DD2424 Assignment1 Mandatory Part

Shuyuan Zhang   shuyuanz@kth.se

March 2019

1 Brief Introduction

In this assignment, we were asked to implement a one layer network to classify cifar-10 data set. I successfully implemented the following functions to compute the gradients analytically.

LoadBatch: Reads the data set and generate one-hot matrix.

EvaluateClassifier: Computes the forward pass and evaluates the network function.

ComputeCost: Returns a loss/cost value given current network structure and parameters.

ComputeAccuracy: Computes the successful prediction rate of our model on data sets.

ComputeGradients: Computes the backward pass and figures out gradients analytically according to:

- Complete the **forward pass**

$$\mathbf{P}_{\text{batch}} = \text{SoftMax}\left(W\mathbf{X}_{\text{batch}} + \mathbf{b}\mathbf{1}_{n_b}^T\right)$$

- Complete the **backward pass**

  1. Set

  $$\mathbf{G}_{\text{batch}} = -\left(\mathbf{Y}_{\text{batch}} - \mathbf{P}_{\text{batch}}\right)$$

  2. Then

  $$\frac{\partial L}{\partial W} = \frac{1}{n_b}\mathbf{G}_{\text{batch}}\mathbf{X}_{\text{batch}}^T, \quad \frac{\partial L}{\partial \mathbf{b}} = \frac{1}{n_b}\mathbf{G}_{\text{batch}}\mathbf{1}_{n_b}$$

Figure 1: Efficient Gradient-computing Equation

MiniBatchGD: Performs the actual GD process.

Details of these functions can be seen in the .m file attached together with this report.

2 Gradient Check

To check whether the gradient was computed correctly, I compared my gradients with results of the provided function *ComputeGradsNumSlow*. The discrepancy between results(W,b)given by different functions is measured by:

$$\frac{|g_a - g_n|}{max(1e-15, |g_a| + |g_n|)}$$

I performed three independent tests on mini-batch gradients of 100 data points and the differences are shown in table below:

Test No.	W difference	b difference
1	1.3678e-07	1.7735e-08
2	6.2204e-08	1.0422e-08
3	1.2941e-07	2.5246e-07

Table 1: Gradient difference between two methods

From the table we can notice that discrepancies are rather small. So we may safely say that the gradient was calculated correctly.

3 Results, plots and figures

In this part I will show some results of my classifier with four different parameter settings.

(1) lambda=0, n_epochs=40, n_batch=100, eta=.1

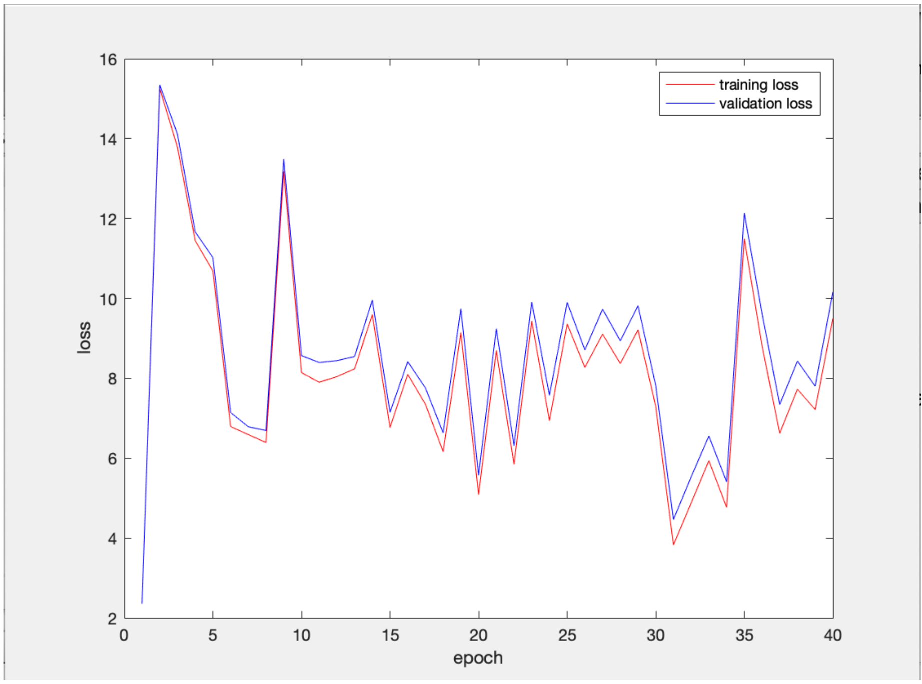

Figure 2: Train/validation loss of parameter setting (1)

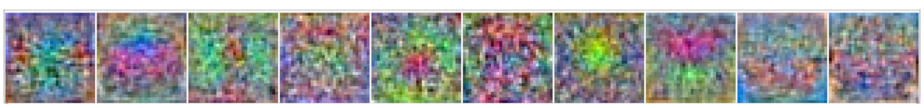

Figure 3: The learnt W matrix of paramete setting (1)

With $train\_accuracy = 0.2511$ and $test\_accuracy = 0.2274$

(2) lambda=0, n_epochs=40, n_batch=100, eta=.01

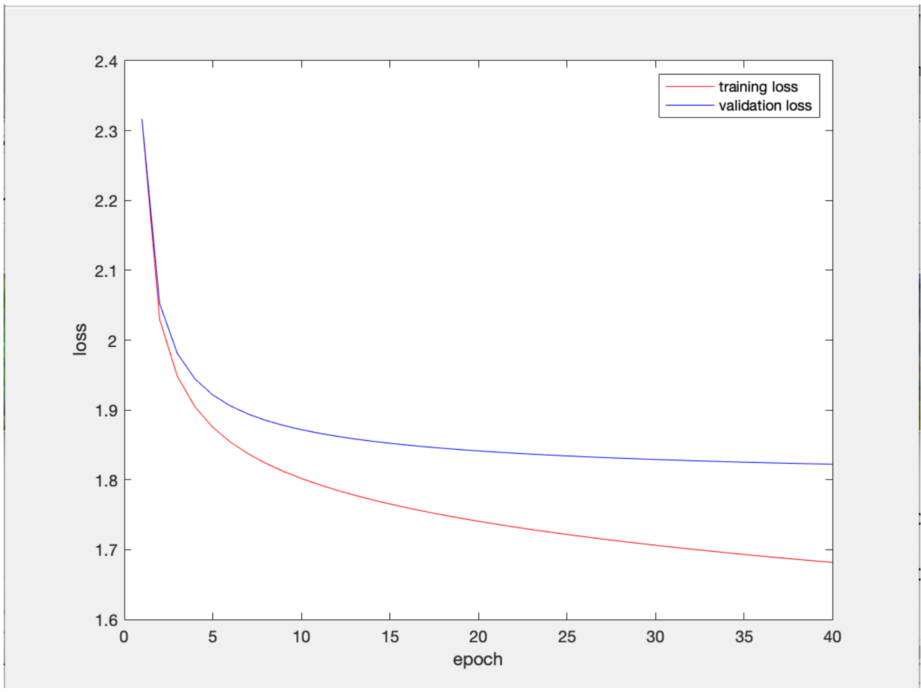

Figure 4: Train/validation loss of parameter setting (2)

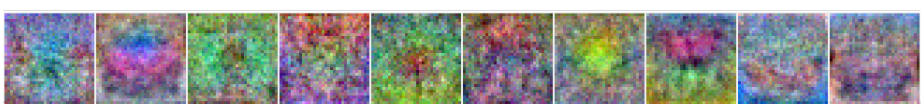

Figure 5: The learnt W matrix of paramete setting (2)

With $train\_accuracy = 0.4177$ and $test\_accuracy = 0.3685$

(3) lambda=.1, n_epochs=40, n_batch=100, eta=.01

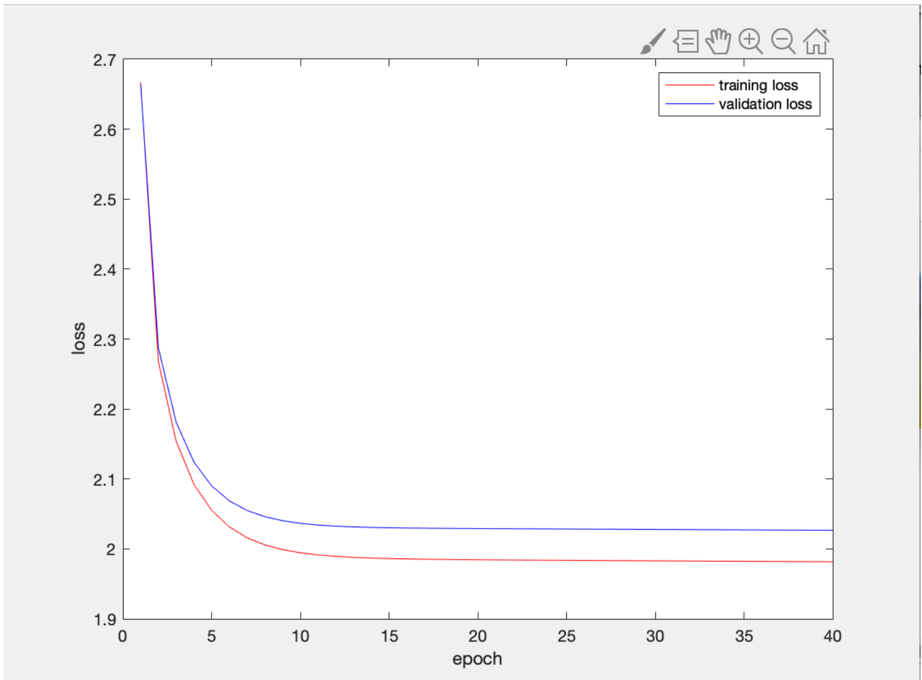

Figure 6: Train/validation loss of parameter setting (3)

Figure 7: The learnt W matrix of paramete setting (3)

With $train\_accuracy = 0.3419$ and $test\_accuracy = 0.3338$

(4) lambda=1, n_epochs=40, n_batch=100, eta=.01

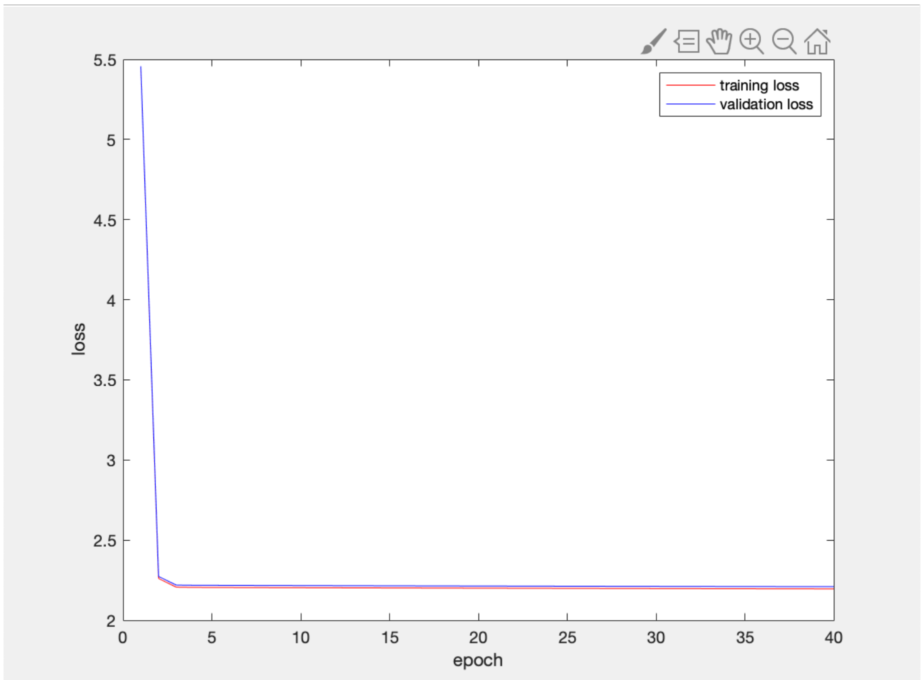

Figure 8: Train/validation loss of parameter setting (4)

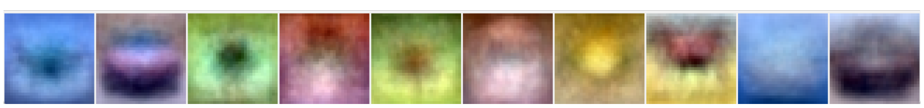

Figure 9: The learnt W matrix of paramete setting (4)

With $train\_accuracy = 0.2227$ and $test\_accuracy = 0.2192$

4 Comments

According to the performance of each test above, we can draw some conclusions about different choices of lambda and learning rate.
A high regularization term (lambda) can efficiently decrease the degree of over-fitting. As we may notice in case (2), (3) and (4), the discrepancy between loss functions of training and validation set decreases as lambda increases.
But as we increase lambda, the bias of the model also grows. Accuracy becomes low when lambda is too big. So it is important to choose a appropriate lambda and find a balance point between over-fitting and high-bias.

A large learning rate, meanwhile, may cause the model to oscillate when training because a step is too long and there is a zig-zag around the optimal point when updating.
A rather small learning rate guarantees a smoother training process, but also leads to a slower training.