ESE 546, FALL 2019

PROBLEM SET 2

JIAXIAO CAI [MIAJXCAI@SEAS.UPENN.EDU]

Solution 1. Your solution goes here.

(a) The result of gradient descent with Nesterov's updates is shown in figure 1.

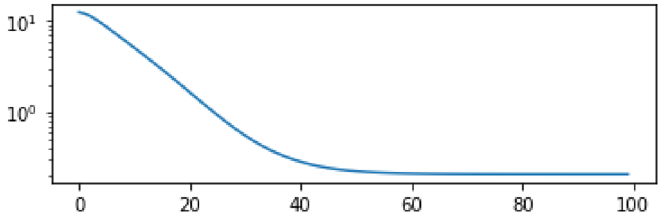

FIGURE 1. Training Loss vs No. Weight Updates for Nesterov's Acceleratio

(b) The result of SGD without Nesterov's acceleration with batch size = 4 is shown in figure 2.

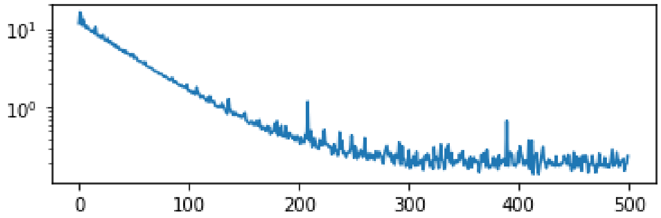

FIGURE 2. Training Loss vs No. Weight Updates for SGD without Nesterov's Acceleratio (batch size = 4)

The result of SGD without Nesterov's acceleration with batch size = 128 is shown in figure 3.

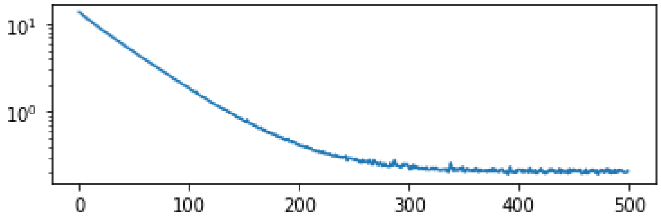

FIGURE 3. Training Loss vs No. Weight Updates for SGD without Nesterov's Acceleratio (batch size = 128)

(c) The result of SGD with Nesterov's acceleration with batch size = 4 is shown in figure 4.

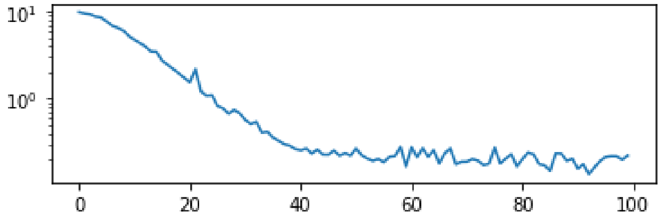

FIGURE 4. Training Loss vs No. Weight Updates for SGD with Nesterov's Acceleratio (batch size = 4)

The result of SGD with Nesterov's acceleration with batch size = 128 is shown in figure 5.

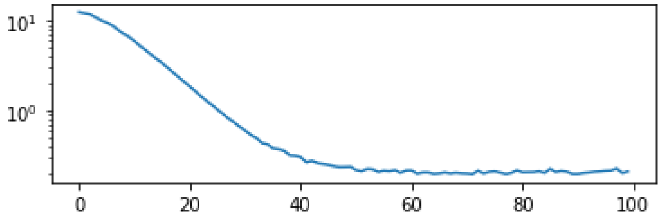

FIGURE 5. Training Loss vs No. Weight Updates for SGD with Nesterov's Acceleratio (batch size = 4)

- Convergence Rate of (ii) and (iii): The convegence rate of (iii) is faster than (ii). Theoretically, the convergence rate of SGD without Nesterov's and SGD with Nesterov's should be the same, because Nesterov's Acceration doesn't work in SGD. However, the result of (iii) is faster than (ii), I think this is because the direction of SGD is different from the direction of global minimum. Under this case, NAG could help to speeds up movement along directions of strong improvement (loss decrease). Momentum is essentially a small change to the SGD parameter update so that movement through the parameter space is averaged over multiple time steps.
- Convergence Rate of (i) and (ii): We could see from the above figures that when the batch size is small, updates are noisier. In Stochastic Gradient Descent, the we converge is expressed as $EiU(1,...,n)[f_i(x_t)] = f(x_t)$, therefore, the expected iterate of SGD converges to the optimum. The downside is that stochasticity brings variance. As shown in result, even if SGD is shown to converge, the variance can seriously handicap the convergence rate. This is especially true the smaller the batch size is, as variance is inversely proportional to the number of examples used to compute the gradient at every step.