CSCI-6960 ML & Optimization HW4

Dong Hu

1 Question 1

1. define $h(t) = \max(0, t)$, we have

$$\partial h(t) = \begin{cases} 0, & t < 0 \\ [0, 1], & t = 0 \\ 1, & t > 0 \end{cases}$$

Here when t = 0, we just pick 1/2 as a subgradient. Notice that if we have F convex, G convex and non-decreasing, then $\partial(G \circ F)(t) = \partial G(F(t))\partial F(t)$. Therefore, let $g(\mathbf{w}, b) = 1 - y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right)$ we have

$$\partial(h \circ g)(\mathbf{w}, b) = \partial h(g(\mathbf{w}, b))\partial g(\mathbf{w}, b) = \partial h \left(1 - y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right)\right) \cdot \left(-y_i \begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix}\right)$$

and

$$\partial f(\mathbf{w}, b) = \frac{1}{n} \sum_{i=1}^{n} -y_i \cdot \partial h \left(1 - y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right)\right) \cdot \begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix} + \lambda \begin{bmatrix} \mathbf{w} \\ 0 \end{bmatrix}$$

2. We first consider the SGD case (m = 1). We first need to choose a good a, by only looking at the first 1000 rows of training data, and track the objective value after 5000 iterations (since 1000 iterations take 0.8 sec, we think 5000 is the appropriate number here). Since this is stochastic, we will run 3 realizations of each choice. We first consider $a \in 10, 1, 0.1, 0.01, 0.001$

. The experimental result is as follows (Table 1):

Table 1: SGD Training Objective

T	100	1000	5000
$a = 10$	1.265	1.177	1.110
$a = 10$	1.042	0.927	0.862
$a = 10$	4.878	3.891	3.210
$a = 1$	0.679	0.545	0.501
$a = 1$	0.533	0.500	0.483
$a = 1$	1.050	0.527	0.491
$a = 0.1$	0.737	0.400	0.364
$a = 0.1$	0.486	0.387	0.354
$a = 0.1$	0.642	0.466	0.399
$a = 0.01$	1.407	0.582	0.412
$a = 0.01$	1.698	0.529	0.380
$a = 0.01$	0.853	0.561	0.388
$a = 0.001$	1.390	1.585	0.733
$a = 0.001$	1.134	1.095	0.546
$a = 0.001$	1.158	2.615	0.665

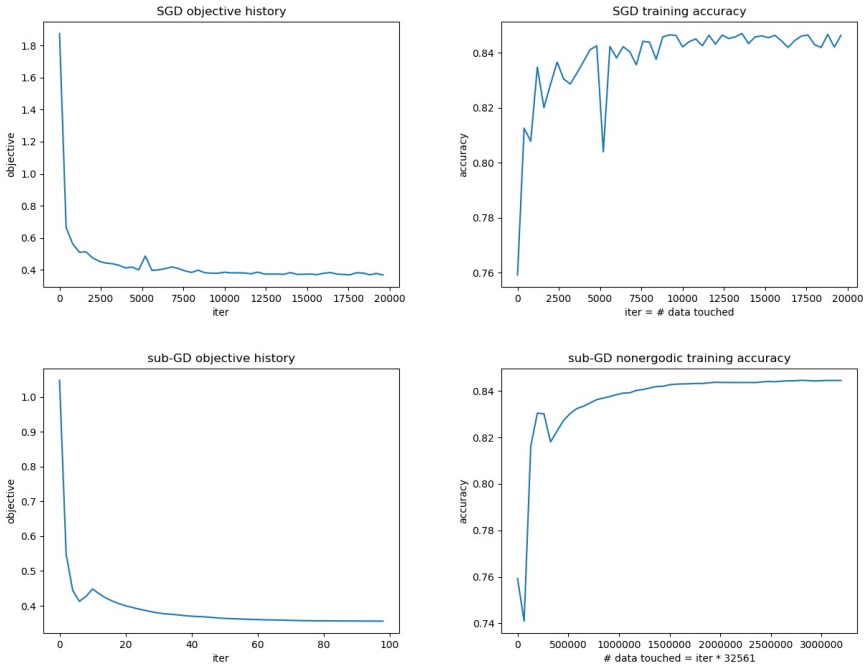
By picking the lowest average objective at the last iterate, We notice optimal a should be between 0.1 and 0.01. Then we do a finer search in that interval, and increase T to 20000. We have Table 2:

Table 2: SGD Training Objective, finer search

T	100	1000	5000	20000
$a = 0.01$	1.523	0.614	0.372	0.330
$a = 0.01$	1.253	0.522	0.446	0.342
$a = 0.01$	1.050	0.529	0.383	0.338
$a = 0.02$	0.733	0.491	0.387	0.350
$a = 0.02$	0.714	0.493	0.393	0.351
$a = 0.02$	0.954	0.467	0.374	0.343
$a = 0.1$	0.657	0.438	0.373	0.351
$a = 0.1$	0.588	0.440	0.377	0.353
$a = 0.1$	0.622	0.444	0.394	0.373

We conclude a = 0.01 is a good choice for SGD. Now with full data set, we run the algorithm until objective curve asymptotes out (which occurs before T = 20000, so we just pick T = 20000). We attached 3 plots below (Figure 1, 2, 3, 4), corresponding to objective history, training accuracy.

3.



4. I had a little bit problem showing the model that average the weights, so they are not included in the plot. attempts are made in the code.

5. Observations are: we notice that stochastic sub gradient descent converges faster than subgradient descent with respect to the time it goes over the data one time; however, knowing that the stochastic gradient descent only uses minibatch to update the weight,it fluctuates a lot(see right hand size of figure 2). The subgradient descent is deterministic method, its more stable.

2 Question 2

Recall each iteration of Projected Subgradient method we perform:

$$\begin{aligned}
\mathbf{x}_{t+1} &= \mathrm{Proj}_C\left(\mathbf{x}_t - \alpha_t \mathbf{g}\right) \\
&= \arg\min_{\mathbf{u}\in C} \frac{1}{2\alpha_t}\left\|\mathbf{u} - \left(\mathbf{x}_t - \alpha_t \mathbf{g}\right)\right\|^2 \\
&= \arg\min_{\mathbf{u}\in C} \frac{1}{2\alpha_t}\left\|\mathbf{u} - \mathbf{x}_t\right\|^2 + \left\langle \mathbf{g}, \mathbf{u} - \mathbf{x}_t\right\rangle + \frac{\alpha_t}{2}\|\mathbf{g}\|^2 \\
&= \arg\min_{\mathbf{u}\in C} f\left(\mathbf{x}_t\right) + \left\langle \mathbf{u}, \mathbf{u} - \mathbf{x}_t\right\rangle + \frac{1}{2\alpha_t}\left\|\mathbf{u} - \mathbf{x}_t\right\|^2
\end{aligned}$$

Thus this algorithm is exactly the projected subgradient method with stepsize given by α_t