Imbalance Trouble: Revisiting Neural-Collapse Geometry

Anonymous Author(s)
Affiliation
Address
email

Abstract

Neural Collapse refers to the remarkable structural properties characterizing the geometry of class embeddings and classifier weights, found by deep nets when trained beyond zero training error. However, this characterization only holds for balanced data. Here we thus ask whether it can be made invariant to class imbalances. Towards this end, we adopt the unconstrained feature model (UFM), a recent theoretical model for studying neural collapse, and introduce *Simplex-Encoded-Labels Interpolation* (SELI) as an invariant characterization of the neural collapse phenomenon. Specifically, we prove for the UFM with cross-entropy loss and vanishing regularization that, irrespective of class imbalances, the embeddings and classifiers always interpolate a simplex-encoded label matrix and that their individual geometries are determined by the SVD factors of this same label matrix. We then present extensive experiments on synthetic and real datasets that confirm convergence to the SELI geometry. However, we caution that convergence worsens with increasing imbalances. We theoretically support this finding by showing that unlike the balanced case, when minorities are present, ridge-regularization plays a critical role in tweaking the geometry. This defines new questions and motivates further investigations into the impact of class imbalances on the rates at which first-order methods converge to their asymptotically preferred solutions.

1 Introduction

What are the unique structural properties of models learned by training deep nets to zero train-error? Is there an *implicit bias* towards solutions of certain geometry? How does this vary across train instances, architectures, and data? These questions are at the core of understanding the optimization landscape of deep-nets. Also, they are naturally informative about the role of models since different parameterizations might affect preferred geometries. Ultimately, such understanding makes progress towards explaning generalization.

Recently, remarkable new progress in answering these questions has been made by Papyan et al. [18], who empirically discover and formalize the so-called *Neural-collapse* (NC) phenomenon. NC describes geometric properties of the learned embeddings (aka last-layer features) and of the classifier weights of deep-nets, trained with cross-entropy (CE) loss and *balanced* data far into the zero train-error regime. The NC phenomenon produces a remarkably simple description of a particularly symmetric geometry: (i) The embeddings of each class collapse to their class mean (see **(NC)** property); and (ii) The class means align with the classifier weights and they form a simplex equiangular tight frame (see **(ETF)** property). Importantly, as noted by the authors of Papyan et al. [18], this simple geometry appears to be "cross-situational *invariant*" across different architectures and different *balanced* datasets.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Geometry of global minimizers for UFM

	UF-RidgeCE (Eqn. (1))	**UF-SVM** (Eqn. (2))
Bal.	**ETF** [14, 4, 24, 3]	**ETF** [9], **[Cor. 3]**
Imbal.	$\forall\lambda$: **NO SELI [Prop. 1]** $\lambda \to 0$: **SELI [Prop. 2]** $\lambda < \frac{1}{k}$: **NO minority collapse [SM]**	**SELI [Thm. 1]**

Table 1: Contributions and comparison to related work.

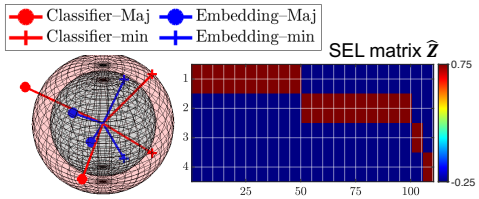


Figure 1: SELI geometry.

In this paper, we study Neural collapse with imbalanced classes: *Is there a (ideally equally simple) description of the geometry that is invariant across class-imbalanced datasets?*

Contributions. We propose a new description capturing the geometric structure of learned-embeddings and classifier-weights on possibly *class-imbalanced data*, which we call the Simplex-Encoded-Labels Interpolation **(SELI)** geometry. This new geometry is a generalization of the ETF geometry. We show that it recovers the latter when data are balanced. Also, unlike ETF, we argue that the new characterization remains invariant across different *imbalance levels*. Importantly, it too, has a simple description: The matrix of learned logits interpolates a simplex-encoded label (SEL) matrix $\hat{\mathbf{Z}}$, and, the individual geometries of the embeddings and classifiers are determined by the SVD factors of this same SEL matrix. Because the particular arrangement of columns of the SEL matrix changes with the imbalance level, this also impacts the geometric arrangement of embeddings / classifiers.

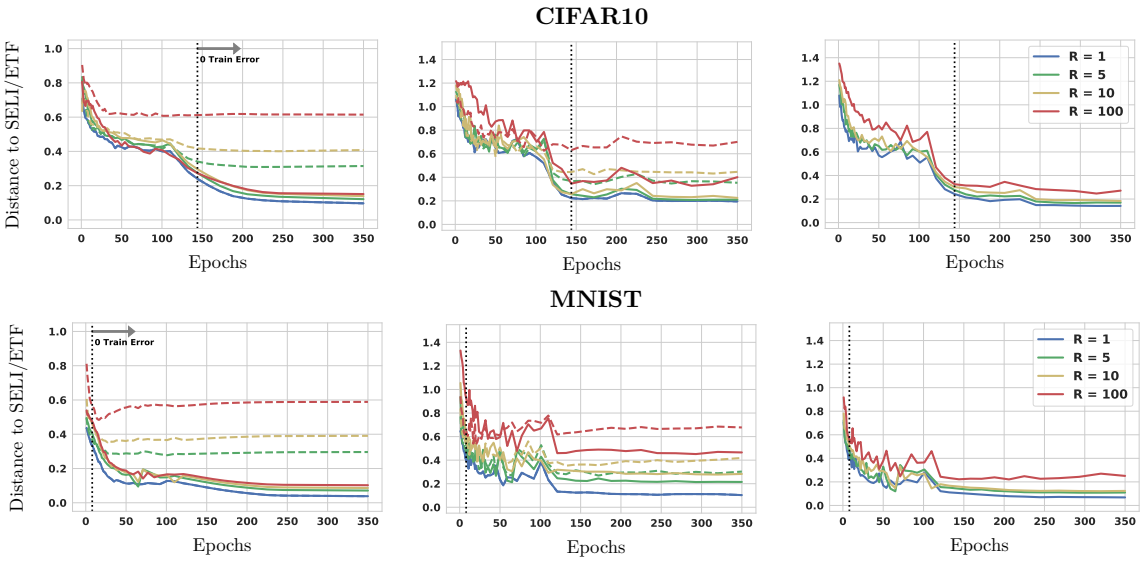
We use an example to illustrate this. Fig. 1 depicts the SEL matrix $\hat{\mathbf{Z}}$ for a STEP-imbalanced $k = 4$-class dataset with two majority classes of 50 examples each, and, two minority classes of 5 examples each. Each column of $\hat{\mathbf{Z}}$ includes the k learned logits for each example in the dataset. Each such column has exactly one entry equal to $1 - 1/k = 0.75$ and three entries equal to $-1/k = -0.25$. The corresponding geometry of the embeddings and classifiers, shown in the 3D plot, is found by an SVD of $\hat{\mathbf{Z}}$: the left eigenvectors determine the classifiers and the right ones the embeddings. Note that $\hat{\mathbf{Z}}$ is rank 3, hence the geometry is 3D. Since embeddings collapse to their class means (see **(NC)** property), we only show the four class mean-embeddings and the corresponding four classifiers. Two of each correspond to majorities ("•" marker) and two to minorities ("+" marker). Note the classifiers are not aligned with the embeddings as in balanced case, and norms are also different.

Below we explain the conception of this geometry and our contributions in detail. The initial major challenge was: *Assuming a class-imbalance-invariant geometry exists, how to find it?*

To answer this question, we adopted the *Unconstrained Feature Model (UFM)* previously introduced in the literature as a two-layer proxy model to theoretically justify neural collapse [17, 3, 24]. Motivated by deep-learning practice and by studies on implicit bias of GD for unregularized CE minimization, we analyze the geometry of solutions to an unconstrained-features Support Vector Machines (UF-SVM) problem. We prove, for STEP-imbalanced data, any solution of the UF-SVM follows the SELI geometry. Thus, the learned end-to-end model always interpolates a simplex labels encoding. We show that **(ETF)** $\Longrightarrow$ **(SELI)**. However, **(SELI)** $\not\Longrightarrow$ **(ETF)** unless classes are balanced or just two ($k = 2$).

Next, we analyze training of the UFM with ridge-regularized CE. Unlike previous studies, we find in the presence of imbalances regularization matters as it changes the geometry of solutions. In fact, we show that there is *no* finite regularization that leads to the SELI geometry. However, we also show that as regularization vanishes, the solutions do interpolate the SEL matrix (after appropriate normalization). Finally, we show that the SELI geometry differs from the minority-collapse phenomenon [3], since the latter does not correspond to solutions with zero training error. In fact, we show that minority collapse does *not* occur for small regularization. (See Fig. 3b and SM due to space constraints.)

We numerically test convergence to the proposed SELI geometry in both synthetic and real class-imbalanced datasets. For different imbalance levels, the learned geometries approach the SELI geometry significantly faster compared to the ETF geometry; Fig. 2 for illustration. However, we also observe that convergence *worsens* with increasing level of imbalance. A plausible theoretical justification is that as we show regularization plays critical role under imbalances. We also consistently get better convergence rate for the classifiers. We believe



(a) Classifiers convergence **(b)** Embeddings convergence **(c)** Logits convergence

Figure 2: Convergence of learned classifiers, embeddings and corresponding logits to the SELI (solid lines) vs ETF (dashed lines) geometries for **STEP-Imbalanced** CIFAR10 and MNIST data trained far beyond zero training error, for different imbalance levels R; see Sec. 5 for metrics and discussion.

our observations strongly motivate further investigations regarding potential frailties of "asymptotic" implicit bias characterizations.

1.1 Related works

The original contribution by Papyan et al. [18] has attracted lots of attention resulting in numerous followups within short time period, e.g., [24, 9, 3, 7, 14, 17, 4, 23]. (See also [7, Sec. E] for a review of the recent literature.) Several works have proposed and/or used the UFM with CE training to analyze theoretical abstractions of NC [24, 9, 3, 4]. We drove particular inspiration from Zhu et al. [24], who presented a particularly transparent and complete analysis of the optimization landscape of ridge-regularized CE minimization for the UFM under balanced data; see Remark 2. In the same spirit, we also relied on the UFM. However, our work is, to the best of our knowledge, the first explicit geometry analysis for class-imbalanced data. See also Table 1 for a comparison.

The only previous work on neural collapse with imbalances is [3], which was the first to note that collapse of the embeddings is preserved, but otherwise the geometry might skew away from ETF. Also, Fang et al. [3] first proposed studying the new geometry using the UFM and appropriate convex relaxations. With this setup, they presented an intriguing finding, which they termed *minority collapse*: as the imbalance level increases, the minorities' classifiers collapse to the same vector. As mentioned above, our focus is on the, particularly relevant for deep-learning practice, zero-training error scenarios. This excludes minority collapse by definition. More importantly, we derive an explicit geometric characterization of *both* embeddings and classifiers for *both* majorities and minorities and for *all* imbalance levels.

Our results and techniques also relate to the literatures on implicit bias and matrix factorization. Due to space constraints, we defer discussions to the SM.

2 Problem setup

We adopt the *unconstrained feature model* (UFM) [17, 3] in a k-class classification setting. Let $\mathbf{W}_{d\times k} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_k]$ be the matrix of classifier weights corresponding to the k classes. Here, d is the feature dimension. We assume throughout that $d \geq k-1$. Next, we let $\mathbf{H}_{d\times n} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n]$ denote a matrix of n feature embeddings, each corresponding to a different example in the training set. We assume each class $c \in [k]$ has $n_c \geq 1$ examples (thus, n_c embeddings) so that $\sum_{c\in[k]} n_c = n$. Without loss of generality, we assume examples are ordered. Formally, defining $n_0 = 0$, we assume that examples $i = 1, \ldots, n_1$ have labels $y_i = 1$, examples $i = n_1+1, \ldots, n_1+n_2$ have labels $y_i = 2$, and so on. The UFM trains the

features $\mathbf{h}_i, i \in [n]$ (jointly with the weights $\mathbf{w}_c, c \in [k]$) without any further constraints, i.e., by minimizing the Ridge-regularized cross-entropy (CE) loss as follows [24]:

$$(\hat{\mathbf{W}}_\lambda, \hat{\mathbf{H}}_\lambda) := \arg\min_{\mathbf{W},\mathbf{H}} \ \mathcal{L}(\mathbf{W}^T\mathbf{H}) + \frac{\lambda}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda}{2}\|\mathbf{H}\|_F^2, \tag{1}$$

where $\mathcal{L}(\mathbf{W}^T\mathbf{H}) := \sum_{i\in[N]} \log\left(1 + \sum_{c\neq y_i} e^{-(\mathbf{w}_{y_i}-\mathbf{w}_c)^T\mathbf{h}_i}\right)$ is the CE loss.

UFM as two-layer linear net. The formulation above does not explicitly specify inputs for each example. An alternative view follows by considering training a 2-layer linear net with hidden dimension d, first layer $\mathbf{H}$, and second layer $\mathbf{W}$, over n examples with inputs $\mathbf{x}_i = \mathbf{e}_i, \in [n]$ and labels y_i as above: $y_i = 1$ for $i = 1, \ldots, n_1$, $y_i = 2$ for $i = n_1 + 1, \ldots, n_1 + n_2$, ...

2.1 Unconstrained-features SVM (UF-SVM)

Since neural-collapse is observed when training with small / vanishing regularization, it is reasonable to consider an unregularized version of (1). In this special case, gradient descent (with sufficiently small step size) on (1) produces iterates that diverge in norm, but converge in direction [15, 9]. In fact, it has been recently shown that the GD solutions converge in direction to a KKT point of the following max-margin classifier [15, 9]:

$$(\hat{\mathbf{W}}, \hat{\mathbf{H}}) \in \arg\min_{\mathbf{W},\mathbf{H}} \ \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{1}{2}\|\mathbf{H}\|_F^2 \qquad \text{sub. to} \quad (\mathbf{w}_{y_i} - \mathbf{w}_c)^T\mathbf{h}_i \geq 1, \ i \in [n], c \neq y_i. \tag{2}$$

For convenience, we refer to the optimization problem in (2) as Unconstrained-features SVM (UF-SVM). Note that this minimization (unlike 'standard' SVM) is non-convex. Hence, KKT points (thus, GD convergence directions) are not necessarily global minimizers; see Sec. 6.

2.2 Class-imbalance model

To streamline the presentation, we focus on the so-called STEP imbalances. This includes balanced data as special case by setting $R = 1$.

Definition 1 ((R,ρ)-STEP imbalance)**.** *In a (R,ρ)-STEP imbalance setting with label-imbalance ratio $R \geq 1$ and minority fraction $\rho \in (0,1)$, the following hold. All minority (resp. majority) classes have the same sample size n_{min} (resp. Rn_{min}). There are ρk minority classes. Without loss of generality, we assume classes $\{1, \ldots, (1-\rho)k\}$ are majorities.*

Notation. For matrix $\mathbf{V} \in \mathbb{R}^{m\times n}$, $\mathbf{V}[i,j]$ denotes its (i,j)-th entry, $\mathbf{v}_j$ denotes the j-th *column*, and $\mathbf{V}^T$ its transpose. We denote $\|\mathbf{V}\|_F, \|\mathbf{V}\|_2$, and, $\|\mathbf{V}\|_*$ the Frobenius, spectral, and, nuclear norms of $\mathbf{V}$. $\mathrm{tr}(\mathbf{V})$ denotes the trace of $\mathbf{V}$. $\odot$ denotes Hadammard product. $\nabla_{\mathbf{V}}\mathcal{L} \in \mathbb{R}^{m\times n}$ is the gradient of a scalar function $\mathcal{L}(.)$ with respect to $\mathbf{V}$. We use $\mathbb{1}_m$ to denote an m-dimensional vector of all ones and $\mathbb{I}_m$ for the m-dimensional identity matrix. For vectors/matrices with all zero entries, we simply write 0, as dimensions are easily understood from context. $\mathbf{e}_j$ denotes a column with a single non-zero entry of 1 in the j-th entry.

3 Global structure of the UF-SVM

In this section, we characterize the global minimizers of the non-convex program in (2). Perhaps surprisingly, we show that they take a particularly simple form that is best described in terms of a *simplex-encoding* of the labels.

Definition 2 (SEL matrix)**.** *The* simplex-encoding label (SEL) matrix $\hat{\mathbf{Z}}_{k\times n}$ *is such that* $\forall c \in [k], i \in [n]: \ \hat{\mathbf{Z}}[c,i] = \begin{cases} 1 - 1/k & , c = y_i \\ -1/k & , c \neq y_i \end{cases}$. *Let $\hat{\mathbf{Z}}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ be the compact SVD of $\hat{\mathbf{Z}}^T$; $\mathbf{\Lambda}$ is a positive $(k-1)$-dim. diagonal matrix and $\mathbf{U}_{n\times(k-1)}$, $\mathbf{V}_{k\times(k-1)}$ have orthonormal columns.*

Column $\hat{\mathbf{z}}_i \in \mathbb{R}^k$ of $\hat{\mathbf{Z}}$ represents class-membership encoding of datapoint $i \in [n]$. This differs from the vanilla one-hot encoding $\hat{\mathbf{y}}_i = \mathbf{e}_{y_i}$; indeed, $\hat{\mathbf{z}}_i = \hat{\mathbf{y}}_i - \frac{1}{k}\mathbb{1}_k$. $\hat{\mathbf{Z}}$ has exactly k different and affinely independent columns, which together with the zero vector form a k-dimensional simplex, motivating the SEL name. Finally, note that $\hat{\mathbf{Z}}^T\mathbb{1}_k = 0$; thus, $\mathrm{rank}(\hat{\mathbf{Z}}) = k - 1$.

Theorem 1 (Structure of the UF-SVM minimizers)**.** *Suppose $d \geq k-1$ and a (R,ρ)-STEP imbalance setting. Let $(\hat{\mathbf{W}}, \hat{\mathbf{H}})$ be any solution and p_* the optimal cost of the the UF-SVM in (2). Then, $\mathrm{p}_* = \|\hat{\mathbf{Z}}\|_* = \|\hat{\mathbf{H}}\|_F^2 = \|\hat{\mathbf{W}}\|_F^2$. Moreover, the following statements characterize the geometry of global minimizers in terms of the the SEL matrix and its SVD.*

(i) For the optimal logits it holds that $\hat{\mathbf{W}}^T\hat{\mathbf{H}} = \hat{\mathbf{Z}}$.

(ii) The Gram matrices satisfy $\hat{\mathbf{H}}^T\hat{\mathbf{H}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ and $\hat{\mathbf{W}}^T\hat{\mathbf{W}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$.

(iii) For some partial orthonormal matrix $\mathbf{R} \in \mathbb{R}^{(k-1)\times d}$, $\hat{\mathbf{W}} = \mathbf{R}^T\mathbf{\Lambda}^{1/2}\mathbf{V}^T$ and $\hat{\mathbf{H}} = \mathbf{R}^T\mathbf{\Lambda}^{1/2}\mathbf{U}^T$.

We outline the proof in Sec. 3.2 (SM for details). The theorem provides an explicit characterization of the geometry of optimal embeddings and classifiers centered around the the optimal logit matrix being equal to the SEL matrix. We highlight the following key features.

Simplicity. The lack of symmetry in the imbalanced setting, makes it a priori unclear whether a simple geometry description is still possible, as in the balanced case. But, the theorem shows this to be the case. The key observation is that the optimal logit matrix $\hat{\mathbf{W}}^T\hat{\mathbf{H}}$ equals $\hat{\mathbf{Z}}$ (cf. Statement (i)). Then, the gram matrices of embeddings and classifiers are given simply in terms of the singular factors of the SEL matrix (cf. Statements (ii),(iii)).

Invariance to imbalances. The theorem's characterization is valid for all types of (R,ρ)-STEP imbalances. In particular, equality of the optimal logit matrix to the SEL matrix is the key invariant characterization across changing imbalances. This also implies that at optimality all margins are equal irrespective of the imbalance type. The description of Gram matrices in terms of the SVD of $\hat{\mathbf{Z}}$ is also invariant. Of course, the particular arrangement of columns of $\hat{\mathbf{Z}}$ itself depends on the (R,ρ). In turn, the singular factors determining the geometry of embeddings and classifiers depend implicitly on the same parameters. Thus, as we show next, the geometry differs for different imbalance levels; see Fig. 1 for an example.

3.1 Invariant properties: NC and SELI

Here, we further discuss the geometry of embeddings and classifiers induced by the SVD of the SEL matrix. The first realization is that the embeddings collapse under *all* settings.

Corollary 1. *The UF-SVM solutions satisfy the following property irrespective of imbalance:*

(NC) *Embeddings collapse to their class means:* $\hat{\mathbf{h}}_i = \hat{\boldsymbol{\mu}}_c := \frac{1}{n_c}\sum_{j:y_j=c}\hat{\mathbf{h}}_j$, $\forall c \in [k]$, $i : y_i = c$.

This statement can be inferred from Theorem 1 (specifically from (iii) and that $\mathbf{U}$ has repeated columns). A more straightforward argument is possible by directly inspecting the UF-SVM optimization in (2). For any fixed (say optimal) $\hat{\mathbf{W}}$, the minimization over $\mathbf{h}_i$ is: (i) separable and identical for all $i : y_i = c$ in same class c; (ii) strongly convex. Hence, for all $i : y_i = c$, there is unique minimizer corresponding to the fixed $\hat{\mathbf{W}}$; this must be their class mean. Beyond (NC), Thm. 1 specifies the exact geometry of solutions.

Definition 3 (SELI geometry)**.** *The embedding and classifier matrices $\mathbf{H}_{d\times n}$ and $\mathbf{W}_{d\times k}$ follow the simplex-encoded-labels interpolation geometry when for some scaling $\alpha > 0$:*

$$\textbf{(SELI)} \quad \begin{bmatrix}\mathbf{W}^T\\ \mathbf{H}^T\end{bmatrix}[\mathbf{W} \quad \mathbf{H}] = \alpha\begin{bmatrix}\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T & \hat{\mathbf{Z}}\\ \hat{\mathbf{Z}}^T & \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\end{bmatrix}, \quad \textit{where } \hat{\mathbf{Z}} = \mathbf{V}\mathbf{\Lambda}\mathbf{U}^T \textit{ is the SEL matrix.}$$

The following is essentially a restatement of Thm. 1.

Corollary 2. *The UF-SVM solutions follow the SELI geometry, irrespective of imbalance.*

For the UFM, $\hat{\mathbf{W}}^T\hat{\mathbf{H}}$ is the learned end-to-end model. Since SELI geometry implies $\mathbf{W}^T\mathbf{H} = \alpha\hat{\mathbf{Z}}$, it means the learned model *interpolates* the SEL matrix motivating our name in Defn. 3.

3.1.1 Special cases

For the special cases of balanced or binary data we show Thm. 1 recovers the ETF structure; thus **(SELI)**≡**(ETF)**. Let $\hat{\mathbf{M}} = [\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_k]$ denote the matrix of mean embeddings.

Corollary 3. *Assume balanced data ($R = 1$) or binary classification ($k = 2$). Then, any UF-SVM solution $(\hat{\mathbf{W}}, \hat{\mathbf{H}})$ follows the ETF geometry as defined in [18]:*

(ETF) $\hat{\mathbf{W}} = \hat{\mathbf{M}}$ *and* $\hat{\mathbf{M}}^T\hat{\mathbf{M}} = \hat{\mathbf{W}}^T\hat{\mathbf{W}} = \mathbb{I}_k - \frac{1}{k}\mathbb{1}_k\mathbb{1}_k^T$.

For all other cases, **(SELI)** $\not\Longrightarrow$ **(ETF)**. In those cases it is in fact possible to obtain explicit closed-form formulas for the magnitudes and angles of the SELI geometry in terms of $R, k, n_{\min}$. The following lemma is an example. See SM for additional details.

Lemma 1. *Assume $(R, 1/2)$-STEP imbalance. Suppose $(\mathbf{W}, \mathbf{H})$ satisfies the* **(SELI)** *property. Let $\mathbf{w}_{maj}, \mathbf{h}_{maj}$ (resp. $\mathbf{w}_{min}, \mathbf{h}_{min}$) denote majority (resp. minority) classifiers and embeddings, respectively. Then,*

$$\frac{\|\mathbf{w}_{maj}\|_2^2}{\|\mathbf{w}_{min}\|_2^2} = \frac{(1-2/k)\sqrt{R} + \frac{\sqrt{(R+1)/2}}{k}}{(1-2/k) + \frac{\sqrt{(R+1)/2}}{k}} \quad \textit{and} \quad \frac{\|\mathbf{h}_{maj}\|_2^2}{\|\mathbf{h}_{min}\|_2^2} = \frac{\frac{1}{\sqrt{R}}(1-2/k) + \frac{1}{k\sqrt{(R+1)/2}}}{(1-2/k) + \frac{1}{k\sqrt{(R+1)/2}}}. \tag{3}$$

Thus, $\|\mathbf{w}_{maj}\|_2 \geq \|\mathbf{w}_{min}\|_2$ and $\|\mathbf{h}_{maj}\|_2 \leq \|\mathbf{h}_{min}\|_2$, with equality if and only if $R = 1$ or $k = 2$.

The fact that CE learns larger-norm classifiers for majorities has been empirically observed in the deep imbalanced-learning literature [13, 12] and Lemma 1 provides a plausible theoretical justification for it. Moreover, it specifies the norm-ratio between majorities and minorities, not only for classifiers, but also for the learned embeddings.

3.2 Proof sketch

We start from the following (by now, relatively standard) convex relaxation of the UF-SVM [20, 6, 24, 3]:

$$\min_{\mathbf{Z}\in\mathbb{R}^{k\times n}} \|\mathbf{Z}\|_* \quad \text{subj. to} \quad \mathbf{Z}[y_i, i] - \mathbf{Z}[c, i] \geq 1,\ \forall c \neq y_i, i \in [n]. \tag{4}$$

The relaxation follows by setting $\mathbf{Z} = \mathbf{W}^T\mathbf{H}$, thus $\mathbf{Z}$ is the logit matrix (also, the end-to-end model) of the non-convex UF-SVM. Our key technical innovation is proving that $\hat{\mathbf{Z}}$ is the unique minimizer of (4). There are three key ingredients in this. First, is a clever re-parameterization of the dual program to (4), introducing the SEL matrix $\hat{\mathbf{Z}}$ in the dual:

$$\max_{\mathbf{B}\in\mathbb{R}^{n\times k}} \quad \mathrm{tr}(\hat{\mathbf{Z}}\mathbf{B}) \quad \text{sub. to } \|\mathbf{B}\|_2 \leq 1, \quad \mathbf{B}\mathbb{1}_k = 0, \quad \mathbf{B} \odot \hat{\mathbf{Z}}^T \geq 0\,. \tag{5}$$

Second, we prove that $\hat{\mathbf{B}} = \mathbf{U}\mathbf{V}^T$ is the unique maximizer of the re-parameterized dual problem in (5). While it is not hard to check that $\hat{\mathbf{B}}$ optimizes a relaxation of (5), it is far from obvious that $\hat{\mathbf{B}}$ is unique, and, even more that it satisfies the third constraint. The key technical challenge here is that the third constraint acts entry-wise on $\mathbf{B}$. In fact, to proceed with the proof we need that the constraint is *not* active, i.e. $\hat{\mathbf{B}} \odot \hat{\mathbf{Z}}^T > 0$, or equivalently, that the sign pattern of the entries of $\hat{\mathbf{B}}$ agrees with the sign pattern of the transpose SEL matrix $\hat{\mathbf{Z}}^T$. We prove this by an explicit construction of the singular factors $\mathbf{U}, \mathbf{V}$ exploiting the structure of the SEL matrix. Once we have shown that $\hat{\mathbf{B}}$ is the unique maximizer and is strictly feasible, we use the KKT conditions to prove that $\hat{\mathbf{Z}}$ is the unique minimizer of the nuclear-norm relaxation in (4). To do this, we leverage that strict feasibility of $\hat{\mathbf{B}}$ implies by complementary slackness all constraints in the primal (4) must be active at the optimum. The proof of the theorem completes by arguing that the relaxation (5) is tight when $d \geq k-1$ allowing us to connect the UF-SVM minimizers to the SEL matrix $\hat{\mathbf{Z}}$.

Remark 1. *The common analysis strategy in* all *previous related works is deriving tight bounds on the CE loss (or related quantities, such as the minimum margin), and, then identifying the structure in the parameters that achieves those bounds. For example, [24, 4, 3] lower bound the CE loss and [9] upper bounds the minimum margin, all using a similar elegant argument based on Cauchy-Schwartz and Jensen inequalities. The ETF geometry is then uncovered by recognizing that it uniquely achieves those bounds. It is not clear how to employ such exercises in the presence of imbalances, due to the absence of symmetry properties (e.g. alignment of classifiers with embeddings). Our proof of Theorem 1 is more direct and is in large enabled by identifying the key role played by the logit matrix.*

4 The role of regularization

In this section, we focus on the regularized CE loss minimization in (1). Specifically, we study how the geometry of solutions $(\hat{\mathbf{W}}_\lambda, \hat{\mathbf{H}}_\lambda)$ depends on the imbalance and on the regularization.

4.1 Global minimizers as solutions to a convex relaxation

The regularized CE minimization in (1) is non-convex. Yet, its landscape is benign and the global solution can be described in terms of the solution to a convex relaxation program.

Theorem 2 (Reformulated from [24])**.** *Let $\lambda > 0$, $d \geq k-1$ and a (R,ρ)-STEP imbalance setting. Let $\hat{\mathbf{Z}}_\lambda \in \mathbb{R}^{k\times n}$ be the unique minimizer of the nuclear-norm regularized loss, i.e.,*

$$\hat{\mathbf{Z}}_\lambda := \arg\min_{\mathbf{Z}} \mathcal{L}(\mathbf{Z}) + \lambda \|\mathbf{Z}\|_* \,, \tag{6}$$

and, denote $\hat{\mathbf{Z}}_\lambda = \mathbf{V}_\lambda \mathbf{\Lambda}_\lambda \mathbf{U}_\lambda^T$ its SVD. Any stationary point of (1) satisfies exactly one of the following two. Either it is a strict saddle, or, it is a global minimizer $(\hat{\mathbf{W}}_\lambda, \hat{\mathbf{H}}_\lambda)$ and satisfies

$$\begin{bmatrix} \hat{\mathbf{W}}_\lambda^T \\ \hat{\mathbf{H}}_\lambda^T \end{bmatrix} \begin{bmatrix} \hat{\mathbf{W}}_\lambda & \hat{\mathbf{H}}_\lambda \end{bmatrix} = \begin{bmatrix} \mathbf{V}_\lambda \\ \mathbf{U}_\lambda \end{bmatrix} \mathbf{\Lambda}_\lambda \begin{bmatrix} \mathbf{V}_\lambda^T & \mathbf{U}_\lambda^T \end{bmatrix}. \tag{7}$$

The theorem describes the structure of the global minimizers of (1) in terms of $\hat{\mathbf{Z}}_\lambda$, the solution to the *convex* minimization in (6) (see SM for uniqueness of $\hat{\mathbf{Z}}_\lambda$). Structurally, the above result (cf. (7)) resembles the characterization in Theorem 1 regarding the UF-SVM. However, in Theorem 1 we went a step further and gave an *explicit* form for the logit matrix, namely the SEL matrix $\hat{\mathbf{Z}}$. Instead, $\hat{\mathbf{Z}}_\lambda$ in Thm. 2 is given implicitly as the solution to a convex program. In the remaining of this section, we ask: *how does $\hat{\mathbf{Z}}_\lambda$ compare to $\hat{\mathbf{Z}}$ for different values of the regularizer? How does the answer depend on the imbalance level?*

Remark 2. *Although not stated explicitly in this form, Thm. 2 is essentially retrieved from the proof of [24, Thm. 3.2] with two adjustments. First, Zhu et al. [24] only considers balanced data. Here, we realize their proof goes through in the imbalanced setting with minor adjustments. Second, we relax their assumption $d \geq k$ to $d \geq k-1$ thanks to a simple observation: $\mathbb{1}_k^T \nabla_{\mathbf{Z}} \mathcal{L}(\mathbf{Z}) = \mathbf{0}$, hence the CE gradient drops rank (see SM for details).*

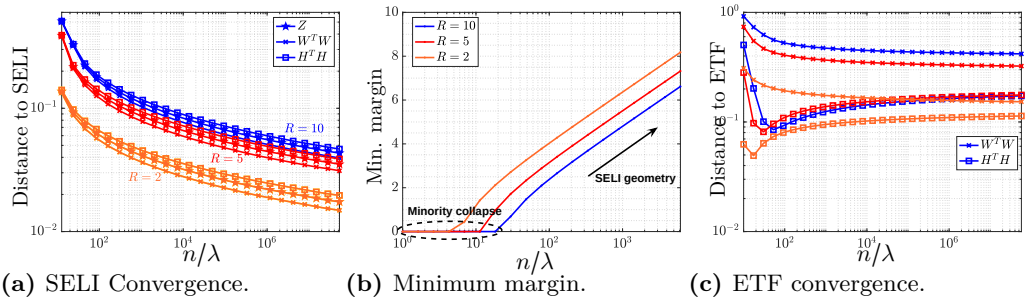
4.2 Regularization matters

For balanced data, previous works have shown that the minimizers $\hat{\mathbf{W}}_\lambda, \hat{\mathbf{H}}_\lambda$ of (1) satisfy the **(NC)** and **(ETF)** properties (up to scaling by a constant) for *every* value of the regularization parameter $\lambda > 0$, e.g. [24]. In our language, **for all $\lambda > 0$, there exists** scalar α_λ such that a scaling $(\alpha_\lambda \hat{\mathbf{W}}_\lambda, \alpha_\lambda \hat{\mathbf{H}}_\lambda)$ of any global solution of the regularized CE minimization in (1) satisfies the ETF geometry. Thus, *for balanced data, up to a global scaling, the geometry is insensitive to $\lambda > 0$ and same as that of the UF-SVM minimizers.*

Here, we show that the situation changes drastically with imbalances: the regularization now plays a critical role and the solution is never the same as that of UF-SVM for finite λ.

Proposition 1 (Imbalanced data: Regularization matters)**.** *Assume imbalanced data* and $k > 2$. *There **does not exist** finite $\lambda > 0$ and corresponding scaling α_λ such that the scaled solution $(\alpha_\lambda \hat{\mathbf{W}}_\lambda, \alpha_\lambda \hat{\mathbf{H}}_\lambda)$ of (1) follows the* **(SELI)** *geometry. Equivalently, there does not exist λ and α_λ such that a scaling of the UF-SVM solution solves (1).*

The proof relies on Thm. 1 as follows. For the sake of contradiction assume for some λ and $\alpha_\lambda > 0$ that the scaled UF-SVM minimizer $(\alpha_\lambda \hat{\mathbf{W}}, \alpha_\lambda \hat{\mathbf{H}}_\lambda)$ solves (1). Then, $\nabla_{\mathbf{W}} \mathcal{L}(\alpha_\lambda^2 \hat{\mathbf{W}}^T \hat{\mathbf{H}}) + \lambda \alpha_\lambda \mathbf{W} = 0 \implies \hat{\mathbf{H}} \left(\nabla_{\mathbf{Z}} \mathcal{L}(\alpha_\lambda^2 \hat{\mathbf{W}}^T \hat{\mathbf{H}})\right)^T = -\lambda \hat{\mathbf{W}}$. But, by Thm. 1: $\hat{\mathbf{W}}^T \hat{\mathbf{H}} = \hat{\mathbf{Z}}$ and $\hat{\mathbf{W}}^T \hat{\mathbf{W}} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T$. Moreover, thanks to the special structure of $\hat{\mathbf{Z}}$ we can check that $\nabla_{\mathbf{Z}} \mathcal{L}(\alpha_\lambda^2 \hat{\mathbf{Z}}) = -\alpha'_\lambda \hat{\mathbf{Z}}$ for $\alpha'_\lambda := k \big/ (\exp(\alpha_\lambda^2) + k - 1)$; see SM. With these, we arrive at the following contradiction: $\alpha'_\lambda \hat{\mathbf{Z}} \hat{\mathbf{Z}}^T = \lambda \hat{\mathbf{W}} \hat{\mathbf{W}}^T \implies \alpha'_\lambda \mathbf{V} \mathbf{\Lambda}^2 \mathbf{V}^T = \lambda \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T \implies \alpha'_\lambda \mathbf{\Lambda}^2 = \lambda \mathbf{\Lambda} \implies \mathbf{\Lambda} = \left(\lambda / \alpha'_\lambda\right) \mathbb{I}_{k-1}$, since for imbalanced data and $k > 2$, *not* all eigenvalues of the SEL matrix are same; see SM.



(a) SELI Convergence. **(b)** Minimum margin. **(c)** ETF convergence.

Figure 3: Numerical investigation of global solutions $(\hat{\mathbf{W}}_\lambda, \hat{\mathbf{H}}_\lambda)$ of (1) across the regularization path.

4.3 Vanishing regularization

As λ vanishes, it is not hard to check that minimizers diverge in norm and the relevant question becomes: where do they converge in direction?

Proposition 2 (Regularization path leads to UF-SVM)**.** *Suppose $d \geq k-1$ and (R,ρ)-STEP imbalance. It holds that* $\lim_{\lambda\to 0} \hat{\mathbf{W}}_\lambda^T \hat{\mathbf{H}}_\lambda / (\|\hat{\mathbf{W}}_\lambda\|_F^2/2 + \|\hat{\mathbf{H}}_\lambda\|_F^2/2) = \hat{\mathbf{Z}}/\|\hat{\mathbf{Z}}\|_\star$

Put together with the content of the previous section: For balanced data, the solution is always the same up to global scaling. However, for imbalanced data, the solution changes with λ and only in the limit of $\lambda \to 0$ does it align with that of the UF-SVM. Regarding the proof of the proposition, we note that thanks to Thms. 1 and 2, it suffices that the solution $\hat{\mathbf{Z}}_\lambda$ of (6) converges in direction to the SEL matrix $\hat{\mathbf{Z}}$. To show this, we critically use from Thm. 1 that $\hat{\mathbf{Z}}$ is *unique* minimizer of (4). The most closely related results are those of [19, 11], who studied the regularization path of p-norm regularized CE.

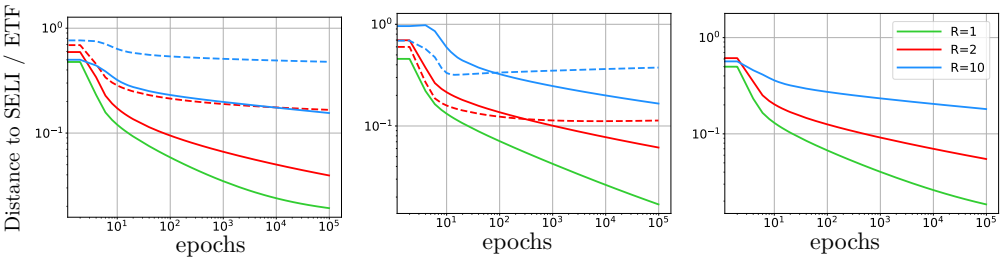
4.4 Imbalance emphasizes the impact of non-convexity

Recall interpreting the UFM as a two-layer linear net trained on the standard basis $\mathbf{e}_i \in \mathbb{R}^n$. Suppose instead that we train a simple k-class linear classifier $\mathbf{\Xi}_{k\times n}$ on the same data by minimizing ridge regularized CE: $\min_{\mathbf{\Xi}} \mathcal{L}(\mathbf{\Xi}) + \frac{\lambda}{2}\|\mathbf{\Xi}\|_F^2$. It is easy to check that (after scaling) $\hat{\mathbf{Z}}$ satisfies first-order optimality conditions. Thus, the optimal linear classifier is such that for all $\lambda > 0$, there exists α_λ such that $\hat{\mathbf{\Xi}}_\lambda = \alpha_\lambda \hat{\mathbf{Z}}$. Contrasting this to Prop. 1, we find that the end-to-end models minimizing ridge-regularized CE for a linear versus a two-layer linear network are the same (in direction) when data are balanced, but differ under imbalances.

5 Experiments

In all our experiments we choose $(R, 1/2)$-STEP imbalances with varying R. Also, in all cases we measure convergence to either the **(SELI)** or the **(ETF)** geometries, in terms of the following three metrics. Denote $\underline{\mathbf{A}} = \mathbf{A}/\|\mathbf{A}\|_F$ the Euclidean normalization and $\mathbf{G}_\mathbf{A} = \mathbf{A}^T\mathbf{A}$ the Gram matrix of matrix $\mathbf{A}$. ◻ **Classifiers:** We measure $\|\underline{\mathbf{G}_\mathbf{W}} - \underline{\hat{\mathbf{G}}_\mathbf{W}}\|_F$, where $\hat{\mathbf{G}}_\mathbf{W}^{\text{ETF}} = \mathbb{I}_k - \frac{1}{k}\mathbb{1}_k\mathbb{1}_k^T =: \mathbf{G}^\star$ and $\hat{\mathbf{G}}_\mathbf{W}^{\text{SELI}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ (see Defn. 3). ◻ **Embeddings:** Because of the NC property, we compute the matrix $\mathbf{M}$ of mean embeddings and measure $\|\underline{\mathbf{G}_\mathbf{M}} - \underline{\hat{\mathbf{G}}_\mathbf{M}}\|_F$, where $\hat{\mathbf{G}}_\mathbf{M}^{\text{ETF}} = \mathbf{G}^\star$ and $\hat{\mathbf{G}}_\mathbf{M}^{\text{SELI}}$ is computed from the n-dimensional $\hat{\mathbf{G}}_\mathbf{H}^{\text{SELI}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ by only keeping the k columns/rows corresponding to the first example of each class. ◻ **Logits:** We measure $\|\underline{\mathbf{W}^T\mathbf{M}} - \underline{\mathbf{G}^\star}\|_F$, which assuming NC holds is same as $\|\underline{\mathbf{W}^T\mathbf{H}} - \underline{\hat{\mathbf{Z}}}\|_F$.

UFM: Global minimizers. Fig. 3 numerically investigates the behavior of global minimizers of regularized CE (1) for the UFM and $k = 4$. Thanks to Thm. 2, we obtained those minimizers by solving the convex program (6) with CVX [5], and then, using (7) to infer the the Gram matrices $\mathbf{G}_\mathbf{W}, \mathbf{G}_\mathbf{M}$ and logits $\mathbf{W}^T\mathbf{H}$. Fig. 3c shows that the distance to ETF is large and not approaching zero for any value of λ. On the other hand, Fig. 3a numerically validates Props. 1 and 2: the distance to SELI for all three metrics is non-zero for any finite $\lambda > 0$, but converges to zero as $\lambda \to 0$. However, this convergence is slow and the rate becomes



(a) Classifiers **(b)** Embeddings **(c)** Logits

Figure 4: Geometry of SGD solutions training unregularized CE on UFM; SELI(Solid)/ETF(Dashed).

even worse as R increases. Finally, Fig. 3b depicts the minimum margins of solutions across λ. Note that for all sufficiently small λ values, the minimum margin is strictly positive. Thus, the "minority collapse" of [3] can only possibly occur for large λ (see SM for a proof).

UFM: SGD solutions. In Fig. 4 we investigate whether the solutions found by SGD are consistent with the prediction of Thm. 1 about global minimizers. In the experiments, we fix $k = 4$ and implement the UFM using the 2-layer network perspective in Sec.. 2. The weights are optimized using SGD with learning rate 0.5, batch size 32 and no weight decay. We train for 10^5 epochs, much beyond 0 training error; see SM for more details. We make three observations regarding the figures. **(i)** SGD iterates favor the SELI, instead of the ETF geometry, suggesting an implicit bias towards global minimizers of the UF-SVM. **(ii)** However, convergence is rather slow and rates get worse with increasing imbalance. **(iii)** Also, the embeddings convergence is more elusive compared to that of the classifiers. Interestingly, the last two observations are reminiscent of the trends we observed in Fig. 3a, suggesting a connection between regularization path and (S)GD iterates, worth investigating further.

Deep-learning experiments. In Fig. 2 we test our proposed SELI geometry in imbalanced MNIST and CIFAR10 training for varying values of R. For both datasets we keep the same total number of $n = 100 \times 50 \times 5 + 50 \times 5 = 25250$ examples across all imbalance ratios $R = 1, 5, 10, 100$. A Resnet-32 model [8] is used and CE loss is optimized with SGD over 350 epochs. Initial learning rate is set to 0.1 and annealed by a factor of 10 at epochs 120 and 240. Weight decay is set to 5×10^{-4}. Models are trained on a single GPU with a dataloader batchsize of 128. For the deep-learning experiments, following [18], we perform a global centering of the mean-embeddings by subtracting their global average; see SM. We observe that convergence to the SELI geometry is better compared to the ETF. But, convergence worsens for increasing imbalance (see $R = 100$) and also for the embeddings compared to classifiers and logits. Additional ablation studies are reported in the SM.

6 Outlook: Imbalance troubles and opportunities

We proposed **(SELI)** as a class-imbalance-invariant geometry. We arrive at it after showing that the UF-SVM global minimizers follow this geometry. Subsequently, we conjecture that: (A) GD on the UFM leads to solutions asymptotically approaching the SELI geometry; (B) training of deep-nets beyond zero train-error learns models that approach the SELI geometry. Note (A) is a conjecture despite our Thm. 1. This is because GD is only known to converge to KKT points of the UF-SVM [15, 9]. While Prop. 2 and our experiments suggest its validity, the conjecture remains to be further investigated. [1] We also note that convergence rates appear slower with increasing imbalance levels; a feature which is interesting to study further. Regarding conjecture (B), we deem our experiments encouraging and hope they motivate further investigations, especially since data imbalances appear frequently across applications. Beyond that, our results could offer new perspectives on generalization, since they uncover different geometries (aka SELI for different R values), each leading to different generalization (worse for increasing R [1]). Relatedly, we envision that further such studies lead to algorithmic contributions in deep imbalanced learning as they can facilitate studying the implicit bias effect of CE adjustments tailored to imbalanced data [2, 16, 22, 13, 12].

[1] Relying on [10], Vardi et al. [21] showed gradient flow finds global minimizers in deep linear nets. However, these only apply to binary classification. We show that interesting behaviors might occur (e.g. in the role of regularization and geometry) when $k > 2$ and $R > 1$ calling further studies.

References

[1] J. Byrd and Z. Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881. PMLR, 2019.

[2] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1567–1578, 2019.

[3] 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*, 118(43), 2021.

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

[5] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, Mar. 2014.

[6] B. D. Haeffele and R. Vidal. Structured low-rank matrix factorization: Global optimality, algorithms, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 42(6):1468–1482, 2019.

[7] X. Han, V. Papyan, and D. L. Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition." computer vision and pattern recognition (2015). *Google Scholar There is no corresponding record for this reference*, pages 770–778, 2015.

[9] 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.

[10] Z. Ji and M. Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

[11] Z. Ji, M. Dudík, R. E. Schapire, and M. Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136. PMLR, 2020.

[12] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis. Decoupling representation and classifier for long-tailed recognition, 2020.

[13] B. Kim and J. Kim. Adjusting decision boundary for class imbalanced learning. *IEEE Access*, 8:81674–81685, 2020. doi: 10.1109/ACCESS.2020.2991231.

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

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

[16] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

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

[18] 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*, 117 (40):24652–24663, 2020.

[19] S. Rosset, J. Zhu, and T. Hastie. Margin maximizing loss functions. In *NIPS*, pages 1237–1244, 2003.

[20] N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. *Advances in neural information processing systems*, 17, 2004.

[21] G. Vardi, O. Shamir, and N. Srebro. On margin maximization in linear and relu networks. *arXiv preprint arXiv:2110.02732*, 2021.

[22] H.-J. Ye, H.-Y. Chen, D.-C. Zhan, and W.-L. Chao. Identifying and compensating for feature deviation in imbalanced deep learning, 2020.

[23] 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.

[24] 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*, 34, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We include concrete theoretical and experimental results substantiating our claims
   (b) Did you describe the limitations of your work? [Yes] We discuss these throughout the paper. For example, we discuss that convergence to the proposed geometry is slower in certain settings. See also Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We do not foresee any potential negative societal impact of this work
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have reviewed the guidelines and stand by the same
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We state our assumptions clearly
   (b) Did you include complete proofs of all theoretical results? [Yes] While main body includes sketches of some of the proofs, more detailed proofs have been deferred to the supplementary material
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Our code scripts and instructions are included in a shared URL
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We include these details with experimental results
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We show results over multiple runs to confirm invariance of the learned geometries across training instances.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We report these in Sec. 5 and in the SM.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Sec. 5
   (b) Did you mention the license of the assets? [Yes] See Sec. 5
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]