TTIC 31230 Problem Set 2
Win 2017

Hao Jiang

January 19, 2017

Problem 1

$$||\nabla_w \ell(w, x, y)|| < b \implies \forall i, n, |(\nabla_w \ell(w, x_i, y_i))_n - (\ell_{\text{generalize}}(w))_n| < 2b$$

Using the concentration inequality for gradient estimation mentioned in the slides,

$$||\nabla_w \ell_{\text{train}}(w) - \nabla_w \ell_{\text{generalize}}(w)|| \leq \frac{2b\left(1 + \sqrt{2\ln(1/\delta)}\right)}{\sqrt{N}}$$

with probability $1 - \delta$

Problem 2

a

The experiment result of 2a is shown in Table 1. The following things can be observed.

- The provided optimal learning rate η^* yields a better test accuracy on batch size 10 and 100, but not on batch size 50.
- For the same learning rate(0.37), larger batch size does not always yield better performance result

Conclusion: There is no single best learning rate for different batch sizes. The learning rate should be setup as a function relying on batch size to provide best performance.

Setting	Training Loss	Test Accuracy
B = 10, $\eta = \eta^*(B)$	0.0099	0.9969
B = 10, $\eta = 0.37$	0.0817	0.9780
B = 50, $\eta = \eta^*(B)$	0.0097	0.9976
B = 50, $\eta = 0.37$	0.0086	0.9980
B = 100, $\eta = \eta^*(B)$	0.0110	0.9977
B = 100, $\eta = 0.37$	0.0213	0.9935

Table 1: Experiment Result for Problem 2a

Setting	Training Loss	Test Accuracy
B = 10, $\eta = 0.37$	0.1193	0.9745
B = 10, $\eta = \eta^*$	0.0076	0.9970
B = 50, $\eta = 0.37$	0.0098	0.9970
B = 50, $\eta = \eta^*$	0.0097	0.9975
B = 100, $\eta = 0.37$	0.0199	0.9947
B = 100, $\eta = \eta^*$	0.0108	0.9965

Table 2: Experiment Result for Problem 2b

b

The experiment result of 2b is demonstrated in Table 2. In the case of momentum, η^* does always yield a better result. When batch size is small, the difference is more obvious. When batch size is larger, the difference become less.

Similar to what is observed in standard SGD, for the given constant learning rate $\eta = 0.37$, batch size 50 gives the best performance.

c

The result of Adam algorithm is demonstrated in Table 3.

For batch size 50, minimal train loss and test accuracy can be obtained around $\eta = 0.0016$.

For batch size 100, the minimal training loss appears at $\eta = 0.0015$, but the highest test accuracy appears at $\eta = 0.0014$

For batch size 10, the minimal training loss appears at $\eta = 0.0014$, and the highest test accuracy appears at $\eta = 0.00145$

In addition, it can also be noticed that a smaller batch size tends to yield a higher training loss and lower test accuracy.

Setting	Training Loss	Test Accuracy
B = 50, $\eta = 0.0014$	0.0110	0.9974
B = 50, $\eta = 0.0015$	0.0099	0.9973
B = 50, $\eta = 0.0016$	0.0088	0.9979
B = 50, $\eta = 0.0017$	0.0090	0.9973
B = 100, $\eta = 0.0013$	0.0116	0.9972
B = 100, $\eta = 0.00135$	0.0104	0.9976
B = 100, $\eta = 0.0014$	0.0093	0.9979
B = 100, $\eta = 0.00145$	0.0096	0.9978
B = 100, $\eta = 0.0015$	0.0089	0.9978
B = 100, $\eta = 0.0016$	0.0100	0.9975
B = 10, $\eta = 0.0013$	0.0223	0.9925
B = 10, $\eta = 0.00135$	0.0136	0.9946
B = 10, $\eta = 0.00138$	0.0190	0.9938
B = 10, $\eta = 0.0014$	0.0124	0.9946
B = 10, $\eta = 0.00145$	0.0143	0.9955
B = 10, $\eta = 0.0015$	0.0191	0.9937
B = 10, $\eta = 0.0016$	0.0229	0.9918

Table 3: Experiment Result for Problem 2c

d

One of the most obvious observation is the relationship between batch size and running time. The running time will in general decrease when the batch size increase. However, the decrease rate will become slower when batch size is too large. When computation is applied to a larger batch group, the efficiency of numpy's vector operations will save some time, but the average result need to be computed on a larger group. Thus there should exist an optimal batch size for specific given problem.

Also, as mentioned in previous section, a smaller batch size tends to lead to both a larger training loss and a smaller test accuracy. This is possibly because in a smaller batch group, outliers can have large impact to the average gradient and lead to inaccurate result.

Finally, larger training loss in general yield to a worse test accuracy. But a smaller training loss does not always means a higher test accuracy. This is due to the skew in the sampling process when choosing training set and test set from the distribution that includes all data.