# Assignment 3: Proof

## Advanced Topics In Machine Learning (CS6360)

Rahul Vigneswaran

CS23MTECH02002

**Paper presented as part of assignment 1**

Neural Collapse Terminus: A Unified Solution for Class Incremental Learning and Its Variants

We consider the following problem,

$$\min_{\mathbf{M}^{(t)}} \quad \frac{1}{N^{(t)}} \sum_{k=1}^{K^{(t)}} \sum_{i=1}^{n_k} \mathcal{L}\left(\mathbf{m}_{k,i}^{(t)}, \hat{\mathbf{W}}_{\mathrm{ETF}}\right), \; 0 \leq t \leq T,$$

$$s.t. \quad \|\mathbf{m}_{k,i}^{(t)}\|^2 \leq 1, \quad \forall 1 \leq k \leq K^{(t)}, \; 1 \leq i \leq n_k,$$

where,

$\mathbf{m}_{k,i}^{(t)} \in \mathbb{R}^d$ : $i$-th sample of class $k$ in session $t$ feature

$n_k$ : no. of samples in class $k$

$K^{(t)}$ : no. of classes in session $t$

$N^{(t)} = \sum_{k=1}^{K^{(t)}} n_k$

$\mathbf{M}^{(t)} \in \mathbb{R}^{d \times N^{(t)}}$ : collection of $\mathbf{m}_{k,i}^{(t)}$

$K = \sum_{t=0}^{T} K^{(t)}$

$\hat{\mathbf{W}}_{\mathrm{ETF}} \in \mathbb{R}^{d \times K}$ : neural collapse terminus all $K$

A simplex equiangular tight frame (ETF) refers to a collection of vectors $\{\mathbf{e}_i\}_{i=1}^{K}$ in $\mathbb{R}^d$, $d \geq K - 1$, that satisfies:

$$\mathbf{e}_{k_1}^T \mathbf{e}_{k_2} = \frac{K}{K-1}\delta_{k_1,k_2} - \frac{1}{K-1}, \qquad \text{DEF1}$$

$$\forall k_1, k_2 \in [1, K],$$

where $\delta_{k_1,k_2} = 1$ when $k_1 = k_2$, and 0 otherwise. All vectors have the same $\ell_2$ norm and any pair of two different vectors has the same inner product of $-\frac{1}{K-1}$, which is the minimum possible cosine similarity for $K$ equiangular vectors in $\mathbb{R}^d$.

## Theorem

Let $\hat{\mathbf{M}}^{(t)}$ denotes the global minimizer by optimizing the model incrementally from $t = 0$, and we have $\hat{\mathbf{M}} = [\hat{\mathbf{M}}^{(0)}, \cdots, \hat{\mathbf{M}}^{(T)}] \in \mathbb{R}^{d \times \sum_{t=0}^{T} N^{(t)}}$. No matter if $\mathcal{L}$ is CE or misalignment loss, for any column vector $\hat{\mathbf{m}}_{k,i}$ in $\hat{\mathbf{M}}$ whose class label is $k$, we have:

$$\|\hat{\mathbf{m}}_{k,i}\| = 1, \ \ \hat{\mathbf{m}}_{k,i}^T \hat{\mathbf{w}}_{k'} = \frac{K}{K-1}\delta_{k,k'} - \frac{1}{K-1},$$

for all $k, k' \in [1, K]$, $1 \leq i \leq n_k$, where $K = \sum_{t=0}^{T} K^{(t)}$ denotes the total number of classes of the whole label space, $\delta_{k,k'} = 1$ when $k = k'$ and 0 otherwise, and $\hat{\mathbf{w}}_{k'}$ is the class prototype in $\hat{\mathbf{W}}_{\mathrm{ETF}}$ for class $k'$.

# KKT Conditions

1. Stationarity condition: $\nabla f(x^*) = -u^* \nabla g(x^*)$
2. Complimentary slackness: $u^* \; g(x) = 0$
3. Primal feasibility: $g(x^*) \leq 0$
4. Dual feasibility: $u^* \geq 0$

$$L(\mathbf{m}_{k,i}^{(t)}, \hat{\mathbf{W}}_{\text{ETF}}) = -\log \frac{\exp(\hat{\mathbf{w}}_k^T \mathbf{m}_{k,i})}{\sum_{j=1}^{K} \exp(\hat{\mathbf{w}}_j^T \mathbf{m}_{k,i})}$$

Lagrangian function,

$$\tilde{L} = \frac{1}{N^{(t)}} \sum_{k=1}^{K^{(t)}} \sum_{i=1}^{n_k} -\log \frac{\exp(\hat{\mathbf{w}}_k^T \mathbf{m}_{k,i})}{\sum_{j=1}^{K} \exp(\hat{\mathbf{w}}_j^T \mathbf{m}_{k,i})} + \sum_{k=1}^{K^{(t)}} \sum_{i=1}^{n_k} \lambda_{k,i} \left( ||\mathbf{m}_{k,i}||^2 - 1 \right)$$

Now lets do $\frac{\partial \tilde{L}}{\partial \mathbf{m}_{k,i}}$,

$$\frac{\partial \tilde{L}}{\partial \mathbf{m}_{k,i}} = -\frac{(1-p_k)}{N^{(t)}}\hat{\mathbf{w}}_k + \frac{1}{N^{(t)}}\sum_{j \neq k}^{K} p_j \hat{\mathbf{w}}_j + 2\lambda_{k,i}\mathbf{m}_{k,i}$$

where $1 \leq i \leq n_k, 1 \leq k \leq K^{(t)}$.

$$p_j = \frac{\exp(\hat{\mathbf{w}}_j^T \hat{\mathbf{m}}_{k,i})}{\sum_{j'=1}^{K} \exp(\hat{\mathbf{w}}_{j'}^T \mathbf{m}_{k,i})}$$

Now lets do $\frac{\partial \tilde{L}}{\partial \mathbf{m}_{k,i}} = 0$. According to to KKT's dual feasibility condition,

$$\lambda_{k,i} \geq 0$$

So we will try out both cases,

- Case 1 : $\lambda_{k,i} = 0$
- Case 2 : $\lambda_{k,i} > 0$

$$\sum_{j \neq k}^{K} p_j = (1 - p_k)$$

Also, from DEF1,

$$\hat{\mathbf{w}}_k^T \hat{\mathbf{w}}_{k'} = \frac{K}{K-1} \delta_{k,k'} - \frac{1}{K-1},$$
$$\forall k, k^{'} \in [1, K],$$

# Case 1: $\lambda_{k,i} = 0$

$$\frac{\partial \tilde{L}}{\partial \mathbf{m}_{k,i}} = 0$$

$$-\frac{(1-p_k)}{N^{(t)}}\hat{\mathbf{w}}_k + \frac{1}{N^{(t)}}\sum_{j \neq k}^{K} p_j \hat{\mathbf{w}}_j + 2\lambda_{k,i}\mathbf{m}_{k,i} = 0$$

$$\sum_{j \neq k}^{K} p_j \hat{\mathbf{w}}_j = (1-p_k)\hat{\mathbf{w}}_k$$

# Case 1: $\lambda_{k,i} = 0$

$$\sum_{j \neq k}^{K} p_j \hat{\mathbf{w}}_j = (1 - p_k)\hat{\mathbf{w}}_k$$

From EQN1, EQN2 and multiply by $\hat{\mathbf{w}}_k$,

$$\begin{aligned}
\sum_{j \neq k}^{K} p_j \hat{\mathbf{w}}_j^T \hat{\mathbf{w}}_k &= (1 - p_k)\hat{\mathbf{w}}_k \hat{\mathbf{w}}_k \\
(1 - p_k)\left(\frac{-K}{(K-1)}\right) &= (1 - p_k)x1 \\
-(1 - p_k)\frac{K}{(K-1)} &= 0 \\
p_k &= 1
\end{aligned}$$

But we already know that $0 < p_k < 1$. $p_k = 1$, only if all other $p = 1$.
That can never be the case because $\|\hat{\mathbf{w}}\| = 1$, $\|\hat{\mathbf{m}}\| < 1$

# Case 1: $\lambda_{k,i} \geq 0$

Now based on KKT's complimentary slackness condition,

$$\begin{aligned} \lambda_{k,i}(||\hat{\mathbf{m}}_{k,i}||^2 - 1) = 0 \\ ||\hat{\mathbf{m}}_{k,i}||^2 - 1 = 0 \\ ||\hat{\mathbf{m}}_{k,i}||^2 = 1 \end{aligned}$$

Now lets do $\frac{\partial \tilde{L}}{\partial \hat{\mathbf{m}}_{k,i}} = 0$

$$\sum_{j \neq k}^{K} p_j(\hat{\mathbf{w}}_j - \hat{\mathbf{w}}_k) + 2N^{(t)}\lambda_{k,i} = 0 \qquad \text{EQN1}$$

EQN1 $* \, \hat{\mathbf{w}}_{j'}, j' \neq k$:

$$\sum_{j \neq k}^{K} p_j(\hat{\mathbf{w}}_j^T \hat{\mathbf{w}}_{j'} - \hat{\mathbf{w}}_k^T \hat{\mathbf{w}}_{j'}) + 2N^{(t)}\lambda_{k,i}\hat{\mathbf{m}}_{k,i}^T \hat{\mathbf{w}}_{j'} = 0$$

$$\sum_{j \neq k}^{K} p_j(\hat{\mathbf{w}}_j^T \hat{\mathbf{w}}_{j'}) - \sum_{j \neq k}^{K} (p_j \hat{\mathbf{w}}_k^T \hat{\mathbf{w}}_{j'}) + 2N^{(t)}\lambda_{k,i}\hat{\mathbf{m}}_{k,i}^T \hat{\mathbf{w}}_{j'} = 0$$

$$
\begin{aligned}
\sum_{j \neq k}^{K} p_j(\hat{\mathbf{w}}_j^T \hat{\mathbf{w}}_{j'}) &= p_{j'}(\hat{\mathbf{w}}_{j'}^T \hat{\mathbf{w}}_{j'}) + \sum_{j \neq k, j \neq j'}^{K} p_j(\hat{\mathbf{w}}_j^T \hat{\mathbf{w}}_{j'}) \\
&= p_{j'} + \frac{-1}{K-1}(1 - p_{j'} - p_k) \\
&\frac{K}{K-1} p_{j'} + \frac{p_k}{K-1} - \frac{1}{k-1}
\end{aligned}
$$

$$
\begin{aligned}
\sum_{j \neq k}^{K} p_j (\hat{\mathbf{w}}_k^T \hat{\mathbf{w}}_{j'}) &= \hat{\mathbf{w}}_k^T \hat{\mathbf{w}}_{j'} \sum_{j \neq k}^{K} p_j \\
&= \frac{-1}{K-1}(1 - p_k)
\end{aligned}
$$

# Putting all together

$$\frac{K}{K-1}p_{j'} + \frac{p_k}{K-1} - \frac{1}{k-1} + \frac{-1}{K-1}(1-p_k) + 2N^{(t)}\lambda_{k,i}\hat{\mathbf{m}}_{k,i}^T\hat{\mathbf{w}}_{j'} = 0$$

$$\frac{K}{K-1}p_{j'} + 2N^{(t)}\lambda_{k,i}\hat{\mathbf{m}}_{k,i}^T\hat{\mathbf{w}}_{j'} = 0 \qquad \text{EQN2}$$

$$p_j^{'} = -\frac{2N^{(t)}\lambda_{k,i}\hat{\mathbf{m}}_{k,i}^T\hat{\mathbf{w}}_{j'}(K-1)}{K}$$

$$\frac{p_{j_1}}{p_{j_2}} = \frac{\hat{\mathbf{m}}_{k,i}^T\hat{\mathbf{w}}_{j_1}}{\hat{\mathbf{m}}_{k,i}^T\hat{\mathbf{w}}_{j_2}} = \frac{\exp(\hat{\mathbf{m}}_{k,i}^T\hat{\mathbf{w}}_{j_1})}{\exp(\hat{\mathbf{m}}_{k,i}^T\hat{\mathbf{w}}_{j_2})}$$

The function $f(x) = \exp(x)/x$ is monotonically increasing when $x < 1$. So, $p_{j_1} = p_{j_2}$, $\hat{\mathbf{m}}_{k,i}^T\hat{\mathbf{w}}_{j_1} = \hat{\mathbf{m}}_{k,i}^T\hat{\mathbf{w}}_{j_2}$, $\forall j_1, j_2 \neq k$

Because, $p_{j_1} = p_{j_2}, \forall j_1, j_2 \neq k$

$$\sum_{j \neq k}^{K} p_j = (1 - p_k)$$

$$(K-1)p_j = (1 - p_k)$$

$$p_j = \frac{(1 - p_k}{K - 1}$$

Now lets rewrite EQN2,

$$\frac{K}{K-1} p_{j'} + 2N^{(t)} \lambda_{k,i} \hat{\mathbf{m}}_{k,i}^T \hat{\mathbf{w}}_{j'} = 0$$

$$\frac{K}{K-1} p_j + 2N^{(t)} \lambda_{k,i} \hat{\mathbf{m}}_{k,i}^T \hat{\mathbf{w}}_j = 0$$

$$\frac{1 - p_k}{K - 1} \frac{K}{(K-1)} = -2N^{(t)} \lambda_{k,i} \hat{\mathbf{m}}_{k,i}^T \hat{\mathbf{w}}_j$$

$$-\frac{K}{(K-1)}(1 - p_k) = 2N^{(t)} \lambda_{k,i} \hat{\mathbf{m}}_{k,i}^T \hat{\mathbf{w}}_j (K-1) \qquad \text{EQN3}$$

Now lets do EQN1 $\times\ \hat{\mathbf{w}}_k$,

$$\sum_{j \neq k}^{K} p_j(\hat{\mathbf{w}}_j^T \hat{\mathbf{w}}_k - \hat{\mathbf{w}}_k^T \hat{\mathbf{w}}_k) + 2N^{(t)} \lambda_{k,i} \hat{\mathbf{m}}_{k,i}^T \hat{\mathbf{w}}_k = 0$$

$$\sum_{j \neq k}^{K} p_j(\frac{-1}{K-1} - 1) + 2N^{(t)} \lambda_{k,i} \hat{\mathbf{m}}_{k,i}^T \hat{\mathbf{w}}_k = 0$$

$$\frac{-K}{K-1}(1 - p_k) + 2N^{(t)} \lambda_{k,i} \hat{\mathbf{m}}_{k,i}^T \hat{\mathbf{w}}_k = 0 \qquad \text{EQN4}$$

Combine EQN3 and EQN4,

$$
\begin{aligned}
2N^{(t)}\lambda_{k,i}\hat{\mathbf{m}}_{k,i}^{T}\hat{\mathbf{w}}_{j}(K-1) + 2N^{(t)}\lambda_{k,i}\hat{\mathbf{m}}_{k,i}^{T}\hat{\mathbf{w}}_{k} &= 0 \\
\hat{\mathbf{m}}_{k,i}^{T}\hat{\mathbf{w}}_{j}(K-1) + \hat{\mathbf{m}}_{k,i}^{T}\hat{\mathbf{w}}_{k} &= 0 \qquad \text{EQN5}
\end{aligned}
$$

From DEF1 we know,

$$\hat{\mathbf{w}}_k^T \hat{\mathbf{w}}_{k'} = \frac{K}{K-1}\delta_{k,k'} - \frac{1}{K-1}$$

We have $\hat{\mathbf{W}}_{\text{WTF}} \cdot \mathbf{1}_K = \mathbf{0}_d$, where $\mathbf{1}_k$ is an all-ones vector in $\mathbb{R}^K$, and $\mathbf{0}_d$ is an all-zeros vector in $\mathbb{R}^d$. Then we have,

$$\sum_{k=1}^{K} \hat{\mathbf{w}} = \mathbf{0}_d$$

Now going back to EQN1 and using the above result,

$$\sum_{j \neq k}^{K} p_j(\hat{\mathbf{w}}_j - \hat{\mathbf{w}}_k) + 2N^{(t)}\lambda_{k,i} = 0$$

$$\sum_{j \neq k}^{K} \frac{(1-p_k)}{K-1}(\hat{\mathbf{w}}_j - \hat{\mathbf{w}}_k) + 2N^{(t)}\lambda_{k,i} = 0$$

$$\frac{(1-p_k)}{K-1}\left[-\hat{\mathbf{w}}_k - \hat{\mathbf{w}}_k(K-1)\right] + 2N^{(t)}\lambda_{k,i} = 0$$

$$-\frac{K}{K-1}(1-p_k)\hat{\mathbf{w}}_k + 2N^{(t)}\lambda_{k,i} = 0$$

which means $\hat{\mathbf{m}}_{k,i}$ is aligned with $\hat{\mathbf{w}}_k$. So we have,

$$\hat{\mathbf{m}}_{k,i}^T\hat{\mathbf{w}}_k = 1$$

Now, we can rewrite EQN5,

$$\hat{\mathbf{m}}_{k,i}^T\hat{\mathbf{w}}_j(K-1) + \hat{\mathbf{m}}_{k,i}^T\hat{\mathbf{w}}_k = 0$$

$$\hat{\mathbf{m}}_{k,i}^T\hat{\mathbf{w}}_j = -\frac{1}{K-1}, \forall j \neq k.$$

Therefore for any column vector $\hat{\mathbf{m}}_{k,i}$ in $\hat{\mathbf{M}}$, we have,

$$\hat{\mathbf{m}}_{k,i}^T\hat{\mathbf{w}}_{k'} = \frac{K}{K-1}\delta_{k,k'} - \frac{1}{K-1},$$

$$||\hat{\mathbf{m}}_{k,i}|| = 1 \forall k, k^{'} \in [1, K], 1 \leq i \leq n_k$$

That concludes the proof.