Adaptive Step sizes weekly report

Anubhav Mittal
Project Supervisors : Martin Jaggi, Aymeric Dieuleveut

November 2, 2018

1 Week of October 29th, 2018

I implemented and compared the performance of different types of SGD, on artificially generated dataset with the following parameters:

- Feature dimension $p = 10$, $SNR = 2$ where SNR=var(x)/(p*var(y given x)).
- weight vector w is fixed as $w_j = 10 * exp(-0.75j)$.
- number of data points N=5000, with each $x_i \sim N(0, I), y_i \sim N(w^T x_i, \sigma^2)$

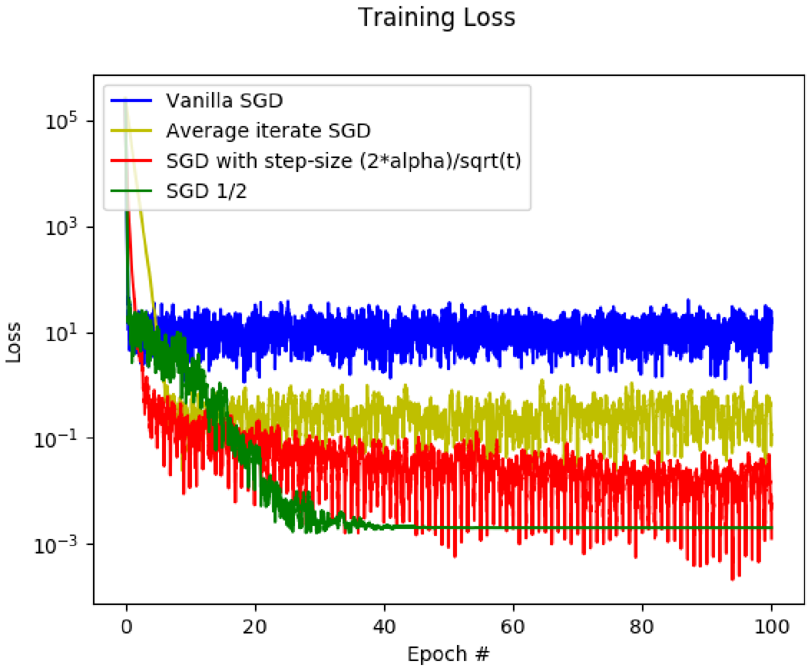

Figure 1:

As clear from the figure, performance of SGD 1/2 is comparable to SGD with $1/\sqrt{t}$ step size.

We know that the sum of dot product of gradients will eventually get a negative value, but in the initial epochs, the value of the dot product of the gradients is very large, and it so happens that if I start adding the dot product of gradients from the first epoch, the value becomes so large that it does not go negative for even 200 epochs. I chose to start summing from epoch 3 for this reason in the code.