Question.3-13

Linear regression을 위한 dataset이 다음과 같이 주어졌다.

$$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ \dots \ , (x^{(n)}, y^{(n)})\}$$

이때, dataset은 $y = ax$에서부터 만들어졌다.

따라서 linear regression을 통해 predictor를 학습시킬때,

model은 $\hat{y} = \theta x$, loss는 square error, cost는 MSE를 사용할 수 있다.

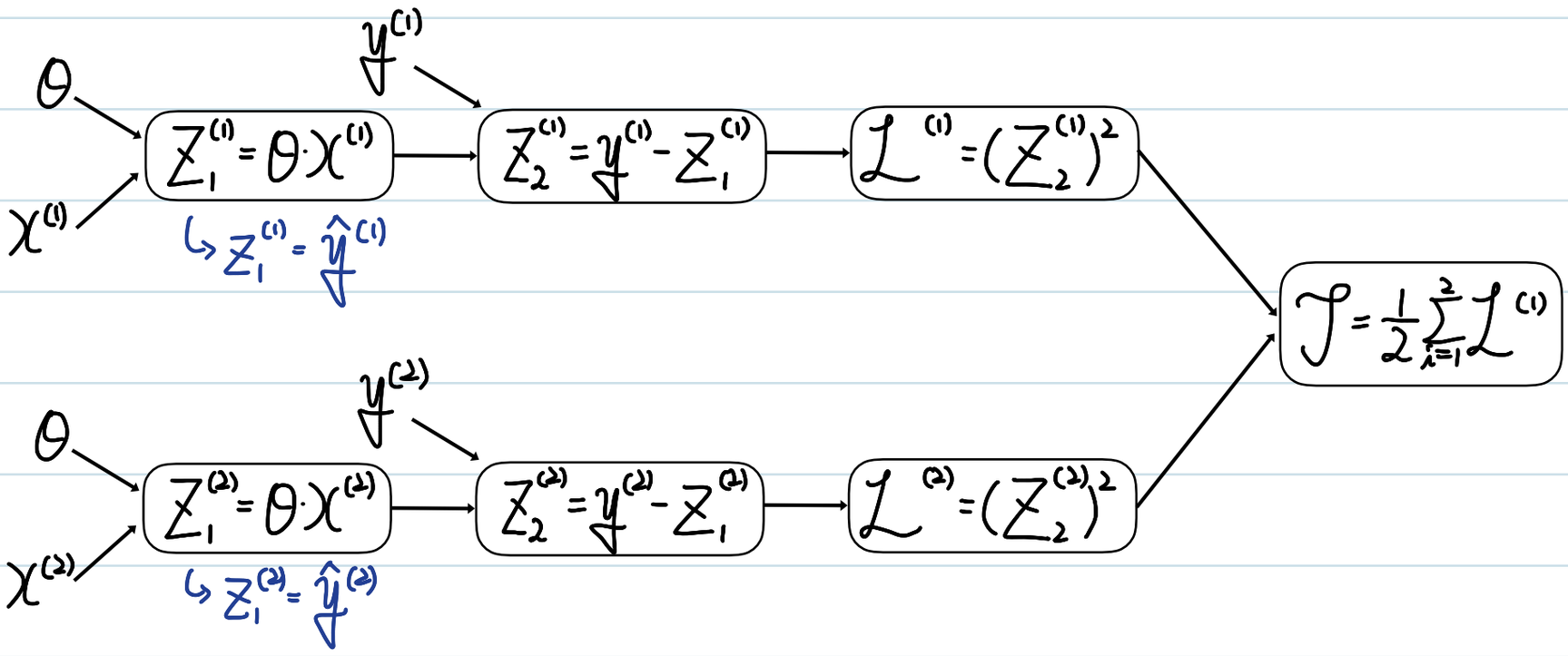
θ를 update하기 위해 2개의 data sample를 이용할때, 1번의 iteration에 대해 θ가 dataset을 잘 표현하는

θ로 update되는 과정을 설명하시오.

단, forward/backward propagation을 설명하기 위해 각 연산은 basic building node들을 이용하시오.

① model setting

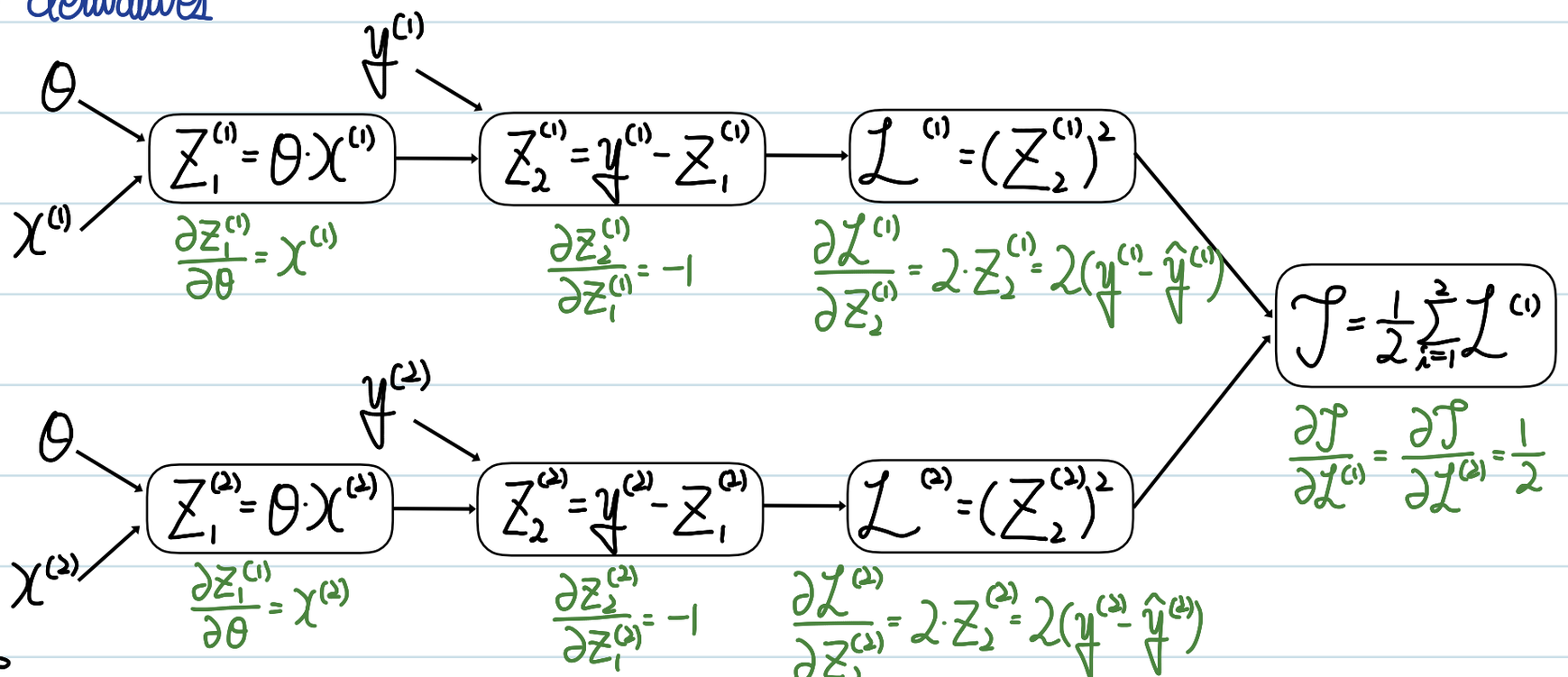
위의 상황에서 θ를 update하기 위해 2개의 data sample을 이용하기 때문에 cost에 대한 gradient descent method를 사용해야한다. 따라서 주어진 상황을 basic building node로 표현하면 다음과 같다.



이때 $Z_1^{(1)}, Z_1^{(2)}$는 각각 $\hat{y}^{(1)}, \hat{y}^{(2)}$이고 $Z_2^{(1)}, Z_2^{(2)}$는 각각 $y^{(1)} - \hat{y}^{(1)}$, $y^{(2)} - \hat{y}^{(2)}$가 된다.

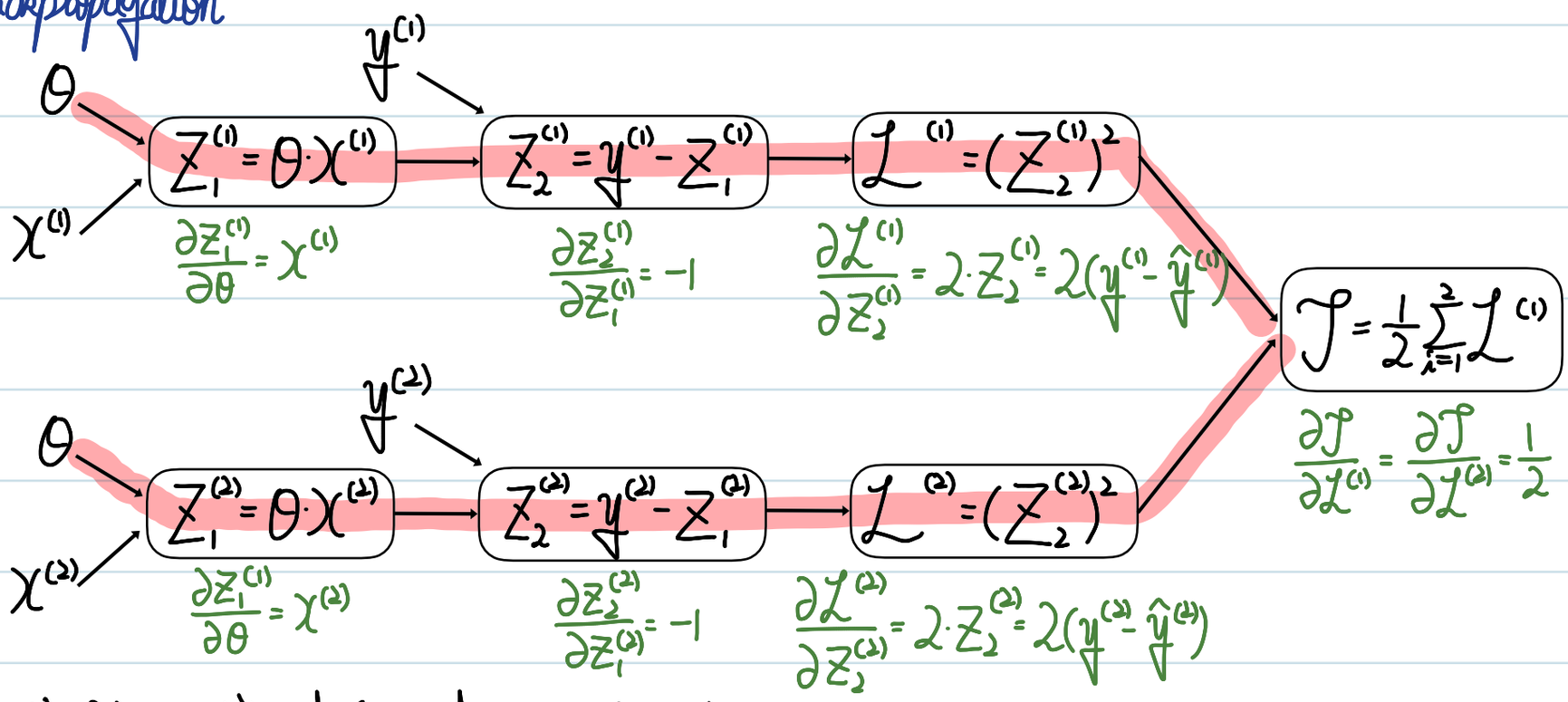
θ의 update에 필요한 partial derivative들을 구하면 다음과 같다.

② partial derivatives



따라서 $\frac{\partial J}{\partial \theta}$는 다음과 같이 partial derivative가 전달된다.

③Backpropagation



따라서 각 노드에 대한 chain rule은 다음과 같다.

$$\frac{\partial J}{\partial Z_2^{(1)}} = \frac{\partial J}{\partial L^{(1)}} \cdot \frac{\partial L^{(1)}}{\partial Z_2^{(1)}} = \frac{1}{2} \cdot 2(y^{(1)} - \hat{y}^{(1)}) \qquad \frac{\partial J}{\partial Z_2^{(2)}} = \frac{\partial J}{\partial L^{(2)}} \cdot \frac{\partial L^{(2)}}{\partial Z_2^{(2)}} = \frac{1}{2} \cdot 2(y^{(2)} - \hat{y}^{(2)})$$

$$\frac{\partial J}{\partial Z_1^{(1)}} = \frac{\partial J}{\partial Z_2^{(1)}} \cdot \frac{\partial Z_2^{(1)}}{\partial Z_1^{(1)}} = \frac{1}{2} \cdot (-2(y^{(1)} - \hat{y}^{(1)})) \qquad \frac{\partial J}{\partial Z_1^{(2)}} = \frac{\partial J}{\partial Z_2^{(2)}} \cdot \frac{\partial Z_2^{(2)}}{\partial Z_1^{(2)}} = \frac{1}{2} \cdot (-2(y^{(2)} - \hat{y}^{(2)}))$$

$$\frac{\partial J}{\partial \theta} = \frac{\partial J}{\partial Z_1^{(1)}} \cdot \frac{\partial Z_1^{(1)}}{\partial \theta} = \frac{1}{2}(-2x^{(1)}(y^{(1)} - \hat{y}^{(1)})) \qquad \frac{\partial J}{\partial \theta} = \frac{\partial J}{\partial Z_1^{(2)}} \cdot \frac{\partial Z_1^{(2)}}{\partial \theta} = \frac{1}{2}(-2x^{(2)}(y^{(2)} - \hat{y}^{(2)}))$$

위의 결과에 의해

$$\frac{\partial J}{\partial \theta} = \frac{1}{2}(-2x^{(1)}(y^{(1)} - \hat{y}^{(1)})) + \frac{1}{2}(-2x^{(2)}(y^{(2)} - \hat{y}^{(2)}))$$

$$= \frac{1}{2}\left[(-2x^{(1)}(y^{(1)} - \hat{y}^{(1)})) + (-2x^{(2)}(y^{(2)} - \hat{y}^{(2)}))\right]$$

가 된다. 여기서 각 term들은 $(x^{(1)}, y^{(1)})$, $(x^{(2)}, y^{(2)})$에 대해 $\frac{\partial L^{(1)}}{\partial \theta}$, $\frac{\partial L^{(2)}}{\partial \theta}$는

$$\frac{\partial L^{(1)}}{\partial \theta} = -2x^{(1)}(y^{(1)} - \theta \cdot x^{(1)}) \qquad \frac{\partial L^{(2)}}{\partial \theta} = -2x^{(2)}(y^{(2)} - \theta \cdot x^{(2)})$$

이므로

$$\frac{\partial J}{\partial \theta} = \frac{1}{2}\left[\frac{\partial L^{(1)}}{\partial \theta} + \frac{\partial L^{(2)}}{\partial \theta}\right]$$

이다. 즉, cost를 이용하여 θ를 update하는 $(x^{(1)}, y^{(1)})$, $(x^{(2)}, y^{(2)})$에 대해 θ가 update되는 양을 평균적으로 이용한다.

④gradient descent method

$\frac{\partial J}{\partial \theta}$와 gradient descent method를 이용하여 θ에 대한 update식을 표현하면

$$\theta := \theta - \alpha \cdot \frac{\partial J}{\partial \theta}$$

이 되고 ③에서 구한 $\frac{\partial J}{\partial \theta}$를 대입하면

$$\theta := \theta + \frac{\alpha}{2} \sum_{i=1}^{2} \frac{\partial L^{(i)}}{\partial \theta}$$

$$= \theta + \frac{\alpha}{2} \sum_{i=1}^{2} \left[2x^{(i)}(y^{(i)} - \hat{y}^{(i)})\right]$$

가 된다. 이때 $\frac{\partial J}{\partial \theta}$는 $\frac{\partial L^{(i)}}{\partial \theta}$들의 평균값이므로 각 $\frac{\partial L^{(i)}}{\partial \theta}$에 대해 Question. 3-09에서 설명한 θ의 학습원리는 2개도 적용된다.