IIT Hyderabad **Rahul Vigneswaran K, CS23MTECH02002**
Advanced Topics in Machine Learning (CS6360) **Assignment #7**
Instructor: Dr. Vineeth N Balasubramanian **May 01, 2024**

---

1 Introduction

In this paper, we extend the pivotal work introduced in UTCIL [Yan+23], which addressed an optimization challenge pertaining to the cross-entropy loss. The focus is on a constrained optimization problem described as follows:

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

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

where:

- $\mathbf{m}_{k,i}^{(t)} \in \mathbb{R}^d$: Represents the feature vector of the i-th sample in class k during session t.
- n_k: Number of samples in class k.
- $K^{(t)}$: Number of classes in session t.
- $N^{(t)}$: Total number of samples in session t.
- $\mathbf{M}^{(t)} \in \mathbb{R}^{d \times N^{(t)}}$: Matrix collecting all feature vectors $\mathbf{m}_{k,i}^{(t)}$.
- $\hat{\mathbf{W}}_{\mathrm{ETF}} \in \mathbb{R}^{d \times K}$: Represents the endpoint of neural collapse for the entire label space.

The goal is to align each feature vector $\mathbf{m}_{k,i}^{(t)}$ with a predefined fixed class-specific vector $\mathbf{w}_k$, where the fixed vectors of each class k is equiangular to each other.

2 Theoretical Statement

The UTCIL framework, as discussed in [Yan+23], posits that the phenomenon of neural collapse could be leveraged to align models effectively in scenarios that involve continual learning and long-tail distributions. However, empirical results presented in the same study reveal a substantial discordance between theoretical expectations and actual outcomes. This significant disparity underscores the limitations of relying exclusively on neural collapse in such contexts.

In an attempt to bridge this gap, the concept of *"flying to collapse"* is introduced by UTCIL, but without sufficient integration into their existing theoretical model. Our study aims to rectify this lack of integration by refining the UTCIL framework to better address the challenges specific to long-tailed datasets. We hypothesize that, under our revised framework, the original claims of UTCIL regarding neural collapse will not hold.

2.1 Preliminaries

Definition 1. 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:

$$\mathbf{e}_{k_1}^T \mathbf{e}_{k_2} = \frac{K}{K-1}\delta_{k_1,k_2} - \frac{1}{K-1}, \forall k_1, k_2 \in [1, K] \tag{2}$$

where $\delta_{k_1,k_2} = 1$ when $k_1 = k_2$, and 0 otherwise. All vectors have the same ℓ_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 $\mathbb{R}^d$.

3 Proofs

In our analysis, we aim to relax two key assumptions from Equation 1:

1. The initial formulation assumes equal weight across classes by dividing the total loss by N. We propose modifying this by replacing N with N_k, the number of samples in class k, and integrating this term within the summation across classes.
2. The original UTCIL model considers only the final feature output $m_{k,i}$ for alignment, neglecting the model's weights. We redefine $m_{k,i}$ as $\theta^T\phi(x_{k,i})$, where $x_{k,i}$ is the i-th sample from class k, ϕ maps inputs to activations at the penultimate layer, and θ is the weight matrix. For simplicity, we denote this as $\theta^T\phi_{k,i}$, and assume θ is orthonormal. We omit the time step (t) for clarity.

Thus, our revised problem statement becomes:

$$\min_{\theta} \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i=1}^{n_k} L(\theta^T\phi_{k,i}, \hat{\mathbf{W}}_{\text{ETF}}), \quad 0 \leq t \leq T, \tag{3}$$

subject to:

$$\|\theta\phi_{k,i}\|^2 \leq 1, \quad \forall 1 \leq k \leq K, \quad 1 \leq i \leq n_k.$$

To convert Equation 3 into an unconstrained optimization problem, we formulate the Lagrangian as follows:

$$\tilde{L} = \sum_k \frac{1}{N_k} \sum_i -\log\left(\frac{\exp(w_k^T(\theta^T\phi_{k,i}))}{\sum_j \exp(w_j^T(\theta^T\phi_{k,i}))}\right) + \sum_k \sum_i \lambda_{k,i}\left(\|\theta^T\phi_{k,i}\|^2 - 1\right). \tag{4}$$

Differentiating $\tilde{L}$ with respect to $\phi_{k,i}$ and setting it to zero, we obtain:

$$\theta w_k + \sum_j p_j \theta w_j + 2N_k\lambda_{k,i}\theta\theta^T\phi_{k,i} = 0, \tag{5}$$

where p_j is defined as:

$$p_j = \frac{\exp(w_j^T(\theta^T\phi_{k,i}))}{\sum_j \exp(w_j^T(\theta^T\phi_{k,i}))}. \tag{6}$$

Given the KKT conditions' requirement for primal feasibility, $\lambda_{k,i} \leq 0$ implies two cases:

- **Case 1:** $\lambda = 0$
- **Case 2:** $\lambda < 0$

3.1 Case 1: $\lambda = 0$

In **Case 1**, we assume $\lambda = 0$. Under this assumption, Equation 5 simplifies to:

$$\theta w_k + \sum_j p_j \theta w_j = 0 \tag{7}$$

Taking the dot product of Equation 7 with θw_k from the left yields:

$$\begin{aligned}
(\theta w_k)^T \theta w_k + \sum_j p_j (\theta w_k)^T \theta w_j &= 0 \\
w_k^T \theta^T \theta w_k + \sum_j p_j w_k^T \theta^T \theta w_j &= 0 \\
w_k^T w_k + p_k w_k^T w_k + \sum_{j \neq k} p_j w_k^T w_j &= 0 \quad \text{(since } \theta \text{ is orthonormal)}
\end{aligned}$$

Simplifying further, we obtain:

$$p_k = \frac{2}{K} - 1 \tag{8}$$

Given that probabilities must satisfy $0 < p_k < 1$, the condition in Equation 8 implies K would have to be fractional, which contradicts the integer nature of K. Hence, **Case 1** results in a contradiction, invalidating the assumption that $\lambda = 0$.

The KKT conditions' complementarity slackness demands that $\lambda_{k,i}(\|\theta^T \phi_{k,i}\|^2 - 1) = 0$. Given the contradiction in **Case 1**, we must have $\lambda_{k,i} \neq 0$, leading to:

$$\|\theta^T \phi_{k,i}\|^2 = 1 \tag{9}$$

3.2 Case 2: $\lambda < 0$

We now analyze **Case 2**, where $\lambda < 0$. Under this condition, Equation 5 simplifies to:

$$\theta w_k + \sum_j p_j \theta w_j + 2N_k \lambda_{k,i} \theta \theta^T \phi_{k,i} = 0 \tag{10}$$

Taking the dot product of Equation 10 with $\theta w_{j'}$ (where $j' \neq k$) from the left, we can rewrite and simplify the expression as follows:

$$(\theta w_{j'})^T \theta w_k + \sum_j p_j (\theta w_{j'})^T \theta w_j + 2N_k \lambda_{k,i} (\theta w_{j'})^T \theta \theta^T \phi_{k,i} = 0$$

$$w_{j'}^T \theta^T \theta w_k + \sum_j p_j w_{j'}^T \theta^T \theta w_j + 2N_k \lambda_{k,i} w_{j'}^T \theta^T \phi_{k,i} = 0$$

$$w_{j'}^T w_k + p_k w_{j'}^T w_k + \sum_{j \neq k} p_j w_{j'}^T w_j + 2N_k \lambda_{k,i} w_{j'}^T \phi_{k,i} = 0 \quad \text{(since } \theta \text{ is orthonormal)}$$

Applying Definition 2 and simplifying the above equation, we derive:

$$p_{j'} = -\frac{2N_k \lambda_{k,i} w_{j'}^T \phi_{k,i} (K-1)}{K} \tag{11}$$

We already know that $j' \neq k$ and from Equation 6 we know what is p_j. So we take j' to be some j_1 and j_2 and take their ratio,

$$\frac{p_{j_1}}{p_{j_2}} = \frac{w_{j_1}^T \theta^T \phi_{k,i}}{w_{j_2}^T \theta^T \phi_{k,i}} = \frac{\exp(w_{j_1}^T \theta^T \phi_{k,i})}{\exp(w_{j_2}^T \theta^T \phi_{k,i})}$$

$$\frac{p_{j_1}}{p_{j_2}} = \frac{\exp(w_{j_1}^T \theta^T \phi_{k,i})}{w_{j_1}^T \theta^T \phi_{k,i}} = \frac{\exp(w_{j_2}^T \theta^T \phi_{k,i})}{w_{j_2}^T \theta^T \phi_{k,i}}$$

The above is of the form $f(x) = \frac{\exp(x)}{x}$ which is monotonically increasing. The only way the above can be possible is if $p_{j_1} = p_{j_2}$. Unfortunately, the rest of proof follows the same flow as [Yan+23]. Despite relaxations and slight adjustments in the formulation, our findings corroborate the original results outlined in [Yan+23], demonstrating the robustness of the theorem under various analytical frameworks. Thus, we reiterate:

Theorem. Let $\hat{\mathbf{M}}^{(t)}$ be the global minimizer obtained by incremental optimization from $t = 0$ to $t = T$, assembled as $\hat{\mathbf{M}} = \left[\hat{\mathbf{M}}^{(0)}, \cdots, \hat{\mathbf{M}}^{(T)}\right] \in \mathbb{R}^{d \times \sum_{t=0}^{T} N^{(t)}}$. Regardless of the loss type ($\mathcal{L}$), whether it is cross-entropy (CE) or misalignment loss, for any vector $(\theta^T \phi_{k,i})$ corresponding to class label k, the following holds:

$$\|(\theta^T \phi_{k,i})\| = 1, \quad (\theta^T \phi_{k,i})^\top \hat{\mathbf{w}}_{k'} = \frac{K}{K-1} \delta_{k,k'} - \frac{1}{K-1},$$

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

4 Conclusion

Despite our efforts to modify the existing framework to better handle the challenges posed by long-tail datasets, the alterations did not yield a new theorem but instead reaffirmed the existing

theorem from the referenced work. This outcome suggests that our approach of differentiating with respect to a single sample, fixing the class k and sample i, may be inherently limited.

One potential avenue for overcoming this limitation involves embedding class-specific elements, such as $\frac{1}{N_k}$, directly into the softmax function. This modification could fundamentally alter the dynamics of the optimization problem, potentially leading to different outcomes even after differentiation.

Regrettably, we were unable to revise the theoretical framework as initially intended within the timeframe of this submission. Therefore, we identify this as a key area for future research, hoping that subsequent investigations will integrate these class-specific adjustments into the softmax function and explore their impact on the overall theoretical model.

References

[Yan+23] Yibo Yang et al. "Neural Collapse Terminus: A Unified Solution for Class Incremental Learning and Its Variants". In: *ArXiv* abs/2308.01746 (2023). URL: https://api.semanticscholar.org/CorpusID:260438867.