CS 373 Spring 2019: Homework 4

Youngsik Yoon, 0029846135

Late Days Used: 4

April 26, 2019

2 Theory

1. Batch gradient descent is an optimization algorithm that makes updates after going over the entire data set. Although stochastic gradient descent is similar to batch gradient descent, stochastic gradient descent calculates the gradient of for the loss function with just one point instead of all the points.

   Stochastic gradient descent is much faster than batch gradient descent for this reason, but stochastic gradient descent relies on the data point to be representative of data set. Batch gradient descent is good when the error surface is smooth. Stochastic gradient descent is better when the error surface has more local minimum and maximums.

2. We know that the model has converged when the gradient vector is zero. If the gradient vector is zero, that means a local minimum was been found.

3. The bias term in BGD/SGD learning is allowing the classifier to have more expressivity. The bias term also allows offsets for the hyperplane.

4. True; stochastic gradient descent only updates on one data sample rather than the entire data set.

5. Randomly shuffling the training examples before using SGD optimization allows the data point to be more representative of the data set. Additionally, if there are runs of data points in the data set, the patterns can be avoided.

6. Although hinge loss is not differentiable at $x = 1$, the use of subgradient helps cover the differential points. The sub-gradient of a function c at x_0 is any vector v such that the set only contains the gradient at x_0 at differentiable points. The gradient of hinge loss function without

regularization term is ...

$$\begin{aligned}
\partial_w \text{max}\{0, y_n(w \cdot x + b)\} &= \partial_w \begin{cases} 0 & \text{if } y_n(w \cdot x + b) > 1 \\ y_n(w \cdot x_n + b) & \text{otherwise} \end{cases} \\
&= \begin{cases} \partial_w 0 & \text{if } y_n(w \cdot x + b) > 1 \\ \partial_w y_n(w \cdot x_n + b) & \text{otherwise} \end{cases} \\
&= \begin{cases} 0 & \text{if } y_n(w \cdot x + b) > 1 \\ y_n x_n & \text{otherwise} \end{cases}
\end{aligned}$$

7. Gradient of L_2 regularization term ...

$$\partial_w \frac{1}{2}\lambda ||w^2|| = \lambda ||w||$$

Gradient of log loss function ...

$$\begin{aligned}
\partial_z g(z_i) &= \partial_z \frac{1}{1 + e^{-z}} \\
&= -(1 + e^{-z})^{-2} \cdot -e^{-z} \\
&= \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2} \\
&= \frac{1}{(1 + e^{-z})} - \frac{1}{(1 + e^{-z})^2} \\
&= g(z_i) - g(z_i)^2
\end{aligned}$$

$$\begin{aligned}
\partial_w L(x_i, y_i; w) &= \partial_w - \sum_i y_i log(g(z_i)) + (1 - y_i) log(1 - g(z_i)) \\
&= -\sum_i (\frac{y_i(g(z_i) - g(z_i)^2)}{g(z_i)} + \frac{(1 - y_i)(g(z_i)^2 - g(z_i))}{1 - g(z_i)}) x_i \\
&= -\sum_i (y_i(1 - g(z_i)) + \frac{(1 - y_i) g(z_i)(g(z_i) - 1)}{1 - g(z_i)}) x_i \\
&= \sum_i (y_i(g(z_i) - 1) + \frac{(1 - y_i) g(z_i)(g(z_i) - 1)}{g(z_i) - 1}) x_i \\
&= \sum_i (y_i(g(z_i) - 1) + (1 - y_i) g(z_i)) x_i \\
&= \sum_i (y_i - g(z_i)) x_i
\end{aligned}$$

Gradient of hinge loss function is done in question 6. Thus, the gradient of the log loss function with the L_2 regularization term is ...

$$\sum_i (y_i - g(z_i))x_i - \lambda||w||$$

, and the gradient of the hinge loss function with the L_2 regularization term is ...

$$\begin{cases} -\lambda||w|| & \text{if } y_n(w \cdot x + b) > 1 \\ y_n x_n - \lambda||w|| & \text{otherwise} \end{cases}$$

8. Regularization, a form of inductive bias, is used to prevent overfitting. Regularization can help minimize the norm of the weight vector. If the λ hyper parameter can take a negative value, the regularization will actually go in the wrong direction and worsen the optimizer.

3 Batch Gradient Descent

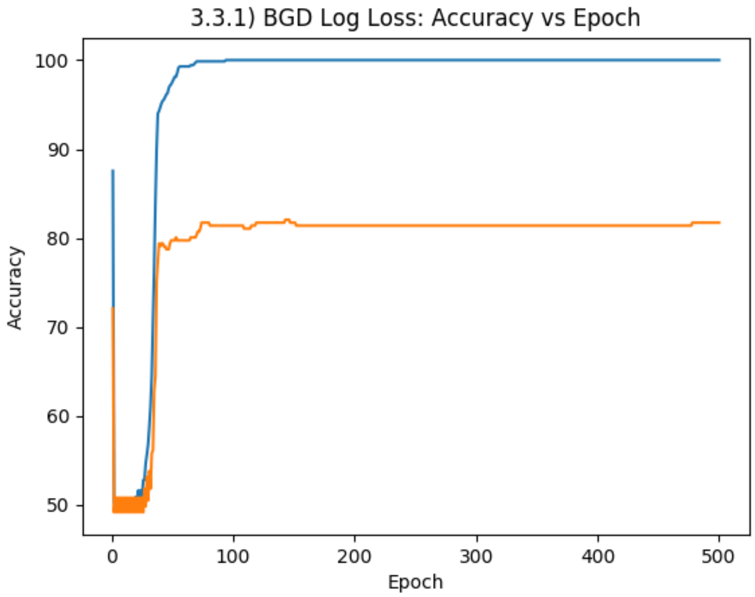
3.1 Algorithm

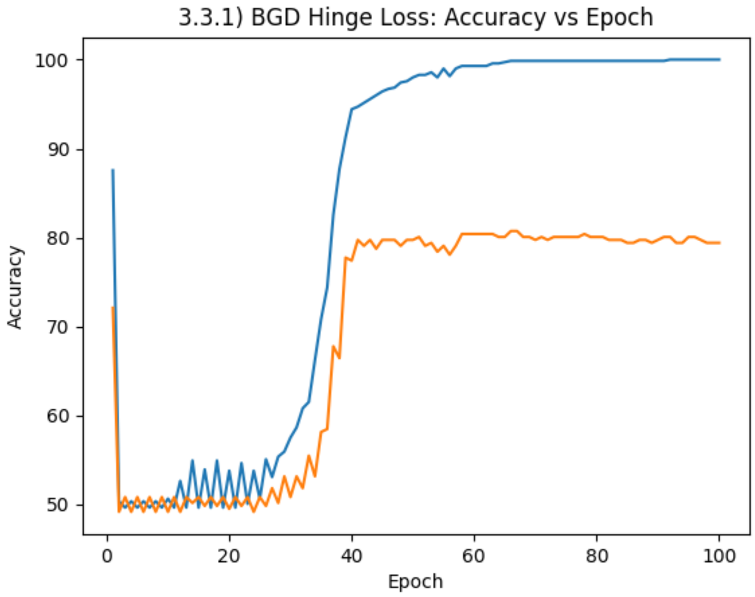
Algorithm 1 LogLossBGD(η, D, MaxIter)

```
$w \leftarrow \langle 0, 0, ..., 0 \rangle$
$b \leftarrow 0$

for iter = 1 ... MaxIter do
    $g_w \leftarrow \langle 0, 0, ..., 0 \rangle$
    $g_b \leftarrow 0$

    for all (x, y) ∈ D do
        $z \leftarrow w \cdot x + b$
        $\sigma \leftarrow \frac{1}{1+e^-z}$

        $g_w \leftarrow g_w + (y - \sigma)x$
        $g_b \leftarrow g_b + (y - \sigma)$
    end for

    if $g_w == \langle 0, 0, ..., 0 \rangle$ then
        break
    end if

    $w \leftarrow w - \eta g_w$
    $b \leftarrow b - \eta g_b$
end for

return $w, b$
```

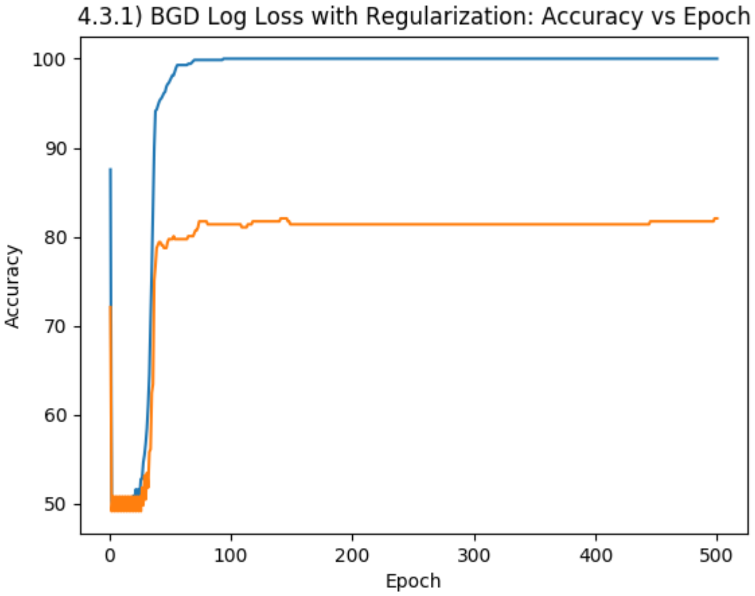
3.3 BGD Analysis

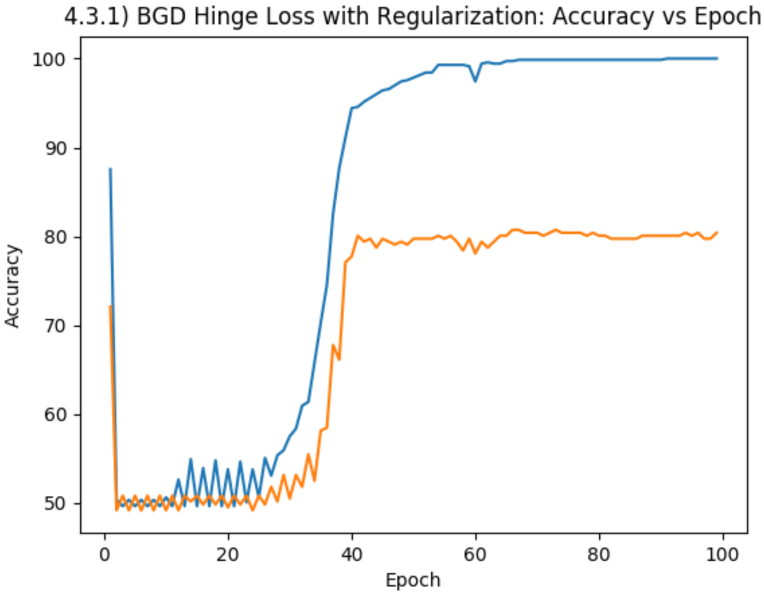
1.





2.


4.3.1) BGD Log Loss with Regularization: Accuracy vs Epoch
Accuracy
100
90
80
70
60
50
0
100
200
300
400
500
Epoch


4.3.1) BGD Hinge Loss with Regularization: Accuracy vs Epoch
Accuracy
100
90
80
70
60
50
0
20
40
60
80
100
Epoch

4 Stochastic Gradient Descent

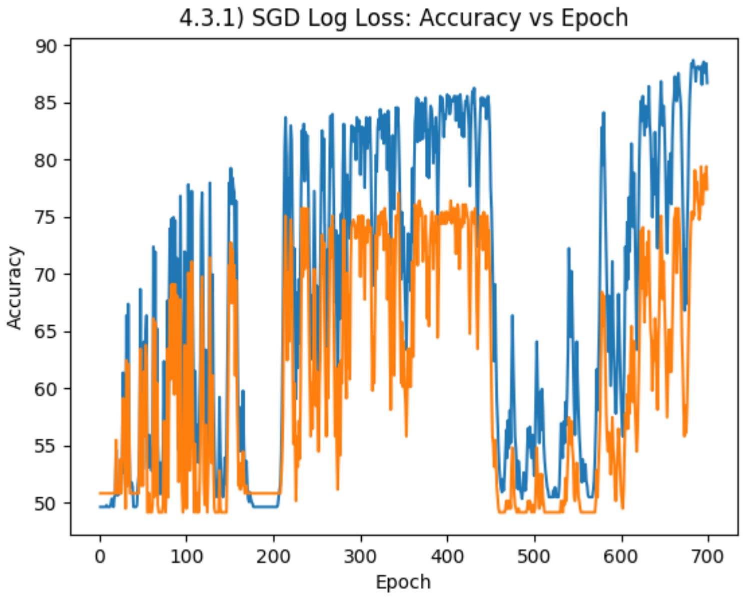
4.1 Algorithm

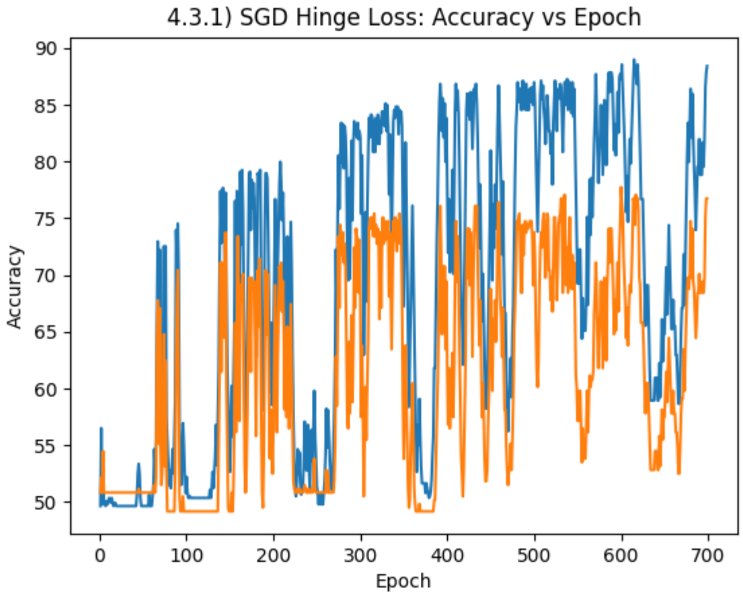
Algorithm 2 HingeLossSGD(η, K, D, MaxIter)

```
$w \leftarrow \langle 0, 0, ..., 0 \rangle$
$b \leftarrow 0$

for iter = 1 ... K do
    $g_w \leftarrow \langle 0, 0, ..., 0 \rangle$
    $g_b \leftarrow 0$

    $D_i \leftarrow$ random data point from $D$
    $x, y \in D_i$

    if $y \cdot (w \cdot x) \leq 1$ then
        $g_w \leftarrow g_w + y \cdot x$
        $g_b \leftarrow g_b + y$
    end if

    $w \leftarrow w - \eta g_w$
    $b \leftarrow b - \eta g_b$
end for

return $w, b$
```

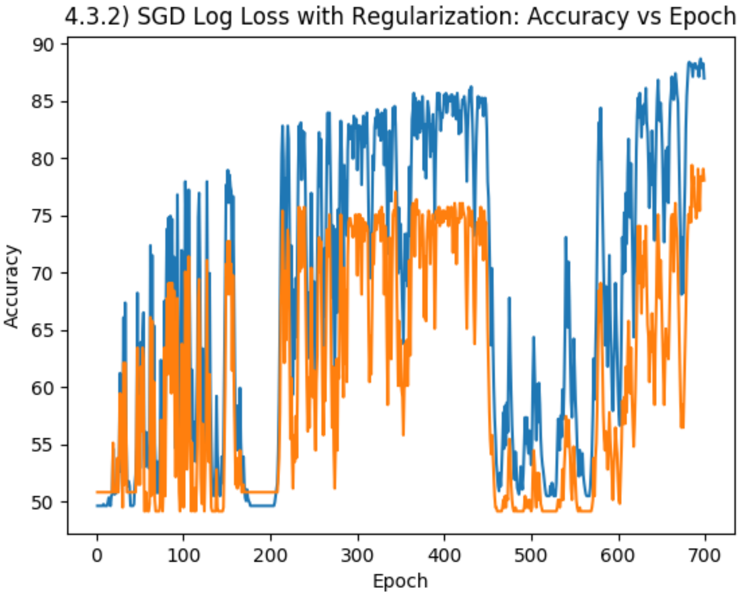
4.3 SGD Analysis

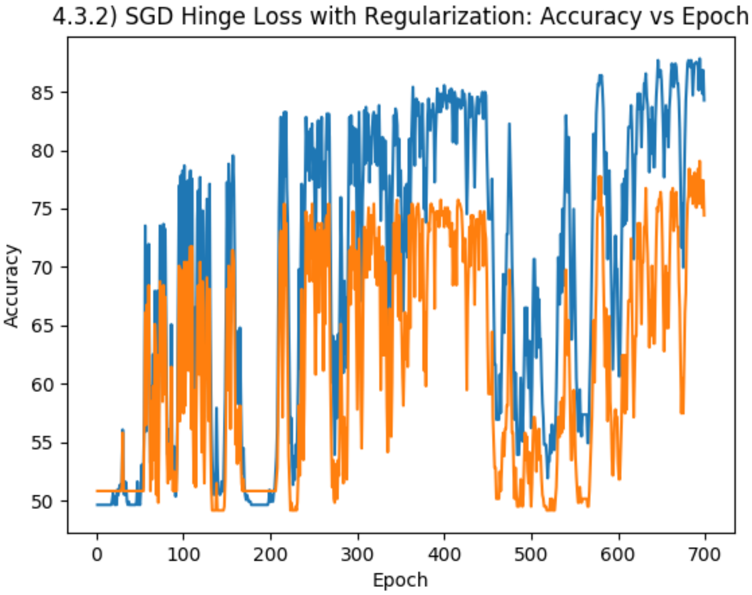
1.





2.


4.3.2) SGD Log Loss with Regularization: Accuracy vs Epoch
Accuracy
90
85
80
75
70
65
60
55
50
0
100
200
300
400
500
600
700
Epoch


4.3.2) SGD Hinge Loss with Regularization: Accuracy vs Epoch
Accuracy
85
80
75
70
65
60
55
50
0
100
200
300
400
500
600
700
Epoch