ML Written Homework 2

Student: R14921A13 鄭皓中

October 17, 2025

1 Layer Normalization

(a) The values we known are $x_i, \mu, \sigma^2, \hat{x}_i, y_i, \frac{\partial \ell}{\partial y_i}$.

$$\frac{\partial \mu}{\partial x_i} = \frac{1}{d}$$

$$\frac{\partial \sigma^2}{\partial x_i} = \frac{1}{d}\sum_{j=1}^{d} 2(x_j - \mu)(\frac{\partial x_j}{\partial x_i} - \frac{\partial \mu}{\partial x_i}) = \frac{2}{d}((x_i - \mu) - \sum_{j=1}^{d}(x_j - \mu)) = \frac{2}{d}(x_i - \mu)$$

$$\frac{\partial \hat{x}_i}{\partial x_i} = \frac{1}{\sqrt{\sigma^2 + \epsilon}}$$

Compute $\frac{\partial \ell}{\partial \hat{x}_i}, \frac{\partial \ell}{\partial \sigma^2}, \frac{\partial \ell}{\partial \mu}$

$$\frac{\partial \ell}{\partial \hat{x}_i} = \frac{\partial \ell}{\partial y_i}\frac{\partial y_i}{\partial \hat{x}_i} = \gamma\frac{\partial \ell}{\partial y_i}$$

$$\frac{\partial \hat{x}_i}{\partial \sigma^2} = \frac{-1}{2}\frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}\frac{1}{\sigma^2 + \epsilon} = -\frac{1}{2}\hat{x}_i\frac{1}{\sigma^2 + \epsilon}$$

$$\frac{\partial \hat{x}_i}{\partial \mu} = -\frac{1}{\sqrt{\sigma^2 + \epsilon}}$$

$$\frac{\partial \sigma^2}{\partial \mu} = -\frac{2}{d}\sum_{j=1}^{d}(x_j - \mu) = 0$$

$$\frac{\partial \ell}{\partial \sigma^2} = \sum_{i=1}^{d}\frac{\partial \ell}{\partial \hat{x}_i}\frac{\partial \hat{x}_i}{\partial \sigma^2} = \sum_{i=1}^{d}\gamma\frac{\partial \ell}{\partial y_i}(-\frac{1}{2}\hat{x}_i\frac{1}{\sigma^2 + \epsilon})$$

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{d}\frac{\partial \ell}{\partial \hat{x}_i}\frac{\partial \hat{x}_i}{\partial \mu} + \frac{\partial \ell}{\partial \sigma^2}\frac{\partial \sigma^2}{\partial \mu} = -\sum_{i=1}^{d}\gamma\frac{\partial \ell}{\partial y_i}\frac{1}{\sqrt{\sigma^2 + \epsilon}}$$

Hence, we have

$$\begin{aligned}
\frac{\partial \ell}{\partial x_i} &= \frac{\partial \ell}{\partial \hat{x}_i}\frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial \ell}{\partial \sigma^2}\frac{\partial \sigma^2}{\partial x_i} + \frac{\partial \ell}{\partial \mu}\frac{\partial \mu}{\partial x_i} \\
&= \gamma\frac{\partial \ell}{\partial y_i}\frac{1}{\sqrt{\sigma^2+\epsilon}} + \sum_{j=1}^{d}\gamma\frac{\partial \ell}{\partial y_j}(-\frac{1}{2}\hat{x}_i\frac{1}{\sigma^2+\epsilon})\frac{2}{d}(x_i-\mu) - \sum_{i=1}^{d}\gamma\frac{\partial \ell}{\partial y_i}\frac{1}{\sqrt{\sigma^2+\epsilon}}\frac{1}{d} \\
&= \frac{1}{\sqrt{\sigma^2+\epsilon}}(\gamma\frac{\partial \ell}{\partial y_i} - \frac{1}{d}\hat{x}_1\sum_{j=1}^{d}\gamma\frac{\partial \ell}{\partial y_j}\hat{x}_j - \frac{1}{d}\sum_{j=1}^{d}\gamma\frac{\partial \ell}{\partial y_j}) \\
\frac{\partial \ell}{\partial \beta} &= \sum_{i=1}^{d}\frac{\partial \ell}{\partial y_i}\frac{\partial y_i}{\partial \beta} = \sum_{i=1}^{d}\frac{\partial \ell}{\partial y_i} \\
\frac{\partial \ell}{\partial \gamma} &= \sum_{i=1}^{d}\frac{\partial \ell}{\partial y_i}\frac{\partial y_i}{\partial \gamma} = \sum_{i=1}^{d}\frac{\partial \ell}{\partial y_i}\hat{x}_i
\end{aligned}$$

(b) After normalization, the distribution of each feature has same mean and std, but if some features need nonzero mean or nonunit std to get the better performance, only normalization is not enough to learn these feature. Hence, we need to use parameter γ, β to make those feature have nonzero mean and nonunit std, after training, if the feature has the best performance with normalization, then it will be $\gamma = 1, \beta = 0$.

2 Classification with Gaussian Mixture Model

(a) (i)

$$\begin{aligned}
L(\theta) = P_\theta(\{(\mathbf{x}_i, y_i)\}_{i=1}^N) &= \prod_{i=1}^{N} P_\theta(\mathbf{X} = \mathbf{x}_i, \mathbf{Y} = y_i) \\
&= \prod_{i=1}^{N}[\pi_1 f_{\mu_1,\Sigma_1}(\mathbf{x}_i)]^{\mathbb{I}(y_i=C_1)}[\pi_2 f_{\mu_2,\Sigma_2}(\mathbf{x}_i)]^{\mathbb{I}(y_i=C_2)} \\
&= \frac{1}{(2\pi)^{Nd}|\mathbf{\Sigma}_1|^{\frac{N}{2}}|\mathbf{\Sigma}_2|^{\frac{N}{2}}}\exp\left(-\frac{1}{2}\sum_{k=1}^{2}\sum_{i=1}^{N}\mathbb{I}(y_i = C_k)\mathbf{S}_{i,k}\right)
\end{aligned}$$

where $\mathbb{I}(y_i = C_k) = 1$ if $y_i = C_k$ else 0 and $\mathbf{S}_{i,k} = (\mathbf{x}_i - \mu_k)^T\Sigma_k^{-1}(\mathbf{x}_i - \mu_k)$

(ii) Consider $I_k = \{i | y_i = C_k\}$, $N_k = |I_k|$ and $\ell(\theta) = \log(L(\theta))$, we have

$$\ell(\theta) = \sum_{k=1}^{2}\sum_{i\in I_k}\log(\pi_k f_{\mu_k,\Sigma_k}(\mathbf{x}_i)) = \sum_{k=1}^{2}(N_k\log\pi_k + \sum_{i\in I_k}\log f_{\mu_k,\Sigma_k}(\mathbf{x}_i))$$

Maximize $N_1\log\pi_1 + N_2\log\pi_2$ constraint on $\pi_1 + \pi_2 = 1$:
Use Lagrange multiplier λ, $J(\pi_1, \pi_2, \lambda) = N_1\log\pi_1 + N_2\log\pi_2 + \lambda(1 - \pi_1 - \pi_2)$

$$\frac{\partial J}{\partial \pi_1} = \frac{N_1}{\pi_1} - \lambda = 0 \Rightarrow \pi_1 = \frac{N_1}{\lambda}$$

$$\frac{\partial J}{\partial \pi_2} = \frac{N_2}{\pi_2} - \lambda = 0 \Rightarrow \pi_2 = \frac{N_2}{\lambda}$$

Then we have $\lambda = N_1 + N_2 = N$, and the maximum happens when $\pi_k^* = \frac{N_k}{N}$.
If we want to maximize $L_k(\mu_k, \Sigma_k) = \sum_{i \in I_k} \log f_{\mu_k, \Sigma_k}(\mathbf{x}_i)$, we only need to consider the case $k = 1$ since $k = 2$ has the same situation as $k = 1$.

$$\begin{aligned}\frac{\partial L_k}{\partial \mu_k} &= -\frac{1}{2} \sum_{i \in I_k} \frac{\partial}{\partial \mu_k} (\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \mu_k) \\ &= -\frac{1}{2} \sum_{i \in I_k} [-2\Sigma_k^{-1}(\mathbf{x}_i - \mu_k)] = \sum_{i \in I_k} \Sigma_k^{-1}(\mathbf{x}_i - \mu_k)\end{aligned}$$

Compute $\frac{\partial L_k}{\partial \mu_k} = 0$, we get $\mu_k^* = \frac{1}{N_k} \sum_{i=1}^{N} \mathbb{I}(y_i = C_k)\mathbf{x}_i$

$$\frac{\partial L_k}{\partial \Sigma_k^{-1}} = \frac{N_k}{2}\Sigma_k - \frac{1}{2} \sum_{i \in I_k} (\mathbf{x_i} - \mu_\mathbf{k}^*)(\mathbf{x_i} - \mu_\mathbf{k}^*)^T$$

Let $\frac{\partial L_k}{\partial \Sigma_k^{-1}} = 0$ we get

$$N_k \Sigma_k^* = \sum_{i \in I_k} (\mathbf{x_i} - \mu_\mathbf{k}^*)(\mathbf{x_i} - \mu_\mathbf{k}^*)^T$$

$$\Sigma_k^* = \frac{1}{N_k} \sum_{i=1}^{N} \mathbb{I}(y_i = C_k)(\mathbf{x_i} - \mu_\mathbf{k}^*)(\mathbf{x_i} - \mu_\mathbf{k}^*)^T$$

(iii)

$$\begin{aligned}P(\mathbf{X} = \mathbf{x} | \mathbf{Y} = C_1) &= \frac{P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = C_1)}{P(\mathbf{Y} = C_1)} \\ &= \frac{\pi_1 f_{\mu_1, \Sigma_1}}{\pi_1} \\ &= f_{\mu_1, \Sigma_1}(\mathbf{x}) \\ &= \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_1|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_1)^T \Sigma_1^{-1} (\mathbf{x} - \mu_1)\right)\end{aligned}$$

This is the Gaussian distribution of class C_1.

$$\begin{aligned}P(\mathbf{Y} = C_1 | \mathbf{X} = \mathbf{x}) &= \frac{P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = C_1)}{P(\mathbf{X} = \mathbf{x})} \\ &= \frac{\pi_1 f_{\mu_1, \Sigma_1}}{\pi_1 f_{\mu_1, \Sigma_1} + \pi_2 f_{\mu_2, \Sigma_2}}\end{aligned}$$

This is the probability of $y = C_1$ with given data $\mathbf{x}$.

(iv)

$$\begin{aligned}P(\mathbf{Y} = C_1 | \mathbf{X} = \mathbf{x}) &= \frac{\pi_1 f_{\mu_1, \Sigma_1}}{\pi_1 f_{\mu_1, \Sigma_1} + \pi_2 f_{\mu_2, \Sigma_2}} \\ &= \frac{1}{1 + \frac{\pi_2 f_{\mu_2, \Sigma_2}}{\pi_1 f_{\mu_1, \Sigma_1}}} \\ &= \frac{1}{1 + \exp(-(-\log \frac{\pi_2 f_{\mu_2, \Sigma_2}}{\pi_1 f_{\mu_1, \Sigma_1}}))}\end{aligned}$$

Then we have $P(\mathbf{Y}=C_1|\mathbf{X}=\mathbf{x})=\sigma(z)$ where $z=-\log\frac{\pi_2 f_{\mu_2,\Sigma_2}}{\pi_1 f_{\mu_1,\Sigma_1}}$. Compute $\log f$

$$\log f_{\mu_k,\Sigma_k}=-\frac{d}{2}\log(2\pi)-\frac{1}{2}\log|\Sigma_k|-\frac{1}{2}\log(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)$$

Hence, we can compute z

$$z=\log(\frac{\pi_1}{\pi_2})-\frac{1}{2}(\log|\Sigma_1|-\log|\Sigma_2|)-\frac{1}{2}(\mathbf{S}_1-\mathbf{S}_2)$$

where $\mathbf{S}_k=(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)$

(b) (i)

$$\begin{aligned}L(\theta)&=P_\theta(\{(\mathbf{x}_i,y_i)\}_{i=1}^N)=\prod_{i=1}^N P_\theta(\mathbf{X}=\mathbf{x}_i,\mathbf{Y}=y_i)\\&=\prod_{i=1}^N[\pi_1 f_{\mu_1,\Sigma_1}(\mathbf{x}_i)]^{\mathbb{I}(y_i=C_1)}[\pi_2 f_{\mu_2,\Sigma_2}(\mathbf{x}_i)]^{\mathbb{I}(y_i=C_2)}\\&=\frac{1}{(2\pi)^{Nd}|\mathbf{\Sigma}|^N}\exp\left(-\frac{1}{2}\sum_{k=1}^2\sum_{i=1}^N\mathbb{I}(y_i=C_k)\mathbf{S}_{i,k}\right)\end{aligned}$$

where $\mathbb{I}(y_i=C_k)=1$ if $y_i=C_k$ else 0 and $\mathbf{S}_{i,k}=(\mathbf{x}_i-\mu_k)^T\Sigma^{-1}(\mathbf{x}_i-\mu_k)$

(ii) The only change is $\Sigma_1=\Sigma_2=\Sigma$, so μ_k^* and π_k^* keeps same.

$$\frac{\partial L_k}{\partial\Sigma^{-1}}=\frac{N}{2}\Sigma-\frac{1}{2}\sum_{k=1}^2\sum_{i=1}^N\mathbb{I}(y_i=C_k)(\mathbf{x_i}-\mu_\mathbf{k}^*)(\mathbf{x_i}-\mu_\mathbf{k}^*)^T$$

Let $\frac{\partial L_k}{\partial\Sigma^{-1}}=0$ we get

$$N\Sigma^*=\sum_{k=1}^2\sum_{i=1}^N\mathbb{I}(y_i=C_k)(\mathbf{x_i}-\mu_\mathbf{k}^*)(\mathbf{x_i}-\mu_\mathbf{k}^*)^T$$

$$\Sigma^*=\frac{1}{N}\sum_{k=1}^2\sum_{i=1}^N\mathbb{I}(y_i=C_k)(\mathbf{x_i}-\mu_\mathbf{k}^*)(\mathbf{x_i}-\mu_\mathbf{k}^*)^T$$

(iii)

$$\begin{aligned}P(\mathbf{X}=\mathbf{x}|\mathbf{Y}=C_1)&=\frac{P(\mathbf{X}=\mathbf{x},\mathbf{Y}=C_1)}{P(\mathbf{Y}=C_1)}\\&=\frac{\pi_1 f_{\mu_1,\Sigma}}{\pi_1}\\&=f_{\mu_1,\Sigma}(\mathbf{x})\\&=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1)\right)\end{aligned}$$

This is the Gaussian distribution of class C_1.

$$P(\mathbf{Y} = C_1|\mathbf{X} = \mathbf{x}) = \frac{P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = C_1)}{P(\mathbf{X} = \mathbf{x})}$$
$$= \frac{\pi_1 f_{\mu_1,\Sigma}}{\pi_1 f_{\mu_1,\Sigma} + \pi_2 f_{\mu_2,\Sigma}}$$

This is the probability of $y = C_1$ with given data $\mathbf{x}$.

(iv) From (a) we have z

$$z = \log(\frac{\pi_1}{\pi_2}) - \frac{1}{2}(\mathbf{S}_1 - \mathbf{S}_2)$$

where $\mathbf{S}_k = (\mathbf{x} - \mu_k)^T \Sigma^{-1} (\mathbf{x} - \mu_k)$.

Compute $\mathbf{S}_1 - \mathbf{S}_2$

$$(\mathbf{x} - \mu_k)^T \Sigma^{-1} (\mathbf{x} - \mu_k) = \mathbf{x}^T \Sigma^{-1} \mathbf{x} - 2\mu_k^T \Sigma^{-1} \mathbf{x} + \mu_k^T \Sigma^{-1} \mu_k$$

$$\mathbf{S}_1 - \mathbf{S}_2 = -2(\mu_1 - \mu_2)^T \Sigma^{-1} \mathbf{x} + \mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2$$

$$z = (\mu_1 - \mu_2)^T \Sigma^{-1} \mathbf{x} + [\log(\frac{\pi_1}{\pi_2}) - \frac{1}{2}(\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2)]$$

3 Multiclass AdaBoost

Apply gradient boosting we have $\mathbf{g}_{T+1}^k = \mathbf{g}_T^k + \alpha_t \mathbf{f}_t(x)$ where $\mathbf{f}_t(x) = (f_t^k(x))_{k=1}^K$ and $f_t^k(x) = \mathbb{I}(f_t(x) = k)$. To minimize loss function L, first we compute $r_{i,k} = -\frac{\partial L}{\partial g_t^k(x_i)}$, assume

$$h_{i,t} = \frac{1}{K-1} \sum_{k \neq \hat{y}_i} g_t^k(x_i) - g_t^{\hat{y}_i}(x_i)$$

Case 1: $k = \hat{y}_i$

$$r_{i,k} = -\frac{\partial L}{\partial g_t^{\hat{y}_i}(x_i)}$$
$$= -\exp(h_{i,t}) \frac{\partial h_{i,t}}{\partial g_t^{\hat{y}_i}(x_i)}$$
$$= -\exp(h_{i,t})(-1)$$
$$= \exp(h_{i,t})$$

Case 2: $k \neq \hat{y}_i$

$$r_{i,k} = -\frac{\partial L}{\partial g_t^k(x_i)}$$
$$= -\exp(h_{i,t}) \frac{\partial h_{i,t}}{\partial g_t^k(x_i)}$$
$$= -\exp(h_{i,t})(\frac{1}{K-1})$$
$$= -\frac{1}{K-1} \exp(h_{i,t})$$

Now consider error rate

$$\text{err}_t = \frac{\sum_{i=1}^m w_i^t \mathbb{I}(f_t(x_i) \neq \hat{y}_i)}{\sum_{i=1}^m w_i^t}$$

where w_i^t is the sample weight, which is given by $r_{i,\hat{y}_i}$

$$w_i^t = \exp(h_{i,t})$$

Observe the following situation

$$\begin{aligned}
\sum_{i=1}^m \sum_{k=1}^K r_{i,k} f_t^k(x_i) &= \sum_{i=1}^m r_{i,f_t(x_i)} \\
&= \sum_{i:correct} r_{i,\hat{y}_i} + \sum_{i:wrong} r_{i,f_t(x_i)} \\
&= (\sum_{i=1}^m w_i^t - \sum_{i:wrong} w_i^t) - \frac{1}{K-1} \sum_{i:wrong} w_i^t \\
&= \sum_{i=1}^m w_i^t - \frac{K}{K-1} \sum_{i=1}^m \mathbb{I}(f_t(x_i) \neq \hat{y}_i) w_i^t
\end{aligned}$$

If we want to maximize $\sum_{i=1}^m \sum_{k=1}^K r_{i,k} f_t^k(x_i)$, it is equal to minimize $\sum_{i=1}^m \mathbb{I}(f_t(x_i) \neq \hat{y}_i) w_i^t$. Hence, we need to find f_t by minimizing the error rate.
Once we find $\mathbf{f}_t$ (the one-hot representation of f_t), we can compute α_t

$$\begin{aligned}
\alpha_t &= \arg\min_\alpha L(\mathbf{g}_t + \alpha \mathbf{f}_t) \\
&= \arg\min_\alpha \sum_{i=1}^m \exp\left(h_{i,t} + \alpha \left(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} f_t^k(x_i) - f_t^{\hat{y}_i}(x_i) \right) \right) \\
&= \arg\min_\alpha \sum_{i=1}^m (\mathbb{I}(f_t(x_i) = \hat{y}_i) w_i^t e^{-\alpha} + \mathbb{I}(f_t(x_i) \neq \hat{y}_i) w_i^t e^{\frac{\alpha}{K-1}})
\end{aligned}$$

$$\begin{aligned}
\text{gradient} &= \frac{\partial}{\partial \alpha} \sum_{i=1}^m (\mathbb{I}(f_t(x_i) = \hat{y}_i) w_i^t e^{-\alpha} + \mathbb{I}(f_t(x_i) \neq \hat{y}_i) w_i^t e^{\frac{\alpha}{K-1}}) \\
&= \sum_{i=1}^m (-\mathbb{I}(f_t(x_i) = \hat{y}_i) w_i^t e^{-\alpha} + \frac{1}{K-1} \mathbb{I}(f_t(x_i) \neq \hat{y}_i) w_i^t e^{\frac{\alpha}{K-1}}) = 0 \\
&\quad -e^{-\alpha} \sum_{i \in I_{\text{correct}}} w_i^t + \frac{1}{K-1} e^{\alpha/(K-1)} \sum_{i \in I_{\text{wrong}}} w_i^t = 0 \\
&\quad e^{-\alpha} \sum_{i \in I_{\text{correct}}} w_i^t = \frac{1}{K-1} e^{\alpha/(K-1)} \sum_{i \in I_{\text{wrong}}} w_i^t \\
&\quad \frac{e^{-\alpha}}{e^{\alpha/(K-1)}} = \frac{1}{K-1} \frac{\sum_{i \in I_{\text{wrong}}} w_i^t}{\sum_{i \in I_{\text{correct}}} w_i^t} \\
&\quad \frac{e^{-\alpha}}{e^{\alpha/(K-1)}} = e^{-\alpha - \frac{\alpha}{K-1}} = e^{-\alpha\left(1 + \frac{1}{K-1}\right)} = e^{-\alpha \frac{K}{K-1}}
\end{aligned}$$

$$
\begin{aligned}
e^{-\alpha \frac{K}{K-1}} &= \frac{1}{K-1} \frac{\sum_{i \in I_{\text{wrong}}} w_i^t}{\sum_{i \in I_{\text{correct}}} w_i^t} \\
&= \frac{1}{K-1} \frac{\text{err}_t}{1-\text{err}_t} \\
-\alpha \frac{K}{K-1} &= \log\left(\frac{\text{err}_t}{1-\text{err}_t}\right) - \log(K-1) \\
\alpha_t &= \frac{K-1}{K}\left[\log\left(\frac{1-\text{err}_t}{\text{err}_t}\right) + \log(K-1)\right]
\end{aligned}
$$

4 Bias–Variance Decomposition for Bregman Divergences

(a) By strictly convexity, we have $\phi(p) > \phi(q) + \langle \nabla\phi(q), p-q \rangle$ for all $p \neq q$. Hence, suppose there exist $p \neq q$ such that $D_\phi(p||q) = 0$. From the strictly convexity, we have $D_\phi(p||q) = \phi(p) - \phi(q) - \langle \nabla\phi(q), p-q \rangle > 0$ is a contradiction. That is $D_\phi(p||q) = 0 \Rightarrow p = q$

Now suppose $p = q$, we have

$$
D_\phi(p||q) = \phi(p) - \phi(q) - \langle \nabla\phi(q), p-q \rangle = \phi(q) - \phi(q) - \langle \nabla\phi(q), q-q \rangle = 0
$$

Then $D_\phi(p||q) = 0 \Leftrightarrow p = q$

(b)

$$
\begin{aligned}
\text{LHS} &= \mathbb{E}[\phi(y) - \phi(\hat{Y}) - \nabla\phi(\hat{Y})(y - \hat{Y})] \\
&= \phi(y) - \mathbb{E}(\phi(\hat{Y})) - \mathbb{E}(\nabla\phi(\hat{Y})(y - \hat{Y}))
\end{aligned}
$$

Consider the part of RHS

$$
D_\phi(y||\bar{Y}) = \phi(y) - \phi(\bar{Y}) - \nabla\phi(\bar{Y})(y - \bar{Y})
$$

$$
\begin{aligned}
\mathbb{E}[D_\phi(\bar{Y}||\hat{Y})] &= \mathbb{E}[\phi(\bar{Y}) - \phi(\hat{Y}) - \nabla\phi(\hat{Y})(\bar{Y} - \hat{Y})] \\
&= \phi(\bar{Y}) - \mathbb{E}[\phi(\hat{Y})] - \mathbb{E}[\nabla\phi(\hat{Y})(\bar{Y} - \hat{Y})] \\
\text{RHS} &= \phi(y) - \nabla\phi(\bar{Y})(y - \bar{Y}) - \mathbb{E}[\phi(\hat{Y})] - \mathbb{E}[\nabla\phi(\hat{Y})(\bar{Y} - \hat{Y})]
\end{aligned}
$$

Then compute the equation LHS=RHS

$$
\phi(y) - \mathbb{E}(\phi(\hat{Y})) - \mathbb{E}(\nabla\phi(\hat{Y})(y - \hat{Y})) = \phi(y) - \nabla\phi(\bar{Y})(y - \bar{Y}) - \mathbb{E}[\phi(\hat{Y})] - \mathbb{E}[\nabla\phi(\hat{Y})(\bar{Y} - \hat{Y})]
$$

$$
\begin{aligned}
\mathbb{E}(\nabla\phi(\hat{Y})(y - \hat{Y})) &= \nabla\phi(\bar{Y})(y - \bar{Y}) + \mathbb{E}[\nabla\phi(\hat{Y})(\bar{Y} - \hat{Y})] \\
\nabla\phi(\bar{Y})(y - \bar{Y}) + \mathbb{E}[\nabla\phi(\hat{Y})(\bar{Y} - \hat{Y})] - \mathbb{E}(\nabla\phi(\hat{Y})(y - \hat{Y})) &= 0 \\
\nabla\phi(\bar{Y})(y - \bar{Y}) + \mathbb{E}[\nabla\phi(\hat{Y})((\bar{Y} - y))] &= 0 \\
(\nabla\phi(\bar{Y}) - \mathbb{E}[\nabla\phi(\hat{Y})])(y - \bar{Y}) &= 0
\end{aligned}
$$

Hence, we have

$$
\left\langle \nabla\phi(\bar{Y}) - \mathbb{E}[\nabla\phi(\hat{Y})], y - \bar{Y} \right\rangle = 0
$$

Since all of the above steps are equivalent, we can say that

$$
\mathbb{E}[D_\phi(y||\hat{Y})] = D_\phi(y||\bar{Y}) + \mathbb{E}[D_\phi(\bar{Y}||\hat{Y})] \Leftrightarrow \left\langle \nabla\phi(\bar{Y}) - \mathbb{E}[\nabla\phi(\hat{Y})], y - \bar{Y} \right\rangle = 0
$$

(c) We want find $\phi = f_1(t)$ such that $D_\phi(y||\hat{Y}) = \text{MSE}(\hat{Y}, y) = (\hat{Y} - y)^2$ Let $\phi = ct^2 + bt + a$ with $c > 0$ to ensure its strictly convexity, we have

$$\begin{aligned} D_\phi(y||\hat{Y}) &= (cy^2 + by + a) - (c\hat{Y}^2 + b\hat{Y} + a) - (2c\hat{Y} + b)(y - \hat{Y}) \\ &= c(y^2 - \hat{Y}^2) - 2cy\hat{Y} + 2c\hat{Y}^2 \\ &= c(y - \hat{Y})^2 \end{aligned}$$

So we can find that when $c = 1$, $D_\phi(y||\hat{Y}) = (\hat{Y} - y)^2$, here we choose $\phi(t) = f_1(t) = t^2$.

Since $D_\phi(p||q) = (p-q)^2$, we have the classical MSE bias-variance identity can be represented as

$$\mathbb{E}[D_\phi(y||\hat{Y})] = D_\phi(y||\bar{Y}) + \mathbb{E}[D_\phi(\bar{Y}||\hat{Y})]$$

That is, when $\phi(t) = f_1(t)$ satisfies $\left\langle \nabla\phi(\bar{Y}) - \mathbb{E}[\nabla\phi(\hat{Y})], y - \bar{Y} \right\rangle = 0$

Check if $\phi(t) = f_1(t)$ satisfies the condition

$$\left\langle \nabla\phi(\bar{Y}) - \mathbb{E}[\nabla\phi(\hat{Y})], y - \bar{Y} \right\rangle = (2\bar{Y} - 2\bar{Y})(y - \bar{Y}) = 0$$

(d) Observe that we need $\phi'(t)$ contain $\log t$ so that we can have $\log(\frac{p_i}{q_i})$. Suppose $\phi(\mathbf{t}) = \mathbf{t}\log\mathbf{t} - \mathbf{t}$

$$\begin{aligned} D_{\phi_i}(p_i||q_i) &= p_i \log p_i - p_i - q_i \log q_i + q_i - (\log q_i)(p_i - q_i) \\ &= p_i \log p_i - p_i + q_i - p_i \log q_i \\ &= p_i \log(\frac{p_i}{q_i}) - p_i + q_i \end{aligned}$$

Choose $f_2(\mathbf{t}) = \sum_{i=1}^d (t_i \log t_i - t_i)$ then we have $D_\phi(p||q) = \sum_{i=1}^d (p_i \log(\frac{p_i}{q_i}) - p_i + q_i)$ If $\sum_{i=1}^d p_i = \sum_{i=1}^d q_i = 1$, the equation will be

$$\begin{aligned} D_\phi(p||q) &= \sum_{i=1}^d (p_i \log(\frac{p_i}{q_i}) - p_i + q_i) \\ &= \sum_{i=1}^d (p_i \log(\frac{p_i}{q_i})) - \sum_{i=1}^d p_i + \sum_{i=1}^d q_i \\ &= \sum_{i=1}^d (p_i \log(\frac{p_i}{q_i})) \end{aligned}$$

5 Gradient Descent Convergence

(a) $g(1) = f(\mathbf{y}), g(0) = f(\mathbf{x})$, from FToC we have

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

From chain rule, $g'(t)$ can be computed by

$$\begin{aligned} g'(t) &= \frac{d}{d\mathbf{z}(t)} g(\mathbf{z}(t)) \frac{d\mathbf{z}(t)}{dt} \\ &= \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^T (\mathbf{y} - \mathbf{x}) \end{aligned} \tag{2}$$

From (1) and (2) we have

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

Since $\nabla f(\mathbf{x})(\mathbf{y} - \mathbf{x})$ is independent on t, we have

$$\nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) = \int_0^1 \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x})\, dt$$

Then

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

By Cauchy-Schwarz inequality,

$$\left|(\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}))^T (\mathbf{y} - \mathbf{x})\right| \leq |\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})|_2 \cdot |\mathbf{y} - \mathbf{x}|_2$$

Let $\mathbf{v} = \mathbf{x} + t(\mathbf{y} - \mathbf{x}), \mathbf{u} = \mathbf{x}$, by β-smoothness definition

$$|\nabla f(\mathbf{v}) - \nabla f(\mathbf{u})|_2 \cdot |\mathbf{y} - \mathbf{x}|_2 \leq \beta t \left\|\mathbf{y} - \mathbf{x}\right\|_2$$

$$\begin{aligned} |f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x})| &\leq \int_0^1 \beta t \left\|\mathbf{y} - \mathbf{x}\right\|_2 \, dt \\ &= \frac{\beta}{2} \left\|\mathbf{y} - \mathbf{x}\right\|_2^2 \end{aligned}$$

Hence we get

$$f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) \leq \frac{\beta}{2} \left\|\mathbf{y} - \mathbf{x}\right\|_2^2$$

(b) Let $\mathbf{y} = \mathbf{x} - \frac{1}{\beta} \nabla f(\mathbf{x})$, the (a) inequality will be

$$f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T \left(-\frac{1}{\beta} \nabla f(\mathbf{x})\right) \leq \frac{\beta}{2} \left(\frac{1}{\beta^2} \left\|\nabla f(\mathbf{x})\right\|_2^2\right)$$

$$\nabla f(\mathbf{x})^T \left(-\frac{1}{\beta} \nabla f(\mathbf{x})\right) = -\frac{1}{\beta} \left\|\nabla f(\mathbf{x})\right\|_2^2$$

$$f(\mathbf{y}) - f(\mathbf{x}) + \frac{1}{\beta} \left\|\nabla f(\mathbf{x})\right\|_2^2 \leq \left(\frac{1}{2\beta} \left\|\nabla f(\mathbf{x})\right\|_2^2\right)$$

$$f(\mathbf{x} - \frac{1}{\beta} \nabla f(\mathbf{x})) - f(\mathbf{x}) \leq -\frac{1}{2\beta} \left\|\nabla f(\mathbf{x})\right\|_2^2$$

Since $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$, we have $f(\mathbf{x}^*) - f(\mathbf{x}) \leq f(\mathbf{y}) - f(\mathbf{x})$. Hence,

$$f(\mathbf{x}^*) - f(\mathbf{x}) \leq -\frac{1}{2\beta} \left\|\nabla f(\mathbf{x})\right\|_2^2$$

(c) Prove the equality:

$$\begin{aligned}\|\theta_{n+1}-\theta^*\|_2^2 &= \|(\theta_n-\eta\nabla f(\theta_n))-\theta^*\|_2^2 \\ &= \|\theta_n-\theta^*\|_2^2+\|\eta\nabla f(\theta_n)\|_2^2-2(\theta_n-\theta^*)^T\eta\nabla f(\theta_n) \\ &= \|\theta_n-\theta^*\|_2^2+\eta^2\|\nabla f(\theta_n)\|_2^2-2\eta(\nabla f(\theta_n))^T(\theta_n-\theta^*)\end{aligned}$$

(d) Now prove the inequality, first by α strongly convexity,

$$\frac{1}{2\beta}\|\nabla f(\theta_n)\|_2^2 \le f(\theta_n)-f(\theta^*) \le \nabla f(\theta_n)^T(\theta_n-\theta^*)-\frac{\alpha}{2}\|\theta^*-\theta_n\|_2^2$$

$$\frac{1}{2\beta}\|\nabla f(\theta_n)\|_2^2+\frac{\alpha}{2}\|\theta^*-\theta_n\|_2^2 \le \nabla f(\theta_n)^T(\theta_n-\theta^*)$$

$$-((\frac{1}{\beta})^2\|\nabla f(\theta_n)\|_2^2+\frac{\alpha}{\beta}\|\theta^*-\theta_n\|_2^2) \ge -\frac{2}{\beta}\nabla f(\theta_n)^T(\theta_n-\theta^*)$$

$$\begin{aligned}\|\theta_{n+1}-\theta^*\|_2^2 &= \|\theta_n-\theta^*\|_2^2+(\frac{1}{\beta})^2\|\nabla f(\theta_n)\|_2^2-2(\frac{1}{\beta})(\nabla f(\theta_n))^T(\theta_n-\theta^*) \\ &\le \|\theta_n-\theta^*\|_2^2+(\frac{1}{\beta})^2\|\nabla f(\theta_n)\|_2^2-((\frac{1}{\beta})^2\|\nabla f(\theta_n)\|_2^2+\frac{\alpha}{\beta}\|\theta^*-\theta_n\|_2^2) \\ &= \|\theta_n-\theta^*\|_2^2-\frac{\alpha}{\beta}\|\theta^*-\theta_n\|_2^2\end{aligned}$$

Hence, we have

$$\|\theta_{n+1}-\theta^*\|_2^2 \le (1-\frac{\alpha}{\beta})\|\theta_n-\theta^*\|_2^2$$

(e) We have $\|\theta_{n+1}-\theta^*\|_2^2 \ge 0$ and the inequality of (d) gives us $\|\theta_n-\theta^*\|_2^2$ is strictly decreasing. By the Monotone convergence theorem

$$\lim_{n\to\infty}\|\theta_n-\theta^*\|_2^2 = 0$$

Hence, θ_n will converge to θ^*.