DD2424
Deep Learning in Data Science
Assignment 1

Franco Ruggeri, fruggeri@kth.se

March 30, 2020

1 Gradients computation

I successfully managed to write the functions to correctly compute the gradients analytically. To check it, I compared the results with the numerical estimations of the gradients given by the function *ComputeGradsNumSlow* for the following test cases:

1. First mini-batch of size 1 and no regularization ($\lambda = 0$).
2. First mini-batch of size 1 and $\lambda = 1$.
3. First mini-batch of size 100, $\lambda = 1$ and first 20 dimensions.
4. First mini-batch of size 100, $\lambda = 1$.

The results are shown in table 1 and 2. As we can see, the differences are really small (the relative ones are more reliable) and this confirms the correctness of the analytical computations.

2 Mini-batch gradient descent

For each parameter setting, the test accuracy is reported in table 3, while the trend of loss and cost function on training and validation data (learning curves) and the class templates are shown in figure 1 and 2, respectively.

Test case	Max absolute difference	Max relative difference
1	3.68e-10	1.88e-06
2	8.98e-10	1.72e-04
3	1.70e-09	2.97e-07
4	2.15e-09	2.35e-04

Table 1: Tests on gradients $\frac{\partial J}{\partial W}$.

Test case	Max absolute difference	Max relative difference
1	2.41e-10	8.38e-10
2	4.34e-10	2.32e-09
3	9.46e-10	4.04e-08
4	6.66e-10	1.18e-08

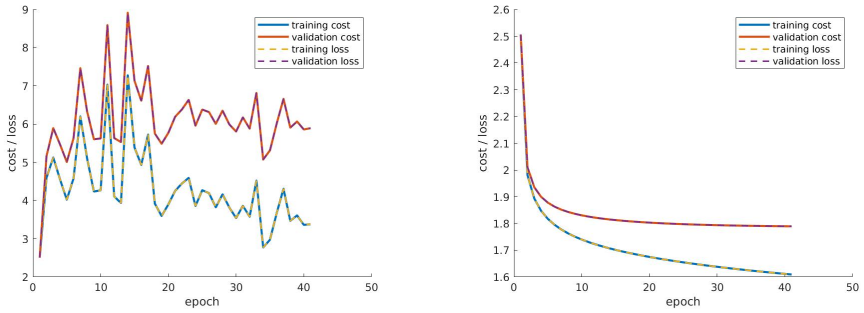
Table 2: Tests on gradients $\frac{\partial J}{\partial b}$.

As we can see in figure 1, increasing the amount of regularization makes the cost function more different (higher) from the loss, because the regularization term gets more important. The benefit of a right amount of regularization is an increase of the generalization capability (table 3). Moreover, the weights are shrunk towards 0 and so are more similar each other, giving more uniform colors in the class templates (figure 2d).

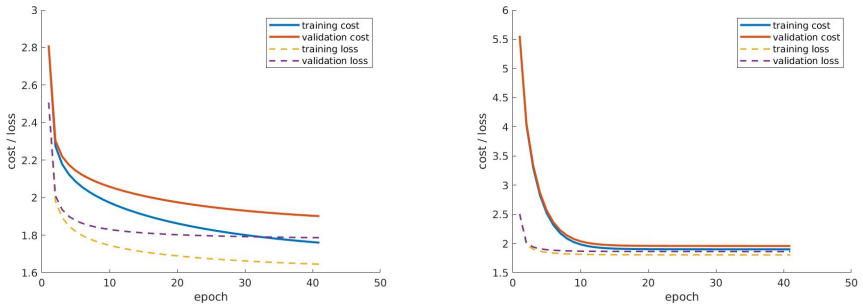
Considering the learning rate, instead, we can notice that a too high value makes the learning unstable (figure 1a), because a single update is too large and we go beyond the local minimum.

Parameter setting	λ	n_{epochs}	n_{batch}	η	Accuracy (%)
1	0	40	100	0.1	29.74
2	0	40	100	0.001	38.69
3	0.1	40	100	0.001	39.08
4	1	40	100	0.001	37.57

Table 3: Accuracy (evaluated on the test set) for several parameter settings.

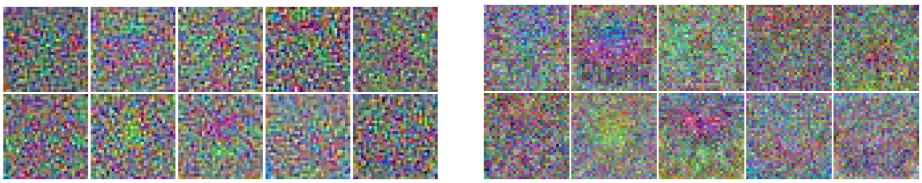


(a) (b)

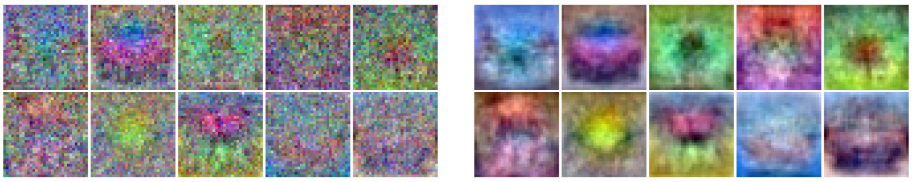


(c) (d)

Figure 1: Learning curves for the same parameter settings as in table 3.



(a) (b)



(c) (d)

Figure 2: Class templates for the same parameter settings as in table 3.