DD2424 Lab 2

Artem Los (arteml@kth.se)

June 20, 2018

Constructing the classifier

Gradient check

Instead of checking the gradients numerically, we used the assumption that if gradients are correct means we should be able to overfit the data and vice versa (this is influenced by the observations in the previous lab, i.e. when gradients were wrong, the accuracy was never higher than 10%).

To test gradients, we set the goal to overfit the classifier on 100 samples using 200 epochs, a batch number of 10 and learning rate of 0.01. The final accuracy on training set was 0.81 and 0.20 on the test set – a clear example of overfitting. So, since we are able to overfit and get good results, the gradient computations should be correct.

Momentum during learning

Momentum helped us to learn faster (and overfit faster), which can be seen in the images below (left image is without momentum and right image is with momentum).

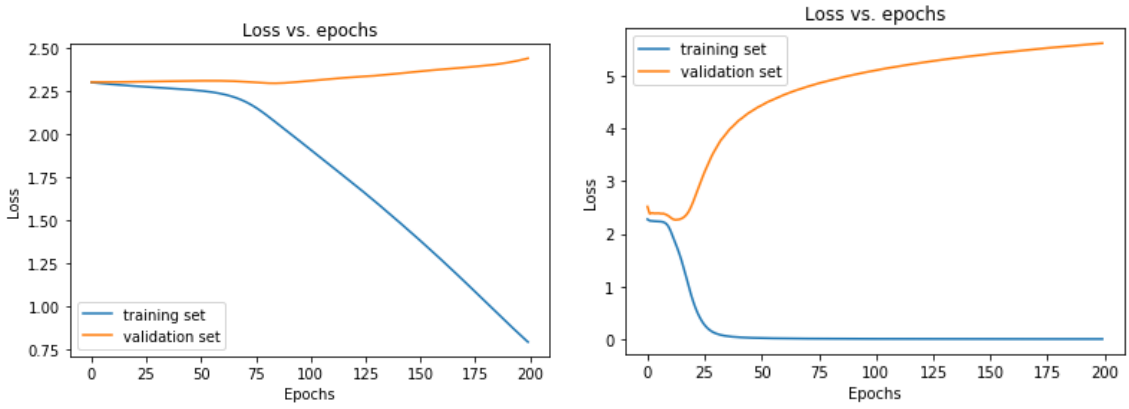

Figure 1: Shows how the training and test errors evolve as the number of epochs increases. The left picture does use momentum learning, whereas the right one does.

Hyperparameter optimization

Coarse to fine search

The initial search of hyper parameters consisted of two steps: hand-picking sensible upper/lower bounds for η and λ and then select 75 random values from uniform distributions formed by those upper bounds. To be more precise, the range used for both η and λ was $[10^{-5}, 10^{-1}]$. For the random search, we used the following network configuration, where λ, η varied.

```
clf = TwoLayerNetwork(n_epochs=3, l=lambdas[i], eta=eta[i], n_batch=100,
    decay_rate=1, rho=0.9, include_train_cost=False)
```

We observed that the best performing networks have the following hyperparameters:

η	λ	accuracy
5.48841991e-02	8.87844412e-04	3.90100000e-01
3.50619259e-02	1.37586799e-04	3.86900000e-01
7.87399444e-02	2.34832466e-03	3.87400000e-01

Figure 2: Top 3 pairs of (η, λ) during training on the first batch of CIFAR-10. Accuracy is computed on the validation set.

From the table above, we can see that the range for η and λ can be narrowed down to $\eta = [10^{-2}, 10^{-1}]$ and $\lambda = [10^{-4}, 10^{-3}]$.

We then performed more narrow searches increasing the number of epochs to 6 and later 10. As a result, the following hyper parameters were observed to give higher accuracy:

η	λ	accuracy
1.60596650e-02	1.49535682e-04	4.33200000e-01
1.43284867e-02	1.64927534e-05	4.32700000e-01
1.61882026e-02	6.09285108e-05	4.32200000e-01

Figure 3: Top 3 pairs of (η, λ) during training on the first batch of CIFAR-10. Accuracy is computed on the validation set (for a narrowed down search with 10 epochs).

Best classifier

For the best performing hyper parameters, we set up two different classifiers: one with learning rate decay of 0.95% and one without. The parameters are shown below:

```
clf = TwoLayerNetwork(n_epochs=30, l=1.49535682e-04, eta=1.60596650e-02,
    n_batch=100, decay_rate=0.95, rho=0.9)
```

The version that used learning rate decay performed better as shown below. In the left picture, where no weight decay is used, we can see a clear sign of overfitting (the validation loss starts to increase). The right picture has less severe overfitting, but we can observe that we could already stop at 15 epochs as no improvement in the loss function occurs.

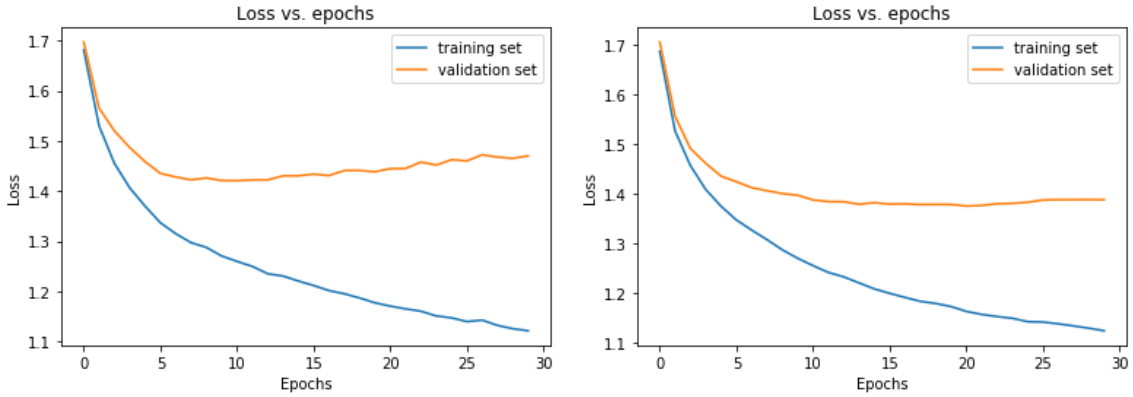

Figure 4: Shows how the training and validation errors evolve as the number of epochs increases. The left picture does not use a weight decay whereas the right one does.

The results of the best performing classifier (with weight decay) are summarized below:

Dataset	Accuracy
Train	0.6031428571428571
Test	0.5018

Figure 5: Performance the best classifier on the training and test sets.