Supervised Contrastive Representation Learning: Landscape Analysis with Unconstrained Features

Tina Behnia, Christos Thrampoulidis

Department of Electrical and Computer Engineering

University of British Columbia

Vancouver, BC, Canada

Email: {tina.behnia, cthrampo}@ece.ubc.ca

Abstract—Recent findings reveal that over-parameterized deep neural networks, trained beyond zero training-error, exhibit a distinctive structural pattern at the final layer, termed as Neural-collapse (NC). These results indicate that the final hidden-layer outputs in such networks display minimal within-class variations over the training set. While existing research extensively investigates this phenomenon under cross-entropy loss, there are fewer studies focusing on its contrastive counterpart, supervised contrastive (SC) loss. Through the lens of NC, this paper employs an analytical approach to study the solutions derived from optimizing the SC loss. We adopt the unconstrained features model (UFM) as a representative proxy for unveiling NC-related phenomena in sufficiently over-parameterized deep networks. We show that, despite the non-convexity of SC loss minimization, all local minima are global minima. Furthermore, the minimizer is unique (up to a rotation). We prove our results by formalizing a tight convex relaxation of the UFM. Finally, through this convex formulation, we delve deeper into characterizing the properties of global solutions under label-imbalanced training data.

I. INTRODUCTION

Contrastive techniques (e.g., [1], [2]) have led to major improvements in the self-supervised representation learning, and recently, [3] introduced the supervised contrastive (SC) loss as an extension of these techniques tailored to the supervised problems. SC loss is applied on the embeddings generated by the network and it penalizes the model by measuring how similar or dissimilar the learned embeddings are for different samples based on their class membership. The objective with the SC loss is to bring the embeddings with the same label close to each other while keeping them apart from those belonging to other classes. In other words, the loss value emphasizes on the interactions of pairs of samples from the training set. This is in contrast to other widely used loss functions like cross-entropy (CE) where the loss is computed on each sample individually, aiming to maximize the class-probability associated with the correct label.

Focusing on the CE loss, [4] recently discovered a common pattern at the final layer of the over-parameterized deep-nets, whose number of parameters far exceeds the number of training samples. According to their observations, in the modern deep learning regime where deep-nets are trained beyond zero-training error, at the last layer, the network learns to compress the training data by forcing the embeddings to collapse to their class-means, exhibiting a low-rank structure. Furthermore, when the training data is balanced, the class-means exhibit a symmetric pattern. In particular, they form a simplex *equiangular tight frame* (ETF), where all the class-mean vectors have equal norms and are maximally-separated. Importantly, this simple geometry appears to be "*cross-situationally invariant*", as it appears at the final layer of various architectures trained on benchmark datasets [4]. Seeking to explain the optimality of this structure, [4] links the feature learning process to an abstract optimal coding task. Treating each embedding as a codeword for the class membership, and the classification task as a decoding of the noisy codeword, [4] demonstrates that this shared geometry yields the optimal codebook. This discovery, termed as Neural collapse (NC), has led to a long line of follow-up works, e.g., [5]–[16]. The majority of these works focus on the CE and mean-squared-error loss, and a few consider analyzing contrastive techniques in this context.

The NC discovery was an empirical attempt to demystify the black-box nature of the deep-nets and to more formally characterize what large networks learn at the training stage. NC also shares connections with the implicit bias literature (e.g., [17]–[19]) where the success of over-parameterization is attributed to the optimization algorithms, arguing that gradient-based algorithms inherently favor solutions with certain properties that generalize well. However, the emergence of NC in deep overparameterized networks remains hard to theoretically explain due to the complex interactions of their parameters and layers. This challenge motivated a series of theoretical studies to justify the NC phenomenon from an optimization perspective (e.g., [5], [7], [10], [12], [15], [16], [20]–[23]). These works rely on a proxy model called *unconstrained features model* (UFM). This model leverages the powerful expressivity of large deep-nets to simplify the mathematical formulation of the training objective. This is made possible by only focusing on the interactions of the last-layer parameters of the network in the loss function. Despite its simplicity, UFM has proved a useful tool for predicting certain empirical patterns learned by deep-nets during training.

A. Related Works

Neural-collapse. After the influential paper [4], important follow-up works introduced UFM [5], [7], [12] as an abstract model for theoretically studying the last-layer of over-parameterized deep-nets. In this model the focus is only on the structure of the parameters in the final layer that optimize the desired objective. Previous works [5], [7], [10], [12], [15], [16], [20], [23] have studied the UFM problem for the CE loss showing that with class-balanced training sets all the global solutions satisfy the NC property and form a simplex ETF as empirically reported by [4]. Despite being a non-convex problem, [15] proves that UFM with CE loss has a benign landscape: all local optimals are global optimal, and there are no spurious saddle points. Thus, any optimization algorithm that can escape saddle points, converges to the global optimal, that is the simplex ETF. While most of these works study the balanced setup, several works use the UFM to predict the learned final-layer embeddings for class-imbalanced training data [5], [20], [21], and support their theoretical findings by experiments on benchmark datasets and architectures. The geometric characterization of the embeddings under the mean-squared error [12], [13], [16] and contrastive loss [5], [7] have also been investigated by adopting the UFM. Several works [6], [9], [24] also explore the potential links of NC to generalization and down-stream tasks.

Supervised contrastive loss. Unlike CE and mean-squared-error loss, there are fewer works on the geometry characterization of the constrastive loss through the lens of NC. The SC loss is known to exhibit more robustness to noise and hyper-parameter values [3]. Its resilience has inspired a series of works where the variants of the SC loss combined with CE are used for training models that tackle distribution shifts in the data with better performance [25]–[29]. In the context of self-supervised learning, [30] investigates the properties of the contrastive feature learning. For supervised classification, [7] is the first to compare how the CE and SC loss encode the training data by studying the UFM. They show that with balanced training data, the global optimizer of the SC loss is ETF similar to CE and demonstrate that deep-nets trained with the SC loss reach this global optimizer. However, unlike CE, they emphasize on the fact that the optimality of this geometry can be violated depending on the mini-batch selection in practical scenarios. Later, [22] suggests adjustments on the training to ensure learning an embedding geometry that is invariant to the randomness in batching and level of class-imbalance. Furthermore, [31] conjecture that as with the CE loss, the UFM trained by the SC loss also has a benign landscape with no spurious local minima despite its non-convex nature. Their claim is motivated by numerical verifications that show the global solution of the UFM can be reached by simple gradient methods applied on the SC loss.

B. Contributions

Following the empirical results on the SC loss in previous works [7], [31], in this paper, we theoretically analyze the landscape of this loss under the UFM assumptions. We first demonstrate that all the stationary points exhibit the collapse property (see Definition 1). Leveraging this low-rank structure at the stationary points, we prove that the SC loss has a benign landscape: we show that although the optimal solution is not unique, all the local solutions only differ up to a rotation and are equivalent to the global solution. Our results rely on introducing a tight convex relaxation of the UFM. As a consequence of this analysis, we are able to simplify finding the global optimizer into a lower-dimensional equivalent program under the STEP imbalanced assumption on the training set. For the special case of balanced training data, this program reveals ETF as the optimal geometry as also previously shown in [7]. However, ETF is not a global solution otherwise.

Notations. We use $[n]$ to denote the set $\{1, 2, ..., n\}$. For matrix $\mathbf{V}_{m \times n}$, $\mathbf{V}_{i,j}$ denotes its (i,j)-th entry, $\mathbf{v}_j$ the j-th *column*, and $\mathbf{V}^\top$ its transpose. When $m = n$, $\mathbf{V} \succeq 0$ is a PSD matrix and $\mathrm{diag}(\mathbf{V})$ is a diagonal matrix that includes the diagonal elements of $\mathbf{V}$. We use $\mathbf{1}_m$ and $\mathbb{I}_m$ to denote the m-dimensional vector of all ones and the identity matrix respectively. Finally, $\otimes$ refers to the Kronecker product.

II. PROBLEM SETUP

Consider a supervised classification setup with k classes and a training set $\mathcal{S} = \{(\mathbf{x}_i, y_i) : i \in [n]\}$ where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in [k]$ are the i-th data sample and its class label. We also let $n_c \geq 1$, $c \in [k]$ to be the number of examples in class c.

We consider the deep-net as a non-linear encoder $\mathbf{h}_{\boldsymbol{\theta}} : \mathbb{R}^p \to \mathbb{R}^d$ mapping the input space to a new embedding space that is parameterized by the trainable variables $\boldsymbol{\theta}$. Define $\mathbf{h}_{\boldsymbol{\theta},i} := \mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_i)$, $i \in [n]$ and let $\mathbf{H}_{\boldsymbol{\theta}} := [\mathbf{h}_{\boldsymbol{\theta},1}, ..., \mathbf{h}_{\boldsymbol{\theta},n}]_{d \times n}$ be the training embeddings learned by the deep-net. With the SC loss, we seek to find the optimal parameters by minimizing $\mathcal{L}_{\mathrm{SC}}(\cdot)$ defined as follows:

$$\mathcal{L}_{\mathrm{SC}}(\mathbf{H}_{\boldsymbol{\theta}}; \tau) = \frac{1}{n} \sum_{i \in [n]} \frac{-1}{n_{y_i} - 1} \sum_{\substack{j \neq i \\ y_j = y_i}} \log \left(\frac{e^{\mathbf{h}_{\boldsymbol{\theta},i}^\top \mathbf{h}_{\boldsymbol{\theta},j}/\tau}}{\sum_{\ell \neq i} e^{\mathbf{h}_{\boldsymbol{\theta},i}^\top \mathbf{h}_{\boldsymbol{\theta},\ell}/\tau}} \right).$$

(1)

Here, the scalar τ is a positive temperature hyper-parameter. This parameter was introduced in the loss by [3] as they observed it can boost the performance of the model by controlling the smoothness of the loss.

By minimizing the SC loss, we favor parameters $\boldsymbol{\theta}$ that pull the examples belonging to the same class closer while simultaneously pushing the samples from different classes away from each other. Training a model with this loss usually involves two stages: first we learn the embeddings $\mathbf{h}_{\boldsymbol{\theta}}(\cdot)$ by minimizing the SC loss, and second, we train a (simple) classifier separately on top of the learned representations. In this work, we focus on the first representation learning stage.

Remark 1. *In the definition of the SC loss, introduced by [3], the embeddings $\mathbf{h}_{\boldsymbol{\theta}}(\cdot)$ are normalized to the unit sphere. In other words, the loss is a function of the cosine similarity of the embeddings learned by the deep-net. Alternatively, we can assume the deep-net at the final layer is followed by*

a normalization layer that maps the embeddings to the unit sphere so that $\|\mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_i)\|_2 = 1$.

A. Unconstrained Features Model (UFM)

In order to develop a theory for the training dynamics of deep-nets, we ought to capture the interplay between several non-linear hidden layers of the network. This makes the theoretical analysis of deep-networks challenging and motivates us to simplify the interactions between layers. UFM [5], [7], [12] serves as an abstract model that gives (partial) theoretical justification of the discovery made by [4]. Focusing on the last layer of the network, UFM, relying on the power of over-parameterization, assumes large deep-nets are expressive enough to (approximately) map any point in the input space to any point in the embedding space. Hence, in this model, we assume deep-nets can learn *unconstrained features* and treat the last-layer embeddings as free optimization parameters. In other words, we relax the dependence of the embeddings $\mathbf{H}_{\boldsymbol{\theta}}$ on the network parameters $\boldsymbol{\theta}$ and optimize the problem directly over $\mathbf{H}$. Thus, under UFM, our problem reduces to:

$$\mathbf{H}^* \in \arg\min_{\mathbf{H}} \mathcal{L}_{\text{SC}}(\mathbf{H}; \tau) \ \text{ subj. to } \|\mathbf{h}_i\|_2^2 = 1, \forall i \in [n]. \quad (2)$$

With this simplification, we study the patterns appearing in large deep-nets, restricting our attention to the interactions of the only element that the NC property is concerned with: the last-layer embeddings. It is also important to highlight that the objective of this optimization problem is still nonconvex as the loss is computed over the non-linear interactions of embeddings, i.e., $\mathbf{H}^\top \mathbf{H}$.

Remark 2. *Throughout this paper, we use a slightly different terminology than the Neural-collapse literature. We borrow the terminology from [21]: we use* Neural-collapse *(NC) particularly when referring to the collapse of the embeddings, and (implicit) geometry when referring to the arrangement of the last-layer feature-embeddings* $\mathbf{H}$ *(e.g., ETF in balanced case). More formally, by NC, we mean the following:*

Definition 1 (NC property). *Consider the embeddings matrix* $\mathbf{H}$. *NC holds for* $\mathbf{H}$ *if* $\mathbf{h}_i = \mathbf{h}_j, \forall i, j : y_i = y_j$.

B. Practices in Contrastive Learning

Contrastive learning involves several heuristics for achieving optimal performance in practice. We choose to overlook some of these common practices to simplify the analysis. We elaborate on these assumptions in this section. While not being fully consistent with the practice, we believe this preliminary analysis can provide useful insights for future works.

Computational cost. Due to the pairwise interaction of the embeddings in the loss function, SC loss has a high computational cost when computed over the whole training set as in (1). As a result, in practice, the loss is approximated over random batches chosen from the training set, instead of the whole set $\mathcal{S}$. As discussed in Section I-A, previous works argue that batch selection might change the optimal embedding geometry [7], [22] as it affects the interaction of embeddings in

the SC loss. To avoid the complication caused by the batching strategy, we focus on the full-batch scenario (1) in our theory.

Projection head. At the training stage, the SC loss is not directly computed on the embeddings. In fact, the embeddings are passed through a projection layer (which can be as simple as a single-layer or shallow MLP), and the loss is minimized on the projected embeddings. This projection layer is discarded at test time, as it is empirically shown that using the pre-projection embeddings yield better performance for downstream tasks. In our analysis, we ignore the projection head of the network, and focus on studying the embeddings on which the loss is directly computed. This is similar to the analysis performed by previous works [5], [7], [30].

Embedding constraints. As discussed in Remark 1, the SC loss is computed over unit norm embeddings. Since, constraining the norms to the sphere yields a non-convex feasibility set, we relax the norm constraints to a ball of radius 1. On the other hand, we can view the temperature parameter τ as a scaling on the embeddings $\mathbf{h}_i$. Specifically, it is easy to see that $\mathcal{L}_{\text{SC}}(\mathbf{H}; \tau) = \mathcal{L}_{\text{SC}}(\widetilde{\mathbf{H}}; 1)$ where $\widetilde{\mathbf{H}} = \mathbf{H}/\sqrt{\tau}$ and as a result $\|\widetilde{\mathbf{h}}_i\|_2^2 = 1/\tau$ for all $i \in [n]$. In other words, after relaxing the constraints, we study the UFM formulated as follows,

$$\mathbf{H}^* \in \arg\min_{\mathbf{H}} \mathcal{L}_{\text{SC}}(\mathbf{H}) \ \text{ subj. to } \|\mathbf{h}_i\|_2^2 \leq \frac{1}{\tau}, \forall i \in [n], \quad (3)$$

where $\mathcal{L}_{\text{SC}}(\mathbf{H}) := \mathcal{L}_{\text{SC}}(\mathbf{H}; \tau = 1)$. Here we recall that, the minimization is over the *unconstrained* embeddings $\mathbf{H}$ that do not depend on the network parameters $\boldsymbol{\theta}$. Although we simplified the problem to deal with a convex feasibility set, the problem (3) still has a non-convex objective function.

III. RESULTS

In this section, we perform an analysis of the local minimizers of the SC loss with the UFM (3). As part of the analysis, we rely on the characterization of a tight convex relaxation of (3) which we use later for taking a closer look at the structure of the global solutions. We first start by stating a general property of the local optimal points of (3).

A. NC Holds at Local Optimal Solutions

Let $n_{\max} = \max_{c \in [k]} n_c$ be the size of the largest class, and recall the necessary condition for local optimality in a general optimization with convex constraints.

Proposition 1 (Necessary stationarity condition [32, Chapter 2]). *Consider the optimization problem* $\min_{\mathbf{x} \in \mathcal{C}} f(\mathbf{x})$ *over a convex set* $\mathcal{C}$. *Let* $\widehat{\mathbf{x}} \in \mathcal{C}$ *be a local minimum. Then,*

$$\nabla f(\widehat{\mathbf{x}})^\top (\mathbf{x} - \widehat{\mathbf{x}}) \geq 0, \qquad \forall \mathbf{x} \in \mathcal{C}. \quad (4)$$

Lemma 1. *Assume* $\tau > 2/\log((n-1)/(n_{\max}-1))$. *Let* $\widehat{\mathbf{H}}$ *be locally optimal in the UFM problem* (3). *The NC property (Definition 1) holds at* $\widehat{\mathbf{H}}$.

Lemma 1 shows all the locally optimal feasible points have a low-rank structure and satisfy the collapse property. Thus, similar to the case of CE [12], [15], it follows that at the global solutions, the embeddings collapse to their class-means.

The proof of this lemma simply follows by showing that condition (4) does not hold for a feasible $\widehat{\mathbf{H}}$ that does not satisfy the NC property (recall that the feasible set in (3) is convex). Specifically, suppose $y_1 = y_2$ and $\widehat{\mathbf{h}}_1 \neq \widehat{\mathbf{h}}_2$. Then, define $\mathbf{H}$ by swapping the first two columns of $\widehat{\mathbf{H}}$, i.e., $\mathbf{H} = [\widehat{\mathbf{h}}_2, \widehat{\mathbf{h}}_1, \widehat{\mathbf{h}}_3, \cdots, \widehat{\mathbf{h}}_n]$. It is then possible to show that if the condition on τ holds, $\nabla \mathcal{L}_{\mathrm{SC}}(\widehat{\mathbf{H}})^\top (\mathbf{H} - \widehat{\mathbf{H}}) < 0$, which contradicts the necessary condition of local optimality (4). The rest of our results in this section, rely on Lemma 1. Thus, in what follows, we assume

$$\tau > 2/\log\left(\frac{n-1}{n_{\max}-1}\right).$$

For balanced data, $n_{\max} = n/k$ and the condition approximately becomes $\tau > 2/\log(k)$ for $n >> k$, which is not restrictive in practical settings. For brevity, we do not re-state this condition in the next sections. We finally note that the condition on τ is a sufficient requirement but not a necessary one. We conjecture that this condition can be relaxed.

B. UFM: Landscape and Optimality Conditions

Here, we study the local and global optimality conditions for the UFM under the SC loss in (3). Since the problem is non-convex, it is not obvious whether (3) has a unique solution or if it has a landscape free from spurious local solutions (which are not globally optimal). In fact, it is easy to see that the global solution is not unique, since rotating the embeddings by an orthonormal matrix $\mathbf{R}$, does not change the value of the loss function, i.e., $\mathcal{L}_{\mathrm{SC}}(\mathbf{R}^\top \mathbf{H}) = \mathcal{L}_{\mathrm{SC}}(\mathbf{H})$. The next theorem characterizes the local and global solutions of (3).

Theorem 1 (UFM landscape with SC loss)**.** *Assume the feature dimension exceeds the number of classes, $d > k$. Then, (i) all the local solutions of the UFM* (3) *are global optima. (ii) if $\mathbf{H}_1^*$ and $\mathbf{H}_2^*$ are global minimizers, $\mathbf{H}_1^{*\top}\mathbf{H}_1^* = \mathbf{H}_2^{*\top}\mathbf{H}_2^*$.*

This theorem proves that while the global optimizer of the non-convex problem (3) is not unique, it has a unique *implicit geometry*. Further, it ensures the loss landscape has no non-global local solutions. We will discuss the proof sketch next.

C. Proof Sketch

The main idea in the proof is to relate the KKT points of the problem (3) to the KKT points of its convex relaxation. We start by formalizing the convex relaxation.

Convex relaxation. To define the convex relaxation, we note that the SC loss is in fact a convex function of the Gram matrix $\mathbf{G} := \mathbf{H}^\top \mathbf{H}$.

$$\mathcal{L}_{\mathrm{SC}}(\mathbf{H}) = \widehat{\mathcal{L}}(\mathbf{G}) := \frac{1}{n} \sum_{i \in [n]} \frac{-1}{n_{y_i} - 1} \sum_{\substack{j \neq i \\ y_j = y_i}} \log\left(\frac{e^{\mathbf{G}_{i,j}}}{\sum_{\ell \neq i} e^{\mathbf{G}_{i,\ell}}}\right).$$

To make the convex loss function symmetric, we define $\mathcal{L}_{\mathrm{cvx}}(\mathbf{G}) := \frac{1}{2}\left(\widehat{\mathcal{L}}(\mathbf{G}) + \widehat{\mathcal{L}}(\mathbf{G}^\top)\right)$ and formulate the relaxed convex problem with the symmetrized objective function:

$$\mathbf{G}^* \in \arg\min_{\mathbf{G}} \mathcal{L}_{\mathrm{cvx}}(\mathbf{G}), \quad \text{subj. to } \mathbf{G} \succeq 0, \ \mathbf{G}_{i,i} \leq \frac{1}{\tau}, \ \forall i \in [n]. \tag{5}$$

This convex program satisfies the Slater's conditions. Hence, strong duality holds and any KKT point of (5) specifies a global solution.

Proof of Theorem 1. Suppose $\widehat{\mathbf{H}}$ is a local optima. Thus, it satisfies the first order and second order necessary condition for local optimality for constrained problems [33, Theorems 12.1 and 12.5]. In particular, consider the Lagrangian function

$$\mathrm{Lag}(\mathbf{H}, \boldsymbol{\lambda}) = \mathcal{L}_{\mathrm{SC}}(\mathbf{H}) + \sum_{i \in [n]} \boldsymbol{\lambda}_i (\mathbf{h}_i^\top \mathbf{h}_i - 1/\tau),$$

where $\boldsymbol{\lambda} \in \mathbb{R}^n$. By the first-order conditions, there exists a dual variable $\widehat{\boldsymbol{\lambda}} \in \mathbb{R}^n$ such that $(\widehat{\mathbf{H}}, \widehat{\boldsymbol{\lambda}})$ are the KKT points of (3). Specifically, $\nabla_{\mathbf{H}} \mathrm{Lag}(\widehat{\mathbf{H}}, \widehat{\boldsymbol{\lambda}}) = 0$, i.e.,

$$2\widehat{\mathbf{H}}(\nabla_{\mathbf{G}} \mathcal{L}_{\mathrm{cvx}}(\widehat{\mathbf{H}}^\top \widehat{\mathbf{H}}) + \mathrm{diag}(\widehat{\boldsymbol{\lambda}})) = 0$$

By the second-order conditions, the Hessian of the Lagrangian should have positive curvature in every direction $\boldsymbol{\Delta}_{d \times n} = [\mathbf{d}_1, ..., \mathbf{d}_n]$ such that

$$\mathbf{d}_i^\top \widehat{\mathbf{h}}_i = 0, \quad \text{if } \|\widehat{\mathbf{h}}_i\|_2^2 = \frac{1}{\tau}, \ \widehat{\boldsymbol{\lambda}}_i > 0,$$

$$\mathbf{d}_i^\top \widehat{\mathbf{h}}_i \geq 0, \quad \text{if } \|\widehat{\mathbf{h}}_i\|_2^2 = \frac{1}{\tau}, \ \widehat{\boldsymbol{\lambda}}_i = 0,$$

Suppose $\mathbf{B} := \nabla_{\mathbf{G}} \mathcal{L}_{\mathrm{cvx}}(\widehat{\mathbf{H}}^\top \widehat{\mathbf{H}}) + \mathrm{diag}(\widehat{\boldsymbol{\lambda}})$ has a negative eigen-value, i.e., there exists $\mathbf{u} \in \mathbb{R}^n$ such that $\mathbf{u}^\top \mathbf{B} \mathbf{u} < 0$. On the other hand, by Lemma 1, the NC property holds at $\widehat{\mathbf{H}}$ and $\mathrm{rank}(\widehat{\mathbf{H}}) \leq k < d$. This implies that for some $\mathbf{v} \in \mathbb{R}^d$, $\widehat{\mathbf{H}}^\top \mathbf{v} = 0$. As a result, it is possible to show that the Hessian of the Lagrangian $\nabla_{\mathbf{H}}^2 \mathrm{Lag}(\widehat{\mathbf{H}}, \widehat{\boldsymbol{\lambda}})$ has a negative curvature in the direction $\boldsymbol{\Delta} := \mathbf{v} \mathbf{u}^\top$. This contradicts the second order condition of local optimality [33, Theorems 12.5] since $\boldsymbol{\Delta}^\top \widehat{\mathbf{H}} = 0$, i.e., $\mathbf{d}_i^\top \widehat{\mathbf{h}}_i = 0, \forall i \in [n]$.

Therefore, it is necessary for $(\widehat{\mathbf{H}}, \widehat{\boldsymbol{\lambda}})$ to satisfy,

$$\nabla \mathcal{L}_{\mathrm{cvx}}(\widehat{\mathbf{H}}^\top \widehat{\mathbf{H}}) + \mathrm{diag}\left(\widehat{\boldsymbol{\lambda}}\right) \succeq 0.$$

From here, it is straightforward to prove $(\widehat{\mathbf{G}} := \widehat{\mathbf{H}}^\top \widehat{\mathbf{H}}, \widehat{\boldsymbol{\lambda}}, \mathbf{B})$ is a KKT point of the convex problem (5). Equivalently, by strong duality, $\widehat{\mathbf{G}}$ is a global optimizer. On the other hand, since (5) is a relaxation of the original UFM (3), for the optimal solution $\mathbf{H}^*$, we have,

$$\mathcal{L}_{\mathrm{SC}}(\widehat{\mathbf{H}}) = \mathcal{L}_{\mathrm{cvx}}(\widehat{\mathbf{G}}) \leq \mathcal{L}_{\mathrm{cvx}}(\mathbf{H}^{*\top} \mathbf{H}^*) = \mathcal{L}_{\mathrm{SC}}(\mathbf{H}^*). \tag{6}$$

Therefore, $\widehat{\mathbf{H}}$ is a global optimum of (3), and the proof is complete. This further shows the relaxation in (5) is tight.

Corollary 1. *Suppose $d > k$. Then, the convex program* (5) *is a tight relaxation of* (3).

We can further show that the optimal solution of (5) is unique. We use this useful result in the next section.

Lemma 2. *The convex UFM problem* (5) *has a unique minimizer.*

For the proof of this lemma, we note two facts: 1) If $\mathbf{G} - \mathrm{diag}(\mathbf{G}) \propto \mathbf{1}_n \mathbf{1}_n^\top - \mathbb{I}_n$, $\mathbf{G}$ is not a global optima, and 2) the objective $\mathcal{L}_{\mathrm{cvx}}(\cdot)$ is strictly convex in any direction $\mathbf{V}$ such that $\mathbf{V} - \mathrm{diag}(\mathbf{V}) \not\propto \mathbf{1}_n \mathbf{1}_n^\top - \mathbb{I}_n$.

D. Global Solutions

By Lemma 1, we know that the global solution of the UFM (3), as a first order stationary point, has the NC property. That is, there exists $\mathbf{M}^*_{d \times k} = [\boldsymbol{\mu}^*_1, \dots \boldsymbol{\mu}^*_k]$ such that $\mathbf{h}^*_i = \boldsymbol{\mu}^*_{y_i}$. Thus, to find the conditions of global optimality, we only need to find the k optimal mean-embedding vectors $\boldsymbol{\mu}^*_c$, $c \in [k]$. To analyze the global solutions, we restrict our attention to a specific training data distribution defined as follows.

Definition 2 ((R, ρ)-STEP imbalanced data). *With $R \geq 1$ as the imbalance ratio $\rho \in (0,1)$ as the minority fraction, the following holds: Let $\overline{\rho} := 1 - \rho$. Classes $c \in \{1, \dots, \overline{\rho}k\}$ are majority classes, all with the same sample size of $n_c = n_{maj}$ and classes $c \in \{\overline{\rho}k+1, \cdots, k\}$ are minority classes, all with the same sample size of $n_c = n_{minor}$, where $n_{\mathrm{maj}} = Rn_{\mathrm{minor}}$.*

Given the NC property at the optimal solution, in the (R, ρ)-STEP imbalanced setting, we can use the symmetry in the size of the classes to even further simplify the structure of the global solution. The following lemma formalizes this argument.

Lemma 3. *Assume an (R, ρ)-STEP imbalanced setting with k classes. Let $\mathbf{H}^*$ be the global solution of (3). Then $\mathbf{G}^* = \mathbf{H}^{*\top}\mathbf{H}^*$ is the unique optimizer of the convex program (5). Furthermore, $\mathbf{G}^*$ has a block structure: there exists scalars $\alpha_{\mathrm{maj/minor}}$, $\beta_{\mathrm{maj/minor}}$, $\theta \in [-1/\tau, 1/\tau]$ such that, $\mathbf{G}^* = \begin{bmatrix} \mathbf{G}^*_{1,1} & \mathbf{G}^*_{1,2} \\ \mathbf{G}^*_{2,1} & \mathbf{G}^*_{2,2} \end{bmatrix}$, and*

$$\mathbf{G}^*_{1,1} = \left((\alpha_{\mathrm{maj}} - \beta_{\mathrm{maj}})\mathbb{I}_{\overline{\rho}k} + \beta_{\mathrm{maj}}\mathbf{1}_{\overline{\rho}k}\mathbf{1}^\top_{\overline{\rho}k} \right) \otimes \mathbf{1}_{n_{\mathrm{maj}}}\mathbf{1}^\top_{n_{\mathrm{maj}}}$$

$$\mathbf{G}^*_{1,2} = \theta\mathbf{1}_{\overline{\rho}kn_{\mathrm{maj}}}\mathbf{1}^\top_{\rho kn_{\mathrm{minor}}}$$

$$\mathbf{G}^*_{2,1} = \theta\mathbf{1}_{\rho kn_{\mathrm{minor}}}\mathbf{1}^\top_{\overline{\rho}kn_{\mathrm{maj}}}$$

$$\mathbf{G}^*_{2,2} = \left((\alpha_{\mathrm{minor}} - \beta_{\mathrm{minor}})\mathbb{I}_{\rho k} + \beta_{\mathrm{minor}}\mathbf{1}_{\rho k}\mathbf{1}^\top_{\rho k} \right) \otimes \mathbf{1}_{n_{\mathrm{minor}}}\mathbf{1}^\top_{n_{\mathrm{minor}}}$$

This lemma follows from a simple symmetry argument that uses the fact that for two classes with equal sizes, their corresponding (mean-)embeddings are weighted similarly in the objective of problem (5):

Lemma 4. *Consider problem (5). If $n_{c_1} = n_{c_2}$ for $c_1 \neq c_2 \in [k]$, then $\forall i,j,\ell \in [n]$, such that $y_i = c_1$, $y_j = c_2$, and $y_\ell \neq c_1, c_2$, $\mathbf{G}^*_{i,\ell} = \mathbf{G}^*_{j,\ell}$.*

Recall from Corollary 1, the relaxation (5) is tight, and finding $\mathbf{G}^*$ characterizes the optimal $\mathbf{H}^*$ up to a rotation. This lemma reduces the dimension of the original problem in the case of STEP imbalanced data. In the specific case of balanced data ($n_c = n/k$), the optimal solution reduces to ETF, as also previously shown [7].

Theorem 2. *Assume that the training set is balanced ($R = 1$). The optimal solution $\mathbf{H}^*$ of the UFM (3) follows an ETF. In other words, there exists $\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k$ such that for $\mathbf{M} = [\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k]$ we have $\mathbf{M}^\top\mathbf{M} \propto \mathbb{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}^\top_k$ and $\mathbf{h}^*_i = \boldsymbol{\mu}_{y_i}$.*

This result follows directly from Lemma 3. Specifically, in the balanced case, $\alpha_{\mathrm{maj}} = \alpha_{\mathrm{minor}}$, $\beta_{\mathrm{maj}} = \beta_{\mathrm{minor}}$, and the optimal $\mathbf{G}^*$ satisfies,

$$\mathbf{G}^* = (\alpha - \beta)\,\mathbb{I}_n + ((\beta - \theta)\mathbb{I}_k + \theta\mathbb{J}_k) \otimes \mathbb{J}_{(\frac{n}{k})},$$

for some scalar α, β, $\theta \in \mathbb{R}$. On the other hand, $\mathbf{G}^*$ is feasible if it is positive semi-definite and its diagonal entries are less than $1/\tau$. Thus, we can show solving UFM (3) reduces to optimizing a linear program as follows

$$(\alpha^*, \beta^*, \theta^*) = \arg\min_{\alpha, \beta, \theta} \theta - \beta, \qquad (7)$$

$$\text{subj. to} \quad \frac{1}{\tau} - \alpha \geq 0,$$
$$\alpha - \beta \geq 0,$$
$$\alpha + \beta(\frac{n}{k} - 1) - \theta\frac{n}{k} \geq 0,$$
$$\alpha + \beta(\frac{n}{k} - 1) + \theta(n - \frac{n}{k}) \geq 0.$$

From above, we can see $\alpha^* = \beta^* = 1/\tau$ and $\theta^* = -1/(\tau(k-1))$, and ETF is the only solution. However, this is not the case if the training set is imbalanced:

Proposition 2. *[22, Lemma D.2] If classes in the training set are not balanced, the global solution of (3) is not an ETF.*

Overall, in balanced scenarios, as also previously shown [5], [7], SC learns the same embeddings as the CE loss. Similarly, the optimal solutions of the SC loss change with imbalances. We leave further investigations for comparing the solutions in the imbalanced case for CE and SC loss for future work.

IV. DISCUSSION

In this paper, complementing the claims and empirical observations in [7], [31], we show that the SC loss has no spurious local optimal solutions under the UFM despite being non-convex. Although we leave the characterization of global solutions for general training distributions to future works, we exploit how the convex relaxation of the problem can be used for uncovering the properties of the unique global minimizer. We believe studying the SC loss is an intermediary step for expanding our understanding of embeddings geometry to the less-understood self-supervised setting. While our focus is on discovering the local/global optimal embeddings geometries in the supervised contrastive learning, it is interesting to study the optimization dynamics analytically as a future direction. Previous works have empirically investigated the degree of convergence of the UFM and benchmark deep models to the optimal geometries for both the CE [4], [15], [20] and the SC [7], [22] loss. Based on these reports, the training setup, such as the imbalance ratio or the loss hyper-parameters, play a significant role on the degree of convergence to the theoretical prediction by UFM. We hypothesize that the training setup can lead to more complex loss landscape, but gradient descent eventually leads to the global solutions asymptotically in the number of epochs. This conjecture remains to be further investigated and formalized.

REFERENCES

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[2] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 776–794.

[3] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[4] V. Papyan, X. Han, and D. L. Donoho, "Prevalence of neural collapse during the terminal phase of deep learning training," *Proceedings of the National Academy of Sciences*, vol. 117, no. 40, pp. 24 652–24 663, 2020.

[5] C. Fang, H. He, Q. Long, and W. J. Su, "Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training," *Proceedings of the National Academy of Sciences*, vol. 118, no. 43, 2021.

[6] T. Galanti, A. György, and M. Hutter, "On the role of neural collapse in transfer learning," *arXiv preprint arXiv:2112.15121*, 2021.

[7] F. Graf, C. Hofer, M. Niethammer, and R. Kwitt, "Dissecting supervised constrastive learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 3821–3830.

[8] X. Han, V. Papyan, and D. L. Donoho, "Neural collapse under mse loss: Proximity to and dynamics on the central path," in *International Conference on Learning Representations*, 2021.

[9] L. Hui, M. Belkin, and P. Nakkiran, "Limitations of neural collapse for understanding generalization in deep learning," *arXiv preprint arXiv:2202.08384*, 2022.

[10] W. Ji, Y. Lu, Y. Zhang, Z. Deng, and W. J. Su, "An unconstrained layer-peeled perspective on neural collapse," *arXiv preprint arXiv:2110.02796*, 2021.

[11] J. Lu and S. Steinerberger, "Neural collapse with cross-entropy loss," *arXiv preprint arXiv:2012.08465*, 2020.

[12] D. G. Mixon, H. Parshall, and J. Pi, "Neural collapse with unconstrained features," *arXiv preprint arXiv:2011.11619*, 2020.

[13] T. Tirer and J. Bruna, "Extended unconstrained features model for exploring deep neural collapse," *arXiv preprint arXiv:2202.08087*, 2022.

[14] L. Xie, Y. Yang, D. Cai, D. Tao, and X. He, "Neural collapse inspired attraction-repulsion-balanced loss for imbalanced learning," *arXiv preprint arXiv:2204.08735*, 2022.

[15] Z. Zhu, T. Ding, J. Zhou, X. Li, C. You, J. Sulam, and Q. Qu, "A geometric analysis of neural collapse with unconstrained features," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[16] J. Zhou, X. Li, T. Ding, C. You, Q. Qu, and Z. Zhu, "On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features," *arXiv preprint arXiv:2203.01238*, 2022.

[17] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, "The implicit bias of gradient descent on separable data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2822–2878, 2018.

[18] Z. Ji and M. Telgarsky, "Risk and parameter convergence of logistic regression," *arXiv preprint arXiv:1803.07300*, 2018.

[19] K. Lyu and J. Li, "Gradient descent maximizes the margin of homogeneous neural networks," *arXiv preprint arXiv:1906.05890*, 2019.

[20] C. Thrampoulidis, G. R. Kini, V. Vakilian, and T. Behnia, "Imbalance trouble: Revisiting neural-collapse geometry," *arXiv preprint arXiv:2208.05512*, 2022.

[21] T. Behnia, G. R. Kini, V. Vakilian, and C. Thrampoulidis, "On the implicit geometry of cross-entropy parameterizations for label-imbalanced data," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 10 815–10 838.

[22] G. R. Kini, V. Vakilian, T. Behnia, J. Gill, and C. Thrampoulidis, "Supervised-contrastive loss learns orthogonal frames and batching matters," *arXiv preprint arXiv:2306.07960*, 2023.

[23] J. Zhou, C. You, X. Li, K. Liu, S. Liu, Q. Qu, and Z. Zhu, "Are all losses created equal: A neural collapse perspective," *arXiv preprint arXiv:2210.02192*, 2022.

[24] T. Galanti, A. György, and M. Hutter, "Generalization bounds for transfer learning with pretrained classifiers," *arXiv preprint arXiv:2212.12532*, 2022.

[25] W. Jitkrittum, A. K. Menon, A. S. Rawat, and S. Kumar, "Elm: Embedding and logit margins for long-tail learning," *arXiv preprint arXiv:2204.13208*, 2022.

[26] D. Samuel and G. Chechik, "Distributional robustness loss for long-tail learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9495–9504.

[27] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," *arXiv preprint arXiv:2011.01403*, 2020.

[28] B. Kang, Y. Li, S. Xie, Z. Yuan, and J. Feng, "Exploring balanced feature spaces for representation learning," in *International Conference on Learning Representations*, 2021.

[29] T. Li, P. Cao, Y. Yuan, L. Fan, Y. Yang, R. S. Feris, P. Indyk, and D. Katabi, "Targeted supervised contrastive learning for long-tailed recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6918–6928.

[30] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9929–9939.

[31] C. Yaras, P. Wang, Z. Zhu, L. Balzano, and Q. Qu, "Neural collapse with normalized features: A geometric analysis over the riemannian manifold," *arXiv preprint arXiv:2209.09211*, 2022.

[32] D. P. Bertsekas, "Nonlinear programming," 1999.

[33] J. Nocedal and S. J. Wright, *Numerical optimization*. Springer, 1999.