ML2024 Fall Assignment 2

Wei-Chen Chang, R12227118

Due: Oct, 18 2024

Problem 1

1. Let $\mathbf{w}_1 = f_1(\mathbf{x}), \mathbf{w}_2 = f_2(\mathbf{w}_1), \mathbf{w}_3 = f_3(\mathbf{w}_2), y = f_4(\mathbf{w}_3)$, So,

$$\mathbf{x} \mapsto \mathbf{w}_1 \mapsto \mathbf{w}_2 \mapsto \mathbf{w}_3 \mapsto y$$

For the forward-mode, we compute:

$$\begin{aligned}
\dot{\mathbf{w}}_1 &= \frac{\partial \mathbf{w}_1}{\partial \mathbf{x}} = \nabla f_1(\mathbf{x}) \\
\dot{\mathbf{w}}_2 &= \frac{\partial \mathbf{w}_2}{\partial \mathbf{x}} = \frac{\partial f_2(\mathbf{w}_1)}{\partial \mathbf{w}_1} \cdot \dot{\mathbf{w}}_1 = \nabla f_2(\mathbf{w}_1) \cdot \dot{\mathbf{w}}_1 \\
\dot{\mathbf{w}}_3 &= \frac{\partial \mathbf{w}_3}{\partial \mathbf{x}} = \frac{\partial f_3(\mathbf{w}_2)}{\partial \mathbf{w}_2} \cdot \dot{\mathbf{w}}_2 = \nabla f_3(\mathbf{w}_3) \cdot \dot{\mathbf{w}}_2 \\
\text{and finally,} \quad \dot{y} &= \frac{\partial y}{\partial \mathbf{x}} = \frac{\partial f_4(\mathbf{w}_3)}{\partial \mathbf{w}_3} \cdot \dot{\mathbf{w}}_3 = \nabla f_4(\mathbf{w}_3) \cdot \dot{\mathbf{w}}_3
\end{aligned}$$

And for the reverse-mode:

$$\begin{aligned}
\overline{\mathbf{w}}_3 &= \frac{\partial y}{\partial \mathbf{w}_3} = \nabla f_4(\mathbf{w}_3) \\
\overline{\mathbf{w}}_2 &= \frac{\partial y}{\partial \mathbf{w}_2} = \frac{\partial y}{\partial \mathbf{w}_3}\frac{\partial \mathbf{w}_3}{\partial \mathbf{w}_2} = \nabla f_3(\mathbf{w}_2) \cdot \overline{\mathbf{w}}_3 \\
\overline{\mathbf{w}}_1 &= \frac{\partial y}{\partial \mathbf{w}_1} = \frac{\partial y}{\partial \mathbf{w}_2}\frac{\partial \mathbf{w}_2}{\partial \mathbf{w}_1} = \nabla f_2(\mathbf{w}_1) \cdot \overline{\mathbf{w}}_2 \\
\text{and finally,} \quad \overline{\mathbf{x}} &= \frac{\partial y}{\partial \mathbf{x}} = \frac{\partial y}{\partial \mathbf{w}_1}\frac{\partial \mathbf{w}_1}{\partial \mathbf{x}} = \nabla f_1(\mathbf{x}) \cdot \overline{\mathbf{w}}_1
\end{aligned}$$

2. Reverse-mode is more efficient. Note that in forward-mode, each $\dot{\mathbf{w}}_i$ is a matrix, and computing the next gradient requires multiplying it by another matrix. But for the reverse-mode, since $y \in \mathbb{R}$, each step is just a matrix multiplies by a vector, which results in lesser scalar multiplications.

Problem 2

First note that:

$$\begin{aligned}
\frac{\partial \mu_B}{\partial x_i} &= \frac{1}{m}\sum_{j=1}^{m}\frac{\partial x_j}{\partial x_i} = \frac{1}{m} \\
\frac{\partial \sigma_B^2}{\partial x_i} &= \frac{1}{m}\left(\frac{\partial}{\partial x_i}\sum_{j=1}^{m}(x_j-\mu_B)^2 + \frac{\partial}{\partial \mu_B}\sum_{j=1}^{m}(x_j-\mu_B)^2\cdot\frac{\partial \mu_B}{\partial x_i}\right) \\
&= \frac{1}{m}\left(2(x_i-\mu_B) + \sum_{j=1}^{m}2(x_j-\mu_B)\cdot\frac{1}{m}\right) = \frac{2}{m}(x_i-\mu_B).
\end{aligned}$$

For the first step of forward mode, compute:

$$\begin{aligned}
\frac{\partial \hat{x}_i}{\partial \mu_B} &= \frac{-1}{\sqrt{\sigma_B^2+\epsilon}} \\
\frac{\partial \hat{x}_i}{\partial \sigma_B^2} &= \frac{-1}{2}\left(\sigma_B^2+\epsilon\right)^{-\frac{3}{2}}(x_i-\mu_B) = \frac{-1}{2(\sigma_B^2+\epsilon)}\hat{x}_i \\
\frac{\partial \hat{x}_i}{\partial x_i} &= \frac{\partial \hat{x}_i}{\partial x_i}\frac{\partial x_i}{\partial x_i} + \frac{\partial \hat{x}_i}{\partial \mu_B}\frac{\partial \mu_B}{\partial x_i} + \frac{\partial \hat{x}_i}{\partial \sigma_B^2}\frac{\partial \sigma_B^2}{\partial x_i} \\
&= \frac{1}{\sqrt{\sigma_B^2+\epsilon}}\cdot 1 - \frac{1}{\sqrt{\sigma_B^2+\epsilon}}\cdot\frac{1}{m} - \frac{1}{2}(\sigma_B^2+\epsilon)^{-\frac{3}{2}}(x_i-\mu_B)\cdot\frac{2}{m}(x_i-\mu_B) \\
&= \frac{1}{\sqrt{\sigma_B^2+\epsilon}}\cdot\left(1-\frac{1}{m}-\frac{(x_i-\mu_B)^2}{m(\sigma_B^2+\epsilon)}\right) \\
&= \frac{1}{m\sqrt{\sigma_B^2+\epsilon}}\left(m-1-\hat{x}_i^2\right) \\
\frac{\partial \hat{x}_i}{\partial x_j} &= \frac{\partial \hat{x}_i}{\partial x_i}\frac{\partial x_i}{\partial x_j} + \frac{\partial \hat{x}_i}{\partial \mu_B}\frac{\partial \mu_B}{\partial x_j} + \frac{\partial \hat{x}_i}{\partial \sigma_B^2}\frac{\partial \sigma_B^2}{\partial x_j} \\
&= 0 - \frac{1}{\sqrt{\sigma_B^2+\epsilon}}\cdot\frac{1}{m} - \frac{1}{2}(\sigma_B^2+\epsilon)^{-\frac{3}{2}}(x_i-\mu_B)\cdot\frac{2}{m}(x_j-\mu_B) \\
&= -\frac{1}{m\sqrt{\sigma_B^2+\epsilon}} - \frac{(x_i-\mu_B)(x_j-\mu_B)}{m\left(\sigma_B^2+\epsilon\right)^{3/2}} \\
&= \frac{-1}{m\sqrt{\sigma_B^2+\epsilon}}\left(1+\hat{x}_i\hat{x}_j\right), \quad (i\neq j).
\end{aligned}$$

And

$$\frac{\partial y_i}{\partial \hat{x}_i} = \gamma, \quad \frac{\partial y_i}{\partial \hat{x}_j} = 0\ (i\neq j), \quad \frac{\partial y_i}{\partial \gamma} = \hat{x}_i, \quad \frac{\partial y_i}{\partial \beta} = 1$$

Hence,

$$
\begin{aligned}
\frac{\partial \ell}{\partial \hat{x}_i} &= \sum_{j=1}^{m} \frac{\partial \ell}{\partial y_j} \frac{\partial y_j}{\partial \hat{x}_i} = \gamma \frac{\partial \ell}{\partial y_i} \\
\frac{\partial \ell}{\partial \mu_B} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \frac{\partial y_i}{\partial \hat{x}_i} \frac{\partial \hat{x}_i}{\partial \mu_B} = \frac{-\gamma}{\sqrt{\sigma_B^2 + \epsilon}} \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \\
\frac{\partial \ell}{\partial \sigma_B^2} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \frac{\partial y_i}{\partial \hat{x}_i} \frac{\partial \hat{x}_i}{\partial \sigma_B^2} = \frac{-\gamma}{2} \left(\sigma_B^2 + \epsilon\right)^{-\frac{3}{2}} \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \left(x_i - \mu_B\right) \\
&= \frac{-\gamma}{2\left(\sigma_B^2 + \epsilon\right)} \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \hat{x}_i
\end{aligned}
$$

And lastly,

$$
\begin{aligned}
\frac{\partial \ell}{\partial x_i} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial \hat{x}_j} \frac{\partial \hat{x}_j}{\partial x_i} + \frac{\partial \ell}{\partial \mu_B} \frac{\partial \mu_B}{\partial x_i} + \frac{\partial \ell}{\partial \sigma_B^2} \frac{\partial \sigma_B^2}{\partial x_i} \\
&= \sum_{j \neq i} \frac{\partial \ell}{\partial \hat{x}_j} \frac{\partial \hat{x}_j}{\partial x_i} + \frac{\partial \ell}{\partial \hat{x}_i} + \frac{\partial \ell}{\partial \mu_B} \frac{\partial \mu_B}{\partial x_i} + \frac{\partial \ell}{\partial \sigma_B^2} \frac{\partial \sigma_B^2}{\partial x_i} \\
&= \sum_{j \neq i} \gamma \frac{\partial \ell}{\partial y_j} \cdot \frac{-1}{m\sqrt{\sigma_B^2 + \epsilon}} \left(1 + \hat{x}_i \hat{x}_j\right) + \gamma \frac{\partial \ell}{\partial y_i} \cdot \frac{1}{m\sqrt{\sigma_B^2 + \epsilon}} \left(m - 1 - \hat{x}_i^2\right) \\
&\quad - \frac{\gamma}{\sqrt{\sigma_B^2 + \epsilon}} \sum_{j=1}^{m} \frac{\partial \ell}{\partial y_i} \cdot \frac{1}{m} - \frac{\gamma}{2(\sigma_B^2 + \epsilon)} \sum_{j=1}^{m} \frac{\partial \ell}{\partial y_i} \hat{x}_i \cdot \frac{2}{m} (x_i - \mu_B) \\
&= -\frac{\gamma}{m\sqrt{\sigma_B^2 + \epsilon}} \left(\sum_{j \neq i} \frac{\partial \ell}{\partial y_j} \left(1 + \hat{x}_i \hat{x}_j\right) + \frac{\partial \ell}{\partial y_i} (\hat{x}_i^2 + 1 - m) + \sum_{j=1}^{m} \frac{\partial \ell}{\partial y_j} \left(1 + \hat{x}_j^2\right) \right) \\
&= -\frac{\gamma}{m\sqrt{\sigma_B^2 + \epsilon}} \left(\sum_{j=1}^{m} \frac{\partial \ell}{\partial y_j} \left(1 + \hat{x}_i \hat{x}_j + 1 + \hat{x}_j^2\right) + \frac{\partial \ell}{\partial y_i} (\hat{x}_i^2 + 1 - m) - \frac{\partial \ell}{\partial y_i} \left(1 + \hat{x}_i^2\right) \right) \\
&= \frac{\gamma}{m\sqrt{\sigma_B^2 + \epsilon}} \left(\frac{\partial \ell}{\partial y_i} m - \sum_{j=1}^{m} \frac{\partial \ell}{\partial y_j} \left(2 + \hat{x}_i \hat{x}_j + \hat{x}_j^2\right) \right) \\
\frac{\partial \ell}{\partial \gamma} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \frac{\partial y_i}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \hat{x}_i \\
\frac{\partial \ell}{\partial \beta} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \frac{\partial y_i}{\partial \beta} = \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \quad \square
\end{aligned}
$$

Problem 3

Let $L_T = L(g_{T+1}^1, \ldots, g_{T+1}^K) = \sum_{i=1}^m \exp\left(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} g_{T+1}^k(x_i) - g_{T+1}^{\hat{y}_i}(x_i)\right)$ and note that

$$
g_{T+1}^k(x_i) = g_T^k(x_i) + \alpha_T f_T^k(x_i) \quad \forall k = 1, \ldots, K, \quad \forall T \in \mathbb{N}.
$$

Hence $L_{T+1} = L(g_T^1 + \alpha_T f_T^1, \ldots, g_T^K + \alpha_T f_T^k)$, which equals to:

$$L_{T+1} = \sum_{i=1}^{m} \exp\left[\frac{1}{K-1}\sum_{k\neq\hat{y}_i}\Big(g_T^k(x_i) + \alpha_T f_T^k(x_i)\Big) - \Big(g_T^{\hat{y}_i}(x_i) + \alpha_T f_T^{\hat{y}_i}(x_i)\Big)\right].$$

In Gradient Boost, first we want to find:

$$f_T^k = \arg\min_{f^k} \frac{\partial L_{T+1}}{\partial \alpha_T}\bigg|_{\alpha_T=0} \qquad k = 1, ..., K.$$

And:

$$\begin{aligned}
\frac{\partial L_{T+1}}{\partial \alpha_T}\bigg|_{\alpha_T=0} &= \sum_{i=1}^{m} \exp\left[\frac{1}{K-1}\sum_{k\neq\hat{y}_i}\Big(g_T^k(x_i) - g_T^{\hat{y}_i}(x_i)\Big) + \frac{\alpha_T}{K-1}\sum_{k\neq\hat{y}_i}\Big(f_T^k(x_i) - f_T^{\hat{y}_i}(x_i)\Big)\right] \\
&\quad \times \frac{1}{K-1}\left(\sum_{k\neq\hat{y}_i} f_T^k(x_i) - f_T^{\hat{y}_i}(x_i)\right)\bigg|_{\alpha_T=0} \\
&= \sum_{i=1}^{m} \exp\left(\frac{1}{K-1}\sum_{k\neq\hat{y}_i} g_T^k(x_i) - g_T^{\hat{y}_i}(x_i)\right) \times \frac{1}{K-1}\sum_{k\neq\hat{y}_i} f_T^k(x_i) - f_T^{\hat{y}_i}(x_i).
\end{aligned}$$

Note that:

$$\frac{1}{K-1}\sum_{k\neq\hat{y}_i} f_T^k(x_i) - f_T^{\hat{y}_i}(x_i) = \begin{cases} -1, & f_T(x_i) = \hat{y}_i \\ \frac{1}{K-1}, & f_T(x_i) \neq \hat{y}_i \end{cases}$$

Hence we can express the sums using indicator function:

$$\begin{aligned}
\frac{1}{K-1}\sum_{k\neq\hat{y}_i} f_T^k(x_i) - f_T^{\hat{y}_i}(x_i) &= \frac{1}{K-1}\mathbf{1}_{f_T(x_i)\neq\hat{y}_i} - \mathbf{1}_{f_T(x_i)=\hat{y}_i} \\
&= \frac{K}{K-1}\mathbf{1}_{f_T(x_i)\neq\hat{y}_i} - 1
\end{aligned}$$

$\frac{1}{K-1}\sum_{k\neq\hat{y}_i} g_T^k(x_i) - g_T^{\hat{y}_i}(x_i)$ can be expressed in similar way. We now define Z_T as:

$$Z_T = \sum_{i=1}^{m} \exp\left(\frac{K}{K-1}\mathbf{1}_{g_T(x_i)\neq\hat{y}_i} - 1\right),$$

and D_T be a r.v. with the density

$$P(D_T = i) = \frac{1}{Z_T}\exp\left(\frac{K}{K-1}\mathbf{1}_{g_T(x_i)\neq\hat{y}_i} - 1\right).$$

Thus,

$$
\begin{aligned}
\left.\frac{\partial L_{T+1}}{\partial \alpha_T}\right|_{\alpha_T=0} &= \sum_{i=1}^{m} \exp\left(\frac{K}{K-1}\mathbf{1}_{g_T(x_i)\neq \hat{y}_i} - 1\right) \times \left(\frac{K}{K-1}\mathbf{1}_{f_T(x_i)\neq \hat{y}_i} - 1\right) \\
&= Z_T \sum_{i=1}^{m} \frac{1}{Z_T} \exp\left(\frac{K}{K-1}\mathbf{1}_{g_T(x_i)\neq \hat{y}_i} - 1\right) \times \left(\frac{K}{K-1}\mathbf{1}_{f_T(x_i)\neq \hat{y}_i} - 1\right) \\
&= Z_T \mathbb{E}_{i\sim D_T} \left(\frac{K}{K-1}\mathbf{1}_{f_T(x_i)\neq \hat{y}_i} - 1\right) \\
&= Z_T \frac{K}{K-1} \mathbb{P}_{i\sim D_T}\left[\mathbf{1}_{f_T(x_i)\neq \hat{y}_i}\right] - Z_T \\
&= Z_T \frac{K}{K-1} \epsilon_T - Z_T,
\end{aligned}
$$

where ϵ_T is the weighted error rate of f_T, and $0 < \epsilon_T < \frac{K-1}{K}$. To minimize $\left.\frac{\partial L_{T+1}}{\partial \alpha_T}\right|_{\alpha_T=0}$, it suffice to minimize ϵ_T, i.e., to optimize f_T.

Second, we need to find

$$
\alpha_T^* = \arg\min_{\alpha_T} \frac{\partial L_{T+1}}{\partial \alpha_T}
$$

$$
\begin{aligned}
\frac{\partial L_T + 1}{\partial \alpha_T} &= Z_T \sum_{i=1}^{m} \frac{1}{Z_T} \exp\left(\frac{K}{K-1}\mathbf{1}_{g_T(x_i)\neq \hat{y}_i} - 1\right) \exp\left[\alpha_T \left(\frac{K}{K-1}\mathbf{1}_{f_T(x_i)\neq \hat{y}_i} - 1\right)\right] \\
&\quad \times \alpha_T \left(\frac{K}{K-1}\mathbf{1}_{f_T(x_i)\neq \hat{y}_i} - 1\right) \\
&= Z_T \mathbb{E}_{i\sim D_T} \left\{\exp\left[\alpha_T \left(\frac{K}{K-1}\mathbf{1}_{f_T(x_i)\neq \hat{y}_i} - 1\right)\right] \alpha_T \left(\frac{K}{K-1}\mathbf{1}_{f_T(x_i)\neq \hat{y}_i} - 1\right)\right\} \\
&= Z_T \left[\exp(-\alpha_T)\cdot(-\alpha_T)\cdot \mathbb{P}_{i\sim D_T}[\mathbf{1}_{f_T(x_i)= \hat{y}_i}] + \exp\left(\frac{\alpha_T}{K-1}\right)\cdot \frac{\alpha_T}{K-1} \cdot \mathbb{P}_{i\sim D_T}[\mathbf{1}_{f_T(x_i)\neq \hat{y}_i}]\right] \\
&= Z_T \left[\exp(-\alpha_T)\cdot(-\alpha_T)\cdot(1-\epsilon_T) + \exp\left(\frac{\alpha_T}{K-1}\right)\cdot \frac{\alpha_T}{K-1}\cdot \epsilon_T\right] \\
&= \frac{\alpha_T}{e^{\alpha_T}} Z_T \left[\exp\left(\frac{2\alpha_T}{K-1}\right)\cdot \frac{1}{K-1}\cdot \epsilon_T - (1-\epsilon_T)\right]
\end{aligned}
$$

Observed that $\frac{\partial L_{T+1}}{\partial \alpha_T}$ increases(decreases) as α_T increasing(decreasing). To attain its minimum, it suffice to set:

$$
\frac{\alpha_T}{e^{\alpha_T}} Z_T \left[\exp\left(\frac{2\alpha_T}{K-1}\right)\cdot \frac{1}{K-1}\cdot \epsilon_T - (1-\epsilon_T)\right] = 0
$$

This implies either

1. $\alpha_T = 0$.
2. $\exp\left(\frac{2\alpha_T}{K-1}\right)\frac{\epsilon_T}{K-1} - (1-\epsilon_T) = 0$

For the second case. Solving the equation, we have

$$
\alpha_T = \frac{K-1}{2}\log\left(\frac{(K-1)(1-\epsilon_T)}{\epsilon_T}\right) \quad \square
$$

Problem 4

1. (a) See Fig.1. Since the circle is the only misclassified object, its weight would be increased.
   (b) 3 iterations. See Fig.2.

Figure 1: Answer to Problem 4 1.(a)

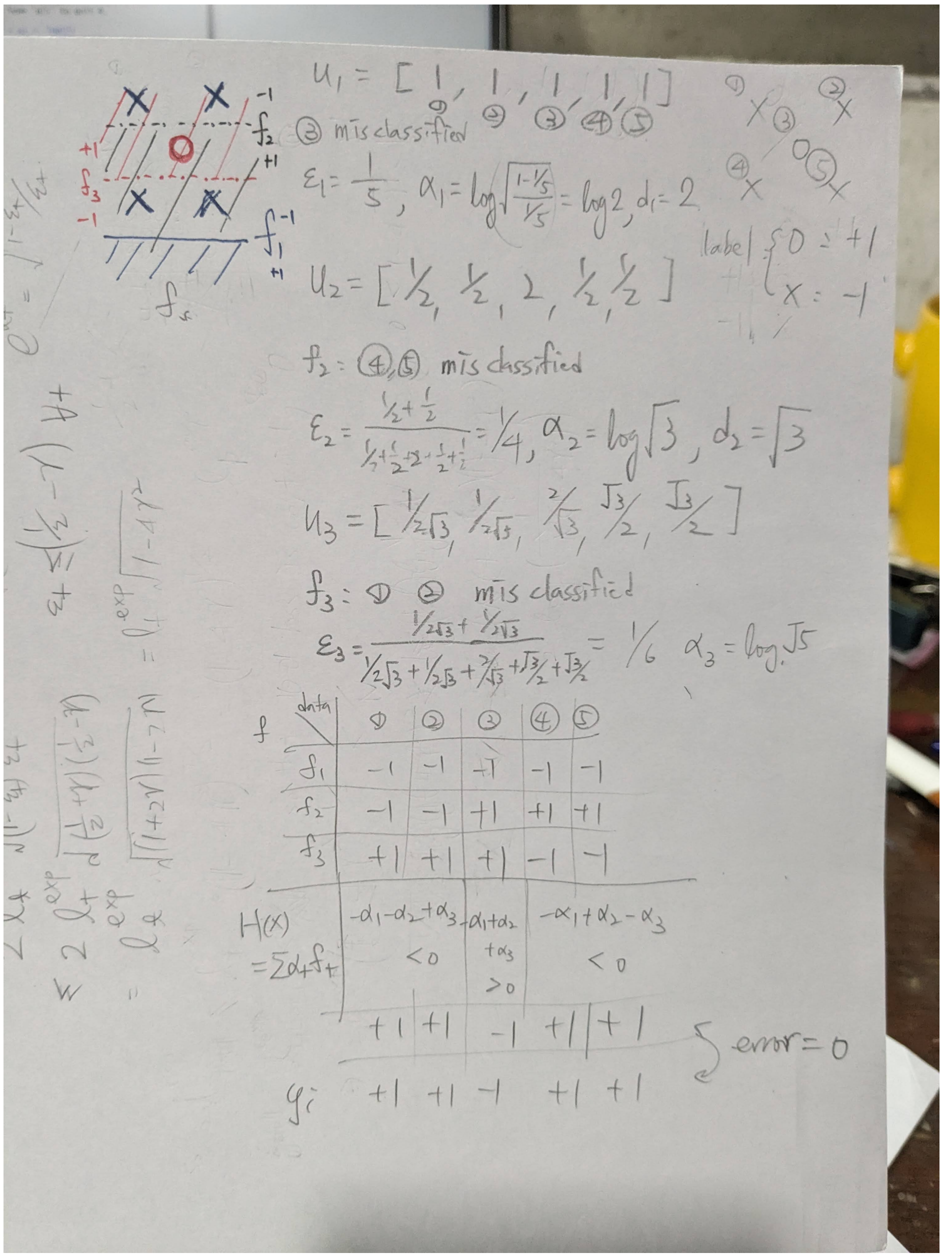

Figure 2: Answer to Problem 4 1.(b)

2. Define the aggregate classifier $H_t(x_i) = \text{sign}(g_t(x_i))$, and

$$g_t(x_i) = \begin{cases} 0, & t = 1 \\ g_{t-1}(x_i) + \alpha_t h_t(x), & t = 2, 3... \end{cases}$$

where h_t is the t'th weak hypothesis with error ϵ_t, and $\alpha_t = \frac{1}{2}\log(\frac{1-\epsilon_t}{\epsilon_t})$ by derivation of AdaBoost algorithm.

Define the exponential loss:

$$\ell_t^{exp} = \frac{1}{N}\sum_{i=1}^{N} e^{-y_i g_t(x_i)},$$

which is the upper bound of 0-1 loss:

$$\ell_t = \frac{1}{N}\sum_{i=1}^{N} \delta(H_t(x_i) \neq y_i).$$

We have:

$$\begin{aligned} \ell_t^{exp} &= \frac{1}{N}\sum_{i=1}^{N} e^{-y_i(g_{t-1}(x_i)+\alpha_t h_t(x_i))} \\ &= \frac{1}{N}\sum_{i=1}^{N} e^{-y_i g_{t-1}(x_i)} e^{-\alpha_t y_i h_t(x_i)}. \end{aligned}$$

Let $Z_t = \sum_{i=1}^{N} e^{-y_i g_t(x_i)}$ and D_t be a r.v. with density $\mathbb{P}(D_t = i) = \frac{1}{Z_t} e^{-y_i g_t(x_i)}$, $\forall t = 1, 2, 3, ...$ Note that $Z_{t-1} = N\ell_{t-1}^{exp}$, thus,

$$\begin{aligned} \ell_t^{exp} &= \frac{1}{N} Z_{t-1} \cdot \mathbb{E}_{i\sim D_{t-1}}\left[e^{-\alpha_t y_i h_t(x_i)}\right] \\ &= \ell_{t-1}^{exp} \cdot \mathbb{E}_{i\sim D_{t-1}}\left[e^{-\alpha_t}\mathbf{1}_{y_i = h_t(x_i)} + e^{\alpha_t}\mathbf{1}_{y_i \neq h_t(x_i)}\right] \\ &= \ell_{t-1}^{exp} \cdot \left[e^{-\alpha_t}(1-\epsilon_t) + e^{\alpha_t}\epsilon_t\right] \\ &= 2\ell_{t-1}^{exp}\sqrt{(1-\epsilon_t)\epsilon_t} \quad (\text{Since } \alpha_t = \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t}) \\ &\leq \ell_{t-1}^{exp}\sqrt{(1-2\gamma)(1+2\gamma)} \quad \forall t = 1, 2, 3, ... \end{aligned}$$

Since $\ell_1^{exp} = 1$,

$$\ell_{T+1}^{exp} \leq \ell_1^{exp}(1-4\gamma^2)^{T/2} = (1-4\gamma^2)^{T/2}$$

To attain zero 0-1 loss after T iterations, its upper-bound (exponential loss) then should be smaller than $\frac{1}{N}$, which is exactly when 1 training sample misclassified. That is,

$$0 = \ell_{T+1} \leq \ell_{T+1}^{exp} \leq (1-4\gamma^2)^{T/2} < \frac{1}{N}$$

Solving the inequality, we have:

$$\frac{T}{2}\log(1-4\gamma^2) < -\log(N)$$

Since $\log(1-4\gamma^2) < 0$, divide both sides with $\frac{\log(1-4\gamma^2)}{2}$ gives:

$$T > \frac{-2\log(N)}{\log(1-4\gamma^2)} \quad \square$$

Problem 5

1. (a) By Fundamental Theorem of Calculus,

$$\int_0^1 g'(t)dt = g(1) - g(0) = f(\mathbf{y}) - f(\mathbf{x}) \quad \square$$

(b) Let $\mathbf{z} = \mathbf{x} + t(\mathbf{y} - \mathbf{x}) = \begin{pmatrix} z_1 & ... & z_n \end{pmatrix}$ and $\mathbf{u} = \Delta t(\mathbf{y} - \mathbf{x}) = \begin{pmatrix} u_1 & ... & u_n \end{pmatrix}$

$$\begin{aligned}
g'(t) &= \lim_{\Delta t \to 0} \frac{f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}) + \Delta t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{f(\mathbf{z} + \mathbf{u}) - f(z_1, z_2 + u_2, ..., z_n + u_n) + ... + f(z_1, z_2, ..., z_n + u_n) - f(\mathbf{z})}{\Delta t} \\
&= \lim_{u_1 \to 0} \frac{f(\mathbf{z} + \mathbf{u}) - f(z_1, z_2 + u_2, ..., z_n + u_n)}{u_1}(y_1 - x_1) + ... \\
&\quad + \lim_{u_n \to 0} \frac{f(z_1, z_2, ..., z_n + u_n) - f(\mathbf{z})}{u_n}(y_n - x_n) \\
&= \sum_{i=1}^{n} \frac{\partial f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))}{\partial u_i}(y_i - x_i) \\
&= \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) \cdot (\mathbf{y} - \mathbf{x}) = \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) \quad \square
\end{aligned}$$

(c) By (a), (b), we have

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^\top (\mathbf{y} - \mathbf{x})\, dt.$$

Adding $-\nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$ and taking absolute value on both sides doesn't change the equality. Thus,

$$\left| f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \right| = \left| \int_0^1 \left(\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}) \right)^\top (\mathbf{y} - \mathbf{x})\, dt \right|$$

Claim that:

$$\left| \int_0^1 f(t)dt \right| \le \int_0^1 |f(t)|dt.$$

proof: Since

$$-|f(t)| \le f(t) \le |f(t)|$$
$$-\int_0^1 |f(t)|dt \le \int_0^1 f(t)dt \le \int_0^1 |f(t)|dt$$

It immediately shows that $\left|\int_0^1 f(t)dt\right| \le \int_0^1 |f(t)|dt$. By claim,

$$\begin{aligned}\left|f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})\right| &= \left|\int_0^1 \left(\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{x})\right)^\top (\mathbf{y}-\mathbf{x})\, dt\right| \\ &\le \int_0^1 \left|\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})\right|\, dt \quad \square\end{aligned}$$

(d) Cauchy-Schwarz Inequality states that for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$:

$$|\mathbf{u}^\top \mathbf{v}| = |\mathbf{u}\cdot\mathbf{v}| \le \|\mathbf{u}\|_2\|\mathbf{v}\|_2$$

Thus:

$$\begin{aligned}\left|f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})\right| &\le \int_0^1 \left|\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{y})^\top(\mathbf{y}-\mathbf{x})\right|\, dt \\ &\le \int_0^1 \left\|\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{x})\right\|_2 \|\mathbf{y}-\mathbf{x}\|_2\, dt \\ &\le \int_0^1 \beta\|\mathbf{x}+t(\mathbf{y}-\mathbf{x})-\mathbf{x}\|_2\, \|\mathbf{y}-\mathbf{x}\|_2\, dt \quad (\beta\text{-smoothness}) \\ &= \int_0^1 \beta t\|\mathbf{y}-\mathbf{x}\|_2^2\, dt \\ &= \beta\|\mathbf{y}-\mathbf{x}\|_2^2 \times \left.\frac{t^2}{2}\right|_{t=0}^1 \\ &= \frac{\beta}{2}\|\mathbf{y}-\mathbf{x}\|_2^2\end{aligned}$$

Hence

$$f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) \le \frac{\beta}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \quad \square$$

2. By 1.(d),

$$\begin{aligned}f\left(\mathbf{x} - \frac{1}{\beta}\nabla f(\mathbf{x})\right) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top\left(-\frac{1}{\beta}\nabla f(\mathbf{x})\right) &\le \frac{\beta}{2}\| - \frac{1}{\beta}\nabla f(\mathbf{x})\|_2^2 \\ \Rightarrow f\left(\mathbf{x} - \frac{1}{\beta}\nabla f(\mathbf{x})\right) - f(\mathbf{x}) + \frac{1}{\beta}\|\nabla f(\mathbf{x})\|_2^2 &\le \frac{1}{2\beta}\|\nabla f(\mathbf{x})\|_2^2 \\ \Rightarrow f\left(\mathbf{x} - \frac{1}{\beta}\nabla f(\mathbf{x})\right) - f(\mathbf{x}) &\le \frac{-1}{2\beta}\|\nabla f(\mathbf{x})\|_2^2.\end{aligned}$$

Also, $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$, so, $f(\mathbf{x}^*) \le f(\mathbf{x}) \quad \forall \mathbf{x} \in \mathbb{R}^n$. Then,

$$f(\mathbf{x}^*) - f(\mathbf{x}) \le \frac{-1}{2\beta}\|\nabla f(\mathbf{x})\|_2^2. \quad \square$$

3.

$$\begin{aligned}\|\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^*\|_2^2 &= \|\boldsymbol{\theta}^n - \boldsymbol{\theta}^* - \eta\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)\|_2^2 \\ &= \|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\|_2^2 + \eta^2\|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)\|_2^2 - 2\eta\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)^\top(\boldsymbol{\theta}^n - \boldsymbol{\theta}^*).\end{aligned}$$

4. Since $\eta = \frac{1}{\beta} > 0$,

$$\begin{aligned}
\frac{-1}{2\beta}\|\nabla f(\boldsymbol{\theta}^n)\|_2^2 &\geq [f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}^n)] \quad \text{(by 2.)} \\
\Rightarrow \frac{1}{\beta^2}\frac{-1}{2\beta}\|\nabla f(\boldsymbol{\theta}^n)\|_2^2 &\geq \frac{1}{\beta^2}[f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}^n)] \\
\Rightarrow \eta^2\|\nabla f(\boldsymbol{\theta}^n)\|_2^2 &\leq \frac{-2}{\beta}[f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}^n)].
\end{aligned}$$

Thus 3. can be rewrite as:

$$\begin{aligned}
\|\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^*\|_2^2 &\leq \|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\|_2^2 - \frac{2}{\beta}[f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}^n)] - \frac{2}{\beta}\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)^\top(\boldsymbol{\theta}^n - \boldsymbol{\theta}^*) \\
&= \|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\|_2^2 - \frac{2}{\beta}\left[f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}^n) + \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)^\top(\boldsymbol{\theta}^* - \boldsymbol{\theta}^n)\right] \\
&\leq \|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\|_2^2 - \frac{2}{\beta}\frac{\alpha}{2}\|\boldsymbol{\theta}^* - \boldsymbol{\theta}^n\|_2^2 \quad (\alpha\text{-strongly convex}) \\
&= (1 - \frac{\alpha}{\beta})\|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\|_2^2. \quad \square
\end{aligned}$$

5.
$$\lim_{n\to\infty}\|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\|_2^2 = \lim_{n\to\infty}(1 - \frac{\alpha}{\beta})^n\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^*\|_2^2 = 0,$$

if $\left|1 - \frac{\alpha}{\beta}\right| < 1$ (i.e., $\alpha < 2\beta$).