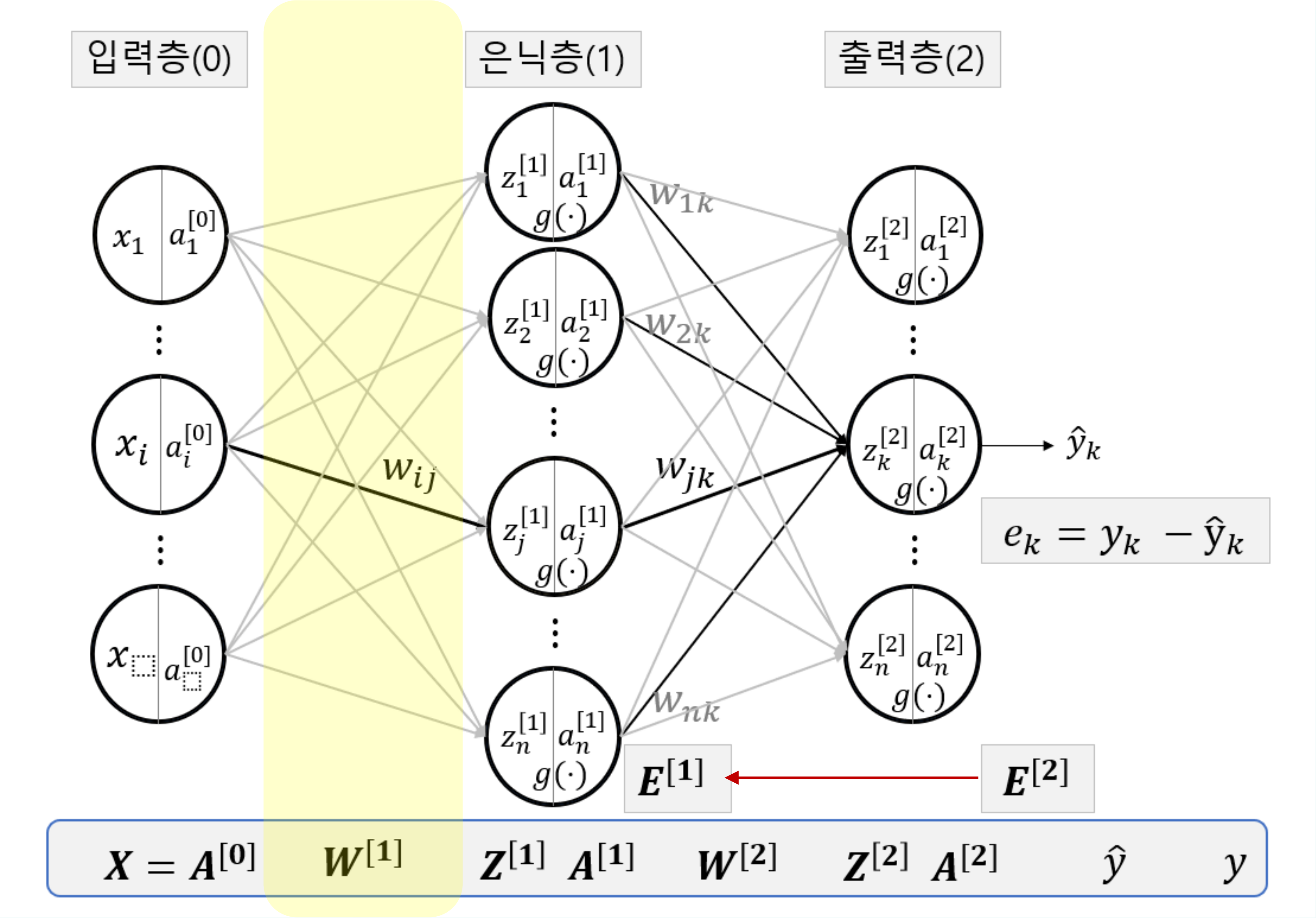


역전파 2: $W^{[2]}$의 오차함수 미분 – 4단계

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

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

역전파 2: $W^{[1]}$의 오차함수 미분



역전파 2: $W^{[1]}$의 오차함수 미분

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

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

역전파 2: 역전파의 가중치 조정

- 최종 결과

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

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

역전파 2: 역전파의 가중치 조정

- 최종 결과

$$
\begin{aligned}
W^{[2]} &:= W^{[2]} - \alpha \Delta W^{[2]} \\
&= W^{[2]} - \alpha \frac{\partial E}{\partial W^{[2]}} \\
&= \mathrm{W}^{[2]} + \boxed{E^{[2]} \cdot g'(Z^{[2]}) \cdot A^{[1]T}}
\end{aligned}
$$

역전파 2: 역전파의 가중치 조정

- 최종 결과

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

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

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

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

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

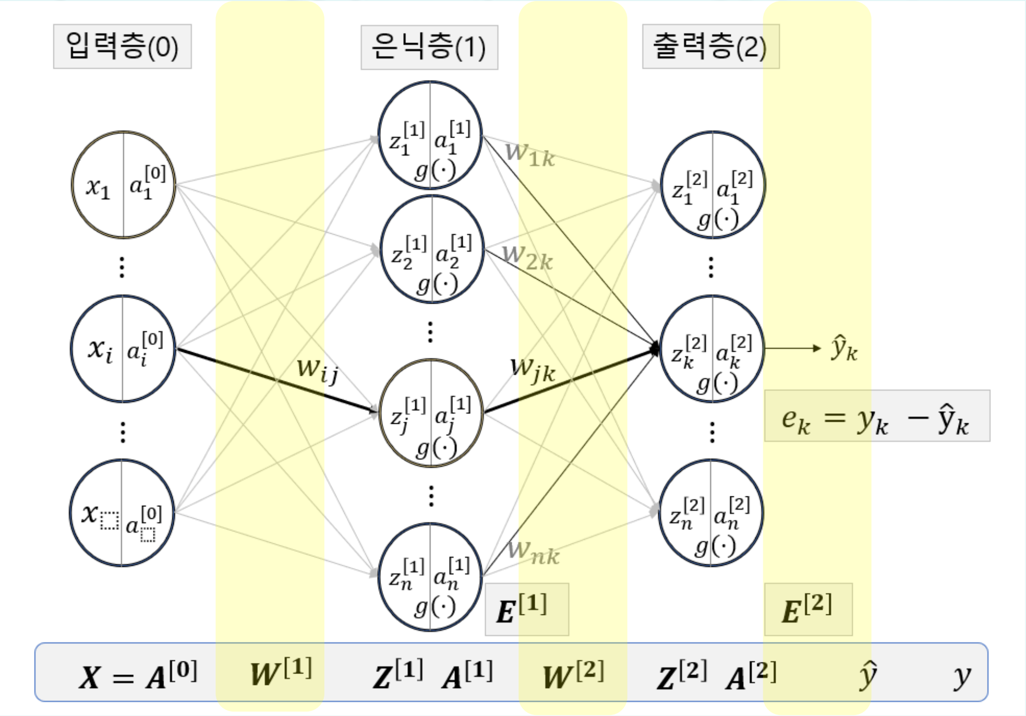
$$= \mathrm{W}^{[1]} + \boxed{\alpha E^{[1]} \cdot g'(Z^{[1]}) \cdot A^{[0]T}}$$

역전파 2: 역전파의 가중치 조정

- 최종 결과

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

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



역전파 2

- 학습 정리;
  - 역전파 과정에서 오차함수 미분하기
  - 미분한 오차함수를 기반으로 신경망의 가중치 조정하기

- **9-2 XOR 신경망 모델링**

10주차(1/3)

역전파 2

파이썬으로 배우는 기계학습

한동대학교
김영섭 교수