MLF - HW3

By: Wu-Jun Pei (B06902029)

Problem 1.

Problem 2.

- PLA

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + [\![y_n \neq \operatorname{sign}(\mathbf{w}_t^T \mathbf{x}_n)]\!](y_n \mathbf{x}_n)$$

- SGD

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \eta(-\nabla \operatorname{err}(\mathbf{w}))$$

When $err(\mathbf{w}) = \max(0, -y\mathbf{w}^T x)$,

- If $y\mathbf{w}^T x \geq 0$, implies $\operatorname{sign}(y) = \operatorname{sign}(\mathbf{w}^T x)$, the err is zero in this case. Since we ignore the points that are not differentiable, the $\nabla err(\mathbf{w})$ is zero, the $\mathbf{w}$ does not change.
- Otherwise, $\operatorname{sign}(y) \neq \operatorname{sign}(\mathbf{w}^T x)$. The gradient of err is $\nabla \operatorname{err}(\mathbf{w}) = -y\mathbf{w}^T$, the $\mathbf{w}$ is going to be updated as

$$
\begin{aligned}
\mathbf{w}_{t+1} &\leftarrow \mathbf{w}_t + \eta(-\nabla \mathrm{err}(\mathbf{w})) \\
&= \mathbf{w}_t + \eta(-(-y\mathbf{w}^T)) \\
&= \mathbf{w}_t + \eta(y\mathbf{w}^T)
\end{aligned}
$$

The SGD results the same as PLA when $\eta = 1$.

Problem 3.

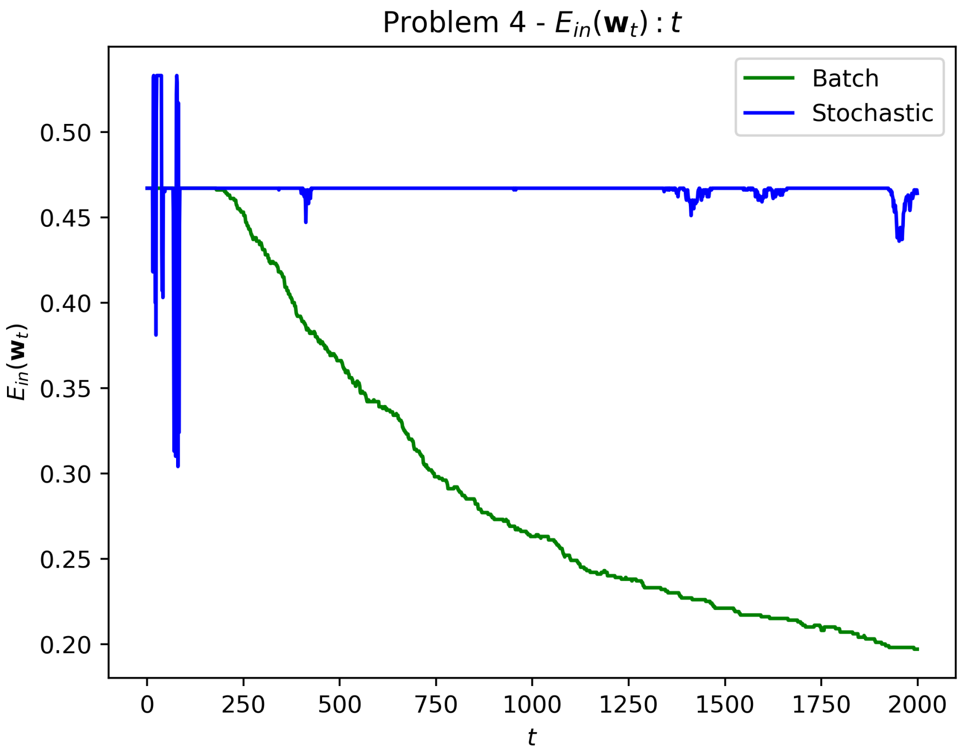
From problem 16 on Coursera, we've derived that

$$
E_{in} = \frac{1}{N}\sum_{n=1}^{N}(\ln(\sum_{i=1}^{K}\exp(\mathbf{w}_i^T x_n)) - \mathbf{w}_{y_n}^T x_n)
$$

Thus, we can have

$$
\begin{aligned}
\frac{\partial E_{in}}{\partial w_i} &= \frac{1}{N}\sum_{n=1}^{N}(\frac{\partial \ln(\sum_{i=1}^{K}\exp(\mathbf{w}_i^T x_n))}{\partial \sum_{i=1}^{K}\exp(\mathbf{w}_i^T x_n)}\frac{\partial \sum_{i=1}^{K}\exp(\mathbf{w}_i^T x_n)}{\partial \mathbf{w}_i} - [\![y_n = i]\!]x_n), && \text{by Chain Rule} \\
&= \frac{1}{N}\sum_{n=1}^{N}(\frac{1}{\sum_{i=1}^{K}\exp(\mathbf{w}_i^T x_n)}\exp(\mathbf{w}_i^T x_n)x_n - [\![y_n = i]\!]x_n), && \text{Derivative} \\
&= \frac{1}{N}\sum_{n=1}^{N}(h_i(x_n)x_n - [\![y_n = i]\!]x_n) \\
&= \frac{1}{N}\sum_{n=1}^{N}((h_i(x_n) - [\![y_n = i]\!])x_n)
\end{aligned}
$$

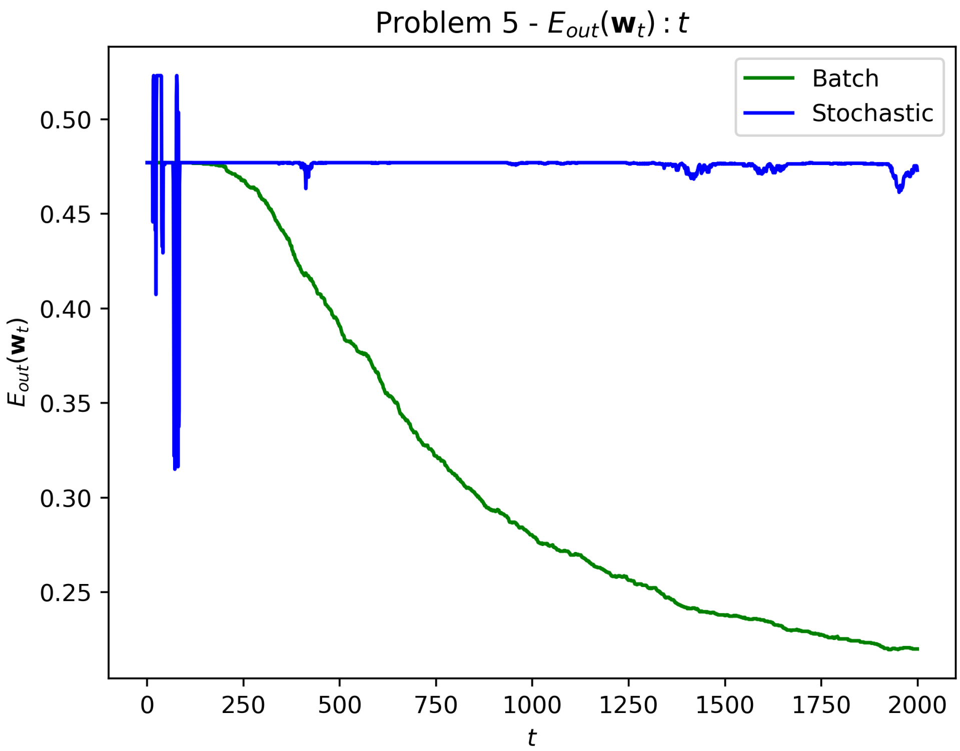
Problem 4.



My findings

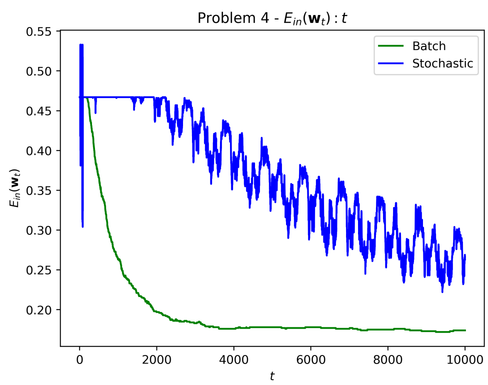
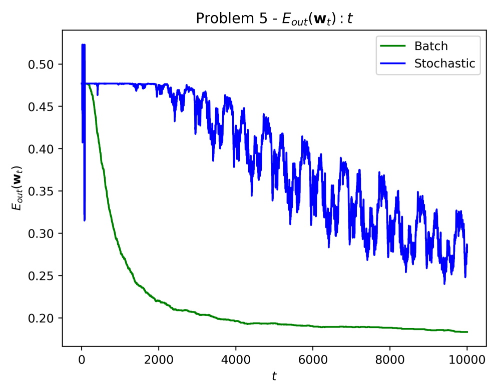
1. The E_{in} of Fixed rate Gradient Descent is monotonic while E_{in} of SGD is not.
2. Generally speaking, the E_{in} of Fixed rate Gradient Descent is smaller than the E_{in} of SGD.
3. I think both *(1)* and *(2)* result from the fact that Fixed rate Gradient Descent takes much more computational resource ($O(N)$ time to update $\mathbf{w}_{t+1}$ with all the testdata) than SGD takes ($O(1)$ time to update $\mathbf{w}_{t+1}$ with one testdata).

Problem 5.



My findings

1. The E_{out} of both the two versions for Gradient Descent are highly correlated with the E_{in} but bigger than E_{in}. It is not a new news because we have the same finding in PLA/pocket algorithm we've learned.
2. I also conducted an experiment by increasing T from 2000 to 10000. The following figures are the result of the experiment

<table>
<tr><th>E_{in}</th><th>E_{out}</th></tr>
<tr><td>
</td><td>
</td></tr>
</table>

1. Both E_{in} and E_{out} of Fixed rate Gradient Descent converges to around 0.20.
2. The E_{in} and E_{out} of SGD finally start to fall. And the curve seems to be periodic, I think the period interval might be the size of the training data (1000).