Γεώργιος Χείρμπος 3130230
1η Εργασία Μηχανική Μάθηση

1 Symbols

Ei	Cost of node i
ai layer	activation value of node i in layer
z i layer	input of node i [w*x]
w ij	weight connecting node i to node j
x	input value

2 W1, W2 equations

Partial derivative of Cost by node weight

$$\frac{\partial E_i}{\partial w_i j^{output}} = (\frac{\partial E_i}{\partial a_i^{output}})(\frac{\partial a_i^{output}}{\partial z_i^{output}})(\frac{\partial z_i^{output}}{\partial w_i j^{output}}), i = outputnode, j = outputweight \tag{1}$$

Partial derivative of Cost by activation value of output node i

$$\frac{\partial E_i}{\partial a_i^{output}} = derivated.of.cost.function \tag{2}$$

Partial derivative of activation function of output node. [derivate of activation(w*x)]

$$\frac{\partial a_i^{output}}{\partial z_i^{output}} = derivative.of.activation.output(zi = wx) \tag{3}$$

Partial derivative of input value [w*x] of output node, by w. = [x -> activated value of hidden]

$$\frac{\partial z_i^{output}}{\partial w_i j^{output}} = output.value.of.hidden \tag{4}$$

Partial derivatives of hidden layer are given below

$$\frac{\partial E_i}{\partial w_i j^{hidden}} = (\frac{\partial E_i}{\partial a_i^{hidden}})(\frac{\partial a_i^{hidden}}{\partial z_i^{hidden}})(\frac{\partial z_i^{hidden}}{\partial w_i j^{hidden}}), i = hiddennode, j = hiddenweight \tag{5}$$

First part consists of the following components

$$\frac{\partial E_i}{\partial a_i^{hidden}} = \sum_{i=0}^{n-1}((\frac{\partial E_i}{\partial a_i^{output}})(\frac{\partial a_i^{output}}{\partial z_i^{output}})(\frac{\partial z_i^{output}}{\partial a_i^{hidden}})) \tag{6}$$

Similar to partial derivatives of Cost. Except for the SUM and

$$\frac{\partial z_i^{output}}{\partial a_i^{hidden}} = W2 \tag{7}$$

which is the partial derivative of w*x value of the input of output node by the activated value of the hidden layer. [activated value of hidden layer = x]. [output layer weight = w].

Second part are:

Same logic as for output layer

$$\frac{\partial a_i^{hidden}}{\partial z_i^{hidden}} = derivative.of.activation.hidden(zi = wx) \tag{8}$$

Same logic as for output layer

$$\frac{\partial z_i^{hidden}}{\partial w_i j^{hidden}} = X_i \tag{9}$$

Combining the equation of partial derivatives of W2 with the above, the final form of partial derivatives of W1 are given by:

$$gradsEW1 = der.activation.hidden(W1 * X) * ((T - Y) * W2)^T * X \tag{10}$$

3 Results

Variable Lambda was not so important in the overall prediction. Batch size, hidden nodes and especially epochs played a bigger role overall. Learning rate after a certain limit the was no point in training as the system overflowed, too small and the system did not move. After crossvalidating various values for the hyperparameters, we list some of the results. Also cifar was grayscaled for overflow avoidance reasons.

3.1 MNIST

Dataset	Lamda	Hidden Nodes	Activation	ETA	Batch Size	Epochs	Accuracy
Mnist	1e-3	100	log(1+exp(a))	1e-3	100	50	0.9722
Mnist	5e-3	100	log(1+exp(a))	1e-3	100	50	0.9713
Mnist	5e-2	200	log(1+exp(a))	1e-3	100	70	0.9747
Mnist	5e-2	100	log(1+exp(a))	1e-3	200	70	0.9746
Mnist	5e-2	200	log(1+exp(a))	1e-3	200	100	0.9754
Mnist	1e-3	100	log(1+exp(a))	1e-3	300	20	0.9658
Mnist	—	—	log(1+exp(a))	>1e-2	—	>2	destabilazation
Mnist	1e-3	200	tanh(a)	1e-3	300	30	0.9756
Mnist	2e-3	200	tanh(a)	1e-3	200	70	0.9798
Mnist	2e-3	300	tanh(a)	1e-3	300	25	0.9744
Mnist	—	—	tanh(a)	>1e-2	—	>2	destabilazation
Mnist	5e-1	100	cos(a)	1e-3	300	50	0.9741
Mnist	1e-3	200	cos(a)	1e-3	300	50	0.9782
Mnist	1e-3	100	cos(a)	1e-2	100	50	0.6712
Mnist	1e-3	100	cos(a)	1e-1	100	10	0.104
Mnist	1e-2	200	cos(a)	1e-2	200	50	0.817

3.2 CIFAR

Dataset	Lamda	Hidden Nodes	Activation	ETA	Batch Size	Epochs	Accuracy
Mnist	5e-3	100	log(1+exp(a))	1e-4	100	50	0.3991
Mnist	5e-3	100	log(1+exp(a))	1e-4	200	50	0.4155
Mnist	1e-2	200	log(1+exp(a))	1e-4	200	50	0.4028
Mnist	1e-3	200	log(1+exp(a))	1e-4	300	25	0.3644
Mnist	1e-3	100	log(1+exp(a))	1e-4	300	70	0.4251
Mnist	1e-3	300	log(1+exp(a))	1e-3	300	70	0.4181
Mnist	—	—	log(1+exp(a))	>1e-3	—	>2	destabilazation
Mnist	2e-3	100	tanh(a)	1e-4	100	70	0.4599
Mnist	1e-3	300	tanh(a)	1e-4	300	70	0.4499
Mnist	1e-3	100	tanh(a)	1e-4	300	50	0.442
Mnist	1e-3	200	tanh(a)	1e-4	100	30	0.4063
Mnist	—	—	tanh(a)	>1e-3	—	>2	destabilazation
Mnist	1e-3	200	cos(a)	1e-4	100	30	0.4079
Mnist	100	200	cos(a)	1e-4	300	30	0.2995
Mnist	1e-1	20	cos(a)	1e-4	200	50	0.4542
Mnist	1e-1	200	cos(a)	1e-1	200	60	0.4614