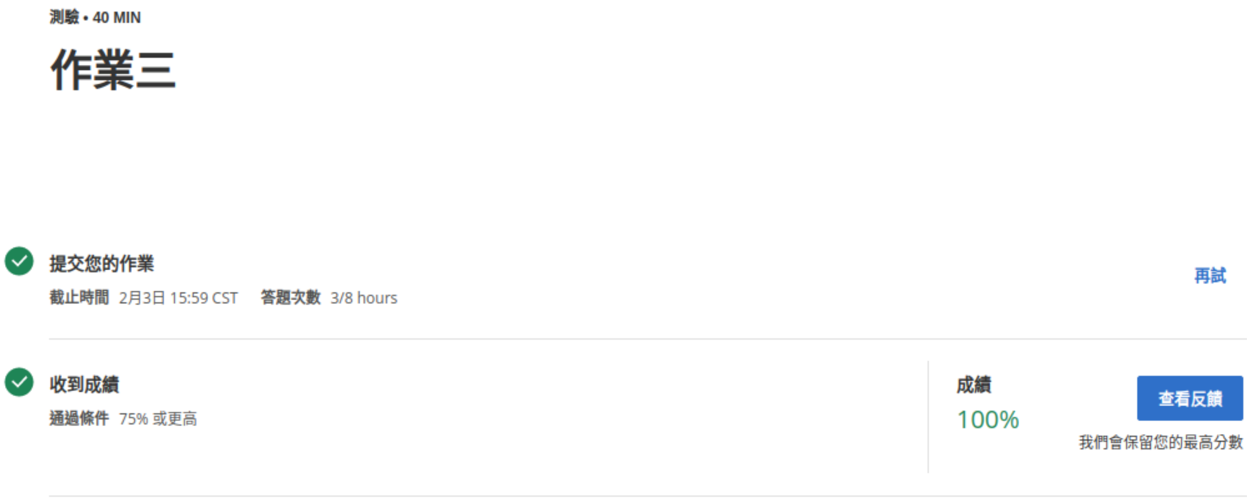
1



2

SGD: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \eta(-\nabla err(\mathbf{w}_t, \mathbf{x}_n, y_n))$

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

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

- **CASE A:**

  If $y\mathbf{w}^T\mathbf{x} \geq 0$, $err(\mathbf{w}) = 0$ and $y = \text{sign}(y) = \text{sign}(\mathbf{w}^T\mathbf{x})$.

  Since we ignore the points that are not differentiable, $\nabla err(\mathbf{w}) = 0$, and $\mathbf{w}_{t+1} = \mathbf{w}_t$.

- **CASE B:**

  Otherwise, $y\mathbf{w}^T x < 0$, $err(\mathbf{w}) = -y\mathbf{w}^T\mathbf{x}$, $\nabla err(\mathbf{w}) = -y\mathbf{x}$. $\mathbf{w}_{t+1}$ will be update as follows:

$$\begin{aligned} \mathbf{w}_{t+1} &= \mathbf{w}_t + \eta(-\nabla err(\mathbf{w}_t, \mathbf{x}_n, y_n)) \\ &= \mathbf{w}_t + \eta(-(-y_n\mathbf{x}_n)) \\ &= \mathbf{w}_t + \eta y_n \mathbf{x}_n \end{aligned}$$

  Since $y\mathbf{w}^T\mathbf{x} < 0$, $y = \text{sign}(y) \neq \text{sign}(\mathbf{w}^T\mathbf{x})$. The SGD result will be same as PLA when $\eta = 1$.

3

we construct the expansion of $f(x+\Delta x)$ by Taylor Theorem:

$$f(x+\Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2$$

To minimize $f(x+\Delta x)$, set the derivative to be zero.

$$0 = \frac{d}{dt}\left(f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2\right) = f'(x) + f''(x)\Delta x$$

thus we can get $\Delta x = -\frac{f'(x)}{f''(x)}$, the iteration will be:

$$x_{k+1} = x_k + \Delta x = x_k - \frac{f'(x)}{f''(x)}$$

According to the question , we will get the result below in high dimension:

$$E(u+\Delta u, v+\Delta v) = E(u,v) - \left(\nabla^2 E(u,v)\right)^{-1} \nabla E(u,v)$$
$$(\Delta u, \Delta v) = -\left(\nabla^2 E(u,v)\right)^{-1} \nabla E(u,v)$$

4

Derive from maximum likelihood solution to negative log likelihood :

$$\begin{aligned}
&\max_h \prod_{n=1}^{N} h_{y_n}(\mathbf{x}_n) \\
=&\max_{\mathbf{w}} \prod_{n=1}^{N} \frac{\exp(\mathbf{w}_{y_n}^T \mathbf{x}_n)}{\sum_{i=1}^{K} \exp(\mathbf{w}_i^T \mathbf{x}_n)} \\
=&\max_{\mathbf{w}} \ln\left(\prod_{n=1}^{N} \frac{\exp(\mathbf{w}_{y_n}^T \mathbf{x}_n)}{\sum_{i=1}^{K} \exp(\mathbf{w}_i^T \mathbf{x}_n)}\right) \\
=&\max_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} \left(\ln\left(\exp(\mathbf{w}_{y_n}^T \mathbf{x}_n)\right) - \ln\left(\sum_{i=1}^{K} \exp(\mathbf{w}_i^T \mathbf{x}_n)\right)\right) \\
=&\min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} \left(\ln\left(\sum_{i=1}^{K} \exp(\mathbf{w}_i^T \mathbf{x}_n)\right) - \mathbf{w}_{y_n}^T \mathbf{x}_n\right)
\end{aligned}$$

5

Find out the optimal $\mathbf{w}$:

$$\begin{aligned}
\min_{\mathbf{w}} E_{in}(\mathbf{w}) &= \min_{\mathbf{w}} \frac{1}{N+K} \left(\sum_{n=1}^{N} (y_n - \mathbf{w}^T \mathbf{x}_n)^2 + \sum_{k=1}^{K} (\tilde{y}_k - \mathbf{w}^T \tilde{\mathbf{x}}_k)^2 \right) \\
&= \min_{\mathbf{w}} \frac{1}{N+K} \ \left(||X\mathbf{w} - \mathbf{y}||^2 + ||\tilde{X}\mathbf{w} - \tilde{\mathbf{y}}||^2 \right) \\
&= \min_{\mathbf{w}} \frac{1}{N+K} \ \left(||\mathbf{w}^T X^T X \mathbf{w} - 2\mathbf{w}^T X^T \mathbf{y} + \mathbf{y}^T \mathbf{y}|| + ||\mathbf{w}^T \tilde{X}^T \tilde{X} \mathbf{w} - 2\mathbf{w}^T \tilde{X}^T \tilde{\mathbf{y}} + \tilde{\mathbf{y}}^T \tilde{\mathbf{y}}|| \right)
\end{aligned}$$

$$\begin{aligned}
\nabla E_{in}(\mathbf{w}) &= \frac{2}{N+K} \left((X^T X \mathbf{w} - X^T \mathbf{y}) + (\tilde{X}^T \tilde{X} \mathbf{w} - \tilde{X}^T \tilde{\mathbf{y}}) \right) = 0 \\
\mathbf{w} &= (X^T X + \tilde{X}^T \tilde{X})^{-1} (X^T \mathbf{y} + \tilde{X}^T \tilde{y})
\end{aligned}$$

6

Use the same way as question 5 to find out the optimal $\mathbf{w}_{reg}$:

$$\mathbf{w}_{reg} = argmin_{\mathbf{w}} \frac{\lambda}{N} ||\mathbf{w}||^2 + \frac{1}{N} ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2$$

$$\begin{aligned}
\min_{\mathbf{w}} E(\mathbf{w}) &= \min_{\mathbf{w}} \frac{\lambda}{N} ||\mathbf{w}||^2 + \frac{1}{N} ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2 \\
&= \min_{\mathbf{w}} \frac{1}{N} \left(||\lambda \mathbf{w}^T \mathbf{w}|| + ||\mathbf{w}^T X^T X \mathbf{w} - 2\mathbf{w}^T X^T \mathbf{y} + \mathbf{y}^T \mathbf{y}|| \right)
\end{aligned}$$

$$\begin{aligned}
\nabla E(\mathbf{w}) &= \frac{2}{N} \left(\lambda \mathbf{w} + (X^T X \mathbf{w} - X^T \mathbf{y}) \right) = 0 \\
\mathbf{w}_{reg} &= (X^T X + \lambda)^{-1} X^T \mathbf{y}
\end{aligned}$$

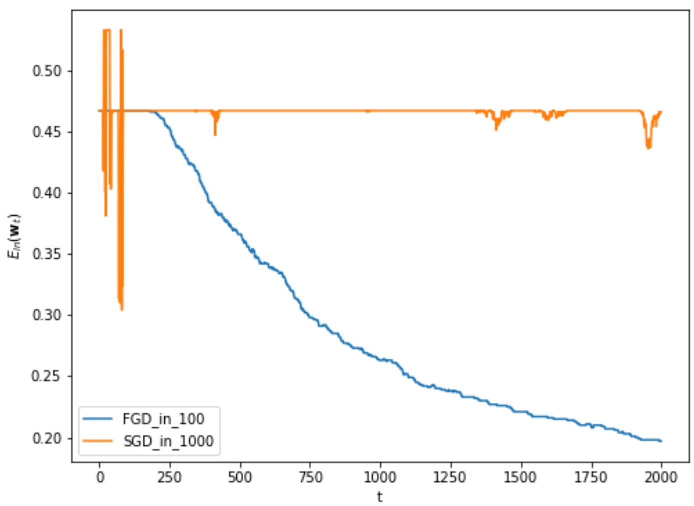
Compare $\mathbf{w}_{reg}$ with $\mathbf{w}$ in question 5, then we can get the equations below:

$$\tilde{X}^T \tilde{X} = \lambda, \ \tilde{X}^T \tilde{\mathbf{y}} = 0 \Longrightarrow \tilde{X} = \sqrt{\lambda} I, \ \tilde{\mathbf{y}} = 0$$

7

My findings

- The orange one is under stochastic gradient descent algorithm with $\eta = 0.001$, and the blue one is under fixed rate gradient descent algorithm with $\eta = 0.01$.
- The E_{in} of fixed rate gradient descent is monotonic, and the E_{in} of stochastic gradient descent is not. The main cause of this difference is the value of η. I have tried two algorithms with same value of η, and we can get similar descent of E_{in}.
- By the introduction of these two algorithms in class, we know that SGD may takes less times on computation than fixed rate gradient.



8

My findings

- The figure is similar to the figure in question 7. For all t, the value of $E_{out}(\mathbf{w}_t)$ is about 0.01 higher than $E_{in}(\mathbf{w}_t)$.