

# Assignment 1 Notes

mn

## Contents

<b>1</b>	<b>Definitions</b>	<b>1</b>
<b>2</b>	<b>Assignment 1</b>	<b>1</b>
2.1	1.a) Softmax function . . . . .	1
2.2	2.a) Derivative of the sigmoid function . . . . .	1
2.3	2.b) Derivative of the Cross Entropy Loss . . . . .	1
2.4	2.c) Gradient of the one-layer network . . . . .	2
2.4.1	Alternative (more layer-systematic) derivation . . . . .	2
2.5	2.d) Number of parameters . . . . .	2

## 1 Definitions

- row vectors:
- 1 layer network:
  - Input dimension:  $D_x$  ( $\mathbf{x} \in [N, D_x]$ )
  - Hidden units:  $H$  ( $\mathbf{W}_1 \in [D_x, H]$ ,  $\mathbf{b}_1 \in [H]$ )
  - Output dimension:  $D_y$  ( $\mathbf{W}_2 \in [H, D_y]$ ,  $\mathbf{b}_2 \in [D_y]$ )

## 2 Assignment 1

Org latex export is not that great yet, and one still has to work on the alignment.

### 2.1 1.a) Softmax function

Softmax function:  $\text{softmax}(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$

$$\text{softmax}(x+c)_i = \frac{e^{x_i+c}}{\sum_j e^{x_j+c}} = \frac{e^c e^{x_i}}{e^c \sum_j e^{x_j}} = \frac{e^{x_i}}{\sum_j e^{x_j}} = \text{softmax}(x)_i$$

### 2.2 2.a) Derivative of the sigmoid function

Sigmoid function:  $\sigma(x) = \frac{1}{1+e^{-x}}$

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

### 2.3 2.b) Derivative of the Cross Entropy Loss

Cross Entropy Loss:  $CE(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_i y_i \log(\hat{y}_i)$

$$\hat{\mathbf{y}} = \text{softmax}(\theta)$$

$$\frac{\partial}{\partial \theta_j} CE(\mathbf{y}, \hat{\mathbf{y}})_j = -\frac{\partial}{\partial \theta_j} \sum_i y_i \log \hat{y}_i = -\frac{\partial}{\partial \theta_j} \sum_i y_i \log(e^{\theta_i} / \sum_l e^{\theta_l}) = -\frac{\partial}{\partial \theta_j} \sum_i y_i (\theta_i - \log \sum_l e^{\theta_l})$$

$\mathbf{y}$  is a *one-hot* label vector with a non-zero value (=1) at index  $k$ .

Given that

$$\frac{\partial}{\partial \theta_j} \log \sum_l e^{\theta_l} = \frac{e_j^\theta}{\sum_l e^{\theta_l}} = \hat{y}_j$$

if  $j \neq k$ :

$$\frac{\partial}{\partial \theta_j} CE(\mathbf{y}, \hat{\mathbf{y}})_j = \frac{\partial}{\partial \theta_j} \log \sum_l e^{\theta_l} = \hat{y}_j$$

if  $j = k$ :

$$\frac{\partial}{\partial \theta_j} CE(\mathbf{y}, \hat{\mathbf{y}})_j = \hat{y}_j - 1$$

Combining (vectorizing) these results:

$$\frac{\partial}{\partial \boldsymbol{\theta}} CE(\mathbf{y}, \hat{\mathbf{y}}) = \hat{\mathbf{y}} - \mathbf{y}$$

## 2.4 2.c) Gradient of the one-layer network

$$\frac{\partial J(x)}{\partial x} = \frac{\partial}{\partial x} CE(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{\partial}{\partial x} \sum_i y_i \log(\text{softmax}(\sigma(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2))$$

With  $\hat{\mathbf{y}} = \text{softmax}(\theta)$ ,  $\theta = \sigma(\mathbf{h})\mathbf{W}_2 + \mathbf{b}_2$  and  $\mathbf{h} = \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1$

$$\frac{\partial J(x)}{\partial x} = \sum_j \frac{\partial J}{\partial \theta_j} \frac{\partial \theta_j}{\partial x} = \sum_j \sigma(\mathbf{h})(1 - \sigma(\mathbf{h}))\mathbf{W}_1\mathbf{W}_{2,j} \times \begin{cases} \hat{y}_j & \text{if } j \neq k \\ \hat{y}_j - 1 & \text{if } j = k \end{cases}$$

### 2.4.1 Alternative (more layer-systematic) derivation

## 2.5 2.d) Number of parameters

Number of parameters:  $(D_x + 1)H + (H + 1)D_y$