Numerical Optimization Assignment 3

David Bulovic

January 2022

We know that input dimension $N_I = 4$ and output dimension $N_O = 3$. Also:

$$\mathbf{x} \in \mathbb{R}^4$$

$$\mathbf{z}^{(1)} = \mathbf{W}^{(0)}\mathbf{x} + \mathbf{b}^{(0)} \in \mathbb{R}^{N_H}$$

We obtain $\mathbf{a}^{(1)}$ from $\mathbf{z}^{(1)}$.

$$\mathbf{z}^{(2)} = \mathbf{W}^{(1)}\mathbf{a}^{(1)} + \mathbf{b}^{(1)} \in \mathbb{R}^{N_O}$$

From this we can conclude that:

$$\mathbf{W}^{(0)} \in \mathbb{R}^{N_H \times 4}$$

$$\mathbf{b}^{(0)} \in \mathbb{R}^{N_H}$$

$$\mathbf{W}^{(1)} \in \mathbb{R}^{3 \times N_H}$$

$$\mathbf{b}^{(1)} \in \mathbb{R}^3$$

To calculate the number of learnable parameters we can use the following function:

$$f(N_H) = 4N_H + N_H + 3N_H + 3 = 8N_H + 3$$

The entirety of the forward pass:

$$\mathbf{z}^{(1)} = \mathbf{W}^{(0)}\mathbf{x} + \mathbf{b}^{(0)}$$

$$\mathbf{a}^{(1)} = h(\mathbf{z}^{(1)}) \iff a_i^{(1)} = ln(1 + exp(z_i^{(1)}))$$

$$\mathbf{z}^{(2)} = \mathbf{W}^{(1)}\mathbf{a}^{(1)} + \mathbf{b}^{(1)}$$

$$\tilde{y} = g(\mathbf{z}^{(2)}) \iff \tilde{y}_i = \frac{exp(z_i^{(2)})}{\sum_{j=1}^{N_O} exp(z_j^{(2)})}$$

The loss function is the cross-entropy loss:

$$l(\tilde{y}^s, y^s) = -\sum_{i=1}^{N_O} y_i^s ln(\tilde{y}_i^s)$$

$$\mathcal{L} = \frac{1}{S}\sum_{s=1}^{S} l(\tilde{y}^s, y^s)$$

We can use the chain rule to calculate the derivatives w.r.t. the learnable parameters:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(1)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{(2)}} \frac{\partial \mathbf{a}^{(2)}}{\partial \mathbf{z}^{(2)}} \frac{\partial \mathbf{z}^{(2)}}{\partial \mathbf{W}^{(1)}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(1)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{(2)}} \frac{\partial \mathbf{a}^{(2)}}{\partial \mathbf{z}^{(2)}} \frac{\partial \mathbf{z}^{(2)}}{\partial \mathbf{b}^{(1)}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(0)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{(2)}} \frac{\partial \mathbf{a}^{(2)}}{\partial \mathbf{z}^{(2)}} \frac{\partial \mathbf{z}^{(2)}}{\partial \mathbf{a}^{(1)}} \frac{\partial \mathbf{a}^{(1)}}{\partial \mathbf{z}^{(1)}} \frac{\partial \mathbf{z}^{(1)}}{\partial \mathbf{W}^{(0)}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(0)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{a}^{(2)}} \frac{\partial \mathbf{a}^{(2)}}{\partial \mathbf{z}^{(2)}} \frac{\partial \mathbf{z}^{(2)}}{\partial \mathbf{a}^{(1)}} \frac{\partial \mathbf{a}^{(1)}}{\partial \mathbf{z}^{(1)}} \frac{\partial \mathbf{z}^{(1)}}{\partial \mathbf{b}^{(0)}}$$

We can see that for all four derivatives the initial derivations is the same, which is what we will calculate first:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{(2)}} \frac{\partial \mathbf{a}^{(2)}}{\partial \mathbf{z}^{(2)}} \iff \frac{\partial \mathcal{L}}{\partial \mathbf{z}_i^{(2)}} = \tilde{y}_i - y_i = \mathbf{e}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(1)}} = (\mathbf{a}^{(1)})^T \mathbf{e}^{(2)}$$

$$\frac{\partial \mathbf{z}^{(2)}}{\partial \mathbf{b}^{(1)}} = 1$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(1)}} = \mathbf{e}^{(2)}$$

For the next two derivatives we need the following:

$$\frac{\partial \mathbf{z}^{(2)}}{\partial \mathbf{a}^{(1)}} = \mathbf{W}^{(1)}$$

$$\frac{\partial \mathbf{a}^{(1)}}{\partial \mathbf{z}^{(1)}} \iff \frac{\partial}{\partial \mathbf{z}_i^{(1)}} ln(1 + exp(z_i^{(1)})) = \frac{exp(z_i^{(1)})}{1 + exp(z_i^{(1)})}$$

$$\frac{\partial \mathbf{z}^{(2)}}{\partial \mathbf{a}^{(1)}} \frac{\partial \mathbf{a}^{(1)}}{\partial \mathbf{z}^{(1)}} = \frac{exp(z_i^{(1)})}{1 + exp(z_i^{(1)})} \mathbf{W}^{(1)} = \mathbf{e}^{(1)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(0)}} = \mathbf{x}^T \mathbf{e}^{(1)} \mathbf{e}^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(0)}} = \mathbf{e}^{(1)} \mathbf{e}^{(2)}$$

In the python file *main.py* the learnable parameters are initialized via normal distribution with $\mu = 0$ and $\sigma = 0.05$.