Homework 2

PENG Qiheng

Student ID 225040065

October 22, 2025

Problem 1.

(a) For a d-dimensional binary linear classifier, its VC dimension is $d + 1$.

(b) Test error is the error on the test dataset, while out-of-sample error is the expected error on the entire distribution. Test error is used to estimate out-of-sample error.

(c) False. If in-sample error is very small, it may be due to overfitting, leading to a large out-of-sample error.

(d) 3.

1) Logistic regression (LR) is designed for classification.

2) LR hasn't a closed-form solution, requiring iterative optimization.

4) LR can be applied to multi-class classification with softmax.

Problem 2.

(a) Let $\mathbf{1}$ denotes all-one vector, we have:

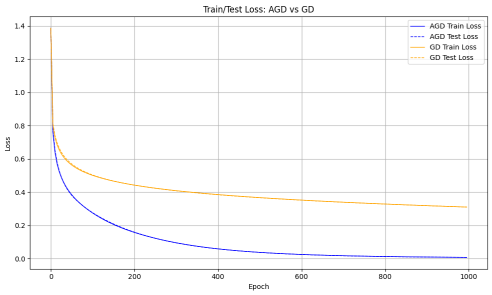
$$\mathcal{L}(\Theta, \mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n}[\log(\mathbf{1}^T e^{\Theta \mathbf{x}_i + b}) - \mathbf{y}^T(\Theta \mathbf{x}_i + \mathbf{b})]$$

$$\mathrm{d}\mathcal{L}(\Theta, \mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n}[\frac{e^{\Theta \mathbf{x}_i + b}}{\mathbf{1}^T e^{\Theta \mathbf{x}_i + b}}\mathbf{x}_i^T - \mathbf{y}^T \mathbf{x}_i^T]d\Theta + \frac{1}{n}\sum_{i=1}^{n}[\frac{e^{\Theta \mathbf{x}_i + b}}{\mathbf{1}^T e^{\Theta \mathbf{x}_i + b}} - \mathbf{y}^T]db$$

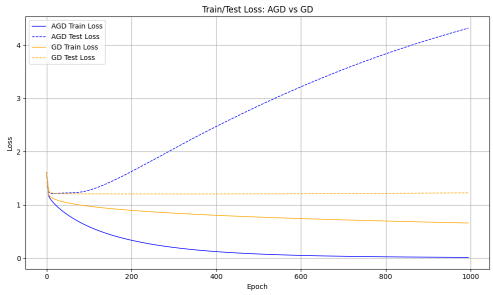
$$\frac{\mathrm{d}\mathcal{L}(\Theta, \mathbf{b})}{\mathrm{d}\Theta} = \frac{1}{n}\sum_{i=1}^{n}[\frac{e^{\Theta \mathbf{x}_i + b}}{\mathbf{1}^T e^{\Theta \mathbf{x}_i + b}} - \mathbf{y}^T]\mathbf{x}_i^T$$

$$\frac{\mathrm{d}\mathcal{L}(\Theta, \mathbf{b})}{\mathrm{d}\mathbf{b}} = \frac{1}{n}\sum_{i=1}^{n}[\frac{e^{\Theta \mathbf{x}_i + b}}{\mathbf{1}^T e^{\Theta \mathbf{x}_i + b}} - \mathbf{y}^T]$$

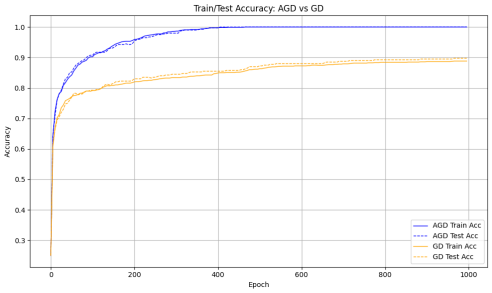
(b) The figure of training loss and test loss of two optimization algorithms are Figure (a) and Figure (b) respectively, while the figure of training accuracy and test accuracy of two optimization algorithms are Figure (c) and Figure (d) respectively.



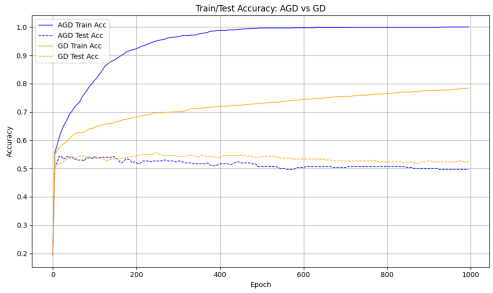
(a) Loss of Coffee Dataset



(b) Loss of Weather Dataset



(c) Accuracy of Coffee Dataset



(d) Accuracy of Weather Dataset

(i) Convergence speed: AGD is faster than GD in both datasets.

(ii) Accuracy: AGD achieves higher training and test accuracy than GD in both datasets.

(iii) Overfitting: AGD is more overfitting than GD in weather dataset, while both algorithms don't have overfitting in coffee dataset.

Problem 3.

(a) The figures of optimality gap $||x_k - x^*||_2$ obtained by the three learning rate schedules are shown in Figure 2.

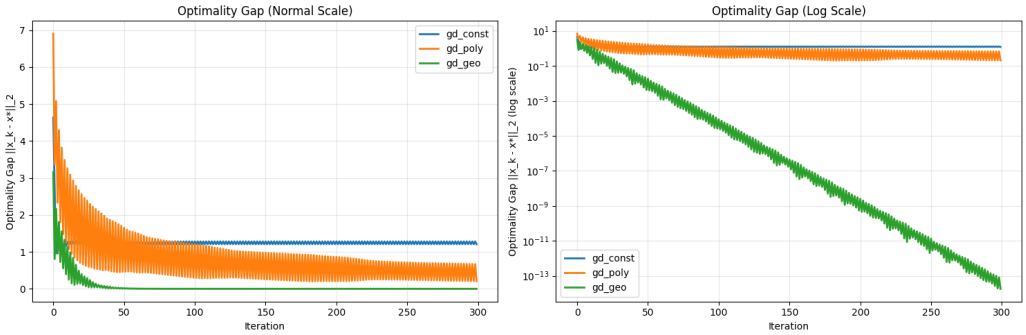


Figure 2: Optimality Gap under Different Learning Rate Schedules

(i) From the figure, we can see that among the three learning rate schedules, the geometrically diminishing learning rate converges to optimal solution, while others do not.

(ii) The speed of constant learning rate and geometrically diminishing learning rate is faster than that of polynomial diminishing learning rate.

(iii) Subgradient descent is similar to gradient descent, but it is more sensitive to learning rate schedule.

Problem 4.

(a) Follow the hint, we have:

$$\begin{aligned}\min \mathrm{Er}_{\mathrm{in}}(f) &= \min \frac{1}{n}\sum_{i=1}^{n}\left(f(x_i)-y_i\right)^2 \\ &= \min_{w} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=0}^{K} w_k L_k(x_i) - y_i\right)^2 \\ &= \min_{w} \frac{1}{n}||Aw-y||_2^2\end{aligned}$$

While A is defined as:

$$A = \begin{bmatrix} L_0(x_1) & L_1(x_1) & \cdots & L_K(x_1) \\ L_0(x_2) & L_1(x_2) & \cdots & L_K(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ L_0(x_n) & L_1(x_n) & \cdots & L_K(x_n) \end{bmatrix}$$

And we have:

$$w^* = (A^T A)^{-1} A^T y$$

For f_2 and f_{10}, we need to compute:

$$w_2^* = A_{:2}^T (A_{:2} A_{:2}^T)^{-1} y$$

$$w_{10}^* = A_{:10}^T (A_{:10} A_{:10}^T)^{-1} y$$

$$A_{:k} = \begin{bmatrix} L_0(x_1) & L_1(x_1) & \cdots & L_k(x_1) \\ L_0(x_2) & L_1(x_2) & \cdots & L_k(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ L_0(x_n) & L_1(x_n) & \cdots & L_k(x_n) \end{bmatrix}$$

(b) We denote that $w_k = 0, \forall k > K, \quad a_q = 0, \forall q > Q_g$, then

$$\begin{aligned}
\mathrm{Er}_{\mathrm{out}}(f_K) &= \mathbb{E}_{x,\epsilon}[(f_K(x) - g(x))^2] \\
&= \mathbb{E}_x \left[\left(\sum_{k=0}^{K} w_k L_k(x) - \frac{1}{C_Q} \sum_{q=0}^{Q_g} a_q L_q(x) \right)^2 \right] \\
&= \mathbb{E}_x \left[\left(\sum_{k=0}^{\max(K,Q_g)} (w_k - \frac{a_k}{C_Q}) L_k(x) \right)^2 \right]
\end{aligned}$$

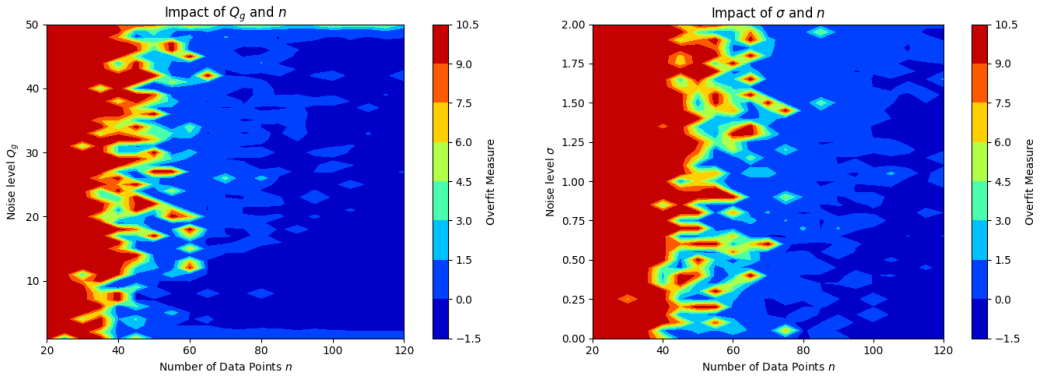
By using the orthogonality of Legendre polynomials, we have:

$$\mathrm{Er}_{\mathrm{out}}(f_K) = \mathbb{E}_x \left[\sum_{k=0}^{\max(K,Q_g)} \left(w_k - \frac{a_k}{C_Q} \right)^2 \right] \mathbb{E}_x[L_k(x)]^2$$

Since $\mathbb{E}_x[L_k(x)]^2 = \frac{1}{2} \int_{-1}^{1} [L_k(x)]^2 dx = \frac{1}{2k+1}$, we have:

$$\mathrm{Er}_{\mathrm{out}}(f_K) = \mathbb{E}_x \left[\sum_{k=0}^{\max(K,Q_g)} \frac{\left(w_k - \frac{a_k}{C_Q} \right)^2}{2k+1} \right]$$

(c) The figures of heat maps are as follows:



The noise level σ_2, the target complexity Q_g, and the number of data points n all affect overfitting.

n affects it most, a large n can efficiently alleviate overfitting, while others affect it less significantly.