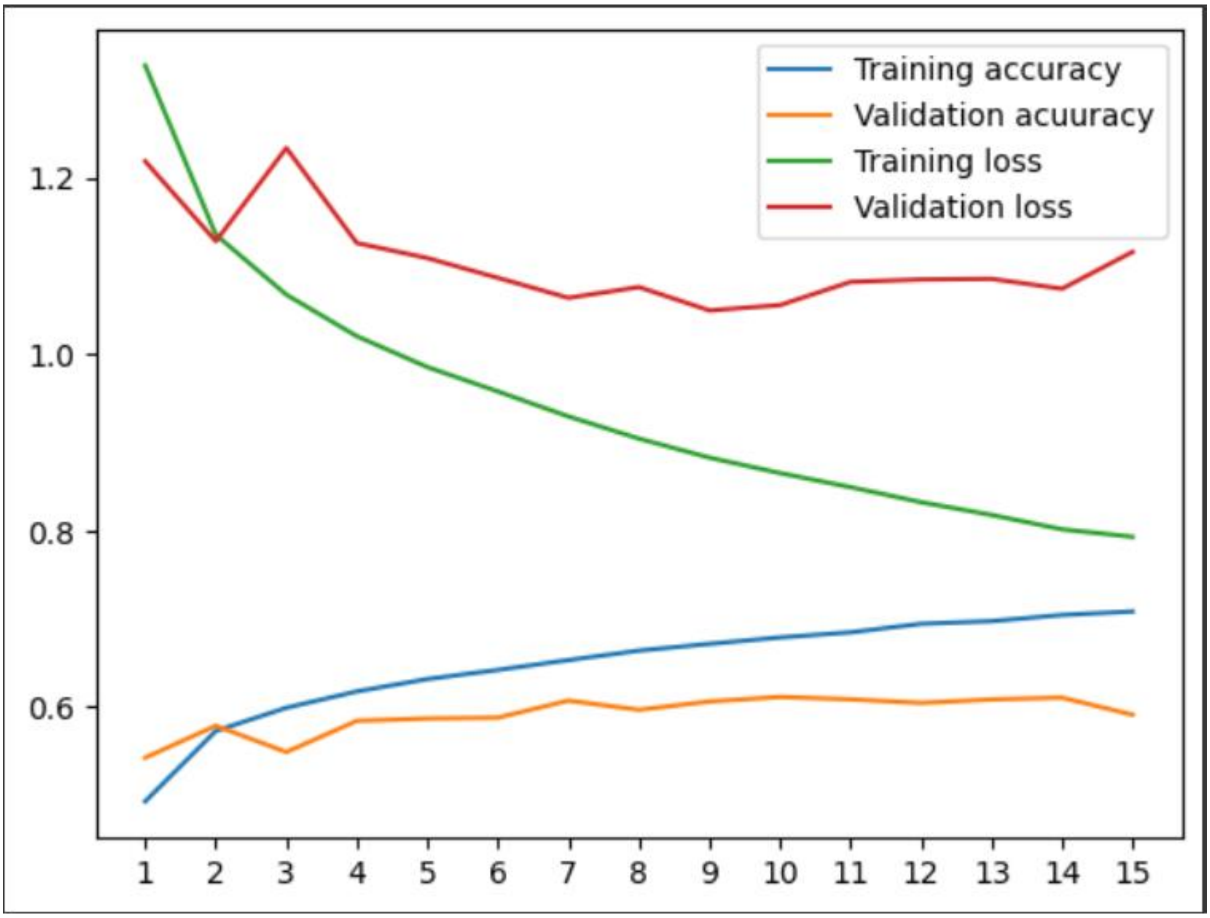
學號：B10705025 系級： 資管三 姓名：彭鈞道

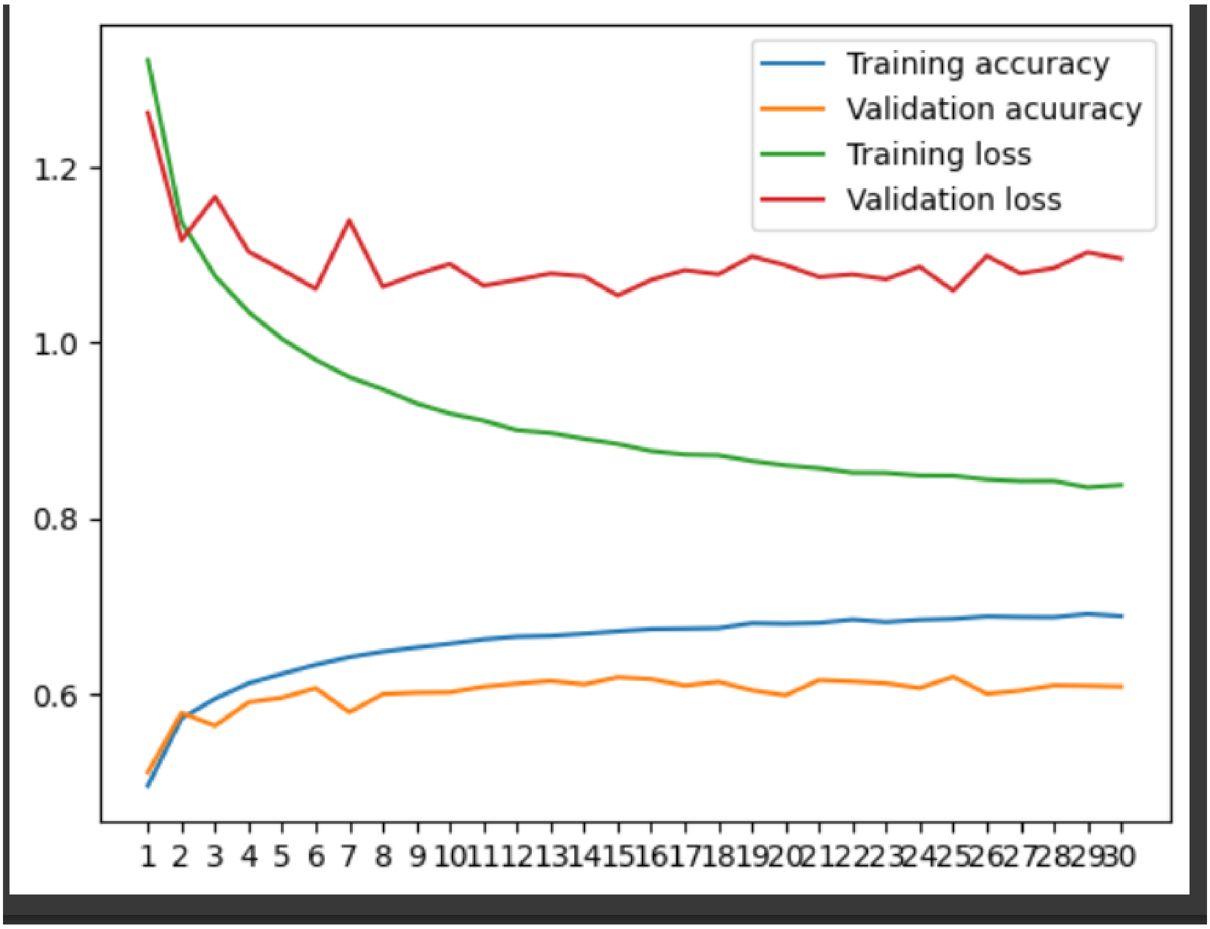
在訓練資料一開始因為助教提供的file路徑有問題，所以我把kaggle上提供的資料下載下來，放到google雲端上在gdown到colab裡，連結如下https://drive.google.com/file/d/1jdh7w55GmpsGzw1y5Z7L0ylg1TACxjlt/view?usp=sharing ，資料跟原本提供的一模一樣。訓練好的模型連結: https://drive.google.com/file/d/1K60k1Lj2Y06ODp7ueK4T8ErvJIr3zX79/view?usp=sharing

1. **(1%) 實作early-stopping，繪製training, validation loss/acc 的 learning curve，比較實作前後的差異，並說明early-stopping的運作機制**

我的early-stopping的運作方式很簡單，只要valid_acc在一定程度內沒有下降，就停止training(5次epoch)



可以發現雖然training loss不斷下降，training_accuracy不斷上升，但valid loss ep validation accuracy在特定epoch術後便停止變好，所以實作early stop.
下面是沒有early stopping的，可以發現validation loss沒有變好，後期甚至上升(為了節省時間，只跑30次，另外跑的時候不小心調到參數，但概念相似)。

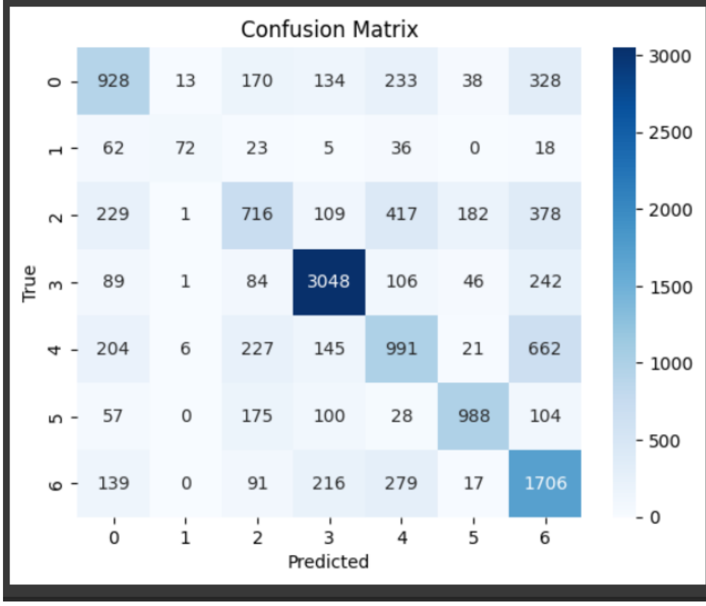


2. (1%) 嘗試使用 **augmentation**, 說明實作細節並比較有無該 **trick** 對結果表現的影響(**validation** 或是 **testing** 擇一即可), 且需說明為何使用這些**augmentation**的原因。
(ref: https://pytorch.org/vision/stable/transforms.html)

使用3種augmentation，分別是水平旋轉，隨機旋轉(-20~+20)，RandomPerspective (distortion_scale=0.3, p=1)，實作的方式是先定義好這三種方式，把train_set、valid_set、test_set都擴充為四倍長度，再用index來判斷，如果index<總長度的1/4，保持原來圖片，如果index>=總長度的1/4但<總長度的2/4，對圖片進行第一種augmentation，以此類推。選用這四種是因為觀察訓練資料集可發現圖片不是保持在正中央，會偏左偏右或有一定程度上的偏差，我希望模型可以處理偏左偏右，角度奇怪的，所以選擇這四種方式。test_set也使用相同的方式augment，但在輸出的時候會有四種預測，所以選擇取四種預測的眾數來當作最後的output。在public test score也發現表現較好。

3. (1%) 畫出 **confusion matrix** 分析哪些類別的圖片容易使 **model** 搞混，找出模型出錯的例子，並分析可能的原因。
(**ref:** https://en.wikipedia.org/wiki/Confusion_matrix)

```
from sklearn.metrics import confusion_matrix
import seaborn as sns
cm = confusion_matrix(true_result, model_result)
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=[0,1,2,3,4,5,6], yticklabels=[0,1,2,3,4,5,6])
plt.xlabel('Predicted')
plt.ylabel('True')
plt.title('Confusion Matrix')
plt.show()
```



模型容易在1號表情出錯，推測是因為在訓練資料集中該類別占比偏少，模型不熟悉，所以無法正確判讀(confusion matrix based on valid_set)

4. (1%) 請統計訓練資料中不同類別的數量比例，並說明：
對 **testing** 或是 **validation** 來說，不針對特定類別，直接選擇機率最大的類別會是最好的結果嗎？
(**ref:** https://arxiv.org/pdf/1608.06048.pdf, **or hints: imbalanced classification**)

```
k = pd.read_csv(LABEL_PATH)['label'].values
df = pd.DataFrame(k, columns = ['label'])
df['label'].value_counts()
```

```
3    7275
6    4955
4    4841
2    4136
0    4009
5    3221
1     450
Name: label, dtype: int64
```

由統計資料可知，最有可能出現的是3號情緒，對validation來說，如果直接選擇3號，正確率應為0.25(validation set，是從training set中隨機選出，分布應類似)，並不會有較好的accuracy.

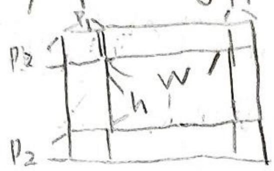
5. (6%)Refer to math problem

Math1

hw 2

1. first Batch does not change, K_1 represents the width of the kernel and K_2 means the height of the kernal

second, by padding, the data look like



the new width would be

$$(w + 2P_1 - K_1)/s_1 + 1$$

and the new height would be

$$(h + 2P_2 - K_2)/s_2 + 1$$

So the new shape will be

$$\left(B, \left(\frac{w+2P_1-K_1}{s_1}+1\right), \frac{h+2P_2-K_2}{s_2}+1, \text{output\_channels}\right)$$

Math2

2.
$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i} \times \frac{\partial y_i}{\partial \gamma}$$

$$= \sum_{i=1}^{m} \frac{\partial l}{\partial y_i} \times \frac{\partial \gamma \hat{x_i} + \beta}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i} \cdot \hat{x_i}$$

γ plays in every y_i

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i} \times \frac{\partial \gamma \hat{x_i} + \beta}{\partial \beta} = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i}$$

$$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x_i}} \cdot \frac{\partial \hat{x_i}}{\partial x_i} + \frac{\partial l}{\partial \sigma_\beta^2} \cdot \frac{\partial \sigma^2 \beta}{\partial x_i} + \frac{\partial l}{\partial \mu_\beta} \cdot \frac{\partial \mu_\beta}{\partial x_i}$$

$$= \frac{\partial l}{\partial \hat{x_i}} \cdot \frac{1}{\sqrt{\sigma_\beta^2 + \varepsilon}} + \frac{\partial l}{\partial \sigma_\beta^2} \cdot \frac{2(x_i - \mu_B)}{m} + \frac{\partial l}{\partial \mu_B} \cdot \frac{1}{m}$$

hw 2

2. $\frac{\partial l}{\partial \hat{x}_i} = \frac{\partial L}{\partial y_i} \cdot \frac{\partial y_i}{\partial \hat{x}_i} = \frac{\partial L}{\partial y_i} \cdot \gamma$

$$\frac{\partial L}{\partial \sigma_\beta^2} = \sum_{i=1}^{m} \frac{\partial L}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \sigma_\beta^2}$$

$$= \sum_{i=1}^{m} \frac{\partial L}{\partial y_i} \cdot \gamma \cdot (x_i - \mu_B) \frac{-1}{2} (\sigma_\beta^2 + \varepsilon)^{-\frac{3}{2}}$$

$$\frac{\partial L}{\partial \mu_B} = \sum_{i=1}^{m} \frac{\partial L}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial \mu_B} + \frac{\partial L}{\partial \sigma_\beta^2} \cdot \frac{\partial \sigma_\beta^2}{\partial \mu_B}$$

$$= \sum_{i=1}^{m} \frac{\partial L}{\partial y_i} \cdot \gamma \cdot \frac{-1}{\sqrt{\sigma_\beta^2 + \varepsilon}} + \sum_{i=1}^{m} \frac{\partial L}{\partial y_i} \cdot \gamma \cdot (x_i - \mu_B) \cdot \frac{-1}{2} \cdot (\sigma_\beta^2 + \varepsilon)^{-\frac{3}{2}} \cdot \frac{\sum_{i=1}^{m} -2(x_i - \mu_B)}{m}.$$

So $\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial y_i} \cdot \gamma \cdot \frac{1}{\sqrt{\sigma_\beta^2 + \varepsilon}}$

$$+ \sum_{i=1}^{m} \frac{\partial L}{\partial y_i} \cdot \gamma \cdot (x_i - \mu_B) \frac{-1}{2} (\sigma_\beta^2 + \varepsilon)^{-\frac{3}{2}} \cdot \frac{2(x_i - \mu_B)}{m}$$

$$+ \frac{1}{m} \cdot \left(\sum_{i=1}^{m} \frac{\partial L}{\partial y_i} \cdot \gamma \cdot \frac{-1}{\sqrt{\sigma_\beta^2 + \varepsilon}} + \sum_{i=1}^{m} \frac{\partial L}{\partial y_i} \cdot \gamma \cdot (x_i - \mu_B) \cdot \frac{-1}{2} \cdot (\sigma_\beta^2 + \varepsilon)^{-\frac{3}{2}} \cdot \frac{\sum_{i=1}^{m} -2(x_i - \mu_B)}{m} \right)$$

Math3.1

hw 2.

3.1 first construct a set of points $B = b_1, b_2 \dots b_n \in \mathbb{R}^m$ with mean = zero vector, covariance matrix $= I = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$, this is quite simple, for every dimension, construct data points with mea=0 and variance = 1.

We let $x_i = A \cdot b_i + \mu$, where μ is the mean

$\frac{1}{n}\sum_{i=1}^{n} x_i = \mathbb{E}[x_i] = \mu + A \cdot \mathbb{E}[b_i]$, since

$\mathbb{E}[b_i] = 0$ (mean = zero vector), $\frac{1}{n}\sum_{i=1}^{n} x_i = \mu$ #

$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$

$= \frac{1}{n}\sum_{i=1}^{n}(A \cdot b_i + \mu - \mu)(A \cdot b_i + \mu - \mu)^T$

$= \frac{1}{n}\sum_{i=1}^{n} A \cdot b_i \cdot b_i^T A^T = A \cdot \mathbb{E}[b_i \cdot b_i^T] \cdot A^T$

$\mathbb{E}[b_i \cdot b_i^T] = \mathbb{E}\begin{bmatrix} b_{i1} b_{i1} & b_{i1} \cdot b_{i2} & \cdots & b_{i1} b_{in} \\ \vdots & & & \vdots \\ b_{in} b_{i1} & \cdots & \cdots & b_{in} b_{in} \end{bmatrix} = I_n$

since b_{ij} has $\mathbb{E}[b_{ij}] = 0$, $var(b_{ij}) = 1$

$cov(b_{ij}, b_{ik}) = 0$, $\mathbb{E}[b_{ij} b_{ik}] = 1$ $i = j$ $(E[X^2] = \mathbb{E}[X]^2 + var(X))$

$\mathbb{E}[b_{ij} b_{ik}] = 0$ $i \neq j$ $(E[XY] = \mathbb{E}[X]E[Y] + 2\,cov(X,Y))$

$= A \cdot A^T = \Sigma$, to let $A \cdot A^T = \Sigma$, prove Σ can be decomposed like this, Σ is a covariance matrix, $\Sigma \in S_n^+$

$\Sigma = PDP^T$ where D is the diagonal matrix with eigenvalues of Σ, let $A = PD^{\frac{1}{2}}$, $A \cdot A^T = PD^{\frac{1}{2}} \cdot (D^{\frac{1}{2}})^T \cdot P^T$, since D is diagonal, $(D^{\frac{1}{2}})^T = D^{\frac{1}{2}}$, $A \cdot A^T = PDP^T = \Sigma$

the relation is $\boxed{x_i = A \cdot b_i + \mu}$, where $A = PD^{\frac{1}{2}}$, b_i is one of the date with zero mean and covariance $= I_n$

hw2 3.2

wlog, take $X = [x_1, x_2 \dots x_N] \in R^{m \times N}$, s.t

Σ is X's covariance matrix, and $\frac{1}{N}\sum_{i=1}^{N} x_i = 0$,

we have $\Sigma = \frac{1}{n} X X^T$

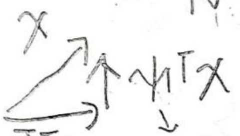
$$trace(\Phi^T \Sigma \Phi) = \frac{1}{n} trace(\Phi^T X X^T \Phi)$$

$$= \frac{1}{N} \|\Phi^T X\|_F^2$$ recall $\|A\|_F^2 = trace(AA^T)$

$$\Phi = [\Phi_1, \dots \Phi_K]$$

to minimize $\frac{1}{N}\|\Phi^T X\|_F^2$ means to maximize

$$\frac{1}{N}\|X - \Phi^T X\|_F^2 = \frac{1}{N}\|\Psi^T X\|_F^2$$



$\Psi = [\psi_1, \dots \psi_{m-K}]$

$\Phi^T x \to$ orthogonal, $x = \psi^T x + \Phi^T x$

recall PCA, we have optimal solutions

for $\Psi = [u_1, u_2, \dots u_{m-K}]$

where u_i is an eigenvector of Σ with eigenvalue λ_i (λ_i in descending order)

$$x = \Psi^T x + \Phi^T x$$

$$\begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_m^T \end{bmatrix} x = \begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_{m-K}^T \end{bmatrix} x + \Phi^T x$$

$$\Rightarrow \Phi^T = \begin{bmatrix} u_{m-K+1}^T \\ \vdots \\ u_m^T \end{bmatrix}, \Phi = [u_{m-K+1}, \dots u_m]$$

#

Math4.1 4.2

hw2 4.1

we now try to minimize

$f(s)=\sum_{i=1}^{m} ||z_i - s||_2$ where s is the point, since L_2-norm is a convex function, the sum of it is also a convex function, so minimum occurs at $\nabla f(s)=0$

$$\nabla_s \sum_{i=1}^{m} (z_i-s)^T(z_i-s)$$

$$=\nabla_s \sum_{i=1}^{m} z_i^T z_i - 2s^T z_i + s^T s$$

$$=\sum_{i=1}^{m} -2z_i + 2s = 0$$

$$\sum_{i=1}^{m} z_i = ms$$

$$s = \frac{\sum_{i=1}^{m} z_i}{m} = \hat{z}$$, so minimum occurs at $\hat{z}$

$\sum_{i=1}^{m} ||z_i - z||_2^2 \geq \sum_{i=1}^{m} ||z_i - \hat{z}||_2^2$ is true

4.2 $L(c^{t+1}, \mu^t)$

$$=\sum_{i=1}^{n} ||x_i - \mu^t_{c^{t+1}(i)}||_2^2$$

$$=\sum_{i=1}^{n} \operatorname{argmin}_{j=1,\dots,k} ||x_i - \mu_j^t||_2^2$$

the corresponding $j = c^t(i)$

$L(c^t, \mu^t)$

$$=\sum_{i=1}^{n} ||x_i - \mu^t_j||_2^2, \quad j=1,\dots,k$$

$\because \operatorname{argmin}_{j=1,\dots,k} ||x_i - \mu^t_j||_2^2 \leq ||x_i - \mu_j^t||_2^2$ $\forall\ 1\leq i\leq n,\ 1\leq j\leq k$

$\Rightarrow \sum_{i=1}^{n} \operatorname{argmin}_{j=1,\dots,k} ||x_i - \mu^t_j||_2^2 \leq \sum_{i=1}^{n} ||x_i - \mu^t_j||_2^2$

$\Rightarrow L(c^{t+1}, \mu^t) \leq L(c^{t+1}, \mu^t)$

1

Math 4.3

hw2 4-3 $L(c^{t+1}, \mu^{t+1})$

$$= \sum_{i=1}^{n} \|x_i - \mu^{t+1}_{c^{t+1}(i)}\|_2^2 = \sum_{q=1}^{k} \sum_{i: c^{t+1}(i)=q} \|x_i - \mu^{t+1}_q\|_2^2$$

$L(c^{t+1}, \mu^t)$

$$= \sum_{i=1}^{n} \| x_i - \mu^t_{c^{t+1}(i)} = \sum_{q=1}^{k} \sum_{i: c^{t+1}(i)=q} \| x_i - \mu^t_q\|_2^2$$

μ^{t+1}_q is the mean of points who label q in c^{t+1}, μ^t_q is the mean of points who label q in c^t

since for each q, $\sum_{i: c^{t+1}(i)=q} \|x_i - z\|_2^2$ has minimum at z equals mean of x_i (by (a)), i.e. the mean of points label q in c^{t+1}, so $z = \mu^{t+1}_q$

since $\sum_{i: c^{t+1}(i)=q} \|x_i - \mu^{t+1}_q\|_2^2$ has minimum

$$\Rightarrow \sum_{i: c^{t+1}(i)=q} \|x_i - \mu^{t+1}_q\|_2^2 \geq \sum_{i: c^{t+1}(i)=q} \|x_i - \mu^t_q\|_2^2$$

$$\Rightarrow \sum_{q=1}^{k} \sum_{i: c^{t+1}(i)=q} \|x_i - \mu^{t+1}_q\|_2^2 \geq \sum_{q=1}^{k} \sum_{i: c^{t+1}(i)=q} \| x_i - \mu^t_q\|_2^2$$

$$\Rightarrow L(c^{t+1}, \mu^{t+1}) \geq L(c^{t+1}, \mu^t)$$

Math 4.4 4.5

hw 2

4.4 by (b), (c)

$\Rightarrow L(C^t, \mu^t) \geq L(C^{t+1}, \mu^t) \geq L(C^{t+1}, \mu^{t+1})$

$\Rightarrow L(C^t, \mu^t) \geq L(C^{t+1}, \mu^{t+1})$

$\Rightarrow l_t \geq l_{t+1} \Rightarrow l_t$ is monotonic decreasing

$$l_t = \sum_{i=1}^{n} \|x_i - \mu_{c(i)}\|_2^2 \geq 0$$

$l_t \geq 0$

since $l_t \geq l_{t+1}$ and $l_t \geq 0$, by monotone divergence theorem, $\{l_t\}$ converges

4.5 We have n points and K classes, the sample space is K^n, since $l_t \geq l_{t+1}$, every time we update, the loss gets small, we will update less than K^n steps, if we calculate more than K^n steps, which shows we must have got the loss before, and it should be the lowest in the K^n steps, so the algorithm stops.

Math 5.1

hw 2. 5-1 (a) $\int_0^1 g'(t)\,dt$

$$= g(t)\Big|_{t=0}^{t=1} = f(y+(x-y)) - f(y)$$
$$= f(x) - f(y)$$

5.1 (b) $g(t) = f(y + t(x-y))$

$$g'(t) = \frac{d}{dt} f(y+t(x-y))$$
$$= \frac{d}{dt} f(y+t(x-y)) \cdot \frac{d}{dt}(y+t(x-y))$$
$$= \nabla_t f(y + t(x-y))^T \cdot (x-y)$$
$$(\because \nabla f(x) = Df(x)^T)$$

5-1 (c) $|f(x) - f(y) - \nabla f(y)^T (x-y)|$

$$= \left|\int_0^1 g'(t)\,dt - \nabla f(y)^T (x-y)\right|$$
$$= \left|\int_0^1 (\nabla f(y+t(x-y)) - \nabla f(y))^T (x-y)\,dt\right|$$

$\left|\int_a^b f(x)\,dx\right| \le \int_a^b |f(x)|\,dx$ (if $f(x)$ are all negative or positive trivially, $\left|\int_a^b f(x)\,dx\right| = \int_a^b |f(x)|\,dx$; but if $f(x)$ change signs in (a,b), then in $\left|\int_a^b f(x)\,dx\right|$, the positive area will minus the negative part, then change to positive. $\int_a^b |f(x)|\,dx$ will change everything to positive and add up)

by applying it, $|f(x) - f(y) - \nabla f(y)^T (x-y)| = \left|\int_0^1 g'(t)\,dt - \nabla f(y)^T (x-y)\right|$

$$\le \int_0^1 \left|(\nabla f(y+t(x-y)) - \nabla f(y)^T)(x-y)\right|\,dt$$

Math 5.2

hw2

5-1

(d)

by (c), we know $|f(x)-f(y)-\nabla f(y)^T(x-y)|$

$\leq \int_0^1 |(\nabla f(y+t(x-y)) - \nabla f(y))^T(x-y)| dt$

apply cauchy-schwarts : $|u^Tv| \leq ||u||_2 \cdot ||v||_2$

$\leq \int_0^1 ||\nabla f(y+t(x-y)) - \nabla f(y)||_2 \cdot ||x-y||_2 dt$

$= ||x-y||_2 \cdot \int_0^1 ||\nabla f(y+t(x-y)) - \nabla f(y)||_2 dt$

$\because ||\nabla f(x) - \nabla f(y)||_2 \leq \beta ||x-y||_2$

$\therefore ||\nabla f(y+t(x-y)) - \nabla f(y)||_2 \leq \beta ||t(x-y)||_2$

$\leq ||x-y||_2 \cdot \int_0^1 \beta \, ||x-y||_2 \cdot t$

$= \beta \cdot ||x-y||_2^2 \cdot \int_0^1 t \, dt$

$= \frac{\beta}{2} \cdot ||x-y||_2^2$

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

5-2 $\forall x, y \in R^n$

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

, x and y are interchangeable

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

$y = x - \frac{1}{\beta}\nabla f(x)$

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

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

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

---

hw2 5-2 $f(x^*) = \text{argmin} f(x)$

$\Rightarrow f(x^*) \leq f(x - \frac{1}{\beta}\nabla f(x))$

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

Math 5.3 5.4

A 5-3 $\|\theta^{n+1}-\theta^*\|_2^2$

$= \|\theta^n - \eta\nabla_\theta f(\theta^n) - \theta^*\|_2^2$

$= ((\theta^n-\theta^*) - \eta\nabla_\theta f(\theta^n))^T((\theta^n-\theta^*)-\eta\nabla_\theta f(\theta^n))$

$= (\theta^n-\theta^*)^T(\theta^n-\theta^*) + \eta^2 \nabla_\theta f(\theta^n)^T \cdot \eta\nabla_\theta f(\theta^n)$

$-2\cdot(\theta^n-\theta^*)\cdot\eta\nabla_\theta f(\theta^n)$

dot

$= \|\theta^n-\theta^*\|_2^2+\eta^2\|\nabla_\theta f(\theta^n)\|_2^2 - 2\eta(\nabla_\theta f(\theta^n))^T(\theta^n-\theta^*)$

5.4 by 5-3, let $y = x - \frac{1}{\beta}\nabla f(x)$, $x^*=\theta$, $x=\theta^n$, $y=\theta^{n+1}$

$\|y-x^*\|_2^2 = \|x-x^*\|_2^2 + \frac{1}{\beta^2}\|\nabla f(x)\|_2^2$

$-2\cdot\frac{1}{\beta}\nabla f(x)^T(x-x^*)$

by α-strongly convex

$f(x^*)-f(x)-\nabla f(x)^T(x^*-x)\ge \frac{\alpha}{2}\|x^*-x\|_2^2$

$-\nabla f(x)^T(x-x^*)\le f(x^*)-f(x)-\frac{\alpha}{2}\|x^*-x\|_2^2$

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

已知 $f(x^*)-f(x)\le -\frac{1}{2\beta}\|\nabla f(x)\|_2$

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

Math 5.5

hw 2 5.5

$$\|\theta^0-\theta^*\|_2^2 \cdot \left(1-\frac{\alpha}{\beta}\right) \geq \|\theta^1-\theta^*\|$$

$$\|\theta^0-\theta^*\|_2^2 \cdot \left(1-\frac{\alpha}{\beta}\right)^2 \geq \|\theta^2-\theta^*\|$$

because right term ≥ 0, left term ≥ 0

$$\Rightarrow \left(1-\frac{\alpha}{\beta}\right) < 1 ,$$

$$\lim_{n\to\infty} \|\theta^n-\theta^*\| \leq \left(1-\frac{\alpha}{\beta}\right)^n \cdot \|\theta^0-\theta^*\|_2^2$$

$$= 0 \cdot \|\theta^0-\theta^*\|_2^2 = 0$$

$\Rightarrow \theta^n$ will converge to θ^*