KTH Royal Institute of Technology

Deep Learning in Datascience
DD2424

Assignment 1

Authors:
Ali Banaei Mobarak Abadi

April 1, 2022

1 Assignment Report

In this section, the results we can get by running the code are presented. For an explanation of the code and implementations, please see section 2. Note that for the rest of this report, the available training data is divided into two parts, validation with a size of 5000 samples and the rest as training.

1.1 Checking the gradients

After implementing the code for the forward path and calculating the gradient, calculated gradients were checked using the provided functions (for the final run, the slower version was used). To make the gradients big enough so that we can use the difference without a need for a division, the weight and bias matrices were initialized with a Gaussian distribution with a mean of 0 and a variance of 4. After calculating the gradients using the two methods, the mean, standard deviation, and maximum values for the absolute values of the difference of gradients were calculated by two methods, and the min, max and the SD of numerical gradients were printed. The results were as follows.

```
For abs of diff of gW: mean: 5.12e-09, std: 3.75e-09, max:2.11e-08, gradient min: -1.72,
  gradient max: 1.53, gradient std: 0.42
For abs of diff of gb: mean: 1.60e-09, std: 1.33e-09, max:4.21e-09, gradient min: -0.1,
  gradient max: 0.064, gradient std: 0.12
```

As we can see, the differences are vary small while the gradients are not. So, we can conclude we are calculating the gradients correctly.

1.2 Training and evaluation

After implementing the remaining functions, the model was trained and tested using the provided hyperparameters. Figure 1 depicts the loss, cost function, and accuracy of training and validation sets during training. Also, the results of the evaluation of the model on test data can be found in Table 1.

Table 1: Evaluation of the model with different values of λ on the test set.

λ	η	Loss	Cost	Accuracy
0	0.1	6.92	6.92	0.27
0	0.001	2.02	2.02	0.32
0.1	0.001	1.74	1.80	0.41
1	0.001	1.85	1.92	0.37

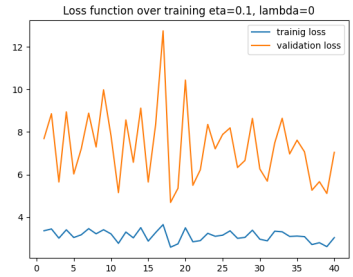
We can see that when we set 0.1 as the learning rate, the network reaches a not-so-bad point after the first iteration; however, it keeps overshooting the optimal values and does not converge to a good set of weights. In this setting, the network jumps around, and it does not have adequate performance in the evaluation. However, when we change the learning rate to 0.001, the network converges, and the final accuracy increases. It is worth mentioning that when we do not use regularization ($\lambda = 0$), the validation and training loss and accuracy have a considerable difference, and the network's performance is better on training data that it has seen. As expected, when we increase λ, this difference decreases, and the generalization performance of the network enhances. However, when we have $\lambda = 1$, the network's performance decreases. It may be because this value would enforce a strong constraint on our network, which structure is very simple. So, the network would not be able to optimize the weights to decrease the loss function effectively.

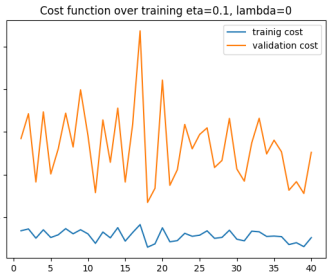
Also, the visualization of the W can be found in Figure 2. As we see, by increasing the λ, we have a smoother matrix. Also, when we do not use regularization, the network does not seem to capture any repetitive pattern in each class, while by using regularization, instead of memorizing the training set, the model learns these patterns.

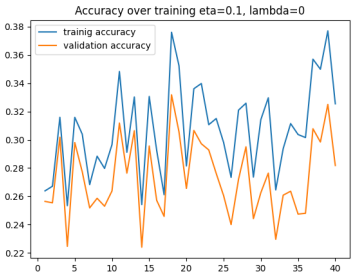
2 Implementation Documentation

blabla

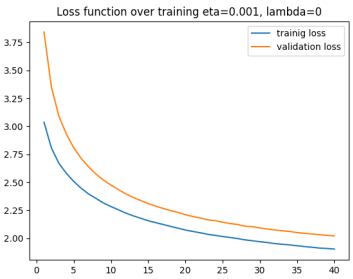
References

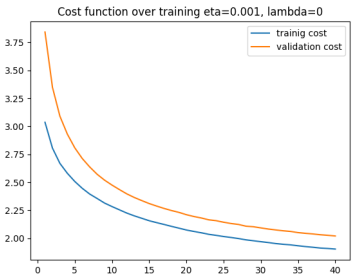


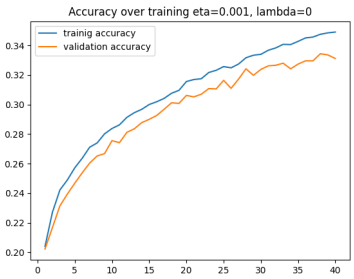




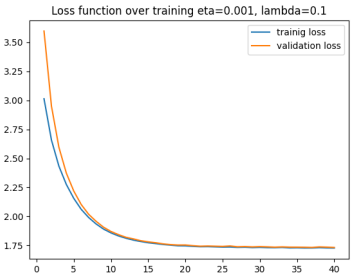
(a) $\lambda = 0,\ \eta = 0.1$

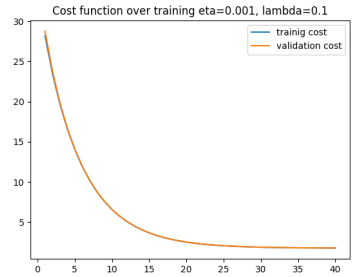


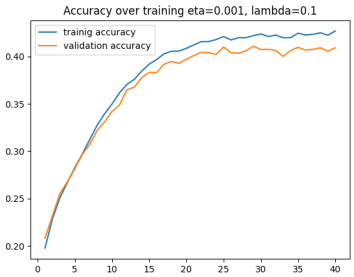




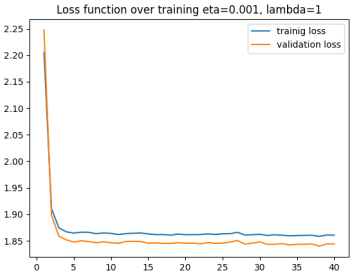
(b) $\lambda = 0,\ \eta = 0.001$

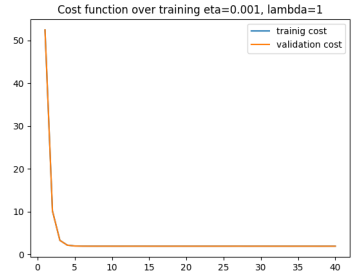


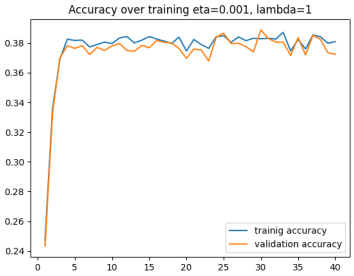




(c) $\lambda = 0.1,\ \eta = 0.001$

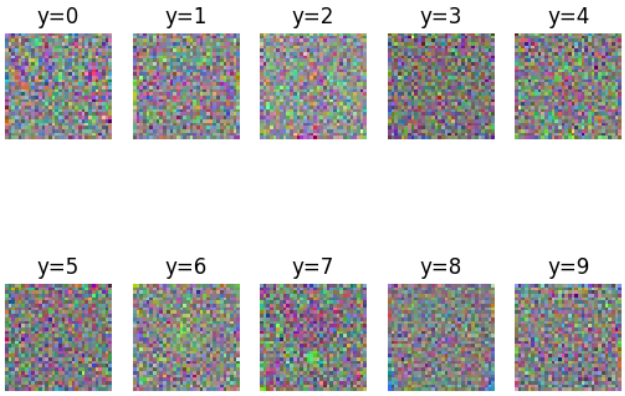




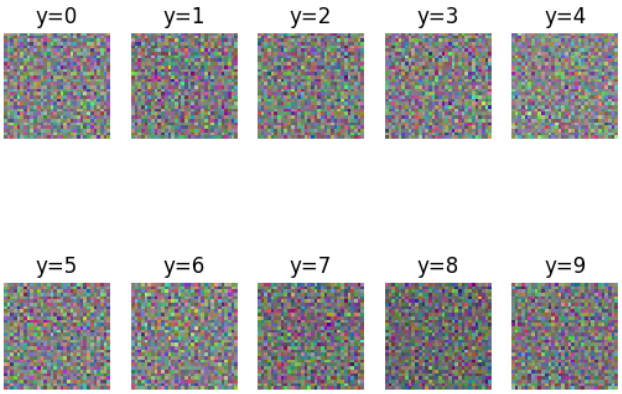


(d) $\lambda = 1,\ \eta = 0.001$

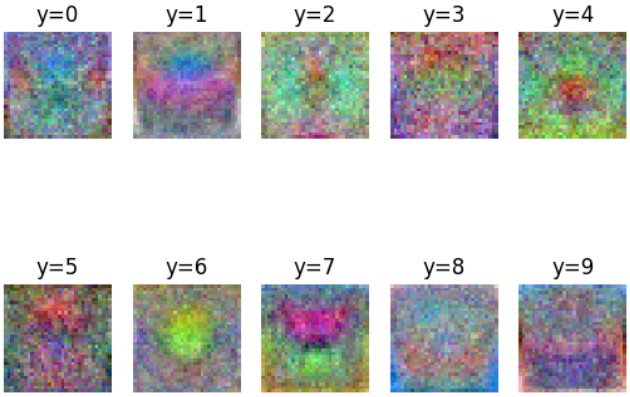
Figure 1: Loss, cost and accuracy of the model during training.



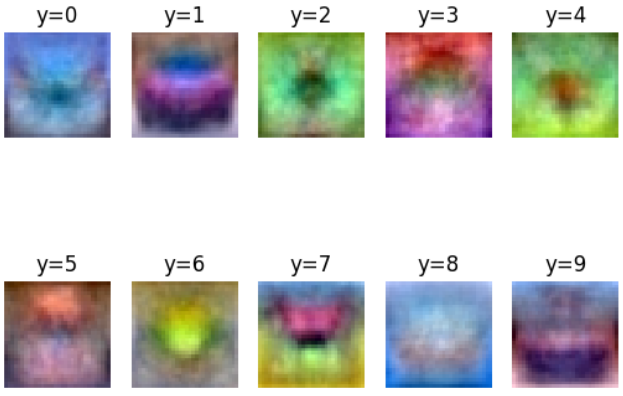
(a) $\lambda = 0,\ \eta = 0.1$



(b) $\lambda = 0,\ \eta = 0.001$



(c) $\lambda = 0.1,\ \eta = 0.001$



(d) $\lambda = 1,\ \eta = 0.001$

Figure 2: Visualization of W after training.