Backpropagation 2

수업 목표

이번 수업의 핵심:

- 복잡한 Computational graph의 Backpropagation
- Local gradient 계산을 통한 Backpropagation path 단순화

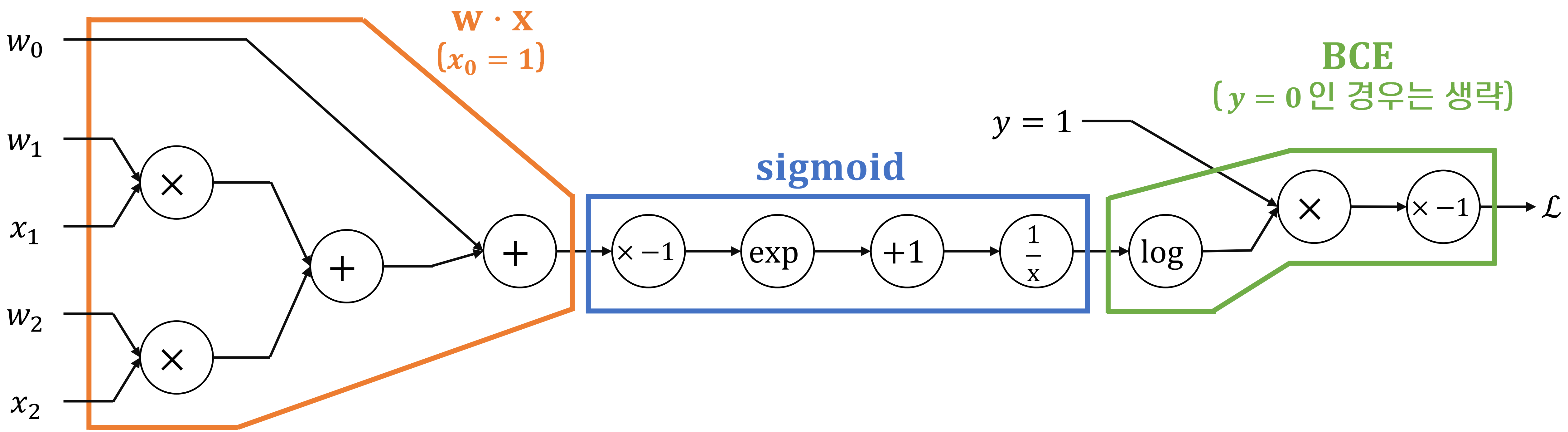
핵심 개념

- Computational graph
- Backpropagation

Logistic regression의 Computational Graph

$$\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1 + e^{-(w_0 + w_1 x_1 + w_2 x_2)}},$$

$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

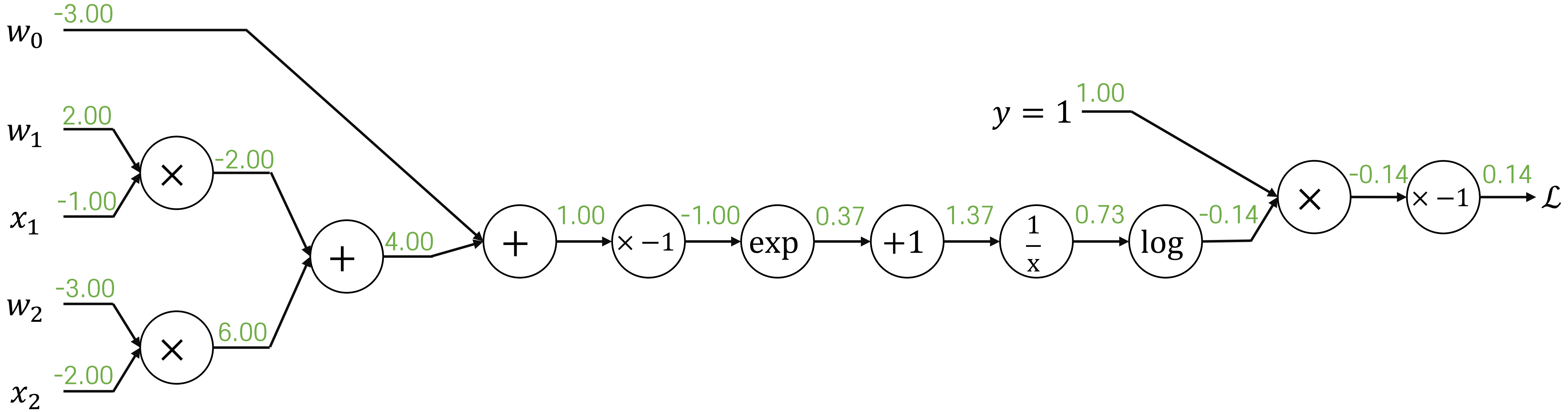


Computational Graph

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y) \log(1-\hat{y})$$

Forward pass: output 계산

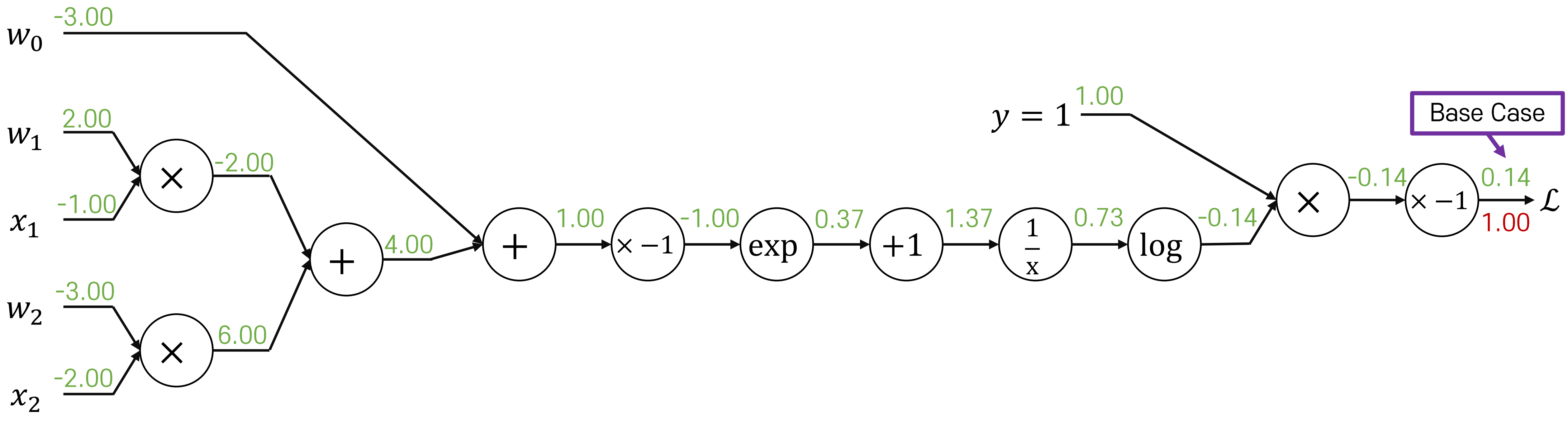


Computational Graph

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$

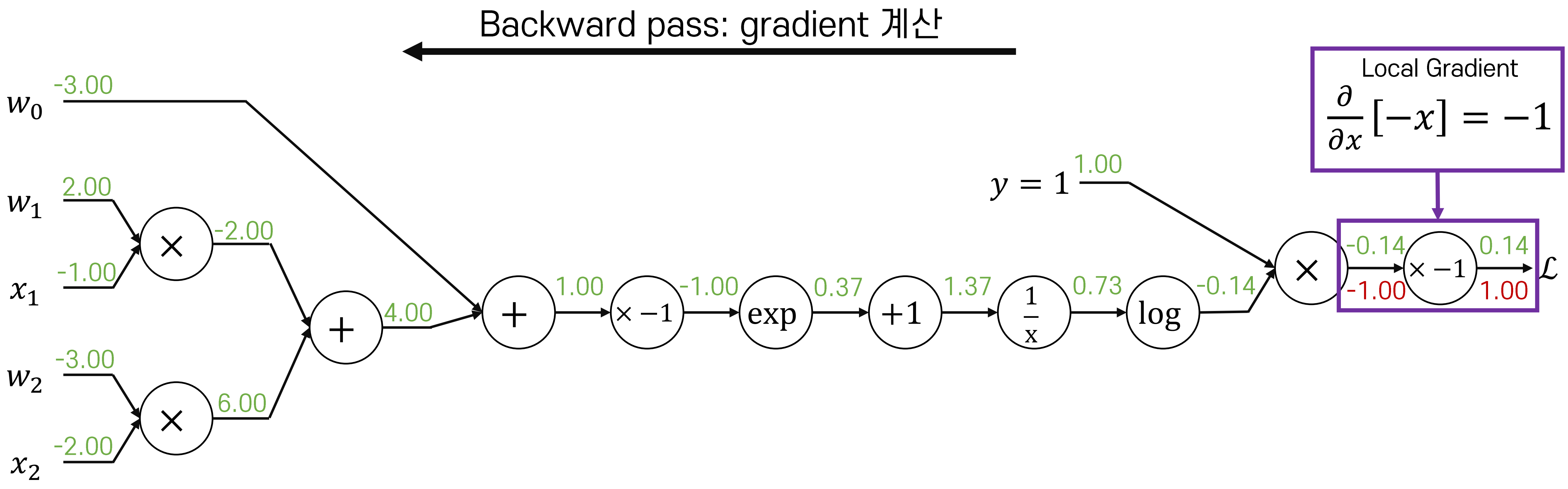
Backward pass: gradient 계산



Computational Graph

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \dfrac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

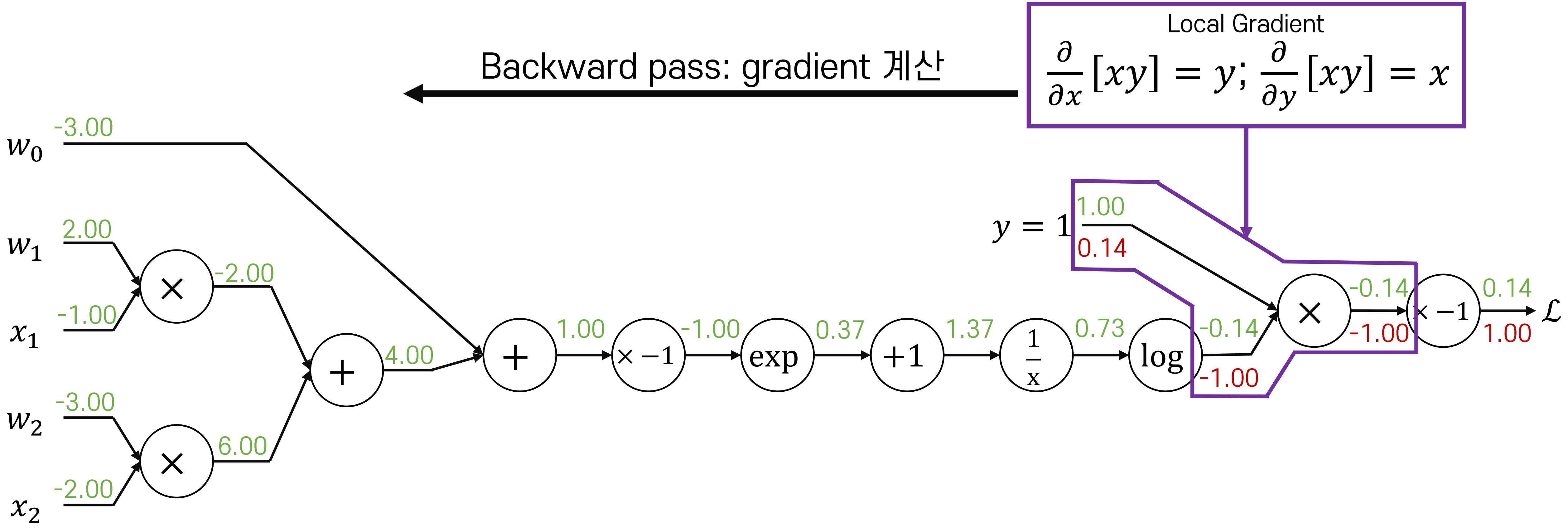
$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$



Computational Graph

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

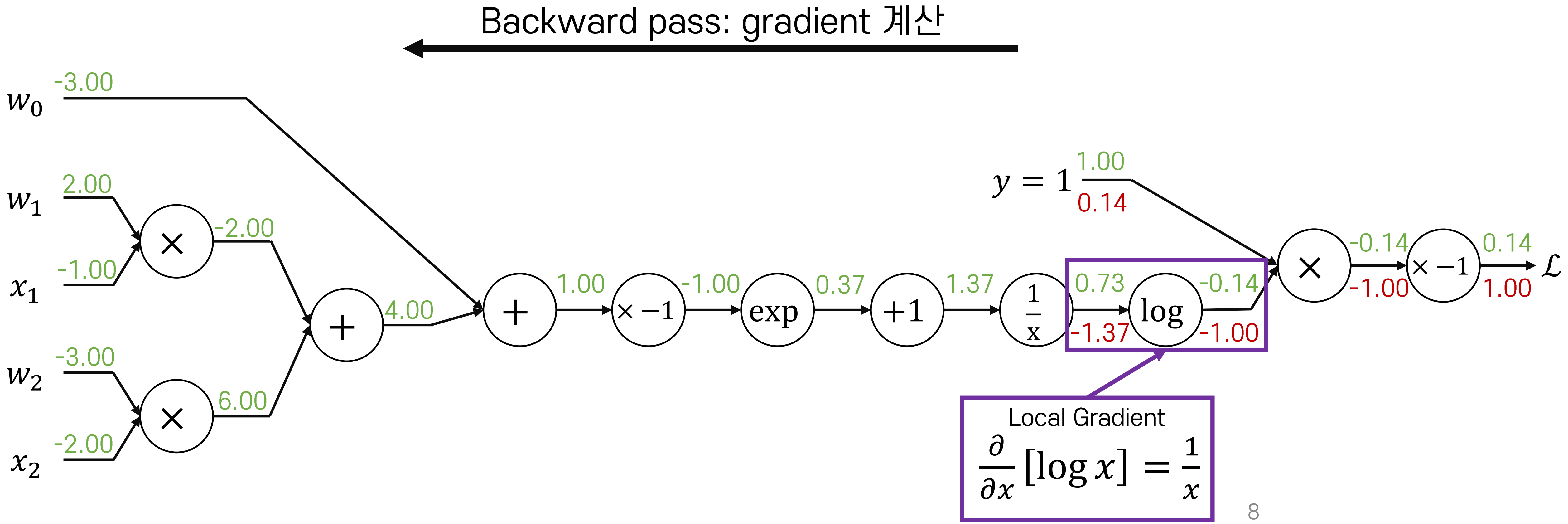
$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y) \log(1-\hat{y})$$



Computational Graph

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$

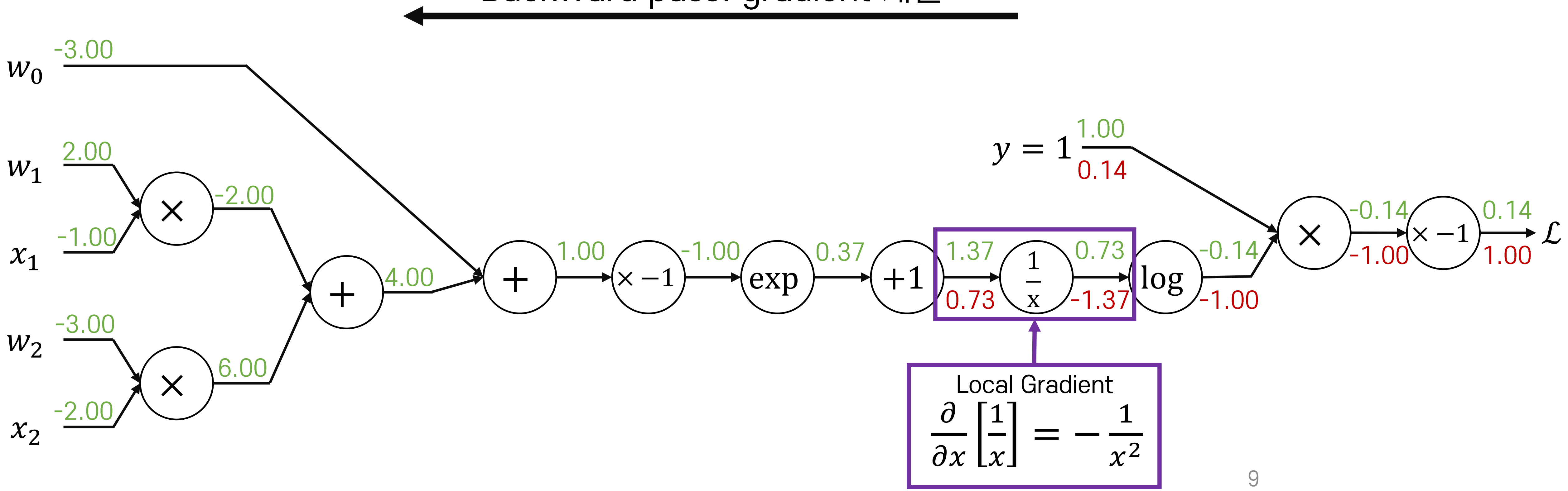


Computational Graph

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$

Backward pass: gradient 계산

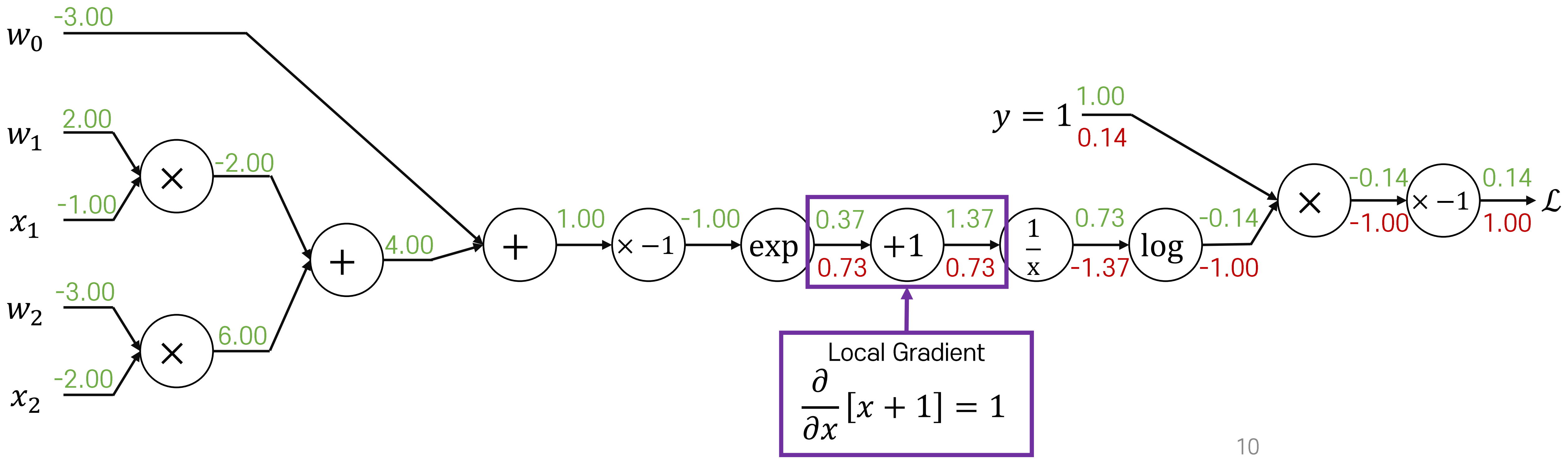


Computational Graph

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$

Backward pass: gradient 계산

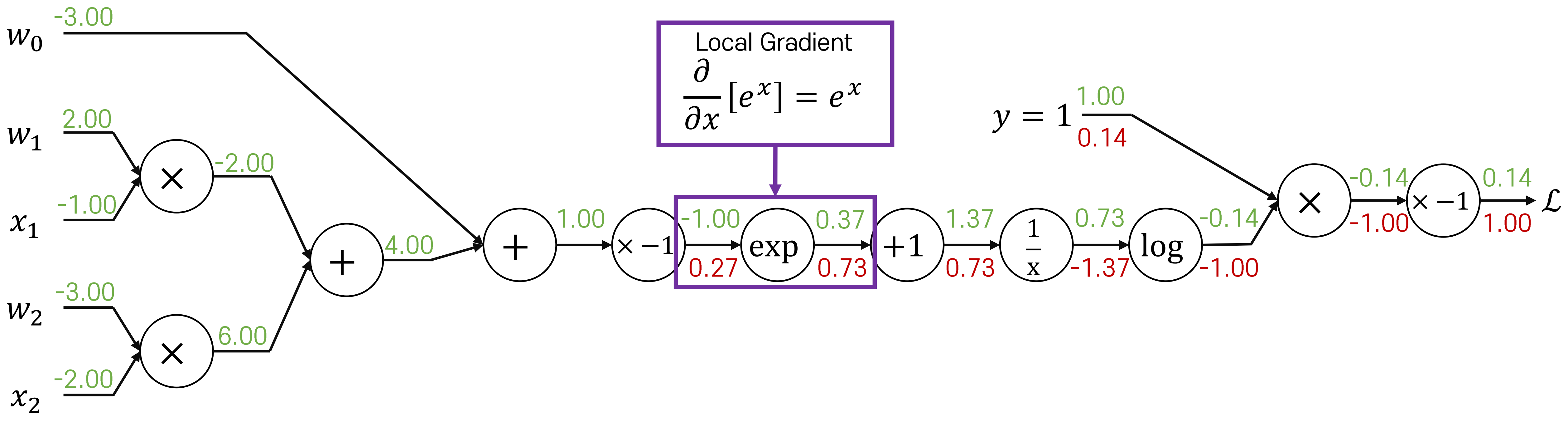


Computational Graph

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$

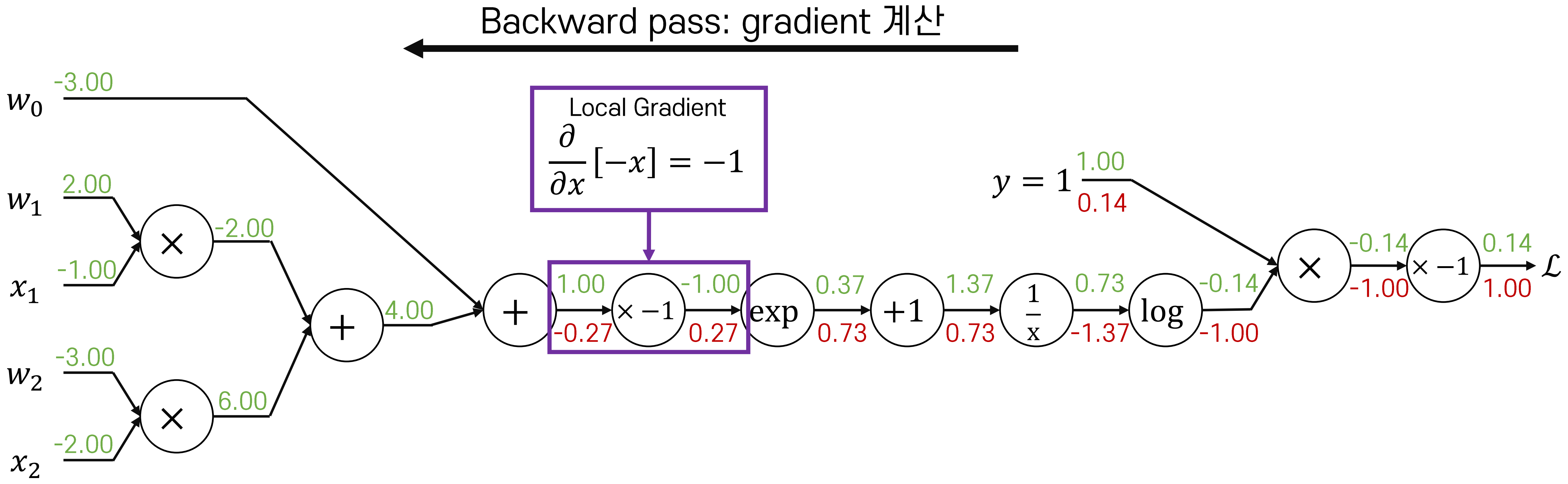
Backward pass: gradient 계산



Computational Graph

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$

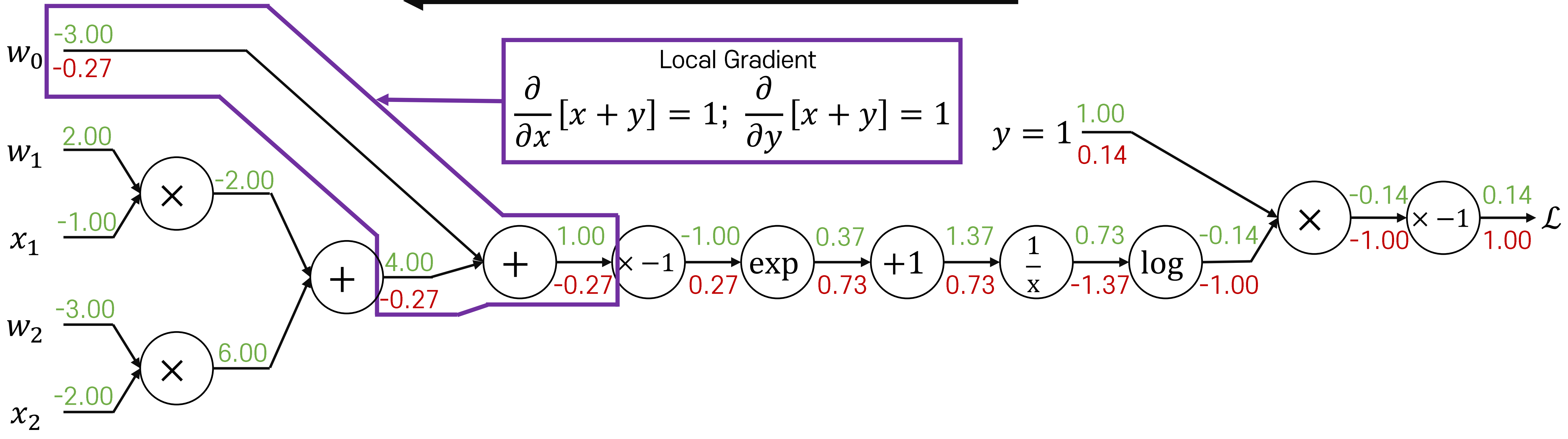


Computational Graph

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$

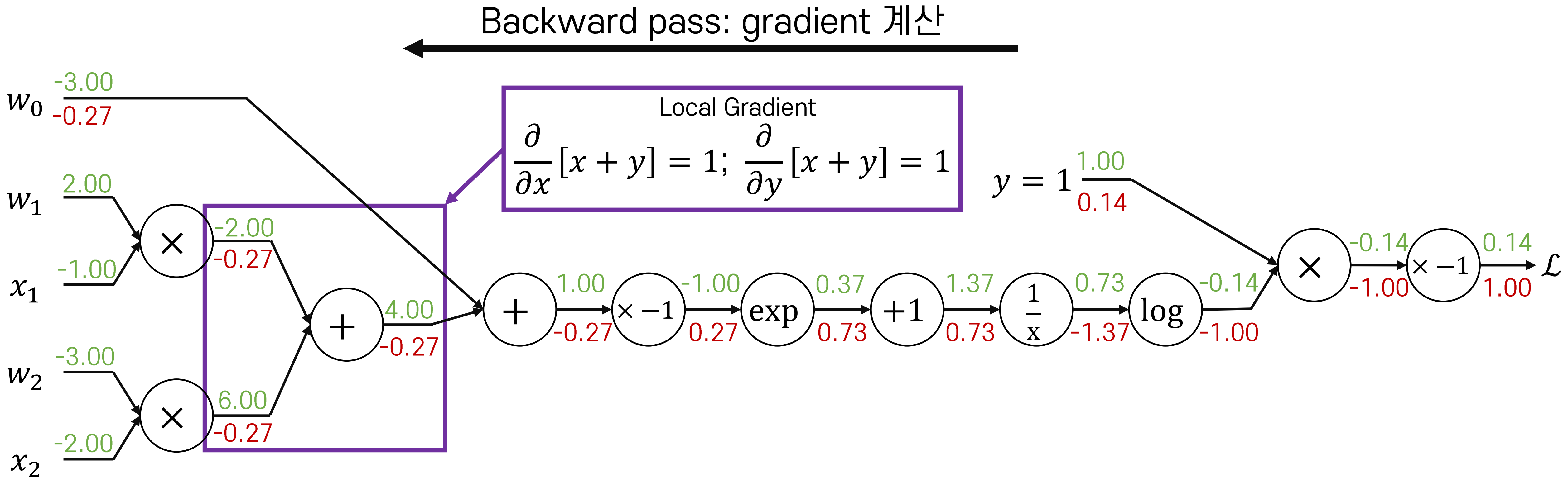
Backward pass: gradient 계산



Computational Graph

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

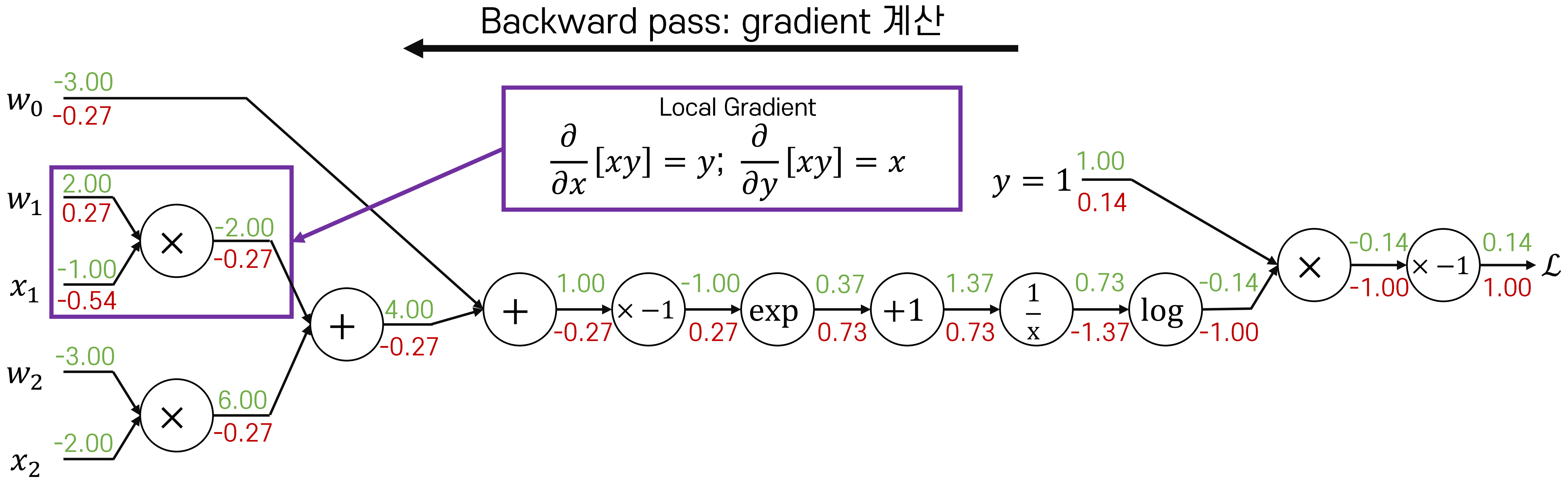
$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$



Computational Graph

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

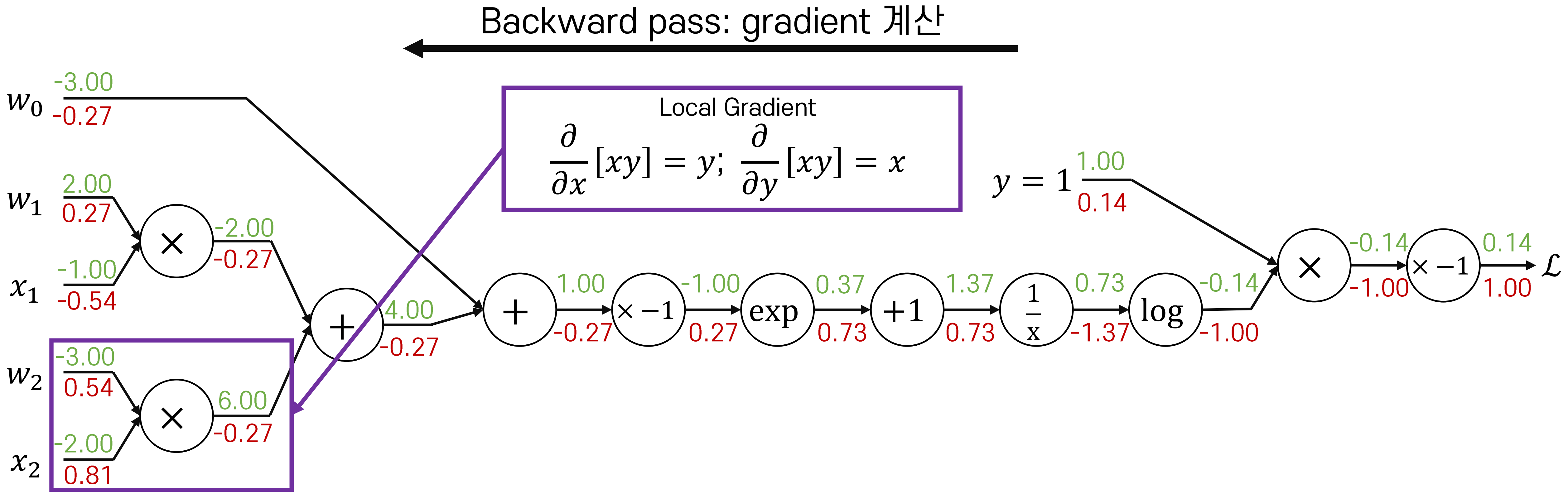
$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$



Computational Graph

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

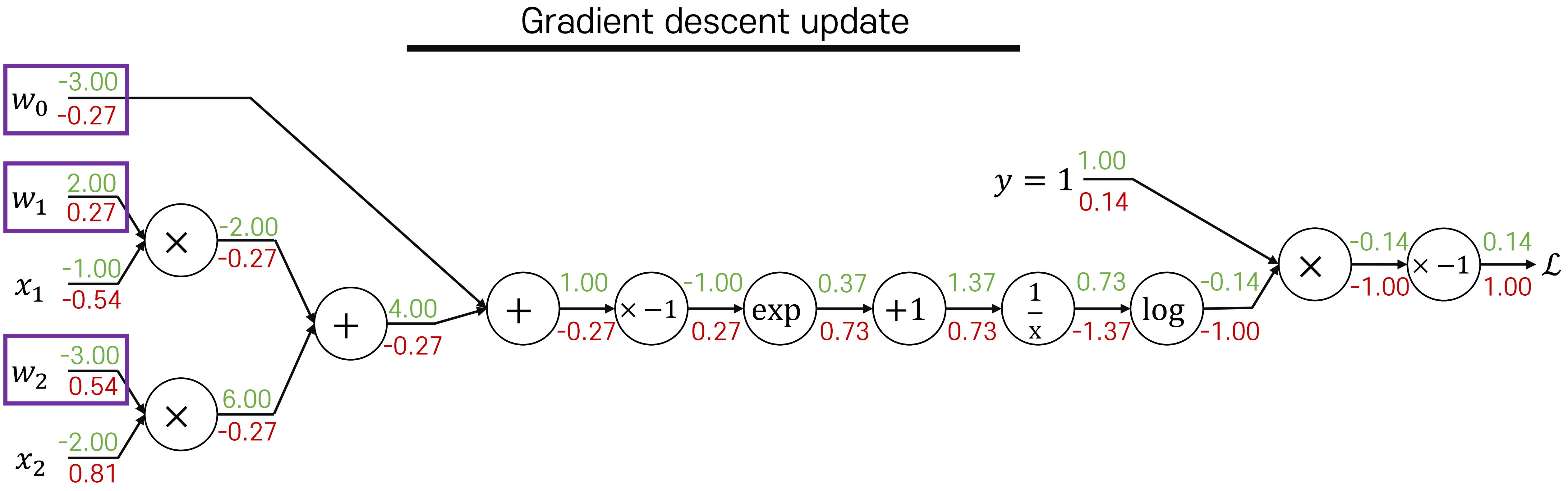
$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$



Computational Graph

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$



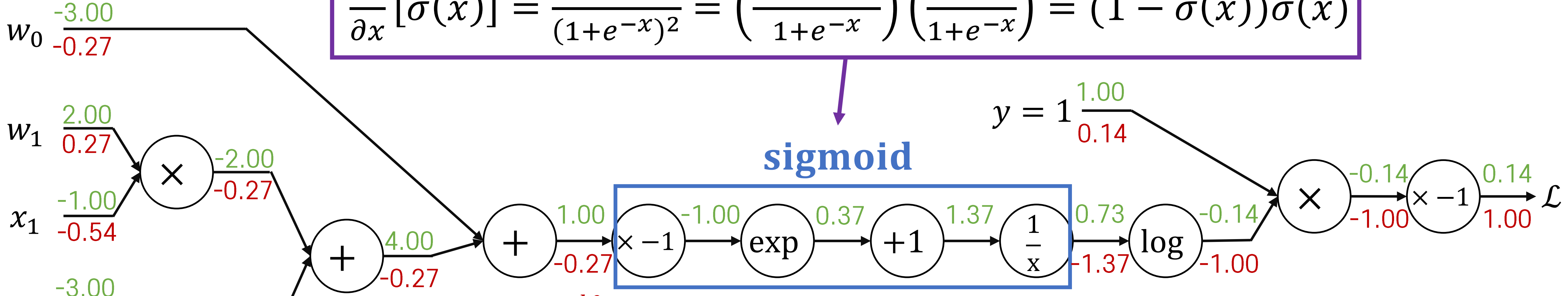
Backpropagation Path의 다양성

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y) \log(1-\hat{y})$$

Sigmoid local gradient

$$\frac{\partial}{\partial x}[\sigma(x)] = \frac{e^{-x}}{(1+e^{-x})^2} = \left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = (1-\sigma(x))\sigma(x)$$



$$\frac{d\mathcal{L}}{dx} = (1 - 0.73) \times 0.73 \times (-1.37) = -0.27$$

Computational graph는 유일하지 않음
긴 graph를 단순한 local gradient로 나타낼 수도 있음

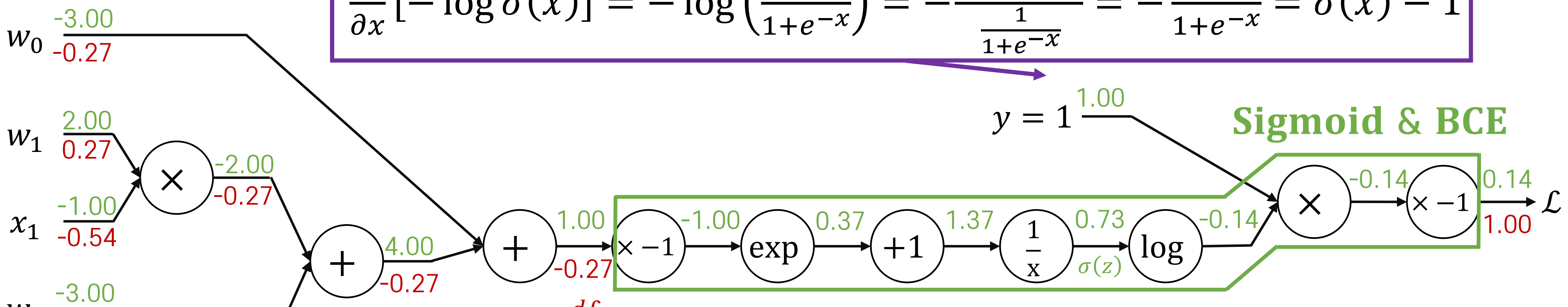
Backpropagation Path의 다양성

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y) \log(1-\hat{y})$$

Sigmoid & BCE local gradient when $y = 1$

$$\frac{\partial}{\partial x}[-\log \sigma(x)] = -\log\left(\frac{1}{1+e^{-x}}\right) = -\frac{\frac{e^{-x}}{(1+e^{-x})^2}}{\frac{1}{1+e^{-x}}} = -\frac{e^{-x}}{1+e^{-x}} = \sigma(x) - 1$$



$$\frac{d\mathcal{L}}{dx} = (1 - 0.73) \times 0.73 \times (-1.37) = -0.27$$

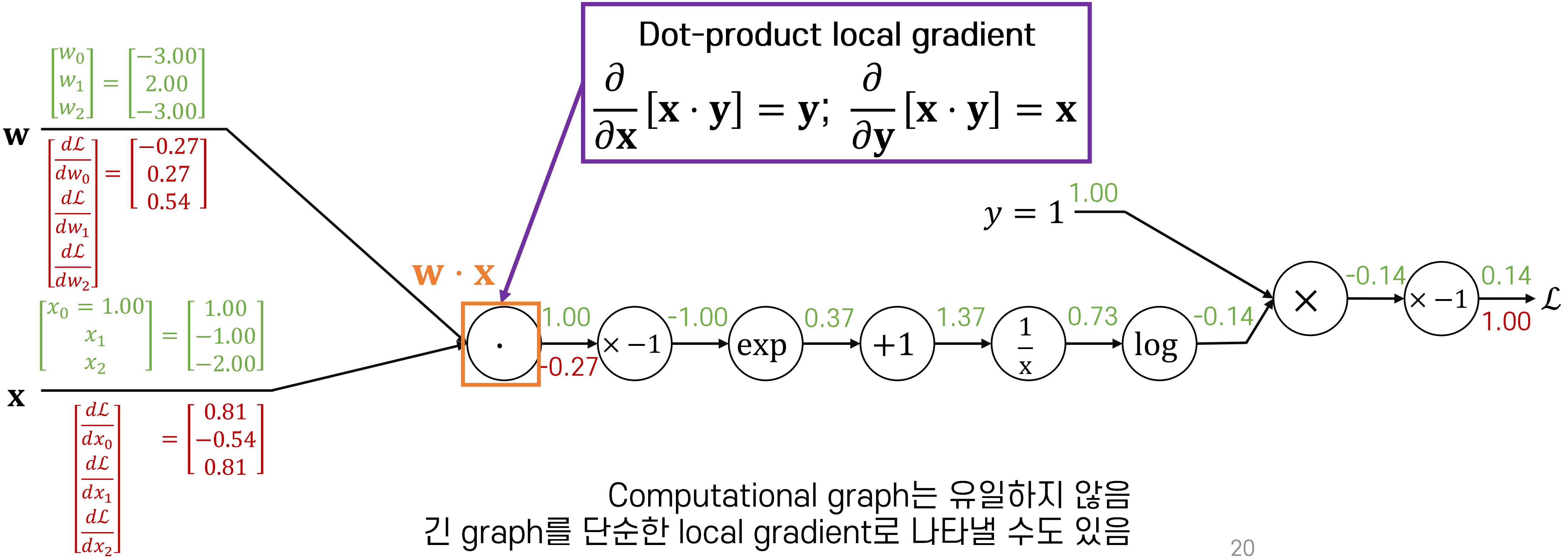
Computational graph는 유일하지 않음

긴 graph를 단순한 local gradient로 나타낼 수도 있음

Backpropagation Path의 다양성

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$



Computational graph는 유일하지 않음
긴 graph를 단순한 local gradient로 나타낼 수도 있음

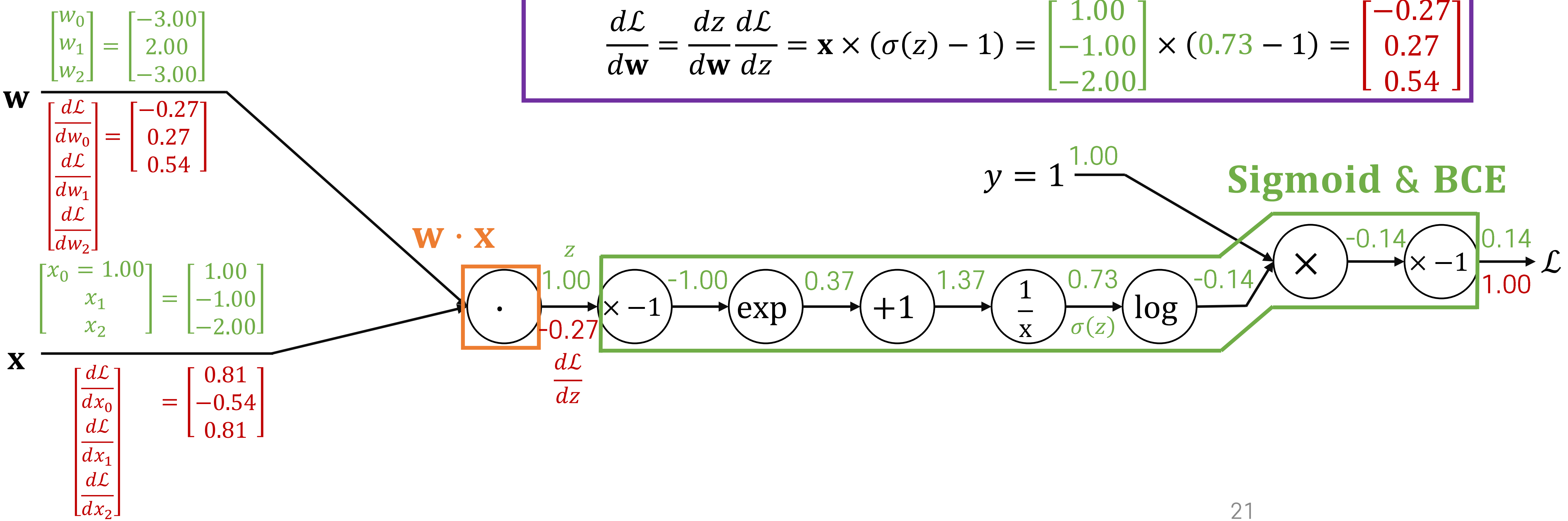
Backpropagation Path와 Chain Rule

다른 예시: $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$,

$$\mathcal{L} = \mathrm{BCE}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$

$z = \mathbf{w} \cdot \mathbf{x}$ 로 가정 →

$$\frac{d\mathcal{L}}{d\mathbf{w}} = \frac{dz}{d\mathbf{w}}\frac{d\mathcal{L}}{dz} = \mathbf{x} \times (\sigma(z) - 1) = \begin{bmatrix} 1.00 \\ -1.00 \\ -2.00 \end{bmatrix} \times (0.73 - 1) = \begin{bmatrix} -0.27 \\ 0.27 \\ 0.54 \end{bmatrix}$$



요약

- Computational graph의 복잡한 예시
- 지수 함수 및 역수 등 다양한 계산에서의 Backpropagation 적용
- 다양한 Local gradient를 활용한 Backpropagation path 단순화