

Assignment 2

Students name: Giuliano Martinelli 1915652, Gabriele Giannotta 1909375, Mario Dhimitri 1910181

Course: *Advanced Machine Learning* – Professor: *Fabio Galasso*

Due date: *November 19th, 2020*

Question 2 - Backpropagation

2. a

Verify that the loss function defined in Eq. (1) has gradient w.r.t. $z^{(3)}$ as Eq. (2):

$$J(\theta, \{x_i, y_i\}_{i=1}^N) = \frac{1}{N} \sum_{i=1}^N -\log \left[\frac{\exp(z_i^{(3)})_{y_i}}{\sum_{j=1}^K \exp(z_i^{(3)})_j} \right] \quad (1)$$

$$\frac{\partial J}{\partial z_i^{(3)}} \left(\theta, \{x_i, y_i\}_{i=1}^N \right) = \frac{1}{N} \left(\psi(z_i^{(3)}) - \delta_{iy_i} \right) \quad (2)$$

Where δ is the Kronecker delta:

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$$

It is possible to verify the initial assumption by calculating the gradient:

$$1. \quad f(x) = \frac{1}{N} \log(x) \rightarrow \frac{\partial f(x)}{\partial x} = -\frac{1}{Nx}$$

$$J(\theta, \{x_i, y_i\}_{i=1}^N) = f(x) = -\frac{1}{N} \log(\psi(z_i^{(3)})_{y_i}), \quad x = \psi(z_i^{(3)})_{y_i}$$

$$\frac{\partial J}{\partial \psi(z_i^{(3)})_{y_i}} = -\frac{1}{N\psi(z_i^{(3)})_{y_i}} = \frac{\partial J}{\partial a_i^{(3)}}$$

$$2. \quad f(x) = \psi(x) \rightarrow \frac{\partial f(x)}{\partial x} = \psi(x)(1 - \psi(x))$$

$$a_i^{(3)} = f(x) = \psi(z_i^{(3)}), \quad x = z_{y_i}^{(3)}$$

$$\frac{\partial a_i^{(3)}}{\partial z_i^{(3)}} = \psi(z_i^{(3)})_{y_i} (1 - \psi(z_i^{(3)})_{y_i})$$

$$3. \quad \frac{\partial J}{\partial z_i^{(3)}} = \frac{\partial J}{\partial a_i^{(3)}} \frac{\partial a_i^{(3)}}{\partial z_i^{(3)}} = -\frac{1}{N\psi(z_i^{(3)})_{y_i}} \psi(z_i^{(3)})_{y_i} (1 - \psi(z_i^{(3)})_{y_i}) = \frac{1}{N} (\psi(z_i^{(3)})_{y_i} - 1)$$

This because $\frac{\partial J}{\partial a_i^{(3)}}$ is the upstream gradient and $\frac{\partial a_i^{(3)}}{\partial z_i^{(3)}}$ is the local gradient.

2. b

To verify that the partial derivative of the loss w.r.t. $W^{(2)}$ is:

$$\begin{aligned}\frac{\partial J}{\partial W^{(2)}} \left(\theta, \{x_i, y_i\}_{i=1}^N \right) &= \sum_{i=1}^N \frac{\partial J}{\partial z_i^{(3)}} \cdot \frac{\partial z_i^{(3)}}{\partial W^{(2)}} \\ &= \sum_{i=1}^N \frac{1}{N} \left(\psi \left(z_i^{(3)} \right) - \delta_{iy_i} \right) a_i^{(2)T}\end{aligned}$$

We can use the property as follows:

$$f(x) = aW, \quad \frac{\partial f(x)}{\partial a} = W, \quad \frac{\partial f(x)}{\partial W} = a$$

Using upstream and local gradient, we can apply the chain rule:

$$\frac{\partial J}{\partial W_2} = \frac{\partial J}{\partial z_i^{(3)}} \frac{\partial z_i^{(3)}}{\partial W_2} \quad \frac{\partial z_i^{(3)}}{\partial W_2} = a_i^{(2)} \quad \text{since } z_i^{(3)} = W_2 a_i^{(2)} + b$$

$$\frac{\partial J}{\partial W_2} = \frac{1}{N} (\psi(z_i^{(3)}) - 1) a_i^{(2)}$$

To verify that the regularized loss in Eq. (3) has the derivative as Eq. (4):

$$\tilde{J} \left(\theta, \{x_i, y_i\}_{i=1}^N \right) = \frac{1}{N} \sum_{i=1}^N -\log \left[\frac{\exp(z_i^{(3)})_{y_i}}{\sum_{j=1}^K \exp(z_i^{(3)})_j} \right] + \lambda \left(\|W^{(1)}\|_2^2 + \|W^{(2)}\|_2^2 \right) \quad (3)$$

$$\frac{\partial \tilde{J}}{\partial W^{(2)}} = \sum_{i=1}^N \frac{1}{N} \left(\psi \left(z_i^{(3)} \right) - \delta_{iy_i} \right) a_i^{(2)T} + 2\lambda W^{(2)} \quad (4)$$

We can do the following:

$$f(x) = \lambda \left(\|W^{(1)}\|_2^2 + \|W^{(2)}\|_2^2 \right) = \lambda \left(\|W^{(1)}\|_2^2 \right) + \lambda \left(\|W^{(2)}\|_2^2 \right) =$$

$$\frac{f}{W_2} = 0 + \lambda \frac{\partial(\sum \sum (W_2)^2)}{\partial W_2} = 2\lambda W_2$$

$$\frac{\partial J}{\partial W_2} = \frac{1}{N} (\psi(z_i^{(3)}) - 1) a_i^T + 2\lambda W_2$$

2. c

We now derive the expressions for the derivatives of the regularized loss in Eq. (3) w.r.t. $W(1)$, $b(1)$, $b(2)$, recalling that we used Δ_i to refer to the Kronecker delta.

- $\frac{\partial J}{\partial z_i^{(3)}} = -\frac{1}{N}(\psi(z_i^{(3)}) - \Delta_i)$
- $z_i^{(3)} = a_i^{(3)}W^{(2)} + b^{(2)}$, so as we have seen in 2.b:

$$\frac{\partial J}{\partial W_2} = \frac{1}{N}(\psi(z_i^{(3)}) - \Delta_i)a_{2i}$$

For the same reason:

- $\frac{\partial J}{\partial a_i^{(2)}} = \frac{1}{N}(\psi(z_i^{(3)}) - \Delta_i)W_2$
- $\frac{\partial J}{\partial b_2} = \frac{\partial J}{\partial z_i^{(3)}} = \frac{1}{N}(\psi(z_i^{(3)}) - \Delta_i)$

Using the chain rule:

- $\frac{\partial J}{\partial z_i^{(2)}} = \frac{\partial J}{\partial a_i^{(2)}} \frac{\partial a_i^{(2)}}{\partial z_i^{(2)}} \rightarrow a_i^{(2)} \begin{cases} 0, & \text{if } z_i^{(2)} < 0 \\ z_i^{(2)}, & \text{if } z_i^{(2)} \geq 0 \end{cases} \rightarrow \frac{\partial a_i^{(2)}}{\partial z_i^{(2)}} = \delta_i = \begin{cases} 0, & \text{if } z_i^{(2)} < 0 \\ 1, & \text{if } z_i^{(2)} \geq 0 \end{cases}$
- $$\frac{\partial J}{\partial z_i^{(2)}} = \frac{1}{N}(\psi(z_i^{(3)}) - \Delta_i)\delta_i$$

- $\frac{\partial J}{\partial b_1} = \frac{\partial J}{\partial z_i^{(2)}} = \frac{1}{N}(\psi(z_i^{(3)}) - \Delta_i)\delta_i$
- $\frac{\partial J}{\partial W_1} = \frac{\partial J}{\partial z_i^{(2)}} a_1^{(1)} = \frac{1}{N}(\psi(z_i^{(3)}) - \Delta_i)\delta_i a_1^{(1)}$
- $\frac{\partial J}{\partial a_i} = \frac{\partial J}{\partial z_i^{(2)}} W_1 = \frac{1}{N}(\psi(z_i^{(3)}) - \Delta_i)\delta_i W_1$