Imbalance Trouble: Revisiting Neural-Collapse Geometry

Christos Thrampoulidis[⋆], **Ganesh R. Kini**[†], **Vala Vakilian**[⋆], **Tina Behnia**[⋆] [⋆]
[⋆]University of British Columbia, [†]University of California, Santa Barbara.

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-features 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. 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 preferred solutions.

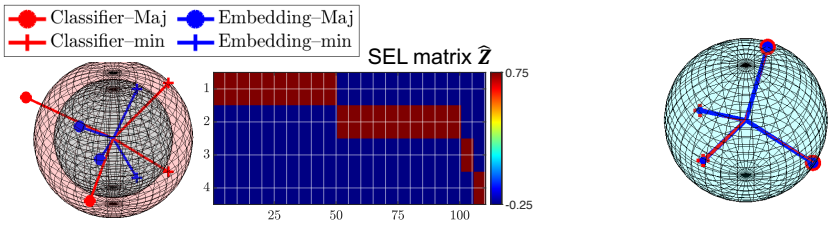
1 Introduction

What are the unique structural properties of models learned by training deep neural networks to zero training error? Is there an *implicit bias* towards solutions of certain geometry? How does this vary across training 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 explaining generalization of overparameterized models.

Recently, remarkable new progress in answering these questions has been made by Papyan et al. [26], 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 training-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 Papyan et al. [26], this simple geometry appears to be "cross-situational *invariant*" across different architectures and different *balanced* datasets.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

[⋆]Supported by an NSERC Undergraduate Student Research Grant, an NSERC Discovery Grant, NSF CCF-2009030, a CRG8-KAUST award, and by UBC Advanced Research Computing services.



(a) SELI geometry; $(10, 1/2)$-STEP imbalance **(b)** ETF geometry; balanced

Figure 1: Visualizing and contrasting the SELI and ETF geometries for $k = 4$ classes.

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?*

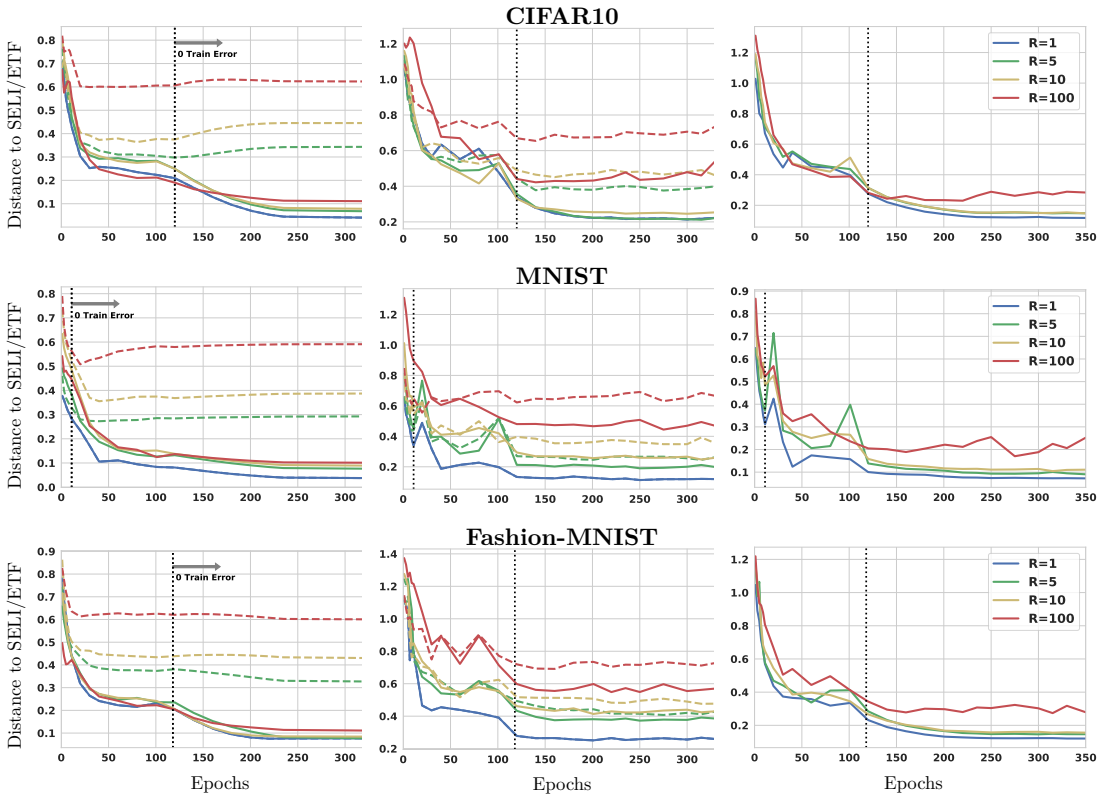
Contributions. We propose a new description capturing the geometric structure of learned-embeddings and classifier-weights on possibly *class-imbalanced data*, which we call the *Simplex-Encoded-Labels Interpolation* **(SELI)** geometry. This new geometry is a generalization of the ETF geometry: It recovers the latter when data are balanced or when there are only two classes, and also, unlike ETF, it 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 the embedding and classifier vectors. Overall, the norms and angles of these vectors admit simple closed-form expressions in terms of the imbalance characteristics and the number of classes.

We use an example to illustrate this. Fig. 1a depicts the SEL matrix $\hat{\mathbf{Z}} \in \mathbb{R}^{4\times 110}$ 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 one of the 110 examples 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). The radii of the two concentric spheres are equal to the norms of the minority classifiers (red sphere) and of the minority embeddings (blue sphere), respectively. Note that the norms of minorities and majorities are different, and so are the angles. Moreover, the classifiers are *not* aligned with the embeddings. Overall, the geometry is different compared to the ETF geometry seen in the balanced case, which is shown in Fig. 1b. What remains invariant across class-imbalances is that the logits (i.e. inner products between classifiers and embeddings) only take values either $1 - 1/k$ or $-1/k$, so that the logit matrix is equal to the SEL matrix. Equivalently, the learned model interpolates the simplex-encoding of the labels.

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 as a two-layer proxy model to theoretically justify neural collapse [24, 3, 41]. Motivated by deep-learning practice and by studies on implicit bias of gradient descent (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 there is just two of them ($k = 2$).

Next, we analyze training of the UFM with ridge-regularized CE. Unlike previous studies, we find in the presence of imbalances that regularization matters as it changes the geometry of solutions. In fact, we show that there is *no* finite regularization that leads to the SELI



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

Figure 2: Convergence of learned classifiers, embeddings and corresponding logits to the SELI (solid lines) vs ETF (dashed lines) geometries, measured using a ResNet-18 model, trained far beyond zero training error on **STEP-Imbalanced** CIFAR10, MNIST and Fashion-MNIST, for different imbalance ratios R; see Sec. 5 for metrics and discussion.

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 minority collapse: (i) does *not* occur for small finite regularization and finite imbalance ratio, and (ii) occurs asymptotically for vanishing regularization, but only asymptotically as the imbalance ratio grows.

We numerically test convergence to the 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). However, 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 our observations strongly motivate further investigations regarding potential frailties of "asymptotic" implicit bias characterizations and how these might vary in multiclass and possibly imbalanced settings.

1.1 Related works

The original contribution by Papyan et al. [26] has attracted lots of attention resulting in numerous followups within short time period, e.g., [41, 12, 3, 9, 21, 24, 5, 40, 32]. (See also [9, 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 [41, 12, 3, 5]. Other works analyze the UFM with square loss [24, 9, 40, 32] and recent extensions accounting for additional layers are studied in [32]. Here, we drove particular inspiration from Zhu et al. [41], who presented a particularly transparent and complete analysis of the optimization landscape of ridge-regularized CE minimization for the UFM under balanced data. In the same spirit, we also relied on the UFM. However, our work is, to the best of our knowledge, the first

Geometry of global minimizers for UFM

	UF-RidgeCE (Eqn. (1))	**UF-SVM** (Eqn. (2))
Balanced	**ETF** [21, 5, 41, 3, 32]	**ETF** [12], **[Cor. 1.3]**
Imbalanced	$\forall\lambda$: **NO SELI [Prop. 1]** $\lambda \to 0$: **SELI [Prop. 2]** $\lambda < \frac{1}{2}$: **NO minority collapse [Sec. H]**	**SELI** **[Thm. 1]**

Table 1: Summary of contributions and comparison to most-closely related work.

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*: for asymptotically large imbalance levels, the minorities' classifiers collapse to the same vector. Instead, we derive an explicit geometric characterization of *both* embeddings and classifiers for *both* majorities and minorities and for *all* imbalance levels. Specializing these findings to vanishing regularization and imbalance ratio growing to infinity recovers and gives new insights to minority collapse. Our results also draw from and relate to the literatures on implicit bias, matrix factorization, and imbalanced deep-learning. We defer a detailed discussion on these to Sec. I of the SM.

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

2 Problem setup

We adopt the *unconstrained feature model* (UFM) [24, 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, 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 [41]:

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

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

UFM as 2-layer linear net. The formulation above does not explicitly specify inputs. Alternatively, consider training a 2-layer linear net with hidden dimension d, first / second layers $\mathbf{H}$ / $\mathbf{W}$, over n examples with n-dim. inputs $\mathbf{x}_i = \mathbf{e}_i \in \mathbb{R}^n, i \in [n]$ and labels as above.

Unconstrained-features SVM (UF-SVM). Since neural-collapse is observed when training with small / vanishing regularization [26], 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 [22, 12]. In fact, it has been recently shown that the GD solutions converge in direction to a KKT point of the following max-margin classifier [22, 12]:

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

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

Class-imbalance model. To streamline the presentation, we focus on a setting with 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 $(1-\rho)k$ majority and ρk minority classes. Without loss of generality, classes $\{1,\dots,(1-\rho)k\}$ are majorities.*

3 Global structure of the UF-SVM: SELI geometry

In this section, we show that the global minimizers of the non-convex program in (2) 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*

$$\forall c \in [k], i \in [n] : \quad \hat{\mathbf{Z}}[c,i] = \begin{cases} 1-1/k & , c = y_i \\ -1/k & , c \neq y_i \end{cases}. \tag{3}$$

Onwards, let $\hat{\mathbf{Z}}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ be the compact SVD of $\hat{\mathbf{Z}}^T$. Specifically, $\mathbf{\Lambda}$ is a positive $(k-1)$-dimensional diagonal matrix and $\mathbf{U}_{n\times(k-1)}$, $\mathbf{V}_{k\times(k-1)}$ have orthonormal columns.

Each column $\hat{\mathbf{z}}_i \in \mathbb{R}^k$ of $\hat{\mathbf{Z}}$ represents a class-membership encoding of datapoint $i \in [n]$. This differs from the vanilla one-hot encoding $\hat{\mathbf{y}}_i = \mathbf{e}_{y_i}$ in that $\hat{\mathbf{z}}_i = \hat{\mathbf{y}}_i - \frac{1}{k}\mathbb{1}_k$. Specifically, $\hat{\mathbf{Z}}$ 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, $\mathrm{rank}(\hat{\mathbf{Z}}) = k-1$. We gather useful properties about the eigenstructure of $\hat{\mathbf{Z}}$ in Sec. A.

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 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 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 theorem's proof in Sec. 3.3 and defer the details to Sec. C.1. The theorem provides an explicit characterization of the geometry of optimal embeddings and classifiers around the key finding that the optimal logit matrix is always equal to the SEL matrix.

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. 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 values of (R,ρ). In turn, the singular factors determining the geometry of embeddings and classifiers depend implicitly on the same parameters. Thus, 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 1.1. *The UF-SVM solutions satisfy the following irrespective of imbalance:*

(NC) *The 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 can be inferred from Theorem 1 (specifically from Statement (iii) and that $\mathbf{U}$ has repeated columns.) A more straightforward argument is by directly inspecting (2) is as follows. 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, and (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), Theorem 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 $\alpha > 0$:*

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

Corollary 1.2. *The UF-SVM solutions follow the SELI geometry, irrespective of imbalance.*

The **(SELI)** geometry specifies (up to a global positive scaling) the Gram matrices $\mathbf{G_W} := \mathbf{W}^T\mathbf{W}$ and $\mathbf{G_H} := \mathbf{H}^T\mathbf{H}$, and the logit matrix $\mathbf{Z} := \mathbf{W}^T\mathbf{H}$: The diagonals of the Gram matrices specify the *norms* [2], and together with their off-diagonal entries, they further specify the *angles* between classifiers and between embeddings. Because of **(NC)**, the norms and angles of the embeddings are uniquely determined in terms of the norms and angles of the mean-embeddings; thus, for all $i \in [n]$, $\mathbf{G_H}[i,j] = \mathbf{G_H}[i,\ell]$ if $y_j = y_\ell$. Finally, the norms together with the entries of the logit matrix determine the angles between the two sets of: (a) the k classifiers and (b) the k mean embeddings. Thus, they specify the degree of alignment between the two sets of vectors. In the next section, we show that it is in fact possible to obtain explicit closed-form formulas describing the norms, angles and alignment of classifier and embedding vectors, in terms of the imbalance characteristics and the number of classes.

Remark 3.1 (Why "SELI"?)**.** *For the UFM, $\mathbf{Z} = \mathbf{W}^T\mathbf{H}$ is the learned end-to-end model. According to its definition, the SELI geometry implies $\mathbf{W}^T\mathbf{H} = \alpha\hat{\mathbf{Z}}$. Thus, the learned model* interpolates *(a scaling of) the SEL matrix, motivates the naming in Definition 3.*

3.1.1 Special case: Balanced or binary data

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

Corollary 1.3 ($R = 1$ or $k = 2$)**.** *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 [26]:*

$$\textbf{(ETF)} \quad \hat{\mathbf{W}} = \hat{\mathbf{M}} \textit{ and } \hat{\mathbf{M}}^T\hat{\mathbf{M}} = \hat{\mathbf{W}}^T\hat{\mathbf{W}} = \mathbb{I}_k - \tfrac{1}{k}\mathbb{1}_k\mathbb{1}_k^T.$$

Thus, when data are balanced or binary: (i) the norms of the classifiers and of the embeddings are all equal; (ii) the angles between any two classifiers or any two embeddings are all equal to $-1/k - 1$; and, (iii) the set of classifiers and the set of embeddings are aligned. See Fig. 1b.

3.2 How the SELI geometry changes with imbalances

For $k > 3$, $R > 1$, **(SELI)** $\not\Longrightarrow$ **(ETF)** and geometry is determined in terms of the SVD factors of the SEL matrix. Sec. A in the SM explicitly characterizes these SVD factors and leads to explicit closed-form formulas for the norms, angles and alignment in terms of R, ρ and $k, n_{\min}$ in Sec. B. For example, the following lemma gives a formula for the ratio of majority and minority norms. For simplicity, assume equal majorities and minorities.

Lemma 3.1 (Norm ratios)**.** *Assume $(R, 1/2)$-STEP imbalance. Suppose $(\mathbf{W}, \mathbf{H})$ satisfies the* **(SELI)** *property. Let $\mathbf{w}_{\text{maj}}, \mathbf{h}_{\text{maj}}$ (resp. $\mathbf{w}_{\text{minor}}, \mathbf{h}_{\text{minor}}$) denote majority (resp. minority) classifiers and embeddings, respectively. Then,*

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

Thus, $\|\mathbf{w}_{\text{maj}}\|_2 \geq \|\mathbf{w}_{\text{minor}}\|_2$ and $\|\mathbf{h}_{\text{maj}}\|_2 \leq \|\mathbf{h}_{\text{minor}}\|_2$, with equalities iff $R = 1$ or $k = 2$.

[2] We assume throughout that the regularization strength is same for embeddings and classifiers. For completeness we treat the general case in Sec. C.3 in the SM, where it is shown that different regularization values do *not* change the SELI geometry as per Definition 3 apart from introducing a (global) relative scaling factor between the norms of the embeddings and classifiers.

The fact that CE learns majority classifiers of larger norm has been empirically observed in the imbalanced deep-learning literature [17, 18]. Lemma 3.1 provides a theoretical justification and precisely quantifies the ratio, not only for classifiers, but also for the learned embeddings. As another example of closed-form formulas, the angles between any two majority (resp. minority) classifiers $\mathbf{w}_{\text{maj}}, \mathbf{w}'_{\text{maj}}$ (resp. $\mathbf{w}_{\text{minor}}, \mathbf{w}'_{\text{minor}}$) satisfy (see Sec. B.1.2):

$$\text{Cos}(\mathbf{w}_{\text{maj}}, \mathbf{w}'_{\text{maj}}) = \frac{-2\sqrt{R} + \sqrt{(R+1)/2}}{(k-2)\sqrt{R} + \sqrt{(R+1)/2}} \text{ and } \text{Cos}(\mathbf{w}_{\text{minor}}, \mathbf{w}'_{\text{minor}}) = \frac{-2 + \sqrt{(R+1)/2}}{k-2+\sqrt{(R+1)/2}}. \tag{4}$$

Both formulas evaluate to $-1/(k-1)$ for $R = 1$. Also, the first is strictly decreasing and the second strictly increasing in R, i.e. with larger R majority classifiers go further away from each other, while minority classifiers come closer; see Fig. 1 and Fig. 6 in the SM.

Remark 3.2 (Asymptotics)**.** *While we focus on finite values of R, computing limits for our formulas gives asymptotic characterizations as $R \to \infty$. As an example, it is easy to see from (4) that the angle between the minority classifiers collapses to zero in that limit. This phenomenon is called "minority collapse" by Fang et al. [3]. Here, we recover it as a special case of Theorem 1 and of SELI. Note also that the rate at which the minority angle collapses is rather slow (see also Fig. 6c in the SM). Additional details are included in Sec. B.*

3.3 Proof sketch

We start from the following standard convex relaxation of the UF-SVM [31, 8, 41, 3]:

$$\min_{\mathbf{Z} \in \mathbb{R}^{k \times n}} \|\mathbf{Z}\|_* \quad \text{subj. to} \quad \mathbf{Z}[y_i, i] - \mathbf{Z}[c, i] \geq 1, \ \forall c \neq y_i, i \in [n]. \tag{5}$$

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 (5). There are three key ingredients in this. First, is a clever re-parameterization of the dual program to (5), introducing the SEL matrix $\hat{\mathbf{Z}}$ in the dual:

$$\max_{\mathbf{B} \in \mathbb{R}^{n \times k}} \quad \text{tr}(\hat{\mathbf{Z}}\mathbf{B}) \quad \text{sub. to } \|\mathbf{B}\|_2 \leq 1, \quad \mathbf{B}\mathbb{1}_k = 0, \quad \mathbf{B} \odot \hat{\mathbf{Z}}^T \geq 0. \tag{6}$$

Second, we prove that $\hat{\mathbf{B}} = \mathbf{U}\mathbf{V}^T$ is the unique maximizer of the re-parameterized dual problem in (6). While it is not hard to check that $\hat{\mathbf{B}}$ optimizes a relaxation of (6), 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 (5). To do this, we leverage that strict feasibility of $\hat{\mathbf{B}}$ implies by complementary slackness all constraints in the primal (5) must be active at the optimum. The proof of the theorem completes by arguing that the relaxation (6) is tight when $d \geq k-1$ allowing us to connect the UF-SVM minimizers to the SEL matrix $\hat{\mathbf{Z}}$. Sec. C.1 for details.

Remark 3.3 (Comparison to literature)**.** *The common analysis strategy in* all *other 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, [41, 5, 3] lower bound the CE loss and [12] 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 study the geometry of solutions $(\hat{\mathbf{W}}_\lambda, \hat{\mathbf{H}}_\lambda)$ of ridge-regularized CE minimization in (1), as a function of both the imbalance and the regularization parameter λ.

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 [41].

Theorem 2 (Reformulated from [41])**.** *Let $\lambda > 0$, $d > k-1$ and (R,ρ)-STEP imbalance. Let $\hat{\mathbf{Z}}_\lambda \in \mathbb{R}^{k\times n}$ be the unique minimizer of the convex nuclear-norm-regularized loss minimization,*

$$\hat{\mathbf{Z}}_\lambda := \arg\min_{\mathbf{Z}} \mathcal{L}(\mathbf{Z}) + \lambda\|\mathbf{Z}\|_* \,, \tag{7}$$

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

$$\begin{bmatrix}\hat{\mathbf{W}}_\lambda^T \\ \hat{\mathbf{H}}_\lambda^T\end{bmatrix}\begin{bmatrix}\hat{\mathbf{W}}_\lambda & \hat{\mathbf{H}}_\lambda\end{bmatrix} = \begin{bmatrix}\mathbf{V}_\lambda \\ \mathbf{U}_\lambda\end{bmatrix}\mathbf{\Lambda}_\lambda\begin{bmatrix}\mathbf{V}_\lambda^T & \mathbf{U}_\lambda^T\end{bmatrix}. \tag{8}$$

This ensures that any first-order method escaping strict saddles finds a stationary point that is a global minimizer [41]. Moreover, it describes the structure of the global minimizers of (1) in terms of $\hat{\mathbf{Z}}_\lambda$, the solution to the *convex* minimization in (7). Structurally, the characterization in (8) resembles the characterization in Theorem 1 regarding the UF-SVM. However, Theorem 1 goes a step further and gives an *explicit* form for the logit matrix, namely the SEL matrix $\hat{\mathbf{Z}}$. Instead, $\hat{\mathbf{Z}}_\lambda$ in Theorem 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?* Also, *how does the answer depend on the imbalance level?*

Remark 4.1. *Although not stated explicitly in this form, Theorem 2 is essentially retrieved from the proof of [41, Theorem 3.2] with two small adjustments. First, Zhu et al. [41] only considers balanced data. Here, we realize their proof actually carries over to the imbalanced setting. Second, we relax their assumption $d > k$ to $d > k-1$ thanks to a simple observation: $\mathbf{1}_k^T \nabla_{\mathbf{Z}} \mathcal{L}(\mathbf{Z}) = \mathbf{0}$, hence the CE gradient drops rank (see Sec. E.1 for details).*

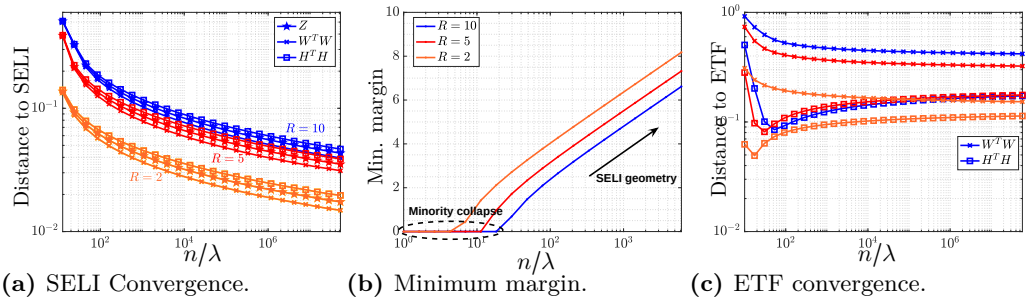
4.2 Regularization matters

For balanced data, 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 $\lambda > 0$ [41] (see also [5, 3, 21].) 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: (i) insensitive to $\lambda > 0$ and (ii) the 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 $\lambda > 0$ and α_λ such that a scaling of the UF-SVM solution solves (1).*

Proof. The proof relies on Theorem 1 as follows. For the sake of contradiction assume there exists $\lambda > 0$ and some $\alpha_\lambda > 0$ such that the scaled UF-SVM minimizer $(\alpha_\lambda \hat{\mathbf{W}}, \alpha_\lambda \hat{\mathbf{H}})$ solves (1). Since then $(\alpha_\lambda \hat{\mathbf{W}}, \alpha_\lambda \hat{\mathbf{H}})$ is a stationary point, it satisfies $\nabla_{\mathbf{W}} \mathcal{L}(\alpha_\lambda^2 \hat{\mathbf{W}}^T \hat{\mathbf{H}}) + \lambda \alpha_\lambda \hat{\mathbf{W}} = 0 \implies \alpha_\lambda \hat{\mathbf{H}} \left(\nabla_{\mathbf{Z}} \mathcal{L}(\alpha_\lambda^2 \hat{\mathbf{W}}^T \hat{\mathbf{H}})\right)^T = -\lambda \hat{\mathbf{W}} \implies \alpha_\lambda \hat{\mathbf{W}}^T \hat{\mathbf{H}} \left(\nabla_{\mathbf{Z}} \mathcal{L}(\alpha_\lambda^2 \hat{\mathbf{W}}^T \hat{\mathbf{H}})\right)^T = -\lambda \hat{\mathbf{W}}^T \hat{\mathbf{W}}$. But, by Theorem 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 $\alpha'_\lambda := k/(\exp(\alpha_\lambda^2) + k - 1)$; see Lemma A.1(v). With these, and denoting $\alpha''_\lambda := \alpha_\lambda \alpha'_\lambda$, we arrive at the following about the singular values of $\hat{\mathbf{Z}}$: $\alpha''_\lambda \hat{\mathbf{Z}}\hat{\mathbf{Z}}^T = \lambda \hat{\mathbf{W}}^T \hat{\mathbf{W}} \implies \alpha''_\lambda \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T = \lambda \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \implies \alpha''_\lambda \mathbf{\Lambda}^2 = \lambda \mathbf{\Lambda} \implies \mathbf{\Lambda} = (\lambda/\alpha''_\lambda)\mathbb{I}_{k-1}$. Thus, all singular values of $\hat{\mathbf{Z}}$ must be the same. However, we show in Lemma A.3 that this is *not* the case unless data are balanced or $k = 2$. Thus, we arrive at a contradiction. □



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

Figure 3: Numerical study 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? The following answers this.

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

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: thanks to Theorems 1 and 2, it suffices that the solution $\hat{\mathbf{Z}}_\lambda$ of (7) converges in direction to the SEL matrix $\hat{\mathbf{Z}}$; see Proposition 3 in Sec. D. To show this, we critically use from Theorem 1 that $\hat{\mathbf{Z}}$ is *unique* minimizer of (5) (see Sec. E.3). Closely related results are [28, 16], who studied the regularization path of p-norm regularized CE.

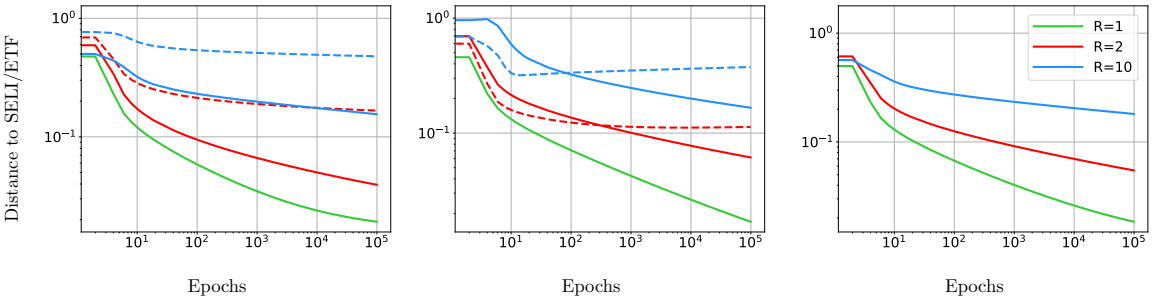
4.4 Imbalance emphasizes the impact of non-convexity

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 Proposition 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 the experiments, we choose $(R, 1/2)$-STEP imbalances with varying R and we measure convergence to either the **(SELI)** or the **(ETF)**, in terms of the three metrics below corresponding to classifiers, embeddings, and logits, respectively. 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 Definition 3). **Embeddings:** Because of **(NC)**, it suffices to work with the matrix $\mathbf{M}$ of mean embeddings (see Sec. 3.1.) Specifically, we 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. In the deep-net experiments, we employ an additional centering of the class means with their (balanced) global mean; see Sec. G. **Logits:** We measure $\|\underline{\mathbf{W}^T\mathbf{M}} - \underline{\mathbf{G}^\star}\|_F$. Note that, when NC holds this metric is essentially analogous to measuring $\|\underline{\mathbf{W}^T\mathbf{H}} - \underline{\hat{\mathbf{Z}}}\|_F$.

UFM: Global minimizers. Fig. 3 investigates the global minimizers of regularized CE (1) for the UFM and $k = 4$ classes. Thanks to Theorem 2, we obtain such minimizers by solving (7) with CVX [6], and then, using (8) to infer 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 Propositions 1 and 2: the distance to SELI for all three metrics is non-zero for any finite $\lambda > 0$, but converges to zero



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

Figure 4: Geometry of SGD solutions on minimizing CE for the UFM; SELI(Solid)/ETF(Dashed).

as $\lambda \to 0$. However, this convergence is slow and the rate becomes even worse as R increases. Finally, Fig. 3b depicts the minimum margins of solutions across λ. For all sufficiently small λ, the minimum margin is strictly positive (see Lemma D.4 in Sec. D.2). As a byproduct, this shows that "minority collapse" [3] can only possibly occur for large λ; see also Sec. H.

UFM: SGD solutions. Fig. 4 investigates whether the solutions found by SGD are consistent with the prediction of Theorem 1 about global minimizers of the UF-SVM. We fix $k = 4$ and, for each R, we select $n_{\min}$ so that $n = ((R+1)/2)kn_{\min} \approx 400$. The weights of the UFM are optimized using SGD with constant learning rate 0.4, batch size 4 and no weight decay. We train for 10^5 epochs, much beyond zero training error and plot the distance to SELI and ETF over time. We observe the following: **(i)** SGD iterates favor the SELI, instead of the ETF geometry. As a matter of fact, the distance to SELI is decreasing with epochs, suggesting an implicit bias of SGD 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 connections between regularization path and (S)GD iterates, worth investigating further. Refer to Sec. F for additional numerical results on the UFM.

Deep-learning experiments. We investigate convergence to SELI in deep-net training of $(R, \rho = 1/2)$-STEP imbalanced MNIST, Fashion-MNIST and CIFAR10 datasets with ResNet-18 [10] and VGG-13 [29]; see Sec. G.1.1 for implementation details. The convergence to SELI and ETF for the classifiers, (centered) mean-embeddings, and logits is illustrated in Fig. 2 for ResNet and Fig. 11 in the SM for VGG. The vertical dashed lines mark the zero-training-error epoch; see Sec. G.1.3. In all plots, the distance to SELI decreases as training evolves and convergence is consistently better compared to the ETF. However, convergence slows down for increasing imbalance (see $R = 100$). Also, convergence is worse for the embeddings compared to classifiers. See Sec. G in the SM for additional results.

6 Outlook: Imbalance troubles and opportunities

We propose **(SELI)** as the class-imbalance-invariant geometry of classifiers and embeddings learnt by overparameterized models when trained beyond zero training error. We arrive at it after showing that the UF-SVM global minimizers follow this geometry. Subsequently, we conjecture and show experiments supporting that: (C1) GD on the UFM leads to solutions approaching the SELI asymptotically in the number of epochs (Sec. I.2 for connection to implicit bias); (C2) training of deep-nets learns models that approach the SELI geometry (Sec. G for additional experiments). We hope our results motivate further theoretical and experimental investigations, especially since data imbalances appear frequently across applications. Beyond that, we believe that further similar studies on identifying geometric structures of learned embeddings and classifiers could offer new perspectives on generalization. Our results could pave that way 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 imbalanced deep-learning as they can facilitate studying the implicit-bias effect of CE adjustments and post-hoc techniques tailored to imbalanced data [2, 23, 39, 18, 17, 19, 20].

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

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

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

[7] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

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

[14] Z. Ji and M. Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798. PMLR, 2019.

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

[16] 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.

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

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

[19] G. R. Kini, O. Paraskevas, S. Oymak, and C. Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34:18970–18983, 2021.

[20] M. Li, X. Zhang, C. Thrampoulidis, J. Chen, and S. Oymak. Autobalance: Optimized loss functions for imbalanced data. *Advances in Neural Information Processing Systems*, 34:3163–3177, 2021.

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

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

[23] 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.

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

[25] M. S. Nacson, J. Lee, S. Gunasekar, P. H. P. Savarese, N. Srebro, and D. Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

[26] 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.

[27] X. Peng, C. Lu, Z. Yi, and H. Tang. Connections between nuclear-norm and frobenius-norm-based representations. *IEEE transactions on neural networks and learning systems*, 29(1):218–224, 2016.

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

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[30] 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*, 19(1): 2822–2878, 2018.

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

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

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

[34] R. Vidal and P. Favaro. Low rank subspace clustering (lrsc). *Pattern Recognition Letters*, 43:47–61, 2014.

[35] K. Wang, V. Muthukumar, and C. Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. *Advances in Neural Information Processing Systems*, 34, 2021.

[36] J. Wright and Y. Ma. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022.

[37] 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.

[38] Y. Yang, L. Xie, S. Chen, X. Li, Z. Lin, and D. Tao. Do we really need a learnable classifier at the end of deep neural network? *arXiv preprint arXiv:2203.09081*, 2022.

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

[40] 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.

[41] 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
   (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]

Roadmap to the Supplementary material

The SM contains the following. In Section A we derive several useful properties of the simplex-encoded-label (SEL) matrix, which we then use in Section B to present closed-form characterizations of the embeddings and classifiers geometries. Notably, these include explicit expressions for the norms and angles of both majority and minority classes, which we accompany with numerical illustrations shedding further light on the features of the proposed **(SELI)** geometry. Next, in Section C, we use the derived properties of the SEL matrix to prove our main Theorem 1 and its corollaries. In Section D we derive several useful properties of the nuclear-norm penalized CE minimization, which we then use in Section E to prove the statements of Section 4. Section F contains additional numerical results on the UFM, such as experiments with varying weight-decay and regularization choices. Additional experiments on real data (such as, isolated minority/majority geometry investigations, and classifiers/embeddings norm ratios), as well as further implementation details, are included in Section G. In Section H we show how results relate to minority collapse providing additional theoretical justifications and novel perspectives to empirical observations reported by previous work. Finally, an elaborate discussion on additional related works is contained in Section I.

Contents

A Properties of the SEL matrix **16**

A.1 Basic facts . . . 16

A.2 Eigen-structure . . . 17

A.2.1 Special case: Balanced data . . . 19

A.2.2 Special case: Equal minorities / majorities ($\rho = 1/2$) . . . 19

A.3 A useful property of the singular spaces . . . 19

B The SELI geometry **20**

B.1 Closed-form expressions . . . 20

B.1.1 Norms . . . 20

B.1.2 Angles . . . 21

B.1.3 (Non)-alignment . . . 22

B.1.4 Centering . . . 23

B.2 Illustrations and discussion on dependence on R and k . . . 23

B.3 Special cases . . . 24

B.3.1 Balanced classes and binary classification . . . 24

B.3.2 Asymptotics . . . 25

C Proofs for UF-SVM Section 3 **26**

C.1 Proof of Theorem 1 . . . 26

C.2 Proofs of Corollaries 1.1, 1.2, 1.3 and of Lemma 3.1 . . . 29

C.3 On different regularization hyperparameters between embeddings and classifiers 30

D Nuclear-norm relaxations of the UFM **30**

D.1 Large λ . . . 32

D.2 Small λ . . . 32

D.3 Vanishing λ . . . 32
D.3.1 Proof of Proposition 3 . . . 33

E Proofs for Regularized CE Section 4 **34**
E.1 Proof of Theorem 2 . . . 34
E.2 Proof of Proposition 1 . . . 35
E.3 Proof of Proposition 2 . . . 35
E.4 Comparison to one-layer linear model . . . 35

F Additional results on UFM **36**
F.1 Experiments with weight-decay . . . 36
F.2 Logit regularization and Ridge-decay . . . 37

G Additional results on real data **38**
G.1 Missing details on experiments of Section 5 . . . 39
G.1.1 Implementation details . . . 39
G.1.2 SELI vs ETF . . . 39
G.1.3 Model accuracies . . . 40
G.1.4 NC property . . . 40
G.1.5 Norms of classifiers / embeddings . . . 41
G.1.6 Non-alignment of classifiers and embeddings . . . 42
G.1.7 Majority vs minority geometry . . . 43
G.2 Capturing weight-decay with regularized UFM . . . 43
G.3 Additional experiments for minority ratios $\rho \neq 1/2$ . . . 46

H Implications on minority collapse **46**

I Additional related work **49**
I.1 Additional works on Neural Collapse . . . 49
I.2 Implicit bias . . . 50
I.3 Matrix factorization and low-rank recovery . . . 50
I.4 Class-imbalanced deep learning . . . 51
I.5 On the Simplex-encoding interpolation . . . 51

A Properties of the SEL matrix

A.1 Basic facts

We gather some basic properties about the SEL matrix below. These properties hold without any assumptions on the number of examples $n_1, \ldots, n_k$ per class other than $n_c \geq 1, \forall c \in [k]$.

Lemma A.1 (SEL matrix — Basic Facts)**.** *The following statements are true.*

(i) Let $\mathbf{Y}$ *be the zero-one hot encoding label matrix, i.e.* $\mathbf{Y}[c,i] = \begin{cases} 1 & , c = y_i \\ 0 & , c \neq y_i \end{cases}, \ \forall c \in [k], i \in [n]$. *Then,* $\hat{\mathbf{Z}} = \mathbf{Y} - \frac{1}{k}\mathbb{1}_k \mathbb{1}_n^T$.

(ii) $\hat{\mathbf{Z}}^T \mathbb{1}_k = 0$ *and* $\operatorname{rank}(\hat{\mathbf{Z}}) = k-1$.

(iii) $\hat{\mathbf{Z}}^T$ *admits a compact SVD* $\hat{\mathbf{Z}}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$, *such that* $\mathbf{\Lambda}$ *is a* $(k-1)$*-diagonal, and* $\mathbf{U} \in \mathbb{R}^{n\times(k-1)}$, $\mathbf{V} \in \mathbb{R}^{k\times(k-1)}$ *are partial orthogonal matrices, i.e.* $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbb{I}_{k-1}$. *Moreover* $\mathbf{V}^T\mathbb{1}_k = 0$, *and the* $k-1$ *columns of* $\mathbf{V}$ *span the subspace orthogonal to* $\mathbb{1}_k$, *i.e.* $\mathbf{V}\mathbf{V}^T = \mathbb{I}_k - \frac{1}{k}\mathbb{1}_k\mathbb{1}_k^T$.

(iv) $\hat{\mathbf{Z}}^T\hat{\mathbf{Z}} = \mathbf{Y}^T\mathbf{Y} - \frac{1}{k}\mathbb{1}\mathbb{1}^T$ *and* $\hat{\mathbf{Z}}\hat{\mathbf{Z}}^T = \operatorname{diag}(\mathbf{n}) + \frac{n}{k^2}\mathbb{1}_k\mathbb{1}_k^T - \frac{1}{k}\mathbf{n}\mathbb{1}_k^T - \frac{1}{k}\mathbb{1}_k\mathbf{n}^T$, *where we defined* $\mathbf{n} = [n_1, n_2, \ldots, n_k]^T$

(v) Let $\mathcal{L}(\mathbf{Z}) = \sum_{i\in[n]} \log\left(1 + \sum_{c\neq y_i} e^{-(\mathbf{Z}[y_i,i]-\mathbf{Z}[c,i])}\right)$. *Then,* $\nabla_{\mathbf{Z}}\mathcal{L}(\alpha\hat{\mathbf{Z}}) = -\frac{k}{e^{\alpha}+k-1}\hat{\mathbf{Z}}$, *for all* $\alpha \in \mathbb{R}$.

Proof. Proof of (i): Follows directly from the definition.

Proof of (ii): The fact that $\hat{\mathbf{Z}}^T\mathbb{1}_k = 0$ is easy to check. Hence, the rank is at most $k-1$. The fact that the rank is exactly $k-1$ follows by noting that the vectors $\mathbf{e}_c - \frac{1}{k}\mathbb{1}_k, c \in [k-1]$ are linearly independent.

Proof of (iii): Follows directly from Statement (ii).

Proof of (iv): Follows easily by direct calculations.

Proof of (v): We show in Lemma A.2 below that $-\nabla_{\mathbf{Z}}\mathcal{L}(\mathbf{Z}) = \mathbf{Y} - \mathbf{A}$ with $\mathbf{A}[c,i] := \frac{e^{-(\mathbf{Z}[y_i,i]-\mathbf{Z}[c,i])}}{1+\sum_{c'\neq y_i} e^{-(\mathbf{Z}[y_i,i]-\mathbf{Z}[c',i])}}$. It is easy to see that for $\mathbf{Z} = \alpha\hat{\mathbf{Z}}$, the matrix $\mathbf{A}$ can be written as $\mathbf{A} = \frac{e^{\alpha}-1}{k-1+e^{\alpha}}\mathbf{Y} + \frac{1}{k-1+e^{\alpha}}\mathbb{1}_k\mathbb{1}_n^T$. The desired then follows by recalling Statement (i). □

Lemma A.2 (Auxiliary result—Gradient of CE)**.** *The negative gradient of the cross-entropy loss* $\mathcal{L}(\mathbf{Z}) = \sum_{i=1}^n \log\left(1 + \sum_{c\neq y_i} e^{-(\mathbf{Z}[y_i,i]-\mathbf{Z}[c,i])}\right)$ *takes the form* $-\nabla_{\mathbf{Z}}\mathcal{L}(\mathbf{Z}) = \mathbf{Y} - \mathbf{A}$ *where* $\mathbf{Y}$ *is the one-hot encoding label matrix and for* $i \in [n], c \in [k]$ *we denote* $\mathbf{A}[c,i] := \frac{e^{-(\mathbf{Z}[y_i,i]-\mathbf{Z}[c,i])}}{1+\sum_{c'\neq y_i} e^{-(\mathbf{Z}[y_i,i]-\mathbf{Z}[c',i])}}$. *Thus, for any* $\mathbf{Z} \in \mathbb{R}^{k\times n}$, *it holds that* $\mathbb{1}_k^T\nabla_{\mathbf{Z}}\mathcal{L}(\mathbf{Z}) = \mathbf{0}$.

Proof. The proof is straightforward. Denote for convenience, $Z_{ci} := \mathbf{Z}[c,i]$ and $s_{ci} := \mathbf{Z}[y_i,i] - \mathbf{Z}[c,i]$. Taking derivatives with respect to $Z_{ci}, c \in [k]$ we have for $c = y_i$,

$$\frac{\partial\mathcal{L}}{\partial Z_{y_i i}} = -\frac{\sum_{c'\neq y_i} e^{-s_{c'i}}}{1+\sum_{c''\neq y_i} e^{-s_{c''i}}}\frac{\partial s_{c'i}}{\partial Z_{y_i i}} = -\frac{\sum_{c'\neq y_i} e^{-s_{c'i}}}{1+\sum_{c''\neq y_i} e^{-s_{c''i}}} = -\left(1-\frac{1}{1+\sum_{c'\neq y_i} e^{-s_{c'i}}}\right) = -(1-\mathbf{A}[y_i,i])$$

and, for $c \neq y_i$,

$$\frac{\partial\mathcal{L}}{\partial Z_{ci}} = -\frac{e^{-s_{ci}}}{1+\sum_{c'\neq y_i} e^{-s_{c'i}}}\frac{\partial s_{ci}}{\partial Z_{ci}} = \frac{e^{-s_{ci}}}{1+\sum_{c'\neq y_i} e^{-s_{c'i}}} = \mathbf{A}[c,i].$$

Thus, $\sum_{c=1}^k \frac{\partial\mathcal{L}}{\partial Z_{ci}} = 0, \forall i \in [n]$. Hence, $\mathbb{1}_k^T\nabla_{\mathbf{Z}}\mathcal{L}(\mathbf{Z}) = \mathbf{0}$ for every $\mathbf{Z}$. □

A.2 Eigen-structure

In this section, we explicitly compute the eigenstructure of the SEL matrix for (R,ρ)-STEP imbalanced data. To simplify the expressions, we assume $n_{\min} = 1$. [3] Also, we need the following definitions. For $m \in [k]$, let $\mathbb{P}_m \in \mathbb{R}^{m\times(m-1)}$ denote an orthonormal basis of the subspace orthogonal to $\mathbb{1}_m$, i.e. $\mathbb{P}_m\mathbb{P}_m^T = \mathbb{I}_m - \frac{1}{m}\mathbb{1}_m\mathbb{1}_m^T$ and $\mathbb{P}_m^T\mathbb{P}_m = \mathbb{I}_{m-1}$. We will also denote $\mathbf{S}_m := \mathbb{I}_m - \frac{1}{k}\mathbb{1}_m\mathbb{1}_m^T \in \mathbb{R}^m$.

[3] It is rather easy to derive all formulas without this requirement. Concretely, the singular values in (9) are multiplied by $\sqrt{n_{\min}}$, the $(1,1)$– and $(2,3)$– blocks of $\mathbf{U}$ in (11) are multiplied by $1/\sqrt{n_{\min}}$ and the dimensions of $\mathbf{U}$ also adjust appropriately. Nevertheless, this does not change the values of $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ and of $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ aside from a global scaling. Besides, the assumption $n_{\min} = 1$ is essentially without loss of generality because the **(NC)** property holds under the SELI property.

Lemma A.3 (SEL matrix — SVD for STEP imbalance)**.** *Assume (R, ρ)-STEP imbalanced data and $n_{\min} = 1$. Also, denote $\overline{\rho} = 1 - \rho$ and recall that the total number of examples is $n = (\rho + R\overline{\rho})k$. Then, the SVD factors of $\hat{\mathbf{Z}} = \mathbf{V}\mathbf{\Lambda}\mathbf{U}^T$ are given as follows:*

$$\mathbf{\Lambda} = \operatorname{diag}\left(\left[\sqrt{R}\mathbb{1}_{(\overline{\rho}k-1)}^T \quad \sqrt{\overline{\rho} + R\rho} \quad \mathbb{1}_{(\rho k-1)}^T\right]\right) \tag{9}$$

$$\mathbf{V} = \begin{bmatrix} \mathbb{P}_{\overline{\rho}k} & -\sqrt{\frac{\rho/\overline{\rho}}{k}}\,\mathbb{1}_{\overline{\rho}k} & 0_{(\overline{\rho}k)\times(\rho k-1)} \\ 0_{(\rho k)\times(\overline{\rho}k-1)} & \sqrt{\frac{\overline{\rho}/\rho}{k}}\mathbb{1}_{\rho k} & \mathbb{P}_{\rho k} \end{bmatrix} \tag{10}$$

$$\mathbf{U} = \begin{bmatrix} \frac{1}{\sqrt{R}}\mathbb{P}_{\overline{\rho}k} \otimes \mathbb{1}_R & -\sqrt{\frac{\rho/\overline{\rho}}{(\overline{\rho}+R\rho)k}}\mathbb{1}_{R\overline{\rho}k} & 0_{(R\overline{\rho}k)\times(\rho k-1)} \\ 0_{(\rho k)\times(\overline{\rho}k-1)} & \sqrt{\frac{\overline{\rho}/\rho}{(\overline{\rho}+R\rho)k}}\mathbb{1}_{\rho k} & \mathbb{P}_{\rho k} \end{bmatrix}. \tag{11}$$

Proof. The challenging part is coming up with the formulas in (9), (10) and (11) for the SVD factors. The lemma already does this for us. Hence, proving that the formulas are correct involves a few tedious calculations, which we present below.

Let us define for convenience:

$$\mathbf{V}_{\text{maj}} := \begin{bmatrix} \mathbb{P}_{\overline{\rho}k} \\ 0_{(\rho k)\times(\overline{\rho}k-1)} \end{bmatrix} \quad \mathbf{v} = \frac{1}{\sqrt{k}}\begin{bmatrix} -\sqrt{\frac{\rho}{\overline{\rho}}}\mathbb{1}_{\overline{\rho}k} \\ \sqrt{\frac{\overline{\rho}}{\rho}}\mathbb{1}_{\rho k} \end{bmatrix} \quad \mathbf{V}_{\min} := \begin{bmatrix} 0_{(\overline{\rho}k)\times(\rho k-1)} \\ \mathbb{P}_{\rho k} \end{bmatrix}$$

so that $\mathbf{V} = [\mathbf{V}_{\text{maj}}, \mathbf{v}, \mathbf{V}_{\min}]$. Recalling for $m = \overline{\rho}k$ or $m = \rho k$ that $\mathbb{P}_m^T\mathbb{P}_m = \mathbb{I}_{m-1}$ and $\mathbb{P}_m^T\mathbb{1}_m = 0$ it is easy to check that $\mathbf{V}^T\mathbf{V} = \mathbb{I}_{k-1}$.

Similarly, let

$$\mathbf{U}_{\text{maj}} := \begin{bmatrix} \frac{1}{\sqrt{R}}\mathbb{P}_{\overline{\rho}k} \otimes \mathbb{1}_R \\ 0_{(\rho k)\times(\overline{\rho}k-1)} \end{bmatrix} \quad \mathbf{u} = \frac{1}{\sqrt{(\overline{\rho} + R\rho)k}}\begin{bmatrix} -\sqrt{\frac{\rho}{\overline{\rho}}}\mathbb{1}_{R\overline{\rho}k} \\ \sqrt{\frac{\overline{\rho}}{\rho}}\mathbb{1}_{\rho k} \end{bmatrix} \quad \mathbf{U}_{\min} := \begin{bmatrix} 0_{(R\overline{\rho}k)\times(\rho k-1)} \\ \mathbb{P}_{\rho k} \end{bmatrix}$$

so that $\mathbf{U} = [\mathbf{U}_{\text{maj}}, \mathbf{u}, \mathbf{U}_{\min}]$. With same argument as above, it is easy to check that $\mathbf{U}^T\mathbf{U} = \mathbb{I}_{k-1}$. Here, we also use that

$$\frac{1}{R}(\mathbb{P}_{\overline{\rho}k} \otimes \mathbb{1}_R)^T(\mathbb{P}_{\overline{\rho}k} \otimes \mathbb{1}_R) = \frac{1}{R}(\mathbb{P}_{\overline{\rho}k}^T \otimes \mathbb{1}_R^T)(\mathbb{P}_{\overline{\rho}k} \otimes \mathbb{1}_R) = \frac{1}{R}\mathbb{P}_{\overline{\rho}k}^T\mathbb{P}_{\overline{\rho}k} \otimes \mathbb{1}_R^T\mathbb{1}_R = \mathbb{I}_{\overline{\rho}k}$$

Thus, it suffices to show that $\mathbf{V}\mathbf{\Lambda}\mathbf{U}^T = \hat{\mathbf{Z}}$. The key observation here is that $\hat{\mathbf{Z}}$ can be written in block-form as follows

$$\hat{\mathbf{Z}} = \begin{bmatrix} \mathbf{S}_{\overline{\rho}k} \otimes \mathbb{1}_R^T & -\frac{1}{k}\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\rho k}^T \\ -\frac{1}{k}\mathbb{1}_{\rho k}\mathbb{1}_{R\overline{\rho}k}^T & \mathbf{S}_{\rho k} \end{bmatrix}. \tag{12}$$

With these, we have the following direct calculations:

$$\begin{aligned}
\mathbf{V}\mathbf{\Lambda}\mathbf{U}^T &= \sqrt{R}\mathbf{V}_{\text{maj}}\mathbf{U}_{\text{maj}}^T + \left(\sqrt{\overline{\rho} + R\rho}\right)\mathbf{v}\mathbf{u}^T + \mathbf{V}_{\min}\mathbf{U}_{\min}^T \\
&= \begin{bmatrix} \mathbb{P}_{\overline{\rho}k}\mathbb{P}_{\overline{\rho}k}^T \otimes \mathbb{1}_R^T & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{k}\begin{bmatrix} \frac{\rho}{\overline{\rho}}\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{R\overline{\rho}k}^T & -\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\rho k}^T \\ -\mathbb{1}_{\rho k}\mathbb{1}_{R\overline{\rho}k}^T & \frac{\overline{\rho}}{\rho}\mathbb{1}_{\rho k}\mathbb{1}_{\rho k}^T \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbb{P}_{\rho k}\mathbb{P}_{\rho k}^T \end{bmatrix} \\
&= \begin{bmatrix} \left(\mathbb{I}_{\overline{\rho}k} - \frac{1}{\overline{\rho}k}\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\overline{\rho}k}^T\right) \otimes \mathbb{1}_R^T & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{k}\begin{bmatrix} \frac{\rho}{\overline{\rho}}\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\overline{\rho}k}^T \otimes \mathbb{1}_R^T & -\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\rho k}^T \\ -\mathbb{1}_{\rho k}\mathbb{1}_{R\overline{\rho}k}^T & \frac{\overline{\rho}}{\rho}\mathbb{1}_{\rho k}\mathbb{1}_{\rho k}^T \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbb{I}_{\rho k} - \frac{1}{\rho k}\mathbb{1}_{\rho k}\mathbb{1}_{\rho k}^T \end{bmatrix} \\
&= \begin{bmatrix} \left(\mathbb{I}_{\overline{\rho}k} - \frac{1}{k}\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\overline{\rho}k}^T\right) \otimes \mathbb{1}_R^T & -\frac{1}{k}\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\rho k}^T \\ -\frac{1}{k}\mathbb{1}_{\rho k}\mathbb{1}_{R\overline{\rho}k}^T & \mathbb{I}_{\rho k} - \frac{1}{k}\mathbb{1}_{\rho k}\mathbb{1}_{\rho k}^T \end{bmatrix} \\
&= \hat{\mathbf{Z}}.
\end{aligned}$$

□

A.2.1 Special case: Balanced data

When classes are balanced, i.e. $R = 1$, the following simple description of the SVD factors is immediate to see from Lemma A.3.

Corollary 2.1. *Assume balanced data and $n_{\min} = 1$. Recall that $\mathbb{P}_k \in \mathbb{R}^{k\times(k-1)}$ denotes an orthonormal basis of the subspace orthogonal to $\mathbb{1}_k$. Then, $\hat{\mathbf{Z}} = \mathbb{P}_k\mathbb{P}_k^T$, that is*

$$\mathbf{\Lambda} = \mathbb{I}_{k-1}, \quad \mathbf{U} = \mathbf{V} = \mathbb{P}_k.$$

A.2.2 Special case: Equal minorities / majorities ($\rho = 1/2$)

Another special case of interest is when the numbers of minorities and majorities are the same, i.e. $\rho = 1/2$. In this case, we get the following simplification of Lemma A.3.

Corollary 2.2. *Consider the setting of step imbalance with even number $k = 2m, m \geq 1$ of classes. Let $n_{\min} = 1$. Then, the SVD of $\hat{\mathbf{Z}}$ is as follows:*

$$\mathbf{\Lambda} = \operatorname{diag}\left(\left[\sqrt{R}\mathbb{1}_{(m-1)}^T \quad \sqrt{(R+1)/2} \quad \mathbb{1}_{(m-1)}^T\right]\right),$$

$$\mathbf{V} = \begin{bmatrix} \mathbb{P}_m & -\frac{1}{\sqrt{k}}\mathbb{1}_m & 0_{m\times(m-1)} \\ 0_{m\times(m-1)} & +\frac{1}{\sqrt{k}}\mathbb{1}_m & \mathbb{P}_m \end{bmatrix}$$

$$\mathbf{U} = \begin{bmatrix} \frac{1}{\sqrt{R}}\mathbb{P}_m \otimes \mathbb{1}_R & -\frac{1}{\sqrt{n}}\mathbb{1}_{Rm} & 0_{(Rm)\times(m-1)} \\ 0_{m\times(m-1)} & +\frac{1}{\sqrt{n}}\mathbb{1}_m & \mathbb{P}_m \end{bmatrix}.$$

A.3 A useful property of the singular spaces

The following result is particularly important for the proof of Theorem 1. It shows that the singular spaces $\mathbf{V}, \mathbf{U}$ of $\hat{\mathbf{Z}}$ are such that the matrix $\mathbf{U}\mathbf{V}^T$ has entries that agree on their sign with the sign of the entries of $\hat{\mathbf{Z}}^T$.

Lemma A.4. *Recall the setting of Lemma A.3 and the SVD $\hat{\mathbf{Z}} = \mathbf{V}\mathbf{\Lambda}\mathbf{U}^T$. The matrix $\hat{\mathbf{B}} = \mathbf{U}\mathbf{V}^T$ satisfies the following element-wise* strict *inequalities: $\hat{\mathbf{B}} \odot \hat{\mathbf{Z}}^T > 0$.*

Proof. From Lemma A.3, we have explicit expressions for the SVD factors $\mathbf{U}$ and $\mathbf{V}$. From these, we can directly compute that

$$\begin{bmatrix} \hat{\mathbf{B}}_{11} & \hat{\mathbf{B}}_{12} \\ \hat{\mathbf{B}}_{21} & \hat{\mathbf{B}}_{22} \end{bmatrix} := \hat{\mathbf{B}} = \mathbf{U}\mathbf{V}^T = \begin{bmatrix} \frac{1}{\sqrt{R}}\mathbb{P}_{\overline{\rho}k}\mathbb{P}_{\overline{\rho}k}^T \otimes \mathbb{1}_R & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{k\sqrt{\overline{\rho} + R\rho}} \begin{bmatrix} \frac{\rho}{\overline{\rho}}\mathbb{1}_{R\overline{\rho}k}\mathbb{1}_{\overline{\rho}k}^T & -\mathbb{1}_{R\overline{\rho}k}\mathbb{1}_{\rho k}^T \\ -\mathbb{1}_{\rho k}\mathbb{1}_{\overline{\rho}k}^T & \frac{\overline{\rho}}{\rho}\mathbb{1}_{\rho k}\mathbb{1}_{\rho k}^T \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbb{P}_{\rho k}\mathbb{P}_{\rho k}^T \end{bmatrix}.$$

To continue, recall again that for any integer m: $\mathbb{P}_m\mathbb{P}_m^T = \mathbb{I}_m - \frac{1}{m}\mathbb{1}_m\mathbb{1}_m^T$. Also note that $\mathbb{1}_{R\overline{\rho}k}\mathbb{1}_{\overline{\rho}k}^T = (\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\overline{\rho}k}^T) \otimes \mathbb{1}_R$. Hence continuing from the display above we find that

$$\hat{\mathbf{B}}_{11} = \left(\frac{1}{\sqrt{R}}\mathbb{I}_{\overline{\rho}k} - \left(\frac{1}{\overline{\rho}k\sqrt{R}} - \frac{\rho/\overline{\rho}}{k\sqrt{\overline{\rho} + R\rho}}\right)\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\overline{\rho}k}\right) \otimes \mathbb{1}_R = \frac{1}{\sqrt{R}}\left(\mathbb{I}_{\overline{\rho}k} - \frac{1}{\overline{\rho}k}\left(1 - \sqrt{\frac{R\rho}{R + \overline{\rho}/\rho}}\right)\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\overline{\rho}k}\right) \otimes \mathbb{1}_R$$

$$\hat{\mathbf{B}}_{12} = -\frac{1}{k\sqrt{\overline{\rho} + R\rho}}\mathbb{1}_{R\overline{\rho}k}\mathbb{1}_{\rho k}^T \qquad \text{and} \qquad \hat{\mathbf{B}}_{21} = -\frac{1}{k\sqrt{\overline{\rho} + R\rho}}\mathbb{1}_{\rho k}\mathbb{1}_{\overline{\rho}k}^T$$

$$\hat{\mathbf{B}}_{22} = \mathbb{I}_{\rho k} - \left(\frac{1}{\rho k} - \frac{\overline{\rho}/\rho}{k\sqrt{\overline{\rho} + R\rho}}\right)\mathbb{1}_{\rho k}\mathbb{1}_{\rho k}^T = \mathbb{I}_{\rho k} - \frac{1}{\rho k}\left(1 - \sqrt{\frac{\overline{\rho}}{1 + R\left(\rho/\overline{\rho}\right)}}\right)\mathbb{1}_{\rho k}\mathbb{1}_{\rho k}^T.$$

Finally, recall from (12) the block-form of $\hat{\mathbf{Z}}$ repeated here for convenience

$$\hat{\mathbf{Z}}^T := \begin{bmatrix} \hat{\mathbf{Z}}_{11}^T & \hat{\mathbf{Z}}_{12}^T \\ \hat{\mathbf{Z}}_{21}^T & \hat{\mathbf{Z}}_{22}^T \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{\overline{\rho}k} \otimes \mathbb{1}_R & -\frac{1}{k}\mathbb{1}_{R\overline{\rho}k}\mathbb{1}_{\rho k}^T \\ -\frac{1}{k}\mathbb{1}_{\rho k}\mathbb{1}_{\overline{\rho}k}^T & \mathbf{S}_{\rho k} \end{bmatrix}.$$

By inspection, the signs of $\hat{\mathbf{B}}_{12}$, $\hat{\mathbf{B}}_{21}$ are negative, same as the signs of $\hat{\mathbf{Z}}_{21}^T$, $\hat{\mathbf{Z}}_{12}^T$. To see that the signs of the diagonal blocks also agree it suffices to check that the following strict inequalities always hold

$$1 > 1 - \sqrt{\frac{R\rho}{R + \overline{\rho}/\rho}} > 0 \qquad \text{and} \qquad 1 > 1 - \sqrt{\frac{\overline{\rho}}{1 + R\left(\rho/\overline{\rho}\right)}} > 0.$$

This completes the proof of the lemma. □

B The SELI geometry

As mentioned the SELI geometry is described in terms of the SVD of the SEL matrix $\hat{\mathbf{Z}}$. In this section, we show that it is in fact possible to get explicit closed-form expressions describing the SELI geometry in terms of the parameters R, ρ, k. Key to this is the explicit construction of the SVD factors in Sec. A.2.

For concreteness, we focus on the case of equal numbers of minorities and majorities (i.e. $\rho = 1/2$) since the formulas are somewhat simpler and the setting is of sufficient interest to convey main messages. Extension to the general case can be done in a similar fashion.

All results in this section hold under the following assumptions (assumed throughout without further explicit reference):

- A $(R, 1/2)$-STEP imbalanced setting.
- The classifiers $\mathbf{w}_c, c \in [k]$ and the embeddings $\mathbf{h}_i, i \in [n]$ follow the SELI geometry in Definition 3.

B.1 Closed-form expressions

B.1.1 Norms

The following two lemmas are essentially restatements of Lemma 3.1.

Lemma B.1 (Norms of classifiers)**.** *The following statements are true about the norms of the classifiers.*

(i) The classifier norms across all majority / minority classes are all the same. That is,

$$\forall c = 1, \dots, k/2\,, \quad \|\mathbf{w}_c\|_2 =: \|\mathbf{w}_{\text{maj}}\|_2 \quad \textit{and} \quad \forall c = k/2+1, \dots, k\,, \quad \|\mathbf{w}_c\|_2 =: \|\mathbf{w}_{\text{minor}}\|_2$$

where we let $\|\mathbf{w}_{\text{maj}}\|_2$ / $\|\mathbf{w}_{\text{minor}}\|_2$ denote the majority / minority norm of an arbitrary class of the corresponding type.

(ii) It holds that

$$\|\mathbf{w}_{\text{maj}}\|_2^2 = \sqrt{R}(1 - 2/k) + \frac{\sqrt{(R+1)/2}}{k} \quad \textit{and} \quad \|\mathbf{w}_{\text{minor}}\|_2^2 = (1 - 2/k) + \frac{\sqrt{(R+1)/2}}{k}. \tag{13}$$

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

Proof. Recall, since the classifiers follow the SELI geometry, it holds that $\mathbf{W}^T\mathbf{W} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$. Hence, it suffices to compute the diagonal of the matrix $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$. We have

$$\begin{aligned}
\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T &= \sqrt{R}\begin{bmatrix} \mathbb{P}_m\mathbb{P}_m^T & 0 \\ 0 & 0 \end{bmatrix} + \frac{\sqrt{(R+1)/2}}{k}\begin{bmatrix} \mathbb{1}_m\mathbb{1}_m^T & -\mathbb{1}_m\mathbb{1}_m^T \\ -\mathbb{1}_m\mathbb{1}_m^T & \mathbb{1}_m\mathbb{1}_m^T \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbb{P}_m\mathbb{P}_m^T \end{bmatrix} \\
&= \begin{bmatrix} \sqrt{R}\mathbb{I}_{k/2} - \frac{1}{k}\left(2\sqrt{R} - \sqrt{(R+1)/2}\right)\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T & -\frac{\sqrt{(R+1)/2}}{k}\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T \\ -\frac{\sqrt{(R+1)/2}}{k}\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T & \mathbb{I}_{k/2} - \frac{1}{k}\left(2 - \sqrt{(R+1)/2}\right)\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T \end{bmatrix}.
\end{aligned} \tag{14}$$

From this, the statements of the lemma follow readily. □

Lemma B.2 (Norms of embeddings)**.** *The following statements are true about the norms of the embeddings:*

(i) The embedding norms across all majority / minority classes are all the same. That is,

$$\begin{aligned}
\forall j \in \{i \in [n] : y_i = 1, \dots, k/2\}\,, \quad \|\mathbf{h}_j\|_2 =: \|\mathbf{h}_{\text{maj}}\|_2, \qquad \textit{and} \\
\forall j \in \{i \in [n] : y_i = k/2+1, \dots, k\}\,, \quad \|\mathbf{h}_j\|_2 =: \|\mathbf{h}_{\text{minor}}\|_2.
\end{aligned}$$

where we let $\|\mathbf{h}_{\text{maj}}\|_2$ / $\|\mathbf{h}_{\text{minor}}\|_2$ denote the majority / minority norm of an arbitrary example of the corresponding type.

(ii) It holds that

$$\|\mathbf{h}_{\text{maj}}\|_2^2 = \frac{1}{\sqrt{R}}(1-2/k) + \frac{1}{k\sqrt{(R+1)/2}} \quad \textit{and} \quad \|\mathbf{h}_{\text{minor}}\|_2^2 = (1-2/k) + \frac{1}{k\sqrt{(R+1)/2}}. \tag{15}$$

Thus, $\|\mathbf{h}_{\text{maj}}\|_2 \le \|\mathbf{h}_{\text{minor}}\|_2$ *with equality if and only if* $R = 1$ *or* $k = 2$.

Proof. Recall, since the classifiers follow the SELI geometry, it holds that $\mathbf{H}^T\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$. Hence, it suffices to compute the diagonal of the matrix $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$. Recalling that $n = k(R+1)/2$ We have

$$\begin{aligned}
\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T &= \sqrt{R}\begin{bmatrix} \frac{1}{R}\mathbb{P}_m\mathbb{P}_m^T \otimes \mathbb{1}_R\mathbb{1}_R^T & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{k\sqrt{(R+1)/2}}\begin{bmatrix} \mathbb{1}_{Rm}\mathbb{1}_{Rm}^T & -\mathbb{1}_{Rm}\mathbb{1}_m^T \\ -\mathbb{1}_m\mathbb{1}_{Rm}^T & \mathbb{1}_m\mathbb{1}_m^T \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbb{P}_m\mathbb{P}_m^T \end{bmatrix} \\
&= \begin{bmatrix} \left(\frac{1}{\sqrt{R}}\mathbb{I}_{k/2} - \frac{1}{k}\left(\frac{2}{\sqrt{R}} - \frac{1}{\sqrt{(R+1)/2}}\right)\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T\right) \otimes \mathbb{1}_R\mathbb{1}_R^T & -\frac{1}{k\sqrt{(R+1)/2}}\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T \\ -\frac{1}{k\sqrt{(R+1)/2}}\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T & \mathbb{I}_{k/2} - \frac{1}{k}\left(2 - \frac{1}{\sqrt{(R+1)/2}}\right)\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T \end{bmatrix}.
\end{aligned} \tag{16}$$

From this, the statements of the lemma follow readily by reading out the diagonal. □

B.1.2 Angles

Lemma B.3 (Angles of classifiers)**.** *The following statements are true about the angles of the classifiers.*

(i) The classifiers' angles across all majority / minority classes are all the same. That is,

$$\begin{aligned}
\forall c \neq c' \in \{1,\dots,k/2\}, &\quad \text{Cos}(\mathbf{w}_c, \mathbf{w}_{c'}) =: \text{Cos}(\mathbf{w}_{\text{maj}}, \mathbf{w}'_{\text{maj}}) \\
\forall c \neq c' \in \{k/2+1,\dots,k\}, &\quad \text{Cos}(\mathbf{w}_c, \mathbf{w}_{c'}) =: \text{Cos}(\mathbf{w}_{\text{minor}}, \mathbf{w}'_{\text{minor}}) \\
\forall c \in \{1,\dots,k/2\}, c' \in \{k/2+1,\dots,k\}, &\quad \text{Cos}(\mathbf{w}_c, \mathbf{w}_{c'}) =: \text{Cos}(\mathbf{w}_{\text{maj}}, \mathbf{w}_{\text{minor}}).
\end{aligned}$$

(ii) It holds that

$$\begin{aligned}
\text{Cos}(\mathbf{w}_{\text{maj}}, \mathbf{w}'_{\text{maj}}) &= \frac{-2\sqrt{R} + \sqrt{(R+1)/2}}{(k-2)\sqrt{R} + \sqrt{(R+1)/2}} \\
\text{Cos}(\mathbf{w}_{\text{minor}}, \mathbf{w}'_{\text{minor}}) &= \frac{-2 + \sqrt{(R+1)/2}}{k - 2 + \sqrt{(R+1)/2}} \\
\text{Cos}(\mathbf{w}_{\text{maj}}, \mathbf{w}_{\text{minor}}) &= -\frac{\sqrt{(R+1)/2}}{k\|\mathbf{w}_{\text{maj}}\|_2\|\mathbf{w}_{\text{minor}}\|_2}.
\end{aligned}$$

Proof. Recall under the SELI geometry that $\mathbf{W}^T\mathbf{W} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$. Hence, inspecting the off-diagonal entries of the matrix computed in Equation (14) gives Statement (i) and the following inner-product relations:

$$\begin{aligned}
\mathbf{w}_{\text{maj}}^T\mathbf{w}'_{\text{maj}} &= -\frac{1}{k}\left(2\sqrt{R} - \sqrt{(R+1)/2}\right) \\
\mathbf{w}_{\text{minor}}^T\mathbf{w}'_{\text{minor}} &= -\frac{1}{k}\left(2 - \sqrt{(R+1)/2}\right) \\
\mathbf{w}_{\text{minor}}^T\mathbf{w}_{\text{maj}} &= -\frac{\sqrt{(R+1)/2}}{k}.
\end{aligned}$$

Combine these with the norm calculations in Lemma B.1 to prove Statement (ii). □

Lemma B.4 (Angles of classifiers)**.** *The following statements are true about the angles of the embeddings.*

(i) The embeddings' angles across all majority / minority classes are all the same. That is,

$$\begin{aligned}
\forall c \neq c' \in \{1, \dots, k/2\}, \quad & \mathrm{Cos}(\mathbf{h}_c, \mathbf{h}_{c'}) =: \mathrm{Cos}(\mathbf{h}_{\mathrm{maj}}, \mathbf{h}'_{\mathrm{maj}}) \\
\forall c \neq c' \in \{k/2+1, \dots, k\}, \quad & \mathrm{Cos}(\mathbf{h}_c, \mathbf{h}_{c'}) =: \mathrm{Cos}(\mathbf{h}_{\mathrm{minor}}, \mathbf{h}'_{\mathrm{minor}}) \\
\forall c \in \{1, \dots, k/2\}, c' \in \{k/2+1, \dots, k\}, \quad & \mathrm{Cos}(\mathbf{h}_c, \mathbf{h}_{c'}) =: \mathrm{Cos}(\mathbf{h}_{\mathrm{maj}}, \mathbf{h}_{\mathrm{minor}}).
\end{aligned}$$

(ii) It holds that

$$\begin{aligned}
\mathrm{Cos}(\mathbf{h}_{\mathrm{maj}}, \mathbf{h}'_{\mathrm{maj}}) &= \frac{-(R+2)}{-(R+2) + k\left(R + 1 + \sqrt{R}\sqrt{(R+1)/2}\right)} \\
\mathrm{Cos}(\mathbf{h}_{\mathrm{minor}}, \mathbf{h}'_{\mathrm{minor}}) &= \frac{1 - \sqrt{2}\sqrt{R+1}}{1 - \sqrt{2}\sqrt{R+1} + k\sqrt{(R+1)/2}} \\
\mathrm{Cos}(\mathbf{h}_{\mathrm{maj}}, \mathbf{h}_{\mathrm{minor}}) &= -\frac{1}{\|\mathbf{h}_{\mathrm{maj}}\|_2 \|\mathbf{h}_{\mathrm{minor}}\|_2 \, k \sqrt{(R+1)/2}}.
\end{aligned}$$

Proof. Recall under the SELI geometry that $\mathbf{H}^T\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$. Hence, inspecting the off-diagonal entries of the matrix computed in Equation (16) gives Statement (i) and the following inner-product relations:

$$\begin{aligned}
\mathbf{h}_{\mathrm{maj}}^T \mathbf{h}'_{\mathrm{maj}} &= -\frac{1}{k}\left(\frac{2}{\sqrt{R}} - \frac{1}{\sqrt{(R+1)/2}}\right) \\
\mathbf{h}_{\mathrm{minor}}^T \mathbf{h}'_{\mathrm{minor}} &= -\frac{1}{k}\left(2 - \frac{1}{\sqrt{(R+1)/2}}\right) \\
\mathbf{h}_{\mathrm{minor}}^T \mathbf{h}_{\mathrm{maj}} &= -\frac{1}{k\sqrt{(R+1)/2}}.
\end{aligned}$$

Combine these with the norm calculations in Lemma B.2 to prove Statement (ii). □

B.1.3 (Non)-alignment

In the previous lemmas we compute the angles between classifiers of different classes and between embeddings of different classes. Here, we also also compute the angles between classifiers and embeddings. Specifically, for each $c \in [k]$, we compute the angle $\mathrm{Cos}(\mathbf{w}_c, \mathbf{h}_i)$ for an example $i : y_i = c$ that belongs to the same class. These values can be thought of as the degree of alignment between classifiers and embeddings, as $\mathrm{Cos}(\mathbf{w}_c, \mathbf{h}_i) = 1$ corresponds to exact alignment between the two.

Lemma B.5 (Alignment of classifiers and embeddings)**.** *The following statements are true about the degree of alignment between class-embeddings and their corresponding classifiers.*

(i) The angles between majority- / minority- class embeddings and their corresponding classifiers are all the same. That is,

$$\begin{aligned}
\forall c \in \{1, \dots, k/2\} \text{ and } i : y_i = c, \quad & \mathrm{Cos}(\mathbf{w}_c, \mathbf{h}_i) = \mathrm{Cos}(\mathbf{w}_{\mathrm{maj}}, \mathbf{h}_{\mathrm{maj}}) \\
\forall c \in \{k/2+1, \dots, k\} \text{ and } i : y_i = c, \quad & \mathrm{Cos}(\mathbf{w}_c, \mathbf{h}_i) = \mathrm{Cos}(\mathbf{w}_{\mathrm{minor}}, \mathbf{h}_{\mathrm{minor}}).
\end{aligned}$$

(ii) It holds that

$$\mathrm{Cos}(\mathbf{w}_{\mathrm{maj}}, \mathbf{h}_{\mathrm{maj}}) = \frac{1 - 1/k}{\|\mathbf{w}_{\mathrm{maj}}\|_2 \|\mathbf{h}_{\mathrm{maj}}\|_2} \tag{17}$$

$$\mathrm{Cos}(\mathbf{w}_{\mathrm{minor}}, \mathbf{h}_{\mathrm{minor}}) = \frac{1 - 1/k}{\|\mathbf{w}_{\mathrm{minor}}\|_2 \|\mathbf{h}_{\mathrm{minor}}\|_2}. \tag{18}$$

Proof. The proof is immediate by recognizing that under the SELI geometry for all $c \in [k]$ and $i : y_i = c$ it holds that $\mathbf{w}_c^T \mathbf{h}_i = 1 - 1/k$. □

B.1.4 Centering

It is easy to see that the classifiers $\mathbf{w}_c, c \in [k]$ following the SELI geometry are centered, i.e. $\sum_{c\in[k]} \mathbf{w}_c = 0$. For example, we can see this from the facts that $\mathbf{W}^T\mathbf{W} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ and $\mathbf{V}^T\mathbb{1}_k = 0$. See also Lemma E.1 for an alternative proof.

On the other hand, the embeddings $\mathbf{h}_i, i \in [n]$ are *not* centered around zero in general. Instead, it holds that

$$\sum_{i\in[n]} \frac{1}{n_{y_i}} \mathbf{h}_i = 0. \tag{19}$$

Note that this reduces to $\sum_{i\in[n]} \mathbf{h}_i$ for balanced data, but is not true otherwise. To see (19) recall first that $\mathbf{H}^T\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$. Second, check that $\sum_{i\in[n]} \frac{1}{n_{y_i}} \hat{\mathbf{z}}_i = 0$, i.e. $\hat{\mathbf{Z}}\boldsymbol{\omega} = 0$ where $\boldsymbol{\omega}[i] = 1/n_{y_i}, i \in [n]$. Thus, $\mathbf{U}^T\boldsymbol{\omega} = 0$. Combining these two it follows that $\mathbf{H}\boldsymbol{\omega} = 0$, which gives the desired.

Now, suppose that the **(NC)** property also holds, i.e. the embeddings collapse to their class means, that is,

$$\forall i \in [n], \quad \mathbf{h}_i = \boldsymbol{\mu}_{y_i} := \frac{1}{n_{y_i}} \sum_{j\,:\,y_j = y_i} \mathbf{h}_j.$$

Then, (19) implies the following about the class means:

$$0 = \sum_{i\in[n]} \frac{1}{n_{y_i}} \mathbf{h}_i = \sum_{c\in[k]} \frac{1}{n_c} \sum_{i\,:\,y_i = c} \mathbf{h}_i = \sum_{c\in[k]} \boldsymbol{\mu}_c. \tag{20}$$

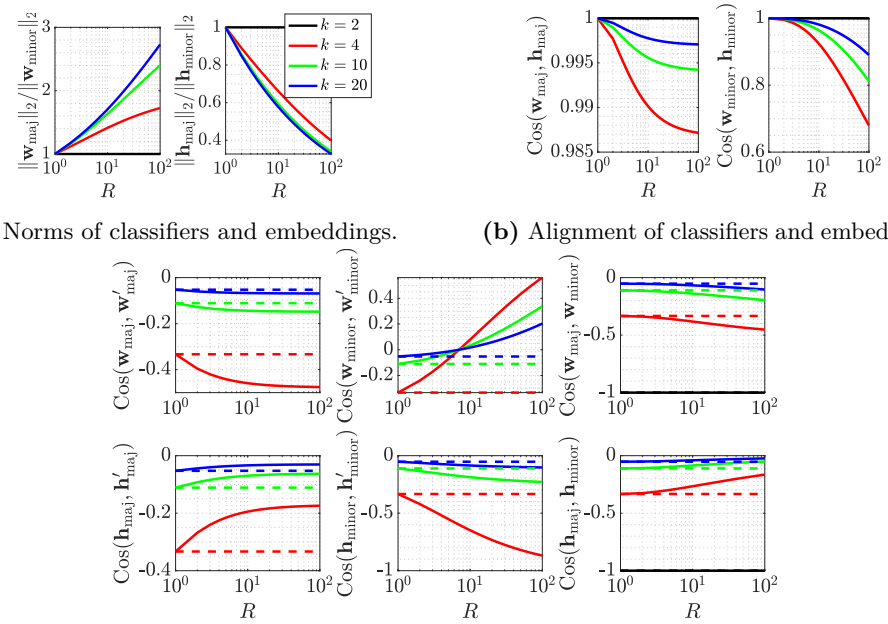
Therefore, the class means are always centered around zero.

B.2 Illustrations and discussion on dependence on R and k

In the previous section, we derived closed-form expressions for the features describing the SELI geometry. Here, we use these expressions to study how varying values of class-number k and imbalance-ratio R change the geometry. We use the numerical illustration in Fig. 5 to guide the discussion. Specifically, in Fig. 5 we compute and plot the norm ratios, alignment and angles between embeddings and classifiers for $k = 2, 4, 10, 20$ classes and imbalance ratio varying from 1 (aka balanced) to 100.

Norms. Fig. 5a shows the ratios of majority vs minority norms for both classifiers and embeddings. The values are computed using Lemmas B.1 and B.2. For binary problems (aka $k = 2$), the norm ratio is always equal to one irrespective of imbalance. For larger values of k, the norm ratio is equal to one only for balanced classes (aka $R = 1$). Recall that equal norm ratios is a feature of the ETF geometry [26]. On the other hand, for $k \neq 2$ and $R > 1$, the majorities have strictly larger norms for the classifiers and strictly smaller norms for the embeddings. Thus, in general the SELI geometry is very different from the ETF geometry. Interestingly, the difference is already evident when going from $k = 2$ to $k = 4$ classes. Also, the change in the ratios for classifiers is more pronounced than that for embeddings, which is changing progressively slower as k increases (see how close are the green and blue curves in the right plot).

Alignment. Fig. 5b shows the degree to which the geometries of classifiers and embeddings are aligned to each other. Specifically, we plot the cosine between any majority (left) / minority (right) classifier and corresponding embeddings belonging to the same class; see also Lemma B.5. For $k = 2$ and $R = 1$ the cosines are equal to one indicating that classifiers and embeddings align for both majorities and minorities. This is consistent with the ETF geometry. On the other hand, the alignment property breaks when $k > 2$ and $R > 1$. The effect is more drastic for the minorities (right plot), while for majorities the alignment is approximately preserved (note the y-axis scale is different in the two plots). Interestingly, alignment is in fact favored for larger number of classes, but deteriorates with increasing R consistently for all values of k.



(a) Norms of classifiers and embeddings.

(b) Alignment of classifiers and embeddings.

(c) Angles of classifiers and embeddings. Dashed lines are are at $-1/(k-1)$.

Figure 5: Illustration of how norms and angles of the SELI geometry vary with the number of classes k and the imbalance ratio R. The minorities fraction is set to $\rho = 1/2$. Unless $R = 1$ or $k = 2$, the geometry differs from the ETF geometry. For example, majority classifiers have larger norms than minorities (see Fig. 6a), minority classifiers are less aligned with their corresponding class mean-embeddings (see Fig. 6b), and the angle between minority classifiers decreases (see Fig. 6c). The depicted values are computed thanks to closed-form formulas.

Angles. Fig. 5c shows the angles between majority/majority (left), minority/minority (center), and minority/majority (right) for both classifiers (top) and embeddings (bottom). The values are computed using Lemmas B.3 and B.4. For binary problems (aka $k = 2$), there is only one majority and one minority class. Thus, we only plot the minority/majority cosines, which are always equal to $-1/(k-1)$ irrespective of imbalance. For larger values of k, the cosines are equal to that same value $-1/(k-1)$ only for balanced classes (aka $R = 1$). Recall that cosine value equal to $-1/(k-1)$ is a unique feature of the ETF geometry. On the other hand, for $k \neq 2$ and $R > 1$, the cosines are different. For reference, we plot the values of $-1/(k-1)$ in dashed lines. For the classifiers, the majority angles increase, while the minority angles decrease. The rate of change is more drastic for minorities. In both cases, the rate of change across R is more pronounced for smaller k. The majority-minority angles also increase with R. The trend is reversed for embeddings. For example, the angles between minority embeddings become larger with increasing R. Again, the effect of imbalance (at least for the values of R shown) is more pronounced here for smaller values of k.

B.3 Special cases

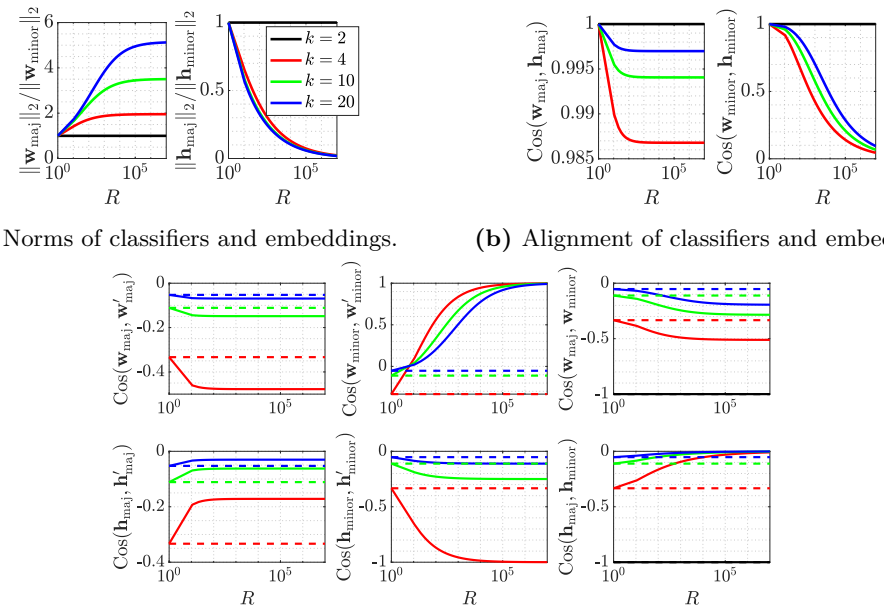
B.3.1 Balanced classes and binary classification

The following result follows directly by combining Lemmas B.1, B.2, B.3, B.4 and B.5.

Corollary 2.3 ($k = 2$ or $R = 1$: **(SELI)≡(ETF)**). *Assume $k = 2$ or $R = 1$. The following hold.*

$$\|\mathbf{w}_1\|_2 = \ldots = \|\mathbf{w}_k\|_2 \quad \textit{and} \quad \|\mathbf{h}_1\|_2 = \ldots = \|\mathbf{h}_n\|_2\,,$$
$$\forall c \neq c' \in [k],\ \ \text{Cos}(\mathbf{w}_c, \mathbf{w}'_c) = -1/(k-1) \quad \textit{and} \quad \forall i \neq i' \in [n],\ \ \text{Cos}(\mathbf{h}_i, \mathbf{h}'_i) = -1/(k-1)\,,$$
$$\forall c \in [k], i : y_i = c,\ \ \text{Cos}(\mathbf{w}_c, \mathbf{h}_i) = 1.$$

Thus, the SELI geometry is same as the ETF geometry in balanced and in binary classification.



(a) Norms of classifiers and embeddings.

(b) Alignment of classifiers and embeddings.

(c) Angles of classifiers and embeddings. Dashed lines show the value $-1/(k-1)$ characterizing the ETF geometry.

Figure 6: Same setting as Figure 5. Here emphasizing asymptotic behaviors as $R \to \infty$. For limit calculations see Corollary 2.4.

B.3.2 Asymptotics

We can also use the results of Sec. B.1 to understand the SELI geometric features asymptotically as R increases. These are included in Corollary 2.4 below. See also Fig. 6 for a numerical illustration of the limiting behavior for large imbalance ratios.

Corollary 2.4 ($R \to \infty$)**.** *Fix even $k > 2$. Then, the following limits hold.*

(i) $\lim_{R\to\infty} \frac{\|\mathbf{w}_{\text{maj}}\|_2^2}{\|\mathbf{w}_{\text{minor}}\|_2^2} = 1 + (k-2)\sqrt{2}$

(ii) $\lim_{R\to\infty} \frac{\|\mathbf{h}_{\text{maj}}\|_2^2}{\|\mathbf{h}_{\text{minor}}\|_2^2} = 0$

(iii) $\lim_{R\to\infty} \text{Cos}(\mathbf{w}_{\text{maj}}, \mathbf{w}'_{\text{maj}}) = \frac{-4+\sqrt{2}}{\sqrt{2}+2(k-2)}$

(iv) $\lim_{R\to\infty} \text{Cos}(\mathbf{w}_{\text{minor}}, \mathbf{w}'_{\text{minor}}) = 1$

(v) $\lim_{R\to\infty} \text{Cos}(\mathbf{w}_{\text{maj}}, \mathbf{w}_{\text{minor}}) = -\frac{1}{\sqrt{\sqrt{2}+2(k-2)}}$

(vi) $\lim_{R\to\infty} \text{Cos}(\mathbf{h}_{\text{maj}}, \mathbf{h}'_{\text{maj}}) = \frac{1}{k(1+\sqrt{2}/2)-1}$

(vii) $\lim_{R\to\infty} \text{Cos}(\mathbf{h}_{\text{minor}}, \mathbf{h}'_{\text{minor}}) = -\frac{2}{k-2}$

(viii) $\lim_{R\to\infty} \text{Cos}(\mathbf{h}_{\text{maj}}, \mathbf{h}_{\text{minor}}) = 0$

(ix) $\lim_{R\to\infty} \text{Cos}(\mathbf{w}_{\text{maj}}, \mathbf{h}_{\text{maj}}) = \frac{k-1}{\sqrt{k+\sqrt{2}-2}\sqrt{k+\sqrt{2}/2-2}}$

(x) $\lim_{R\to\infty} \text{Cos}(\mathbf{w}_{\text{minor}}, \mathbf{h}_{\text{minor}}) = 0$

Proof. The expressions that appear in the lemmas in Sec. B.1 hold for any value of R. Take the limit of $R \to \infty$ to yield the expressions above. We omit the details for brevity. □

Several interesting conclusions are immediate from the formulas above. For example, Statement (iv) shows that, asymptotically in R, the minority classes collapse to the same vector. This is a manifestation of the minority collapse phenomenon discovered by Fang et al. [3]. Asymptotic minority collapse had not been shown before for the UF-SVM. See also Sec. H for an extended discussion. We note that beyond minority classifiers, Corollary 2.4 is further conclusive about the behavior of majority classifiers, as well as, minority and majority embeddings. For example, Statement (viii) suggests that the minority and majority embeddings become orthogonal to each other asymptotically. Further investigations of the validity of such asymptotic conclusions in deep-net training is beyond our scope.

C Proofs for UF-SVM Section 3

C.1 Proof of Theorem 1

We consider a relaxation of the non-convex SVM in (2) by setting

$$\mathbf{X} = \begin{bmatrix} \mathbf{W}^T \\ \mathbf{H}^T \end{bmatrix} [\mathbf{W} \quad \mathbf{H}] = \begin{bmatrix} \mathbf{W}^T\mathbf{W} & \mathbf{W}^T\mathbf{H} \\ \mathbf{H}^T\mathbf{W} & \mathbf{H}^T\mathbf{H} \end{bmatrix} \in \mathbb{R}^{(k+n)\times(k+n)}. \tag{21}$$

With this consider the following semidefinite program:

$$\begin{aligned} \mathrm{q}_* = \min_{\mathbf{X}\succeq 0} \quad & \frac{1}{2}\operatorname{tr}\left(\mathbf{X}\right) \\ \text{sub. to} \quad & \mathbf{X}[y_i, k+i] - \mathbf{X}[c, k+i] \geq 1, \ \ \forall i \in [n], c \neq y_i. \end{aligned} \tag{22}$$

It is not hard to see that $\mathrm{q}_* \leq \mathrm{p}_*$. In what follows, we will compute the optimal set of (22) and use this to show that the relaxation is in fact tight. This will allow us to characterize the solution of the original problem.

Dual of the convex relaxation: The optimization in (22) is convex and satisfies Slater's conditions. (Since constraints are affine, it suffices to check feasibility, which is easily verified.) Hence, strong duality holds and KKT conditions are necessary and sufficient for optimality. The dual of (22) is written as follows:

$$\begin{aligned} \mathrm{d}_* = \max_{\{a_{i,c}\}_{i\in[n], c\neq y_i}} \quad & \sum_{i\in[n]} \sum_{c\neq y_i} \alpha_{ic} \\ \text{sub. to} \quad & \begin{bmatrix} \mathbb{I}_k & \mathbf{A}^T \\ \mathbf{A} & \mathbb{I}_n \end{bmatrix} \succeq 0 \\ & \alpha_{ic} \geq 0, i \in [n], c \neq y_i, \quad \forall i \in [n] : \mathbf{A}[i,c] = \begin{cases} -\sum_{c'\neq y_i} \alpha_{ic'}, & y_i = c, \\ \alpha_{ic}, & y_i \neq c. \end{cases} \end{aligned} \tag{23}$$

Also, the complementary slackness conditions are

$$\begin{bmatrix} \mathbb{I}_k & \mathbf{A}^T \\ \mathbf{A} & \mathbb{I}_n \end{bmatrix} \mathbf{X} = 0 \qquad \text{and} \qquad \forall i \in [n], c \neq y_i \,:\, \alpha_{ic}(1 - \mathbf{X}[y_i, k+i] + \mathbf{X}[c, k+i]) = 0. \tag{24}$$

Instead of working with the dual in the above standard form, it is convenient to work with an alternative representation by (re)-defining dual variables $\beta_{ic}, i \in [n], c \in [k]$ such that [4]

$$\beta_{ic} = \begin{cases} \sum_{c'\neq y_i} \alpha_{ic'}, & y_i = c, \\ -\alpha_{ic}, & y_i \neq c. \end{cases} \tag{25}$$

Specifically, arrange these new dual variable representations in a matrix $\mathbf{B} \in \mathbb{R}^{n\times k}$ with entries $\mathbf{B}[i,c] = \beta_{ic}, i \in [n], c \in [k]$ and recall the SEL matrix $\hat{\mathbf{Z}} \in \mathbb{R}^{k\times n}$ in Definition 2. With

[4] The same re-parameterization trick, but for a simpler (convex) max-margin program was used by Wang et al. [35].

these, we can rewrite the dual in (23) in the following more convenient form:

$$\mathrm{d}_* = \max_{\mathbf{B}\in\mathbb{R}^{n\times k}} \quad \mathrm{tr}(\hat{\mathbf{Z}}\mathbf{B}) \tag{26}$$

$$\text{sub. to} \quad \begin{bmatrix} \mathbb{I}_k & -\mathbf{B}^T \\ -\mathbf{B} & \mathbb{I}_n \end{bmatrix} \succeq 0$$

$$\forall i \in [n] : \mathbf{B}[i, y_i] = -\sum_{c\neq y_i} \mathbf{B}[i, c]$$

$$\mathbf{B} \odot \hat{\mathbf{Z}}^T \geq 0\,. \tag{27}$$

Analogously, the complementary slackness conditions in (24) are equivalent to the following:

$$\begin{bmatrix} \mathbb{I}_k & -\mathbf{B}^T \\ -\mathbf{B} & \mathbb{I}_n \end{bmatrix}\mathbf{X} = 0 \quad \text{and} \quad \forall i \in [n], c \neq y_i : \mathbf{B}[i, c]\,(1 - \mathbf{X}[y_i, k+i] + \mathbf{X}[c, k+i]) = 0. \tag{28}$$

To see the equivalence of the objective of (26) denote $A = \sum_{i\in[n]}\sum_{c\neq y_i}\alpha_{ic}$ the objective in (23) and note that we get simultaneously $A = \sum_{i\in[n]}\beta_{iy_i} = \sum_{i\in[n]}\sum_{c\neq y_i}\big(-\beta_{ic}\big)$ following the definition in Equation (25). Then, we have $A = \frac{k-1}{k}A + \frac{1}{k}A = \sum_{i\in[n]}\sum_{c\in[k]}\hat{z}_{ic}\beta_{ic} = \mathrm{tr}\big(\hat{\mathbf{Z}}\mathbf{B}\big)$.

<u>*Solution to the dual:*</u> To continue, we consider the following relaxation of the dual problem (26) (by removing the constraint in Equation (27)):

$$\hat{\mathrm{d}} = \max_{\mathbf{B}\in\mathbb{R}^{n\times k}} \quad \mathrm{tr}(\hat{\mathbf{Z}}\mathbf{B}) \qquad \text{sub. to} \quad \|\mathbf{B}\|_2 \leq 1 \quad \text{and} \quad \mathbf{B}\mathbb{1}_k = 0. \tag{29}$$

Here, recall that $\|\mathbf{B}\|_2$ denotes the spectral norm, hence the first constraint is equivalent to the first constraint in the maximization in (26) by Schur-complement argument. Using standard arguments, it can be shown that $\hat{\mathrm{d}} \leq \|\hat{\mathbf{Z}}\|_*$ and equality holds by setting

$$\hat{\mathbf{B}} = \mathbf{U}\mathbf{V}^T, \tag{30}$$

where we recalled the compact SVD $\hat{\mathbf{Z}} = \mathbf{V}\mathbf{\Lambda}\mathbf{U}^T$. It is also not hard to see that $\hat{\mathbf{B}}$ is feasible in (29) (recall here that $\mathbf{V}^T\mathbb{1}_k = 0$). Hence, $\hat{\mathbf{B}}$ is a maximizer and $\hat{\mathrm{d}} = \|\hat{\mathbf{Z}}\|_*$. The following lemma, the proof of which we defer to the end of this section, proves something stronger: $\hat{\mathbf{B}}$ is in fact the only maximizer in (29).

Lemma C.1. *The optimal cost of the maximization in* (29) *is* $\|\hat{\mathbf{Z}}\|_*$ *and* $\hat{\mathbf{B}} = \mathbf{U}\mathbf{V}^T$ *is its unique maximizer.*

Next, we use Lemma C.1 to show that $\hat{\mathbf{B}}$ also satisfies the inequality constraints in (27). To do this, we use an explicit construction of the singular factors $\mathbf{U}$ and $\mathbf{V}$ presented in Sec. A.2, which is possible thanks to the special structure of $\hat{\mathbf{Z}}$. Specifically, we prove in Lemma A.3 that

$$\hat{\mathbf{B}} \odot \hat{\mathbf{Z}}^T = (\mathbf{U}\mathbf{V}^T) \odot \hat{\mathbf{Z}}^T > 0. \tag{31}$$

Hence, $\hat{\mathbf{B}}$ is feasible in (26). In fact, the feasibility inequalities are strict, which we will use soon. For now, note that feasibility of $\hat{\mathbf{B}}$ in (26) guarantees that it is its unique maximizer (since it is the unique maximizer of the program's relaxation, as established in Lemma C.1.) Therefore,

$$\mathrm{d}_* = \hat{\mathrm{d}} = \|\hat{\mathbf{Z}}\|_*\,, \tag{32}$$

and $\hat{\mathbf{B}}$ in (30) is dual optimal for the semidefinite program in (22).

<u>*Solution to the primal relaxation:*</u> By strong duality, this implies $\mathrm{q}_* = \mathrm{d}_* = \|\hat{\mathbf{Z}}\|_*$ and that any primal minimizer $\hat{\mathbf{X}} = \begin{bmatrix} \hat{\mathbf{X}}_{11} & \hat{\mathbf{X}}_{12} \\ \hat{\mathbf{X}}_{12}^T & \hat{\mathbf{X}}_{22} \end{bmatrix}$ satisfies the complementary slackness conditions in (28) with $\mathbf{B} = \hat{\mathbf{B}}$, i.e.,

$$\forall i \in [n], c \neq y_i : \hat{\mathbf{X}}_{12}[y_i, i] - \hat{\mathbf{X}}_{12}[c, i] - 1 = 0 \tag{33}$$

$$\hat{\mathbf{X}}_{11} - \hat{\mathbf{B}}^T\hat{\mathbf{X}}_{12}^T = \hat{\mathbf{X}}_{12} - \hat{\mathbf{B}}^T\hat{\mathbf{X}}_{22} = -\hat{\mathbf{B}}\hat{\mathbf{X}}_{12} + \hat{\mathbf{X}}_{22} = 0\,. \tag{34}$$

In (33) we used from (31) that inequalities are strict; thus, $\beta_{ic} > 0$ for all $i \in [n], c \neq y_i$. Now, from (34) it follows that

$$\hat{\mathbf{X}}_{11} = \hat{\mathbf{B}}^T \hat{\mathbf{X}}_{12}^T, \quad \hat{\mathbf{X}}_{22} = \hat{\mathbf{B}} \hat{\mathbf{X}}_{12}, \quad \hat{\mathbf{X}}_{12} = \hat{\mathbf{B}}^T \hat{\mathbf{X}}_{22}.$$

Combining the last two equations shows

$$\begin{aligned}(\mathbb{I}_n - \hat{\mathbf{B}}\hat{\mathbf{B}}^T)\hat{\mathbf{X}}_{22} = 0 &\implies (\mathbb{I}_n - \mathbf{U}\mathbf{U}^T)\hat{\mathbf{X}}_{22} = 0 \implies \hat{\mathbf{X}}_{22} = \mathbf{U}\mathbf{D}\mathbf{U}^T, \\ &\implies \hat{\mathbf{X}}_{12} = \hat{\mathbf{B}}^T \hat{\mathbf{X}}_{22} = \mathbf{V}\mathbf{D}\mathbf{U}^T \qquad \text{and} \qquad \hat{\mathbf{X}}_{11} = \hat{\mathbf{B}}^T \hat{\mathbf{X}}_{12}^T = \mathbf{V}\mathbf{D}\mathbf{V}^T\end{aligned}$$

for some (k-1)–dimensional matrix $\mathbf{D} \in \mathbb{R}^{(k-1)\times(k-1)}$.

Next, we will use $\hat{\mathbf{X}}_{12} = \mathbf{V}\mathbf{D}\mathbf{U}^T$ in (33) to compute $\mathbf{D}$. For convenience denote $\mathbf{e}_{k,c} \in \mathbb{R}^k$ the c-th standard basis vector in $\mathbb{R}^k$ and $\mathbf{e}_{n,i} \in \mathbb{R}^n$ the i-th standard basis vector in $\mathbb{R}^n$. Then, starting with (33), we get the following chain of implications:

$$\begin{aligned}&\forall i \in [n], c \neq y_i : (\mathbf{e}_{k,y_i} - \mathbf{e}_{k,c})^T \mathbf{V}\mathbf{D}\mathbf{U}^T \mathbf{e}_{n,i} = 1 \qquad (35)\\ &\overset{(\Sigma_{c\neq y_i}\cdot)}{\implies} \forall i \in [n] : \big((k-1)\mathbf{e}_{k,y_i} - (\mathbb{1}_k - \mathbf{e}_{k,y_i})\big)^T \mathbf{V}\mathbf{D}\mathbf{U}^T \mathbf{e}_{n,i} = k-1 \\ &\implies \forall i \in [n] : \big(k\mathbf{e}_{k,y_i} - \mathbb{1}_k\big)^T \mathbf{V}\mathbf{D}\mathbf{U}^T \mathbf{e}_{n,i} = k-1 \\ &\overset{(\mathbf{V}^T\mathbb{1}_k=0)}{\implies} \forall i \in [n] : \mathbf{e}_{k,y_i}^T \mathbf{V}\mathbf{D}\mathbf{U}^T \mathbf{e}_{n,i} = 1 - 1/k \\ &\overset{(35)}{\implies} \forall i \in [n] : \mathbf{e}_{k,y_i}^T \mathbf{V}\mathbf{D}\mathbf{U}^T \mathbf{e}_{n,i} = 1 - 1/k \qquad \text{and} \qquad \mathbf{e}_{k,c}^T \mathbf{V}\mathbf{D}\mathbf{U}^T \mathbf{e}_{n,i} = -1/k,\ c \neq y_i\end{aligned}$$

Specifically, the last equation written in matrix form gives [5]

$$\mathbf{V}\mathbf{D}\mathbf{U}^T = \hat{\mathbf{Z}} \implies \mathbf{D} = \mathbf{V}^T \hat{\mathbf{Z}} \mathbf{U} \implies \mathbf{D} = \mathbf{\Lambda}, \qquad (36)$$

where the second and third equalities used $\mathbf{V}^T\mathbf{V} = \mathbf{U}^T\mathbf{U} = \mathbb{I}_{k-1}$.

To conclude, we have shown that any optimal point $\hat{\mathbf{X}}$ of (22) satisfies

$$\hat{\mathbf{X}} = \begin{bmatrix}\mathbf{V}\\ \mathbf{U}\end{bmatrix} \mathbf{\Lambda} \begin{bmatrix}\mathbf{V}^T & \mathbf{U}^T\end{bmatrix}. \qquad (37)$$

Solving the original problem: Now, we show that the convex relaxation in (22) is tight. For some partial orthonormal matrix $\mathbf{R} \in \mathbb{R}^{(k-1)\times d}$ (recall that $d \geq k-1$) with $\mathbf{R}\mathbf{R}^T = \mathbb{I}_{k-1}$, let

$$\hat{\mathbf{W}} = \mathbf{R}^T \mathbf{\Lambda}^{1/2} \mathbf{V}^T \qquad \text{and} \qquad \hat{\mathbf{H}} = \mathbf{R}^T \mathbf{\Lambda}^{1/2} \mathbf{U}^T. \qquad (38)$$

By construnction $\hat{\mathbf{W}}^T\hat{\mathbf{H}} = \hat{\mathbf{X}}_{12}$. Hence, $(\hat{\mathbf{W}}, \hat{\mathbf{H}})$ is feasible in (2). Thus, $p_* \leq \frac{1}{2}\|\hat{\mathbf{W}}\|_F^2 + \frac{1}{2}\|\hat{\mathbf{H}}\|_F^2 = \frac{1}{2}\operatorname{tr}(\hat{\mathbf{X}}) = q_*$. But, we have already argued that $p_* \geq q_*$. Hence, $p_* = q_*$.

Take now any minimizer $(\hat{\mathbf{W}}, \hat{\mathbf{H}})$ of the original problem (2). By feasibility of $(\hat{\mathbf{W}}, \hat{\mathbf{H}})$, the matrix $\mathbf{\Omega} := \begin{bmatrix}\hat{\mathbf{W}}^T\\ \hat{\mathbf{H}}^T\end{bmatrix}\begin{bmatrix}\hat{\mathbf{W}} & \hat{\mathbf{H}}\end{bmatrix}$ is feasible in (22). Also, $\frac{1}{2}\operatorname{tr}(\mathbf{\Omega}) = \frac{1}{2}\|\hat{\mathbf{W}}\|_F^2 + \frac{1}{2}\|\hat{\mathbf{H}}\|_F^2 = p_*$. But, $p_* = q_*$. Hence, $\frac{1}{2}\operatorname{tr}(\mathbf{\Omega}) = q_*$, which implies that $\mathbf{\Omega}$ is optimal. It must then be from (37) that

$$\begin{bmatrix}\hat{\mathbf{W}}^T\\ \hat{\mathbf{H}}^T\end{bmatrix}\begin{bmatrix}\hat{\mathbf{W}} & \hat{\mathbf{H}}\end{bmatrix} = \mathbf{\Omega} = \hat{\mathbf{X}} = \begin{bmatrix}\mathbf{V}\\ \mathbf{U}\end{bmatrix} \mathbf{\Lambda} \begin{bmatrix}\mathbf{V}^T & \mathbf{U}^T\end{bmatrix}.$$

This completes the proof of the theorem.

Proof of Lemma C.1: It only remains to prove Lemma C.1, which we do here. Any feasible $\mathbf{B}$ satisfies $\|\mathbf{B}\|_2 \leq 1$. Hence, also recalling that $\hat{\mathbf{Z}}$ has rank $k-1$ (because $\mathbf{V}^T\mathbb{1}_k = 0$):

$$\operatorname{tr}(\hat{\mathbf{Z}}\mathbf{B}) = \sum_{i=1}^{k-1} \mathbf{\Lambda}[i,i](\mathbf{u}_i^T \mathbf{B} \mathbf{v}_i) \leq \sum_{i=1}^{k-1} \mathbf{\Lambda}[i,i] = \|\hat{\mathbf{Z}}\|_*. \qquad (39)$$

[5] Note that the derivation and conclusion following (35) is general. Specifically, the display above shows that $\hat{\mathbf{Z}}$ is the only $k \times n$ matrix for which it holds simultaneously that $\mathbf{Z}[y_i, i] - \mathbf{Z}[c, i] = 1, \forall i \in [n], c \neq y_i$ and $\mathbb{1}_k^T \mathbf{Z} = 0$.

The inequality above is tight if and only if

$$\forall i \in [k-1] : \epsilon_i := \mathbf{u}_i^T \mathbf{B} \mathbf{v}_i = 1, \tag{40}$$

which is indeed satisfied by $\hat{\mathbf{B}} = \mathbf{U}\mathbf{V}^T$. Clearly, $\hat{\mathbf{B}}$ is also feasible. Hence, $\hat{\mathbf{B}}$ is a maximizer of (29).

Next, we will show that there is no other maximizer, say $\widetilde{\mathbf{B}}$. Indeed, since $\widetilde{\mathbf{B}}$ is optimal, it must satisfy (40). Hence, it has rank at least $k-1$. But since $\widetilde{\mathbf{B}}\mathbb{1}_k = 0$, we find that $\widetilde{\mathbf{B}}$ has rank exactly $k-1$. Let $\widetilde{\mathbf{B}} = \mathbf{U_B}\mathbf{\Sigma_B}\mathbf{V_B}^T$ be its compact SVD and denote $\sigma_i := \mathbf{\Sigma_B}[i,i],\ i \in [k-1]$ its singular values. Finally, define $(k-1)\times(k-1)$ matrices

$$\mathbf{P} = \mathbf{U_B}^T\mathbf{U} \qquad \text{and} \qquad \mathbf{Q} = \mathbf{V_B}^T\mathbf{V} \tag{41}$$

with columns $\mathbf{p}_i, \mathbf{q}_i$, $i \in [k-1]$. By Equation (40) we have the following chain of inequalities for all $i \in [k-1]$:

$$\begin{aligned} 1 = \mathbf{u}_i^T\mathbf{B}\mathbf{v}_i = \mathbf{p}_i^T\mathbf{\Sigma_B}\mathbf{q}_i &= \sum_{j\in[k-1]} \sigma_j \mathbf{p}_i[j]\mathbf{q}_i[j] \le \sum_{j\in[k-1]} \sigma_j |\mathbf{p}_i[j]\mathbf{q}_i[j]| \\ &\le \sum_{j\in[k-1]} |\mathbf{p}_i[j]\mathbf{q}_i[j]| \le \|\mathbf{p}_i\|_2\|\mathbf{q}_i\|_2 \le 1. \end{aligned}$$

Inspecting this, we note that all inequalities must be equalities. The first inequality in the second line follows because $\|\widetilde{\mathbf{B}}\|_2 \le 1 \implies \forall j \in [k-1] : \sigma_j \le 1$. Hence, $\sigma_j = 1$ for all $j \in [k-1]$. Equivalently $\mathbf{\Sigma_B} = \mathbb{I}_{k-1}$. The second inequality in that same line is Cauchy-Schwarz and equality implies

$$\forall i \in [k-1] : \mathbf{p}_i = \pm\mathbf{q}_i. \tag{42}$$

The last inequality follows because

$$\|\mathbf{p}_i\|_2 = \|\mathbf{U_B}^T\mathbf{u}_i\|_2 \le \|\mathbf{U_B}\|_2\|\mathbf{u}_i\|_2 \le 1 \qquad \text{and} \qquad \|\mathbf{q}_i\|_2 = \|\mathbf{V_B}^T\mathbf{v}_i\|_2 \le \|\mathbf{V_B}\|_2\|\mathbf{v}_i\|_2 \le 1. \tag{43}$$

Since $\mathbf{U_B}$ (resp. $\mathbf{V_B}$) has orthonormal columns, equality in (43) holds if and only if for all $i \in [k-1]$, $\mathbf{u}_i$ (resp. $\mathbf{v}_i$) is a column of $\mathbf{U_B}$ (resp. $\mathbf{V_B}$). Then, it must be that $\mathbf{P}$ and $\mathbf{Q}$ are permutation matrices. Combined with (42) this gives

$$\mathbf{P} = \mathbf{Q} = \mathbf{\Pi} \implies \mathbf{U_B}^T\mathbf{U} = \mathbf{V_B}^T\mathbf{V} = \mathbf{\Pi} \tag{44}$$

for some permutation matrix $\mathbf{\Pi}$. Continuing from this,

$$\mathbf{U_B}^T\mathbf{U}\mathbf{\Pi}^T = \mathbf{V_B}^T\mathbf{V}\mathbf{\Pi}^T = \mathbb{I}_{k-1} \implies \mathbf{U_B} = \mathbf{U}\mathbf{\Pi}^T \text{ and } \mathbf{V_B} = \mathbf{V}\mathbf{\Pi}^T.$$

Putting things together, we conclude that

$$\widetilde{\mathbf{B}} = \mathbf{U_B}\mathbf{\Sigma_B}\mathbf{V_B}^T = \mathbf{U_B}\mathbf{V_B}^T = \mathbf{U}\mathbf{\Pi}^T\mathbf{\Pi}\mathbf{V}^T = \mathbf{U}\mathbf{V}^T = \hat{\mathbf{B}}.$$

This concludes the proof of the lemma.

C.2 Proofs of Corollaries 1.1, 1.2, 1.3 and of Lemma 3.1

The proofs of the rest of the results of Sec. 3 are presented in the following sections.

- **Corollary 1.1:** See paragraph below the statement of the corollary in Sec. 3.1.
- **Corollary 1.2:** As already mentioned in Sec. 3.1 this is nothing but a reformulation of Theorem1(i)–(ii) in view of Definition 3 of the SELI property.
- **Corollary 1.3:** See Corollary 2.3 in Sec. B.
- **Lemma 3.1:** See Lemma B.1 in Sec. B.

C.3 On different regularization hyperparameters between embeddings and classifiers

Thus far in our analysis of the UFM, we assumed same regularization strength λ for the embeddings and classifiers; see Eqn. (1). Here, we discuss a slight generalization allowing different regularizations λ_W and λ_H for the classifiers and embeddings, respectively. This is motivated by previous studies of the UFM in the literature, e.g. [41]. As we show, this modification does *not* change the SELI geometry modulo a relative scaling factor between the embeddings and classifiers.

Concretely, consider the following slight generalization version of Eqn. (1):

$$(\hat{\mathbf{W}}_{(\lambda_W,\lambda_H)}, \hat{\mathbf{H}}_{(\lambda_W,\lambda_H)}) \coloneqq \arg\min_{\mathbf{W},\mathbf{H}} \ \mathcal{L}(\mathbf{W}^T\mathbf{H}) + \frac{\lambda_W}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_H}{2}\|\mathbf{H}\|_F^2,$$

which can be more conveniently reparameterized as follows:

$$(\hat{\mathbf{W}}_{(\lambda_W,\lambda_H)}, \hat{\mathbf{H}}_{(\lambda_W,\lambda_H)}) \coloneqq \arg\min_{\mathbf{W},\mathbf{H}} \ \mathcal{L}(\mathbf{W}^T\mathbf{H}) + \sqrt{\lambda_W\lambda_H}\left(\frac{1}{2\beta}\|\mathbf{W}\|_F^2 + \frac{\beta}{2}\|\mathbf{H}\|_F^2\right),$$

where $\beta \coloneqq \sqrt{\frac{\lambda_H}{\lambda_W}}$ and $\lambda_H, \lambda_W > 0$. Now, consider the limit of vanishing regularization $\lambda_W \to 0, \lambda_H \to 0$, with a fixed finite and non-zero ratio $\beta^2 = \lambda_H/\lambda_W$. Entirely analogous to Eqn. (2), this leads to the a β-parameterized UF-SVM as follows:

$$(\hat{\mathbf{W}}_\beta, \hat{\mathbf{H}}_\beta) \in \arg\min_{\mathbf{W},\mathbf{H}} \ \frac{1}{2\beta}\|\mathbf{W}\|_F^2 + \frac{\beta}{2}\|\mathbf{H}\|_F^2 \quad \text{sub. to} \quad (\mathbf{w}_{y_i} - \mathbf{w}_c)^T\mathbf{h}_i \geq 1, \ i \in [n], c \neq y_i. \tag{45}$$

Note that the UF-SVM we study in Eqn. (2) is a special case of the above for $\beta = 1$. For general values of $\beta > 0$, it is not hard to see that there is a one-to-one mapping of global solutions $(\hat{\mathbf{W}}_\beta, \hat{\mathbf{H}}_\beta)$ of (45) to global solutions $(\hat{\mathbf{W}}_\beta, \hat{\mathbf{H}}_\beta)$ of (2) as follows:

$$(\hat{\mathbf{W}}_\beta, \hat{\mathbf{H}}_\beta) = (\sqrt{\beta}\hat{\mathbf{W}}, \frac{1}{\sqrt{\beta}}\hat{\mathbf{H}}_\beta).$$

Hence, from Theorem 1, it follows that the solutions of (45) satisfy for any $\beta > 0$ the following:

$$\hat{\mathbf{W}}_\beta^T\hat{\mathbf{H}}_\beta = \hat{\mathbf{Z}}, \quad \hat{\mathbf{H}}_\beta^T\hat{\mathbf{H}}_\beta = \frac{1}{\beta}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \quad \text{and} \quad \hat{\mathbf{W}}_\beta^T\hat{\mathbf{W}}_\beta = \beta\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T.$$

Therefore, different regularization between embeddings and classifiers only affects the geometry of global minimizers of the corresponding UF-SVM (i.e., at vanishing regularization) up to introducing an extra scaling factor between the Gram matrices of embeddings and classifiers. Specifically, this only affects the relative scaling between the norms of embeddings and classifiers.

D Nuclear-norm relaxations of the UFM

In this section, we gather useful properties of the nuclear-norm-regularized CE minimization (7), repeated here for convenience:

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

These properties are useful in the proof of the results that appear in Sec. 4. As a reminder, the nuclear-norm-regularized CE minimization is relevant to us because of Theorem 2, i.e. it forms a tight convex relaxation of the non-convex ridge-regularized CE-minimization for the UFM.

The following lemma gathers some basic properties, which we use later to study the behavior of the solutions to (7) for different regularization strengths.

Lemma D.1 (Basic properties of (7))**.** *The following statements are true.*

(i) There is a unique minimizer $\hat{\mathbf{Z}}_\lambda$.

(ii) $\hat{\mathbf{Z}}_\lambda$ *satisfies the following necessary and sufficient first-order optimality conditions:*

$$\nabla_{\mathbf{Z}}\mathcal{L}(\hat{\mathbf{Z}}_\lambda)\mathbf{U}_\lambda = -\lambda\mathbf{V}_\lambda, \qquad \left(\nabla_{\mathbf{Z}}\mathcal{L}(\hat{\mathbf{Z}}_\lambda)\right)^T\mathbf{V}_\lambda = -\lambda\mathbf{U}_\lambda, \qquad \|\nabla_{\mathbf{Z}}\mathcal{L}(\hat{\mathbf{Z}}_\lambda)\|_2 \le \lambda. \quad (46)$$

where, $\hat{\mathbf{Z}}_\lambda = \mathbf{V}_\lambda\mathbf{\Lambda}_\lambda\mathbf{U}_\lambda^T$ *is its compact SVD.*

(iii) It holds $\mathbb{1}_k^T\mathbf{V}_\lambda = 0$*. Thus also,* $\mathbb{1}_k^T\hat{\mathbf{Z}}_\lambda = 0$*.*

Proof. We prove each statement separately below.

Proof of (ii): This is straightforward from first-order optimality of the convex program (7). For example, see [41, Lemma C.3] for details.

Proof of (iii): Start from Statement (ii) and use from Lemma A.2 that $\mathbb{1}_k^T\nabla_{\mathbf{Z}}\mathcal{L}(\hat{\mathbf{Z}}_\lambda) = 0$. Then, it must then be that $\mathbb{1}_k^T\mathbf{V}_\lambda = \frac{1}{\lambda}\mathbb{1}_k^T\nabla_{\mathbf{Z}}\mathcal{L}(\hat{\mathbf{Z}}_\lambda)\mathbf{U}_\lambda = 0$, which in turn implies $\mathbb{1}_k^T\hat{\mathbf{Z}}_\lambda = 0$.

Proof of (i): Now, we prove that $\hat{\mathbf{Z}}_\lambda$ is unique. This is a consequence of Lemma D.2 stated below and the fact that $\mathcal{L}(\mathbf{Z}) = \mathcal{L}(\mathbf{Z}) = \sum_{i\in[n]}\log\left(1+\sum_{c\neq y_i}e^{-(\mathbf{Z}[y_i,i]-\mathbf{Z}[c,i])}\right) = \sum_{i\in[n]}\ell_{y_i}(\mathbf{z}_i)$ where $\mathbf{z}_i$ are the columns of $\mathbf{Z}$. Specifically, suppose there were two minimizers $\hat{\mathbf{Z}}_\lambda = [\mathbf{z}_1,\dots,\mathbf{z}_n]$ and $\hat{\mathbf{Z}}_\lambda + \mathbf{\Delta}$ for $\mathbf{\Delta} = [\boldsymbol{\delta}_1,\dots,\boldsymbol{\delta}_n] \neq 0$. From Statement (iii), it must be that $\mathbb{1}_k^T\mathbf{\Delta} = 0$. Thus, there exists $j\in[n]$ such that $(\mathbb{I}_k - \frac{1}{k}\mathbb{1}_k\mathbb{1}_k^T)\boldsymbol{\delta}_j \neq 0$. With these, we have from Lemma D.2 that

$$\begin{aligned}
\mathcal{L}(\hat{\mathbf{Z}}_\lambda + \mathbf{\Delta}) &= \ell_{y_j}(\mathbf{z}_j + \boldsymbol{\delta}_j) + \sum_{i\neq j}\ell_{y_i}(\mathbf{z}_i + \boldsymbol{\delta}_i) \\
&> \ell_{y_j}(\mathbf{z}_j) + \left(\nabla_{\mathbf{z}}\ell_{y_j}(\mathbf{z}_j)\right)^T\boldsymbol{\delta}_j + \sum_{i\neq j}\ell_{y_i}(\mathbf{z}_i) + \sum_{i\neq j}\left(\nabla_{\mathbf{z}}\ell_{y_i}(\mathbf{z}_i)\right)^T\boldsymbol{\delta}_i \\
&= \mathcal{L}(\hat{\mathbf{Z}}_\lambda) + \operatorname{tr}\left(\left(\nabla_{\mathbf{Z}}\mathcal{L}(\hat{\mathbf{Z}}_\lambda)\right)^T\mathbf{\Delta}\right) = \mathcal{L}(\hat{\mathbf{Z}}_\lambda),
\end{aligned}$$

where the last equality uses optimality of $\hat{\mathbf{Z}}_\lambda$. The above display contradicts optimality of $\hat{\mathbf{Z}}_\lambda + \mathbf{\Delta}$ and completes the proof. □

Lemma D.2 (Auxiliary result—Strict convexity of CE)**.** *For some* $y \in [k]$ *and* $\mathbf{z} = [z_1,\dots,z_k]$*, let* $\ell_y(\mathbf{z}) := \log\left(1+\sum_{c\neq y}e^{-(\mathbf{z}[y]-\mathbf{z}[c])}\right)$*. The function* ℓ_y *is strictly convex along any direction on the* $(k-1)$*-dimensional subspace orthogonal to* $\mathbb{1}_k$*. That is, for all* non-zero $\mathbf{v}\in\mathbb{R}^k$ *such that* $(\mathbb{I}_k - \frac{1}{k}\mathbb{1}_k\mathbb{1}_k^T)\mathbf{v} \neq 0$*, it holds for all* $\mathbf{z}\in\mathbb{R}^k$ *that* $\ell_y(\mathbf{z}+\mathbf{v}) > \ell_y(\mathbf{z}) + \left(\nabla_{\mathbf{z}}\ell_y(\mathbf{z})\right)^T\mathbf{v}$*.*

Proof. Define univariate function $g(t) = \ell_y(\mathbf{z}+t\mathbf{v})$. From Taylor's expansion, for some $\theta\in(0,1)$

$$\ell_y(\mathbf{z}+\mathbf{v}) = \ell_y(\mathbf{z}) + \left(\nabla_{\mathbf{z}}\ell_y(\mathbf{z})\right)^T\mathbf{v} + \frac{1}{2}\mathbf{v}^T\nabla_{\mathbf{z}}^2\ell_y(\mathbf{z}+\theta\mathbf{v})\mathbf{v}.$$

Hence, it will suffice showing that $\mathbf{v}^T\nabla_{\mathbf{z}}^2\ell_y(\mathbf{z})\mathbf{v} > 0$ for any $\mathbf{z}$. Denote for convenience vector $\mathbf{a}\in\mathbb{R}^k$ with $\mathbf{a}[c] := \frac{e^{-(\mathbf{z}[y]-\mathbf{z}[c])}}{1+\sum_{c'\neq y}e^{-(\mathbf{z}[y]-\mathbf{z}[c'])}}$ for all $c\in[k]$. From Lemma A.2, we have $\nabla_{\mathbf{z}}\ell_y(\mathbf{z}) = -\mathbf{e}_y + \mathbf{a}$. Thus, a straightforward calculation yields $\nabla_{\mathbf{z}}^2\ell_y(\mathbf{z}) = \operatorname{diag}(\mathbf{a}) - \mathbf{a}\mathbf{a}^T$. For the sake of contradiction assume $\mathbf{v}^T\nabla_{\mathbf{z}}^2\ell_y(\mathbf{z})\mathbf{v} = 0$. Then, since $\sum_{c\in[k]}\mathbf{a}[c] = 1$ and $\mathbf{a}[c]\ge 0$, it must be from Cauchy-Schwartz:

$$\sum_c \mathbf{v}[c]^2\mathbf{a}[c] = \left(\sum_c \mathbf{v}[c]\mathbf{a}[c]\right)^2 \le \left(\sum_c \mathbf{v}[c]^2\mathbf{a}[c]\right)^2\left(\sum_c \mathbf{a}[c]\right)^2 = \sum_c \mathbf{v}[c]^2\mathbf{a}[c]$$

that $\mathbf{v} = C\mathbf{a}$. But then, $\mathbf{v}^T\mathbb{1}_k = 0$, which violates the lemma's assumption. □

D.1 Large λ

Lemma D.3 (Large λ behavior of (7))**.** *The minimizer is zero, i.e.* $\hat{\mathbf{Z}}_\lambda = 0$*, if and only if* $\lambda \geq \|\hat{\mathbf{Z}}\|_2$*, where* $\hat{\mathbf{Z}}$ *is the SEL matrix.*

Proof. From the KKT conditions in Lemma D.1(ii), $\hat{\mathbf{Z}}_\lambda = 0$ is optimal if and only if $\|\nabla_{\mathbf{Z}}\mathcal{L}(0)\|_2 \leq \lambda$. The key observation here is that $\nabla_{\mathbf{Z}}\mathcal{L}(0) = \mathbf{Y} - \frac{1}{k}\mathbb{1}_k\mathbb{1}_n^T = \hat{\mathbf{Z}}$. (See Lemma A.2 for a formula evaluating the gradient.) Thus, $\|\nabla_{\mathbf{Z}}\mathcal{L}(0)\|_2 = \|\hat{\mathbf{Z}}\|_2$, which completes the proof. □

We note that, for (R,ρ)-STEP imbalance the necessary and sufficient condition of the lemma for $\hat{\mathbf{Z}}_\lambda = 0$ becomes $\lambda \geq \|\hat{\mathbf{Z}}\|_2 = \sqrt{R}$ (see Lemma A.3.)

D.2 Small λ

Lemma D.4 (Small λ behavior of (7))**.** *If* $\lambda < \frac{1}{2}$*, then* $\hat{\mathbf{Z}}_\lambda$ *is such that it linearly separates the data, i.e. it holds* $\hat{\mathbf{Z}}_\lambda[y_i,i] - \hat{\mathbf{Z}}_\lambda[c,i] > 0$ *for all* $i \in [n], c \neq y_i$*.*

Proof. For the sake of contradiction assume $\lambda < 1/2$ and there exists $j \in [n]$ and $c \neq y_j$ such that $\hat{\mathbf{Z}}_\lambda[y_j,j] - \hat{\mathbf{Z}}_\lambda[c,j] \leq 0$. Denote $\mathbf{G} = \nabla_{\mathbf{Z}}\mathcal{L}(\hat{\mathbf{Z}}_\lambda)$ the gradient. From KKT conditions in Lemma D.1(ii), it must be that

$$\|\mathbf{G}\|_2 \leq \lambda < 1/2. \tag{47}$$

A simple calculation (e.g., see Lemma A.2) gives $|\mathbf{G}[y_j,j]| = 1 - \frac{1}{1+\sum_{c'\neq y_j} e^{-(\hat{\mathbf{Z}}_\lambda[y_j,j]-\hat{\mathbf{Z}}_\lambda[c',j])}}$.

But, from assumption on j,c:

$$1 + \sum_{c'\neq y_j} e^{-(\hat{\mathbf{Z}}_\lambda[y_j,j]-\hat{\mathbf{Z}}_\lambda[c',j])} = 1 + e^{-(\hat{\mathbf{Z}}_\lambda[y_j,j]-\hat{\mathbf{Z}}_\lambda[c,j])} + \sum_{c'\notin\{y_j,c\}} e^{-(\hat{\mathbf{Z}}_\lambda[y_j,j]-\hat{\mathbf{Z}}_\lambda[c',j])} \geq 2.$$

Hence, $|\mathbf{G}[y_j,j]| \geq 1/2$. Since $\|\mathbf{G}\|_2 \geq |\mathbf{G}[y_j,j]|$, this in turn implies that $\|\mathbf{G}\|_2 \geq 1/2$, which contradicts (47). □

In Sec. H we combine Theorem 2 with Lemma D.4 above to show that the solution $(\hat{\mathbf{W}}_\lambda, \hat{\mathbf{H}}_\lambda)$ of (1) cannot satisfy the minority collapse property.

D.3 Vanishing λ

Proposition 3 (Vanishing λ behavior of (7))**.** *Assume* (R,ρ)*-STEP imbalance. Then*

$$\lim_{\lambda\to 0} \frac{\hat{\mathbf{Z}}_\lambda}{\|\hat{\mathbf{Z}}_\lambda\|_*} = \frac{\hat{\mathbf{Z}}}{\|\hat{\mathbf{Z}}\|_*}, \tag{48}$$

where $\hat{\mathbf{Z}}$ *is the SEL matrix.*

The proposition is an extension of [28, Theorem 3.1] to nuclear-norm regularization. Its proof follows the exact same steps as in Rosset et al. [28] who studied ℓ_p-regularization. The critical observation allowing this is that $\hat{\mathbf{Z}}$ is the unique minimizer of (2) thanks to Theorem 1. Because of that, we can get the following max-margin formulation for (the normalized) $\hat{\mathbf{Z}}$, which is key in the proof of Proposition 3.

Lemma D.5. *Assume* (R,ρ)*-STEP imbalance so that* $\hat{\mathbf{Z}}$ *is the unique solution of* (2) *(see Theorem 1). Then, it holds that*

$$\frac{\hat{\mathbf{Z}}}{\|\hat{\mathbf{Z}}\|_*} := \arg\max_{\|\mathbf{Z}\|_*\leq 1} \min_{i\in[n], c\neq y_i} \mathbf{Z}[y_i,i] - \mathbf{Z}[c,i]. \tag{49}$$

We present the proof of Proposition 3 together with the proof of Lemma D.5 in the next section.

D.3.1 Proof of Proposition 3

First, by KKT conditions (see Lemma D.1(ii)), as $\lambda \to 0$, we have $\|\nabla \mathcal{L}(\hat{\mathbf{Z}}_\lambda)\|_2 \to 0$. This, in turn, implies $\|\nabla \mathcal{L}(\hat{\mathbf{Z}}_\lambda)\|_* \to 0$ since $\|\nabla \mathcal{L}(\hat{\mathbf{Z}}_\lambda)\|_* \le (k-1)\|\nabla \mathcal{L}(\hat{\mathbf{Z}}_\lambda)\|_2$. Moreover, it implies that

$$m_\lambda := \min_{i\in[n], c\neq y_i} \hat{\mathbf{Z}}_\lambda[y_i, i] - \hat{\mathbf{Z}}_\lambda[c, i]$$

diverges. To see this, denote $\mathbf{G}_\lambda := \nabla \mathcal{L}(\hat{\mathbf{Z}}_\lambda)$ and note from Lemma A.2 that for all $i \in [n]$ $\mathbf{G}_\lambda[y_i, i] := -\big(1 + 1/\sum_{c\neq y_i} e^{-(\hat{\mathbf{Z}}_\lambda[y_i,i] - \hat{\mathbf{Z}}_\lambda[c,i])}\big)^{-1}$. But,

$$\begin{aligned}
\|\nabla \mathcal{L}(\hat{\mathbf{Z}}_\lambda)\|_2 &\ge \max_{i\in[n]} |\mathbf{G}_\lambda[y_i, i]| = \max_{i\in[n]} \big(1 + 1/\sum_{c\neq y_i} e^{-(\hat{\mathbf{Z}}_\lambda[y_i,i] - \hat{\mathbf{Z}}_\lambda[c,i])}\big)^{-1} \ge \big(1 + e^{m_\lambda}\big)^{-1} \\
&\implies m_\lambda \ge \log\big(\|\nabla \mathcal{L}(\hat{\mathbf{Z}}_\lambda)\|_2^{-1} - 1\big).
\end{aligned}$$

Thus, $\|\nabla \mathcal{L}(\hat{\mathbf{Z}}_\lambda)\|_2 \to 0 \implies m_\lambda \to +\infty$.

Assume for the sake of contradiction that $\lim_{\lambda\to 0} \frac{\hat{\mathbf{Z}}_\lambda}{\|\hat{\mathbf{Z}}_\lambda\|_*} = \widetilde{\mathbf{Z}}$ for $\widetilde{\mathbf{Z}} \neq \hat{\mathbf{Z}}/\|\hat{\mathbf{Z}}\|_*$. Then, by Lemma D.5, it must be that

$$m = \min_{i\in[n], c\neq y_i} \widetilde{\mathbf{Z}}[y_i, i] - \widetilde{\mathbf{Z}}[c, i] < \frac{1}{\|\hat{\mathbf{Z}}\|_*} \min_{i\in[n], c\neq y_i} \hat{\mathbf{Z}}[y_i, i] - \hat{\mathbf{Z}}[c, i] =: m_*.$$

Moreover, since $m_\lambda \to +\infty$, we also have that $m > 0$. The rest of the argument follows mutatis-mutandis the proof of [28, Theorem 2.1]. We repeat here for completeness. By continuity of the minimum margin in $\mathbf{Z}$, there exists open neighborhood of $\widetilde{\mathbf{Z}}$ on the nuclear-norm sphere:

$$\mathcal{N}_{\widetilde{\mathbf{Z}}} := \{\mathbf{Z} : \|\mathbf{Z}\|_* = 1, \|\mathbf{Z} - \widetilde{\mathbf{Z}}\|_2 \le \delta\}$$

and an $\epsilon > 0$ such that $\min_{i\in[n], c\neq y_i} \mathbf{Z}[y_i, i] - \mathbf{Z}[c, i] < m_* - \epsilon$ for all $\mathbf{Z} \in \mathcal{N}_{\widetilde{\mathbf{Z}}}$. To continue, we use the following lemma

Lemma D.6. *Assume $\mathbf{Z}_1, \mathbf{Z}_2$ such that $\|\mathbf{Z}_1\|_* = \|\mathbf{Z}_2\|_* = 1$ and*

$$0 < m_2 = \min_{i\in[n], c\neq y_i} \mathbf{Z}_2[y_i, i] - \mathbf{Z}_2[c, i] < \min_{i\in[n], c\neq y_i} \mathbf{Z}_1[y_i, i] - \mathbf{Z}_1[c, i] = m_1.$$

Then, there exists $T := T(m_1, m_2)$ such that

$$\forall t > T : \mathcal{L}(t\mathbf{Z}_1) < \mathcal{L}(t\mathbf{Z}_2).$$

By Lemma D.6, there exists $T := T(m_*, m_* - \epsilon)$ such that for all $t > T$ and all $\mathbf{Z} \in \mathcal{N}_{\widetilde{\mathbf{Z}}}$, $\mathcal{L}(t\hat{\mathbf{Z}}/\|\hat{\mathbf{Z}}\|_*) < \mathcal{L}(t\mathbf{Z})$. Therefore, $\widetilde{\mathbf{Z}}$ cannot be a convergence point of $\hat{\mathbf{Z}}_\lambda/\|\hat{\mathbf{Z}}_\lambda\|_*$.

We finish the proof by showing how to get Lemmas D.5 and D.6.

Proof of Lemma D.5. Clearly, $\frac{\hat{\mathbf{Z}}}{\|\hat{\mathbf{Z}}\|_*}$ is feasible in (49). Thus, $\max_{\|\mathbf{Z}\|_*\le 1} \min_{i\in[n], c\neq y_i} \mathbf{Z}[y_i, i] - \mathbf{Z}[c, i] \ge 1/\|\hat{\mathbf{Z}}\|_*$. Now, suppose there exists $\widetilde{\mathbf{Z}} \neq \frac{\hat{\mathbf{Z}}}{\|\hat{\mathbf{Z}}\|_*}$ such that $\|\widetilde{\mathbf{Z}}\|_* \le 1$ and is a maximizer in the max-min problem in (49). Then, $\min_{i\in[n], c\neq y_i} \widetilde{\mathbf{Z}}[y_i, i] - \widetilde{\mathbf{Z}}[c, i] \ge 1/\|\hat{\mathbf{Z}}\|_*$. This means, $\|\hat{\mathbf{Z}}\|_* \widetilde{\mathbf{Z}}$ is feasible in (2). But then it must be optimal therein since $\big\|\|\hat{\mathbf{Z}}\|_* \widetilde{\mathbf{Z}}\big\|_* = \|\hat{\mathbf{Z}}\|_* \|\widetilde{\mathbf{Z}}\|_* \le \|\hat{\mathbf{Z}}\|_*$. From Theorem 1, $\hat{\mathbf{Z}}$ is the unique minimizer of (49). Thus, we have shown that $\|\hat{\mathbf{Z}}\|_* \widetilde{\mathbf{Z}} = \hat{\mathbf{Z}}$, which contradicts the assumption on $\widetilde{\mathbf{Z}}$ and completes the proof.

Proof of Lemma D.6. Define $\epsilon := m_1/m_2 - 1 > 0$ and let T large enough such that $e^{-T\epsilon m_2} n(k-1) < 1/2$. We then have the following chain of inequalities for $t > T$

$$\begin{aligned}
\mathcal{L}(t\mathbf{Z}_1) &= \sum_{i\in[n]} \log\left(1 + \sum_{c\neq y_i} e^{-t(\mathbf{Z}_1[y_i,i]-\mathbf{Z}_1[c,i])}\right) \leq n\log(1+(k-1)e^{-t\,m_1}) \\
&\leq n(k-1)e^{-t\,m_1} = n(k-1)e^{-t\frac{m_1}{m_2}\,m_2} = n(k-1)e^{-t\,\epsilon\,m_2}e^{-t\,m_2} \\
&< n(k-1)e^{-T\,\epsilon\,m_2}e^{-t\,m_2} \\
&< \frac{e^{-t\,m_2}}{2} \leq \frac{e^{-t\,m_2}}{1+e^{-t\,m_2}} \\
&\leq \log(1+e^{-t\,m_2}) \\
&\leq \mathcal{L}(t\mathbf{Z}_2).
\end{aligned}$$

The inequality in the third line used that $t > T$ and $\epsilon > 0, m_2 > 0$. The next inequality follows by our choice of T. Throughout, we also used both sides of the inequality $\frac{x}{1+x} \leq \log(1+x) \leq x, x \geq 0$.

E Proofs for Regularized CE Section 4

E.1 Proof of Theorem 2

As mentioned in Remark 4.1 the theorem is drawn from [41, Theorem 3.2] with the following three small adjustments. Since the main proof argument remains essentially unaltered, we refer the reader to [41, Sec. C] for detailed derivations. Instead here, we only overview the necessary adjustments.

First, Theorem 2 holds for imbalanced classes. Technically, Zhu et al. [41] only consider balanced data. However, a close inspection of their proof shows that such a restriction is not necessary.

Second, Theorem 2 further shows that the nuclear-norm CE minimization in (7) has a unique solution. We prove this in Lemma D.1(i) in Sec. D.

Finally, Theorem 2 relaxes an assumption $d > k$ in [41, Theorem 3.2] to $d > k-1$. (In fact, Zhu et al. [41] conjecture that this relaxation is possible. We close the gap.) The assumption $d > k$ is only used by Zhu et al. [41] to show there exists nonzero $\mathbf{a} \in \mathbb{R}^d$ such that $\hat{\mathbf{W}}_\lambda^T \mathbf{a} = 0$ for a stationary point $(\hat{\mathbf{W}}_\lambda, \hat{\mathbf{H}}_\lambda)$ of (1). (This step is necessary to construct a negative curvature direction at stationary points for which $\|\nabla_{\mathbf{Z}}\mathcal{L}(\hat{\mathbf{W}}_\lambda^T\hat{\mathbf{H}}_\lambda)\|_2 > \lambda$; see [41, Sec. C.1].) Indeed, if $d > k$, then existence of $\mathbf{a}$ is guaranteed because $\hat{\mathbf{W}}_\lambda$ has k columns implying $\text{rank}(\hat{\mathbf{W}}_\lambda) \leq k$. To relax this requirement to $d > k-1$, we show in Lemma E.1 below that $\hat{\mathbf{W}}_\lambda \mathbb{1}_k = 0$. Hence, $\text{rank}(\hat{\mathbf{W}}_\lambda) \leq k-1$.

Lemma E.1. *Let $\hat{\mathbf{W}}_\lambda, \hat{\mathbf{H}}_\lambda$ be a stationary point of (1). Then, $\hat{\mathbf{W}}_\lambda \mathbb{1}_k = 0$. Similarly, any global minimizer $(\hat{\mathbf{W}}, \hat{\mathbf{H}})$ of (2) is such that $\hat{\mathbf{W}}\mathbb{1}_k = 0$.*

Proof. If $(\hat{\mathbf{W}}_\lambda, \hat{\mathbf{H}}_\lambda)$ is a stationary point of (1) and we denote $\hat{\mathbf{Z}}_\lambda = \hat{\mathbf{W}}_\lambda^T\hat{\mathbf{H}}_\lambda$, then by stationarity condition and Lemma A.2:

$$\nabla_{\mathbf{W}}\mathcal{L}(\hat{\mathbf{W}}_\lambda^T\hat{\mathbf{H}}_\lambda) = -\lambda\hat{\mathbf{W}}_\lambda \implies \hat{\mathbf{H}}_\lambda\left(\nabla_{\mathbf{Z}}\mathcal{L}(\hat{\mathbf{Z}}_\lambda)\right)^T = -\lambda\hat{\mathbf{W}}_\lambda \overset{\text{(Lem. A.2)}}{\implies} \hat{\mathbf{W}}_\lambda\mathbb{1} = 0. \tag{50}$$

Next, consider any global minimizer of (2). For the sake of contradiction assume that $\bar{\mathbf{w}} := \frac{1}{k}\sum_{c\in[k]}\hat{\mathbf{w}}_c \neq \mathbf{0}$, that is $\|\bar{\mathbf{w}}\|_2 > 0$. Consider the pair $(\mathbf{V}, \hat{\mathbf{H}})$ where the columns of $\mathbf{V}$ are defined each such that $\mathbf{v}_c = \hat{\mathbf{w}}_c - \bar{\mathbf{w}}$. Clearly, the new pair is feasible, since $(\hat{\mathbf{W}}, \hat{\mathbf{H}})$ is feasible. But,

$$\sum_{c\in[k]}\|\mathbf{v}_c\|^2 = \sum_{c\in[k]}\|\mathbf{w}_c\|^2 + k\|\bar{\mathbf{w}}\|^2 - \langle\bar{\mathbf{w}}, \sum_{c\in[k]}\mathbf{w}_c\rangle = \sum_{c\in[k]}\|\mathbf{w}_c\|^2 - k\|\bar{\mathbf{w}}\|^2 < \sum_{c\in[k]}\|\mathbf{w}_c\|^2,$$

which contradicts optimality of $(\hat{\mathbf{W}}, \hat{\mathbf{H}})$. Hence, it must be that $\bar{\mathbf{w}} = \mathbf{0}$. □

E.2 Proof of Proposition 1

We prove the proposition right below its statement in Sec. 4.2. The only remaining thing to show is that for $R > 1$ and $k > 2$, *not* all eigenvalues of the SEL matrix are same. This follows immediately from Lemma A.3 in Sec. A.2. Specifically, we show in (9), that if $R > 1$ and $k > 2$, then the maximum eigenvalue of $\mathbf{\Lambda}$ is $\sqrt{R}$ and the minimum one is 1 (thus, different).

E.3 Proof of Proposition 2

The proposition follows by combining Propisition 3 and Theorem 2. First, thanks to Theorem 2, for all $\lambda > 0$:

$$\frac{\hat{\mathbf{W}}_\lambda^T \hat{\mathbf{H}}_\lambda}{\|\hat{\mathbf{W}}_\lambda\|_F^2/2 + \|\hat{\mathbf{H}}_\lambda\|_F^2/2} = \frac{\hat{\mathbf{Z}}_\lambda}{\|\hat{\mathbf{Z}}_\lambda\|_*},$$

where $\hat{\mathbf{Z}}_\lambda$ is the solution to (7). Here, we used the fact from Equation (8) that

$$\|\hat{\mathbf{W}}_\lambda\|_F^2 + \|\hat{\mathbf{H}}_\lambda\|_F^2 = \operatorname{tr}(\hat{\mathbf{W}}_\lambda^T \hat{\mathbf{W}}_\lambda) + \operatorname{tr}(\hat{\mathbf{H}}_\lambda^T \hat{\mathbf{H}}_\lambda) = \operatorname{tr}(\mathbf{V}_\lambda \mathbf{\Lambda}_\lambda \mathbf{V}_\lambda^T) + \operatorname{tr}(\mathbf{U}_\lambda \mathbf{\Lambda}_\lambda \mathbf{U}_\lambda^T) = 2\operatorname{tr}(\mathbf{\Lambda}_\lambda) = 2\|\hat{\mathbf{Z}}_\lambda\|_*.$$

Next, from Proposition 3

$$\lim_{\lambda \to 0} \frac{\hat{\mathbf{Z}}_\lambda}{\|\hat{\mathbf{Z}}_\lambda\|_*} = \frac{\hat{\mathbf{Z}}}{\|\hat{\mathbf{Z}}\|_*}.$$

The desired follows by combining the above two displays.

E.4 Comparison to one-layer linear model

In Sec. 4.4, we compared the solution to the non-convex minimization (1), corresponding to a two-layer linear model, to the solution found by an one-layer linear model. Specifically, the one-layer linear model trains k-class linear classifier $\mathbf{\Xi}_{k \times n}$ by solving the following convex ridge-regularized CE:

$$\min_{\mathbf{\Xi}} \ \mathcal{L}(\mathbf{\Xi}) + \frac{\lambda}{2}\|\mathbf{\Xi}\|_F^2. \tag{51}$$

The following lemma computes the solution of this minimization.

Lemma E.2 (Linear model: Ridge-regularized CE)**.** *Let $\hat{\mathbf{\Xi}}_\lambda$ denote the solution of the convex ridge-regularized CE minimization. Irrespective of imbalances, for all $\lambda > 0$, there exists $\alpha_\lambda > 0$ such that $\hat{\mathbf{\Xi}}_\lambda = \alpha_\lambda \hat{\mathbf{Z}}$. Moreover, in the limit of $\lambda \to 0$, $\lim_{\lambda \to 0} \hat{\mathbf{\Xi}}_\lambda / \|\hat{\mathbf{\Xi}}_\lambda\|_F = \hat{\mathbf{\Xi}}_0 / \|\hat{\mathbf{\Xi}}_0\|_F$, where*

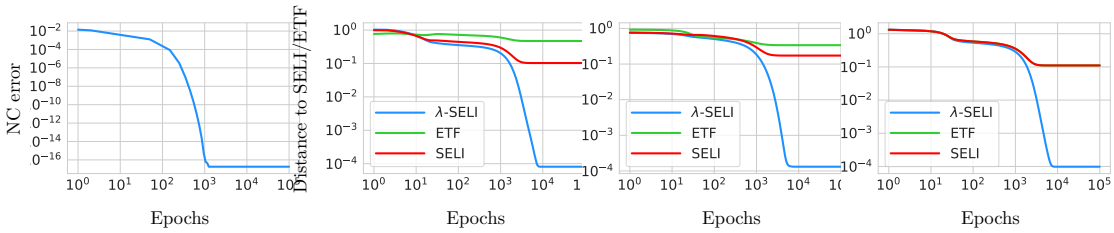
$$\hat{\mathbf{\Xi}}_0 := \arg\min_{\mathbf{\Xi}} \ \|\mathbf{\Xi}\|_2 \qquad \textit{sub. to} \ \ \mathbf{\Xi}[y_i, i] - \mathbf{\Xi}[c, i] \geq 1, \ \ \forall i \in [n], c \neq y_i. \tag{52}$$

In fact, the solution to (52) is the SEL matrix, i.e., $\hat{\mathbf{\Xi}}_0 = \hat{\mathbf{Z}}$.

According to the lemma above, the one-layer linear model always finds the SEL matrix: irrespective of imbalances and for any value of λ (including vanishing ones). On the other hand, by Proposition 1, we know that the end-to-end models minimizing ridge-regularized CE for a two-layer linear network correspond to the SEL matrix only if data are balanced, or there is two classes, or regularization is vanishing. Specifically, the solution is different when $k > 2$, $R = 1$ and $\lambda > 0$.

It is also worth noting the following connection between the one- and two-layer models. Thanks to the convex relaxation of Theorem 2 the end-to-end model $\mathbf{W}^T\mathbf{H}$ found by the two-layer model solves CE minimization with nuclear norm minimization (cf. (7)) compared to the ridge regularization in (51). Correspondingly, for vanishing regularization, the two-layer model corresponds to the "nuclear-norm SVM" in (5) compared to the vanilla SVM in (52). Notably, Theorem 1 proves that the solution to the former is always $\hat{\mathbf{Z}}$, i.e., the same as that of the latter.

Proof of Lemma E.2. The minimization in (51) is convex. Hence, it suffices to prove there exists α_λ such that that $\nabla_{\mathbf{Z}} \mathcal{L}(\alpha_\lambda \hat{\mathbf{Z}}) + \lambda \alpha_\lambda \hat{\mathbf{Z}} = 0$. Thanks to Lemma A.1(v), it suffices that



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

Figure 7: SGD solution with ridge-regularized CE converging to $(\hat{\mathbf{Z}}_\lambda, \hat{\mathbf{W}}_\lambda, \hat{\mathbf{H}}_\lambda)$ for $R = 10$ and $\lambda/n = 10^{-3}$. The iterates clearly converge towards λ-SELI (see Eqn. (8) in Theorem 2) and are far away from ETF. The distance to SELI also stays large as training progresses. However, note that it tracks the curves measuring distance to λ-SELI until $\sim 5e3$ epochs.

$\exists \alpha_\lambda$ such that $\lambda\alpha_\lambda = \frac{k}{e^{\alpha_\lambda}+k-1}$. It is easy to check that this equation always has a positive solution $\alpha_\lambda > 0$ since the LHS is increasing in $(0, \infty)$, the RHS is decreasing in the same interval, and they both take values in $(1, \infty)$. The fact that following the regularization path $\lambda \to 0$ leads (in direction) to the SVM solution follows from [28]. These two combined also show that the solution to (52) is $\hat{\mathbf{Z}}$. □

F Additional results on UFM

F.1 Experiments with weight-decay

In this section, we show experiments on the UFM supporting the claim of Theorem 2: the global solution $(\mathbf{W}_\lambda, \mathbf{H}_\lambda)$ of ridge-regularized CE in (1), call it "λ-SELI" for convenience (with some abuse of the term SELI, as the logit matrix does not represent the simplex encoding anymore for $\lambda > 0$), satisfies (8). Moreover any first-order method that avoids strict saddles converges to that global optimum [41].

Fig. 7 investigates the above claims in a setting of $(R = 10, \rho = 1/2)$-STEP imbalance, $k = 4$ classes and $n_{\min} = 1$, where we ran SGD on the ridge-regularized CE with $\lambda/n = 10^{-3}$. We set learning rate to 1 and implement ridge-regularization as weight-decay on the parameters. We observe the following. Fig. 7b, 7c, and 7d verify convergence to λ-SELI, while Fig. 7a verifies that the **(NC)** property also holds. Also, the solution is clearly away from ETF geometry (see green lines in Fig. 7). This is a noteworthy difference of the behavior of learning with imbalanced data, compared to that with balanced data. With balanced data, the geometry with ridge-regularization $\lambda > 0$ is always ETF. On the contrary, the geometry for learning from imbalanced data is sensitive to λ, as discussed in Sec. 4.2. While the distance from λ-SELI is the least, the distance to SELI is smaller compared to ETF. Thus, while SELI is not the "true" characterization when training with finite ridge-regularization, it is nevertheless a significantly better approximation than the ETF. In addition, compared to the λ-SELI solution, the SELI one admits explicit closed-form expressions (see Sec. B), rather than requiring numerical solution to a nuclear-norm CE minimization. Finally, note that, up to a certain point of time during the training, the distances from SELI and λ-SELI are comparable. Interestingly, the divergence between the two distances becomes more prominent at epoch count that also corresponds to a sharp fall of the NC error-curve in Fig. 7a.

Remark F.1 (A note on λ scaling)**.** *In Equation (1), the CE loss is* not *normalized by the number of examples. On the other hand, in all our experiments, we normalize the CE loss by $1/n$. This is why, when denoting the regularization used in experiments, we write λ/n in the axis-labels of all figures.*

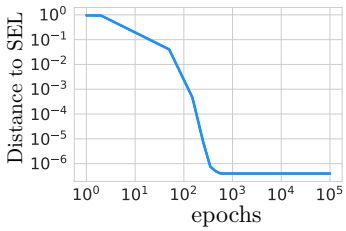
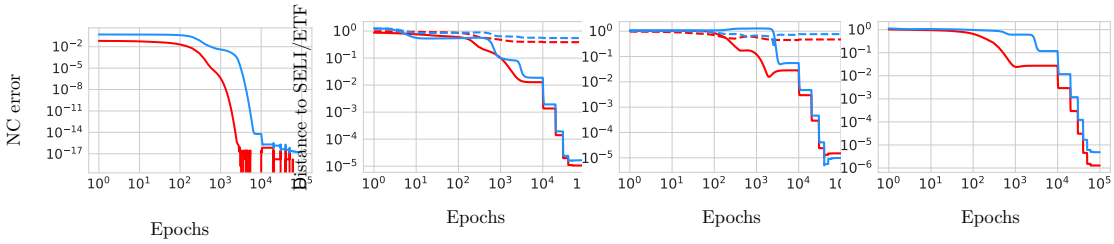


Figure 8: GD iterates on logit-regularized CE without ridge-regularization favor yielding logit matrix in the direction of SEL matrix $\hat{\mathbf{Z}}$.



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

Figure 9: Geometry of SGD solutions minimizing logit-regularized CE with ridge-decay on UFM; SELI(Solid)/ETF(Dashed), $R = 10$(Red)/$R = 100$(Blue); See Sec. F.2 for details.

F.2 Logit regularization and Ridge-decay

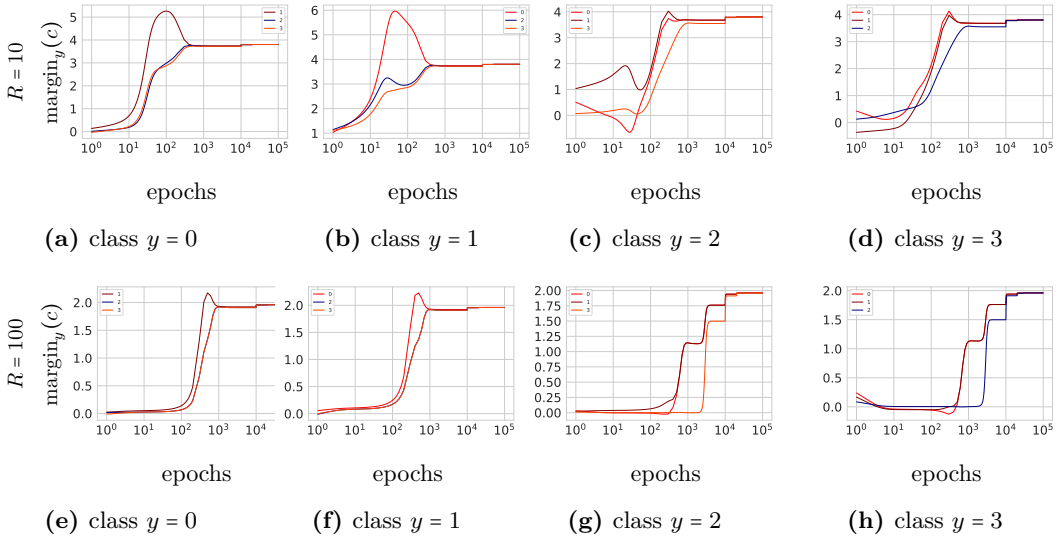
Our Proposition 2 proves that the optimal logits of CE minimization for the UFM with vanishing ridge-regularization is the SEL matrix. However, we also find in Fig. 4 that the convergence of SGD to the SEL matrix is rather slow. Motivated by this slow convergence we discuss here *logit-regularized* CE minimization, i.e., the solution to the following (non-convex) program:

$$\min_{\mathbf{W},\mathbf{H}} \mathcal{L}(\mathbf{W}^T\mathbf{H}) + \frac{\lambda}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda}{2}\|\mathbf{H}\|_F^2 + \lambda_L\|\mathbf{W}^T\mathbf{H}\|_F^2. \tag{53}$$

Note that when $\lambda = 0$, the global solutions of the above minimization give a logit matrix that aligns with the SEL matrix $\hat{\mathbf{Z}}$. This is easy to see by noting the resemblance to the convex program in (51) and invoking Lemma E.2.

Empirically, we observe that the above logit regularization helps achieve SGD convergence to solutions with SEL matrix as logits converge much faster, even without additional ridge-regularization. Specifically, for $(R = 10, \rho = 1/2)$-STEP imbalance, $k = 4$ classes and $n_{\min} = 1$, we run SGD on the logit-regularized CE with $\lambda_L/n = 10^{-3}$ and with zero ridge-regularization ($\lambda = 0$). We also set the learning rate to 1. Fig. 8 depicts convergence of the logit-matrix $\mathbf{Z}$ in the direction of the SEL matrix $\hat{\mathbf{Z}}$. However, we find that this does *not* ensure of convergence for the individual geometries of $\mathbf{W}$ and $\mathbf{H}$ towards SELI, although their inner product $\mathbf{W}^T\mathbf{H}$ aligns well to the SEL matrix.

On the other hand, we also find that logit-regularized ERM on the UFM yields the SELI geometry simultaneously for logits, classifiers and embeddings when we follow a specific decaying schedule for the ridge-regularization parameter. Specifically, in the experiments shown in Fig. 9 and 10 below, we start with a large initial value $\lambda/n = 10^{-2}$ and progressively decay λ by a factor of 10 after every few epochs. We also set a finite strength of logit-regularization $\lambda_L/n = 10^{-3}$, although the convergence direction is not sensitive to that choice. We term this scheduling "ridge-decay". While the exact dynamics followed by the SGD with such a scheme require further analysis, "ridge-decay" can be thought of as emulating the regularization path of ridge-regularization with $\lambda \to 0$.



(a) class $y = 0$ **(b)** class $y = 1$ **(c)** class $y = 2$ **(d)** class $y = 3$

(e) class $y = 0$ **(f)** class $y = 1$ **(g)** class $y = 2$ **(h)** class $y = 3$

Figure 10: Average Margins for 4 classes corresponding to the experiments of $R = 10$ (top row) and $R = 100$ (bottom row) from Fig. 9. We run GD on logit-regularized ($\lambda_L/n = 10^{-3}$) CE with ridge-decay $\lambda_{\text{initial}}/n = 10^{-2}$ (see Sec. F.2 for details).

Fig. 9 show convergence of the GD solution with the above described ridge-decay to the SELI geometry. Here, we choose $k = 4$ classes, with $(R = 10, 100, \rho = 1/2)$-STEP imbalance and $n_{\min} = 1$. The learning rate is again fixed to 1. In this experiment, we use GD instead of SGD since the number of examples is small. This is also useful as it shows that GD is able to drive the solution towards the SELI geometry without stochastic updates being necessary.

In summary, we make the following observations from Fig. 9. First, GD iterates favor the SELI, instead of the ETF geometry, suggesting an implicit bias towards global minimizers of the UF-SVM. Second, with logit-regularization and ridge-decay, convergence towards SELI is achieved at a faster rate than in the case of unregularized CE. Third, while Fig. 8 showed that the logit matrix converges fast in direction of the SEL matrix $\hat{\mathbf{Z}}$ with only logit-regularization, Fig. 9b and 9c show that ridge-decay promotes convergence of classifiers and embeddings to their respective SELI geometries as well. Finally, for completeness, we also present the NC convergence in Fig. 9a. The metric used to measure NC is as described in Sec. G.1.4.

The corresponding margins for the 4 classes are shown in Fig. 10. For examples belonging to class y, we define average margin with respect to another class $c \neq y$ as follows:

$$\text{margin}_y(c) := (\mathbf{w}_y - \mathbf{w}_c)^T \boldsymbol{\mu}_y.$$

Note that this is an average over examples from the class y since by NC property $\mathbf{h}_i \approx \boldsymbol{\mu}_y, \forall i : y_i = c$.

We make the following remarks regarding Fig. 10. First, as training progresses, the (average) margins are positive, thus zero training error is achieved for all classes. Second, the average margins for a class y with respect to classes $c : c \neq y$ converge to a common value, even though their initial values differ. This can be seen from the convergence of the four different colored curves within a plot (e.g. Fig. 10a). Third, all margins for all pairs of classes converge to the same quantity, irrespective of being majority or minority classes. Note, for instance, from Fig. 10a,10b,10c, and 10d that the final value of all graphs is the same. Finally, the value of margins stagnates to a level that is governed by the strength of the logit-regularization λ_L.

G Additional results on real data

This section complements Sec. 5 with additional deep-net experiments.

G.1 Missing details on experiments of Section 5

G.1.1 Implementation details

We investigate convergence to the proposed SELI geometry in deep-net training of (R, ρ)-STEP imbalanced MNIST, Fashion-MNIST and CIFAR10 datasets. No data augmentation was used following [26]. We train two deep architectures, ResNet-18 [10] and VGG-13 [29], and optimize the models using CE loss with SGD over 350 epochs. Models are trained on a single GPU with a dataloader batchsize of 128.

Specifically, our experiments build on the code provided by [26][6]. Following [26], we use batch normalization in place of the dropout layers in VGG-13. For both models, we disable the biases of all the fully-connected layers, similar to the experiments on UFM. We adopt the same training strategy as [26], namely SGD on CE loss, with momentum (0.9), small weight decay (5×10^{-4}), and learning rate 0.1 decayed at two stages (epochs 120 and 240) by a factor of 10. We train the network on a $(R, 1/2)$-STEP imbalance setting. (Additional experimental results for other values of the minority ratio are deferred to Sec. G.3.) To create imbalanced data, we use the data sampler provided by Cao et al. [2] [7]. Following [26] we do not use any data augmentation. In all the experiments, we fix the first 5 classes to be majorities, and the rest as minorities. To have a fair comparison between the models with different imbalance ratios $R = 1, 5, 10$ and 100, we sample the datasets to have $n = 100 \times 50 \times 5 + 50 \times 5 = 25250$ training images in all cases. While the training set is imbalanced, when measuring test performance of a trained model we do so on a balanced test set, e.g. just like [1, 2]. We measure the metrics at certain epochs, and similar to [26], we sample epochs more frequently at the start of the training as the network parameters change more quickly in the beginning.

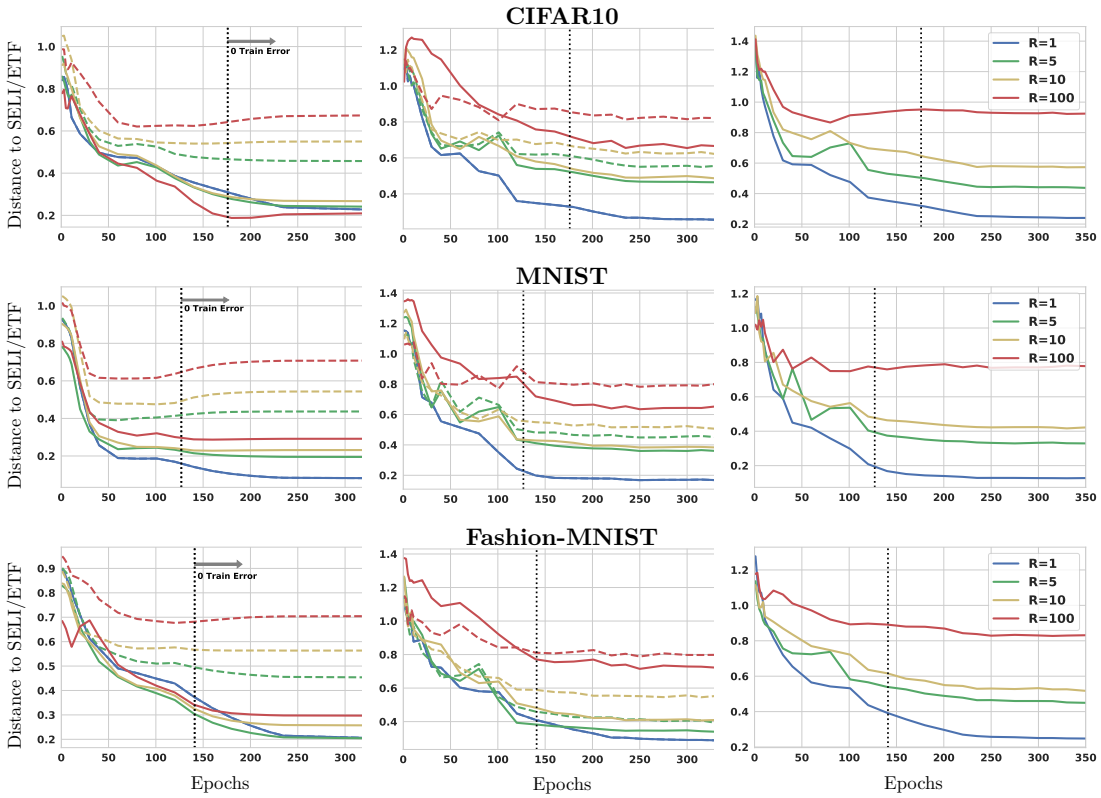
Centering for deep-net experiments. In the deep-net experiments, when investigating the embeddings' geometry, we employ an additional centering of the class means with their (balanced) global mean. Specifically, we compute for each $c \in [k] : \overline{\boldsymbol{\mu}_c} = \boldsymbol{\mu}_c - \boldsymbol{\mu}_G$ with $\boldsymbol{\mu}_G = \frac{1}{k}\sum_{c\in[k]} \boldsymbol{\mu}_c$. Thus, at each epoch, we compute $\overline{\mathbf{M}} = [\overline{\boldsymbol{\mu}_1}, \overline{\boldsymbol{\mu}_2}, ..., \overline{\boldsymbol{\mu}_k}] \in \mathbb{R}^{d\times k}$ and compare, after normalization, $\mathbf{G}_{\mathbf{M}} = \overline{\mathbf{M}}^T\overline{\mathbf{M}}$ to $\hat{\mathbf{G}}_{\mathbf{M}}$, which we calculate as described above for the ETF and SELI geometries, respectively. On the other hand, for the logits calculations, we compute $\mathbf{Z} = \mathbf{W}^T\mathbf{M}$ without centering. The discrepancy between the UFM solutions being already centered, while deep-net embeddings require centering before computation of geometric measures is a common denominator in all previous works on the UFM, e.g. [41, 12, 3, 9, 21, 24, 5, 40]. Here, the chosen centering $\boldsymbol{\mu}_G = \frac{1}{k}\sum_{i\in[n]} \frac{1}{n_{y_i}}\mathbf{h}_i$ is in general different (and the same only when classes are balanced) from the global centering $\mathbf{h}_G := \frac{1}{n}\sum_{i\in[n]} \mathbf{h}_i$ used by Papyan et al. [26]. Our choice of the former is motivated by the fact that the SELI geometry (as predicted by the UFM) satisfies $\sum_{c\in[k]} \overline{\boldsymbol{\mu}_c} = 0$ (see Eqn. (20) in Sec. B.1.4), but not always $\frac{1}{n}\sum_{i\in[n]} \mathbf{h}_i = 0$.

G.1.2 SELI vs ETF

The convergence to SELI and ETF for the classifiers, (centered) mean-embeddings, and logits are illustrated in Fig. 2 and 11 for ResNet and VGG models, respectively. The vertical dashed lines mark the epoch at which the model reaches zero training error under all imbalance ratios; see Sec. G.1.3 for details. Note that in all plots, the distance to SELI geometry decreases as training evolves. Also, convergence to the SELI geometry is consistently better compared to the ETF geometry. However, convergence slows down for increasing imbalance (see $R = 100$). Another interesting observation is that convergence is worse for the embeddings compared to classifiers. In Sec. G.1.7 we compare individual quadrants of the (normalized) $\underline{\mathbf{G}_{\mathbf{W}}}$ and $\underline{\mathbf{G}_{\mathbf{H}}}$ matrices, which facilitates understanding the *individual* behavior of majorities and minorities.

[6] https://colab.research.google.com/github/neuralcollapse/neuralcollapse/blob/main/neuralcollapse.ipynb

[7] https://github.com/kaidic/LDAM-DRW



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

Figure 11: Same setup as Fig. 2 only now training with a VGG-13 model.

G.1.3 Model accuracies

Consistent with the requirements of the neural collapse phenomenon by Papyan et al. [26], all models are trained well beyond zero error. Specifically, as illustrated in Fig. 2 and 11, most of the models achieve zero error around epoch 120, while training continues until epoch 350. Table 2 presents the first epochs at which each model achieves 100 percent accuracy for each value of the imbalance ratio R. For practical purposes, the minimal requirement for majority classes to be declared having achieved zero training error is set to 0.2% error. For minority classes $n_{min} \leq 500$ we set 0.00% for the same requirements.

Zero Error Epoch	R = 1	R = 5	R = 10	R = 100
ResNet (VGG)				
CIFAR10	118 (121)	117 (119)	119 (119)	120 (176)
MNIST	10 (117)	8 (125)	11 (127)	7 (117)
Fashion-MNIST	117 (119)	117 (119)	118 (120)	117 (141)

Table 2: First epoch that zero training error is achieved in deep-net eperiments of Sec. G.

Balanced test errors are reported in Table 3. Majority and minority errors are calculated by averaging the per-class majority and minority errors respectively. The total error is calculated by averaging all per-class accuracies. The averaging is done with equal weights. The test error for the balanced case ($R = 1$) is larger than errors reported in [26] since the model is not trained on the whole dataset, but on $n = 25250$ samples.

G.1.4 NC property

From the **(NC)** property, we expect that the embeddings collapse to their class means. In order to quantify validity of this property, we follow [26]. Specifically, we compute the within-class covariance ($\mathbf{\Sigma}_W$) and between-class covariance ($\mathbf{\Sigma}_B$) as, $\mathbf{\Sigma}_W = \sum_{i\in[n]}(\mathbf{h}_i - \boldsymbol{\mu}_{y_i})(\mathbf{h}_i -$

Test Error	R = 1	R = 5	R = 10	R = 100
CIFAR10 - ResNet (VGG)				
Total	16.47% (17.27%)	26.17% (21.08%)	34.68% (30.03%)	53.89% (53.21%)
Majority	19.18% (20.60%)	11.26% (12.64%)	10.58% (11.30%)	9.68% (10.28%)
Minority	13.76% (13.94%)	41.08% (30.96%)	58.78% (48.76%)	98.1% (96.14%)
MNIST - ResNet (VGG)				
Total	0.55% (0.52%)	0.78% (0.70%)	0.96% (0.81%)	3.04% (4.75%)
Majority	0.38% (0.35%)	0.18% (0.25%)	0.06% (0.06%)	0.02% (0.08%)
Minority	0.73% (0.70%)	1.41% (1.16%)	1.87% (1.60%)	6.18% (9.62%)
Fashion-MNIST - ResNet (VGG)				
Total	7.64% (7.54%)	10.27% (9.26%)	11.63% (10.35%)	16.68% (16.67%)
Majority	8.56% (8.40%)	5.72% (5.50%)	5.12% (4.58%)	4.78% (4.94%)
Minority	6.72% (6.68%)	14.82% (13.02%)	18.14% (16.12%)	28.58% (28.40%)

Table 3: Balanced test error of deep-net experiments in Sec. G.

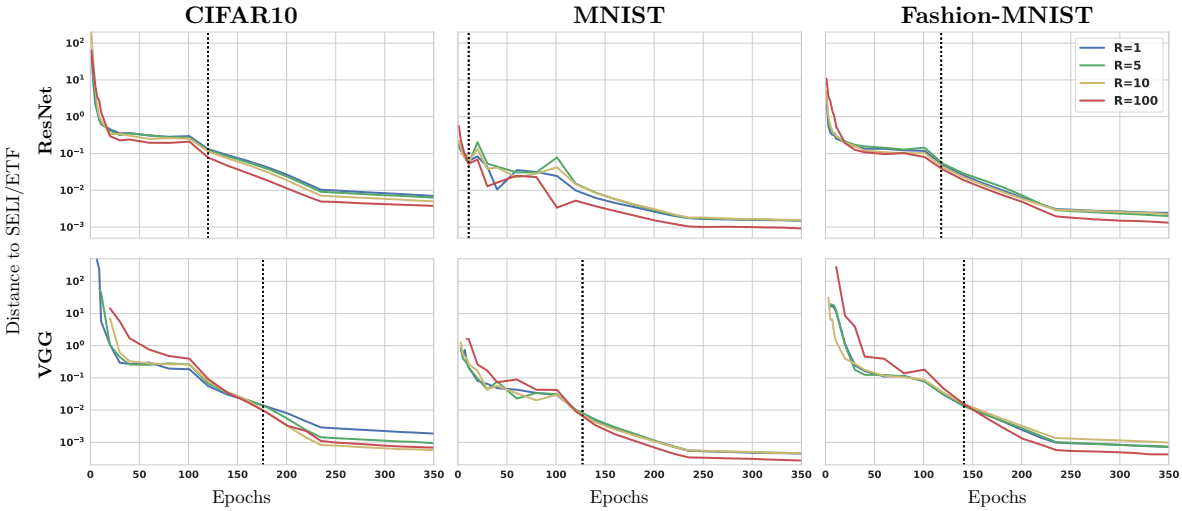


Figure 12: NC property for different imbalance levels R.

$\boldsymbol{\mu}_{y_i})^T \in \mathbb{R}^{d\times d}$, and $\boldsymbol{\Sigma}_B = \sum_{c\in[k]}(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)^T \in \mathbb{R}^{d\times d}$, where $\boldsymbol{\mu}_c = \frac{1}{n_c}\sum_{i:y_i\in[c]} \mathbf{h}_i$ is the mean embedding of class c and $\boldsymbol{\mu}_G = \frac{1}{k}\sum_{c\in[k]} \boldsymbol{\mu}_c$ is their (blanaced) global mean. We can now measure NC by computing $\mathrm{tr}(\boldsymbol{\Sigma}_W \boldsymbol{\Sigma}_B^{\dagger})/k$. Fig. 12 illustrates how this quantity indeed decreases as training evolves. This confirms that feature embeddings converge to their class means, regardless of the imbalance ratio R.

G.1.5 Norms of classifiers / embeddings

Here, we further investigate the geometry of learned embeddings and classifiers, by focusing on their norms. In particular, we study the ratios $\tau_{\mathbf{h}} := \|\mathbf{h}_{\mathrm{maj}}\|_2/\|\mathbf{h}_{\mathrm{minor}}\|_2$ and $\tau_{\mathbf{w}} := \|\mathbf{w}_{\mathrm{maj}}\|_2/\|\mathbf{w}_{\mathrm{minor}}\|_2$. Assuming the classifiers and embeddings follow the SELI geometry, those ratios admit explicit closed-form expressions thanks to Lemmas B.1 and B.2. To determine deviations of the measured norm-ratios $\tau_{\mathbf{w}}$ and $\tau_{\mathbf{h}}$ compared to those reference closed-form expressions, we calculate and report the following quantity:

$$\mathrm{Average}_{c,c'}\big(|\tau_{\mathbf{w}}(c,c') - \hat{\tau}_{\mathbf{w}}|/\hat{\tau}_{\mathbf{w}}\big),$$

where $\hat{\tau}_{\mathbf{w}}$ is given by Lemma B.1 for SELI and is equal to 1 for ETF, and, $\tau_{\mathbf{w}}(c,c') = \|\mathbf{w}_c\|_2/\|\mathbf{w}_{c'}\|_2$ with c being a majority and c' a minority class. Similarly, we compute distances for the norm-ratios of centered mean embeddings $\overline{\boldsymbol{\mu}_c}$. Fig. 13 and 14 depict these metrics during training of the ResNet and VGG networks. The results confirm once more that the SELI geometry accurately captures features of the learned geometries. On the contrary, this is not the case for ETF when data are imbalanced. We observe that convergence to SELI

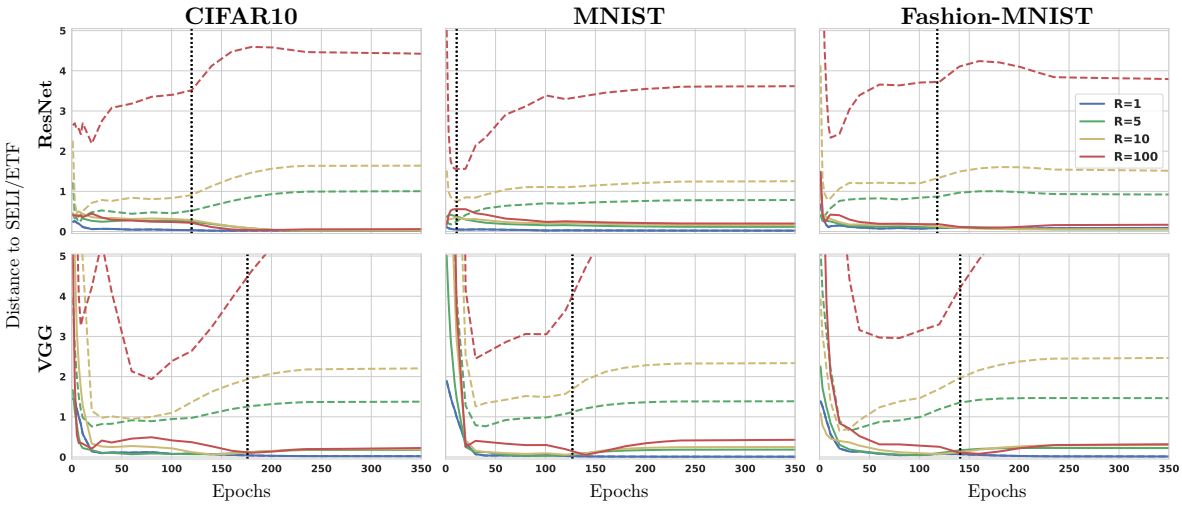


Figure 13: Convergence of classifier norm ratio ($\|\mathbf{w}_{\text{maj}}\|_2/\|\mathbf{w}_{\text{minor}}\|_2$) to SELI (solid) vs ETF (dashed) for different imbalance levels R. The SELI geometry values are computed as per Lemma 3.1.

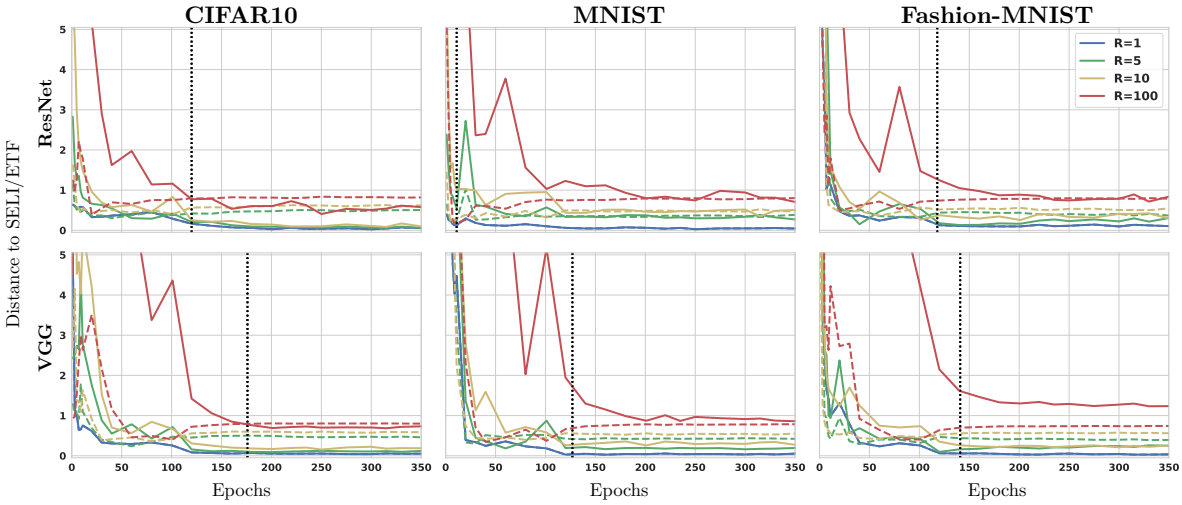


Figure 14: Convergence of embeddings norm ratio ($\|\boldsymbol{\mu}_{\text{maj}}\|_2/\|\boldsymbol{\mu}_{\text{minor}}\|_2$) to SELI (solid) vs ETF (dashed) for different imbalance levels R.

geometry (and respective deviation from ETF) is more pronounced for the classifier weights and becomes elusive for the embeddings particularly for large imbalance ratios ($R = 100$.)

G.1.6 Non-alignment of classifiers and embeddings

While from previous empirical results on balanced datasets [26], we expect an alignment between the classifiers and embeddings, Lemma B.5 suggests these two geometries deviate as data becomes more imbalanced. To verify this property, we compute the angle between mean embeddings and their corresponding classifiers, and measure the deviation from the SELI geometry. Namely, let $\theta_c = \text{Cos}(\mathbf{w}_c, \mathbf{h}_c)$. Then, similar to the previous section, we compute,

$$\text{Average}_c\big(|\theta_c - \hat{\theta}|/\hat{\theta}\big),$$

where c ranges over minority classes and $\hat{\theta}$ is given by (18). Fig. 15 shows how this quantity evolves during training. From Fig. 5b, we know that the embeddings and classifiers of the majority classes remain aligned even for highly imbalanced data, thus we only analyze the impact of imbalance ratio on the minority classes.

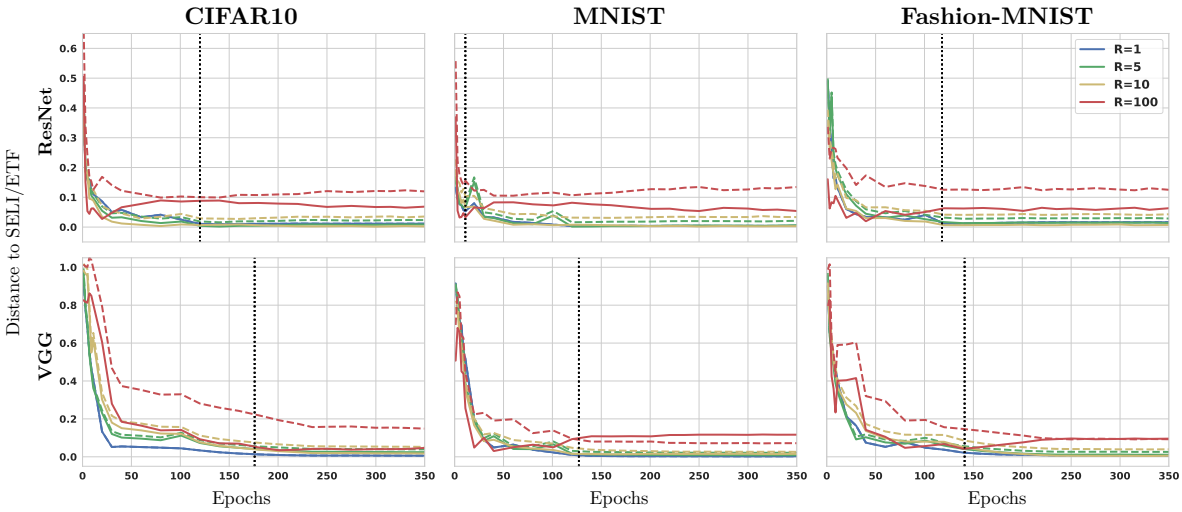


Figure 15: Alignment of minority embeddings and classifiers. Convergence of angles to SELI (solid) and ETF (dashed) for different imbalance levels R.

G.1.7 Majority vs minority geometry

In order to better understand the individual behavior of majorities and minorities, we now compare individual quadrants of the (normalized) $\underline{\mathbf{G_W}}$ and $\underline{\mathbf{G_H}}$ matrices. Concretely, let

$$\underline{\mathbf{G_W}} = \begin{bmatrix} \underline{\mathbf{G_W}}^{\text{maj-maj}} & \underline{\mathbf{G_W}}^{\text{maj-minor}} \\ \left(\underline{\mathbf{G_W}}^{\text{maj-minor}}\right)^T & \underline{\mathbf{G_W}}^{\text{minor-minor}} \end{bmatrix}$$

be a partition of the normalized $\underline{\mathbf{G_W}} = \mathbf{G_W}/\|\mathbf{G_W}\|_F$ to $(k/2)\times(k/2)$ sub-blocks. Comparing quadrants $\underline{\mathbf{G_W}}^{\text{maj-maj}}, \underline{\mathbf{G_W}}^{\text{maj-minor}}$ and $\underline{\mathbf{G_W}}^{\text{minor-minor}}$ to the corresponding quadrants of the reference SELI/ETF matrix $\hat{\mathbf{G}}_\mathbf{W}$ allows us to "zoom-in" the majority-majority, majority-minority and minority-minority structures. Entirely analogous calculations allow the same for the embeddings.

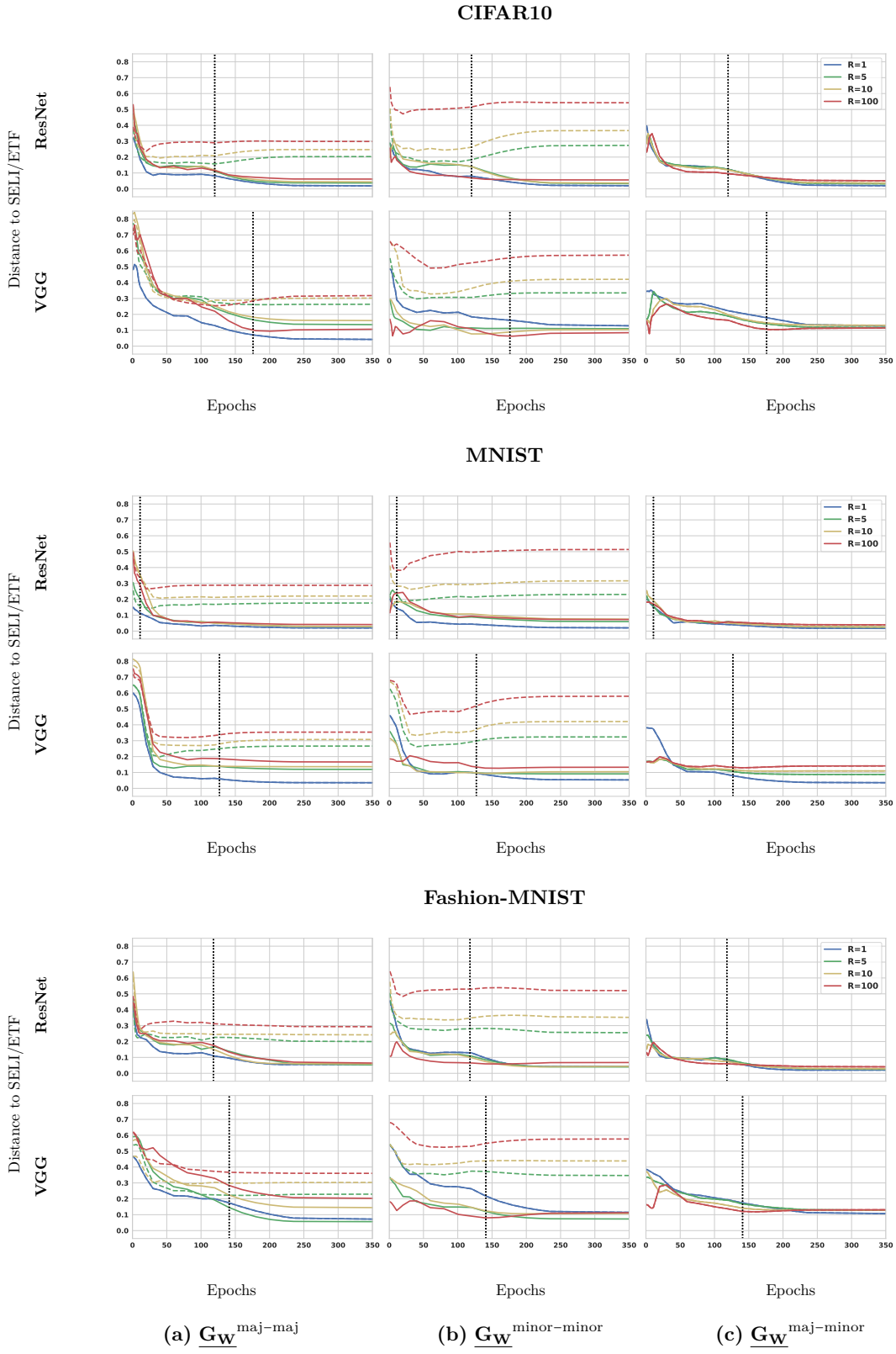
Classifiers Geometry. Fig. 16 confirms that both majority and minority geometries converge to SELI properly. Interestingly, we see that the minorities diverge the most from the equiangular structure of ETF geometry.

Embeddings Geometry. We find thanks to Fig. 17 that the "error" in convergence of embeddings to SELI geometry (compare to the better convergence for classifiers) shown previously in Fig. 2 and 11 is primarily due to the minority class geometries. However, an overall inspection of the subfigures shows that embeddings also tend to align better to SELI compared to ETF. This alignment property is pronounced in the case of majority geometries (see $\underline{\mathbf{G_M}}^{\text{maj-maj}}$).

G.2 Capturing weight-decay with regularized UFM

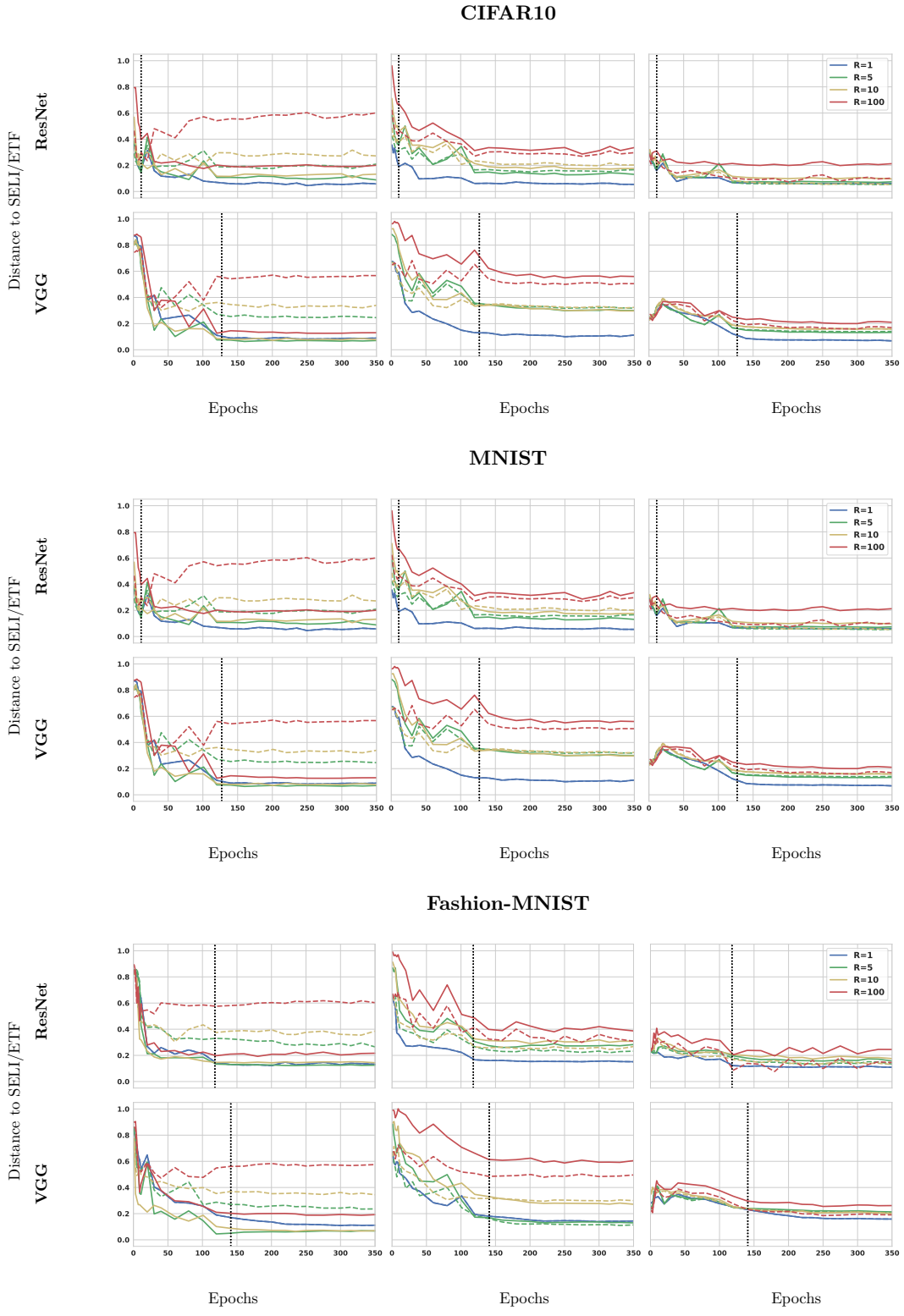
In Fig. 18 we compare for $R = 10$ the distances of $\underline{\mathbf{G_W}}$ and $\underline{\mathbf{G_M}}$ matrices to various λ-SELI geometries for different values of regularization λ. Specifically, for the λ-SELI geometries, we obtained the reference matrices $\hat{\mathbf{G}}^\lambda_\mathbf{W}, \hat{\mathbf{G}}^\lambda_\mathbf{M}$ as follows. For each value of λ, we solve the nuclear-norm CE minimization in (7) to find $\hat{\mathbf{Z}}_\lambda$ for $k = 10$ and $R = 10$ (to match the CIFAR10 settings). We then form $\hat{\mathbf{G}}^\lambda_\mathbf{W}, \hat{\mathbf{G}}^\lambda_\mathbf{M}$ as described in the first paragraph of Sec. 5 only now using $\hat{\mathbf{Z}}_\lambda$ instead of the SELI $\hat{\mathbf{Z}}$. For comparison, we also plot the distances to the SELI and ETF geometries.

Note that we have experimented with a wide range of values for the regularization λ. Among those is the value $5e-4$ that matches with the choice of weight-decay parameters in the deep-net experiments. Recall also from the discussion in Sec. 4 that $\hat{\mathbf{Z}}_\lambda$ is sensitive to λ in this setting where $R = 10, k = 10$.



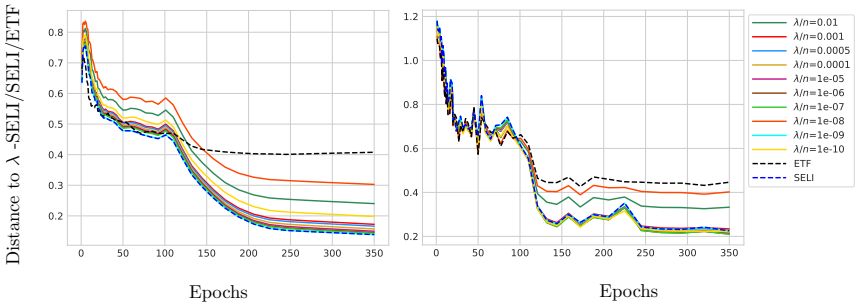
(a) $\underline{\mathbf{G}_{\mathbf{W}}}^{\text{maj}-\text{maj}}$ (b) $\underline{\mathbf{G}_{\mathbf{W}}}^{\text{minor}-\text{minor}}$ (c) $\underline{\mathbf{G}_{\mathbf{W}}}^{\text{maj}-\text{minor}}$

Figure 16: Convergence of learned classifiers' majority and minority individual geometries to the SELI (solid lines) vs ETF (dashed lines) geometries.



(a) $\underline{\mathbf{G_M}}^{\text{maj-maj}}$ (b) $\underline{\mathbf{G_M}}^{\text{minor-minor}}$ (c) $\underline{\mathbf{G_M}}^{\text{maj-minor}}$

Figure 17: Convergence of learned classifiers' majority and minority individual geometries to the SELI (solid lines) vs ETF (dashed lines) geometries.



(a) Classifiers **(b)** Embeddings

Figure 18: Distances of ResNet learned classifiers and mean-embeddings trained on CIFAR10 data with imbalance ratio $R = 10$ from the λ-SELI geometry (aka solution to (7) as defined in Theorem 2) for different values of λ, as well as, from the SELI and ETF geometries.

We make the following interesting observations from Fig. 18. First, note that for classifiers the minimum distance is that to SELI. The distance of the embeddings' geometry to SELI is also among the lowest ones. Specifically, despite being slightly larger than that from λ-SELI for a few values of λ, the difference is very small. This suggests that SELI is indeed a good approximation for the learned geometry even when training with finite weight-decay. Besides, there are two key advantages of the SELI over the λ-SELI geometry. First, it is unclear what the mapping ought to be (if such a mapping exists) between training-implementation choices (such as weight-decay) and λ. Second, even if such a mapping was known, the SELI geometry has the unique advantage of being expressed simply in terms of the (SVD of the) SEL matrix. In fact, as we show in Sec. B it is possible to get closed-form expressions for the norms and angles describing the geometry. This not only makes calculations much easier, but also it allows further analysis of the properties (e.g. quantifying norm-ratios as in Fig. 13).

G.3 Additional experiments for minority ratios $\rho \neq 1/2$

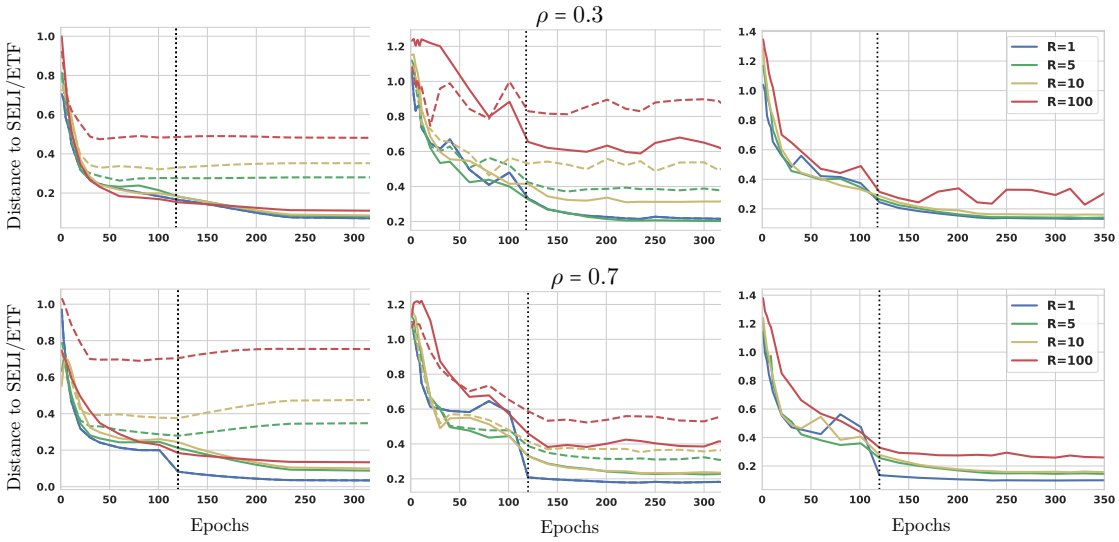
Thus far, in our previous experiments we considered imbalanced data with minority ratio $\rho = 1/2$, i.e. same number of minorities and majorities. However, note that our theoretical results hold for any value of ρ. Specifically, Theorem 1 shows that the solution of the UF-SVM follows the SELI geometry irrespective of ρ. Here, we empirically study convergence to the SELI geometry for a ResNet-18 model on (R,ρ)-STEP imbalanced CIFAR10 data, for two values of minority ratio $\rho = 0.3$ and 0.7. These experiments complement our previous demonstrations for $\rho = 0.5$. Specifically, we create training set of the same size of $n = 15350$ in all experiments for a fair comparison.[8] All other experimental settings are as described in Section G.1.1. Figs. 19, 20 and 21 demonstrate how classifiers and embeddings converge to the proposed SELI geometry. Consistent with the previous experiments, the learnt geometries of deep networks are well-captured by the SELI. We also observe empirically that larger values of ρ exhibit slightly faster convergence.

H Implications on minority collapse

In this section, we further elaborate on how our results relate to the *minority collapse* phenomenon, which is defined by Fang et al. [3] as the phenomenon during which *minority classifiers become completely indistinguishable*. Notably, Fang et al. [3] discover its occurrence both in the UFM and in real datasets trained with deep-nets. Below, we first state their concrete findings and then we discuss how our results extend them.

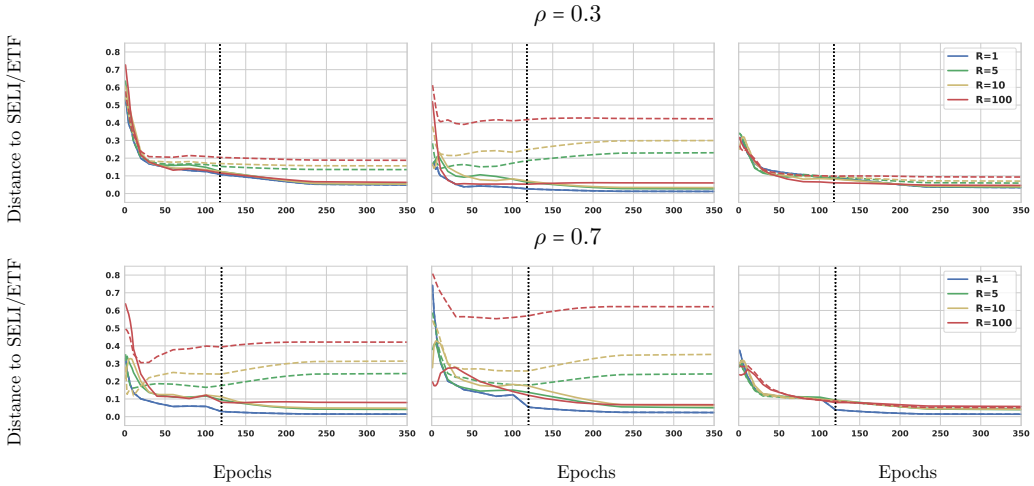
Summary of findings by Fang et al. [3]. Fang et al. [3] make the following key findings.

[8] This is smaller than the total number $n = 25250$ of examples used previously in our experiments for $\rho = 1/2$. The reason is that there is a limited number of 5000 images per class in CIFAR-10 making it impossible to have $(R = 100, \rho = 0.7)$-STEP imbalanced CIFAR-10 data.



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

Figure 19: Impact of ρ on the convergence of ResNet geometry to the SELI, trained on CIFAR10.



(a) $\underline{\mathbf{G_W}}^{\text{maj-maj}}$ **(b)** $\underline{\mathbf{G_W}}^{\text{minor-minor}}$ **(c)** $\underline{\mathbf{G_W}}^{\text{maj-minor}}$

Figure 20: Impact of ρ on the convergence of majority/minority classifiers of ResNet model trained on CIFAR10.

FHLS(1) In [3, Theorem 5] they prove for a constrained UFM, that $\lim_{R\to\infty} \mathrm{Cos}(\mathbf{w}_{\text{minor}}, \mathbf{w}'_{\text{minor}}) = 1$ (here, $\mathbf{w}_{\text{minor}}, \mathbf{w}'_{\text{minor}}$ are any two distinct minority classifiers.)

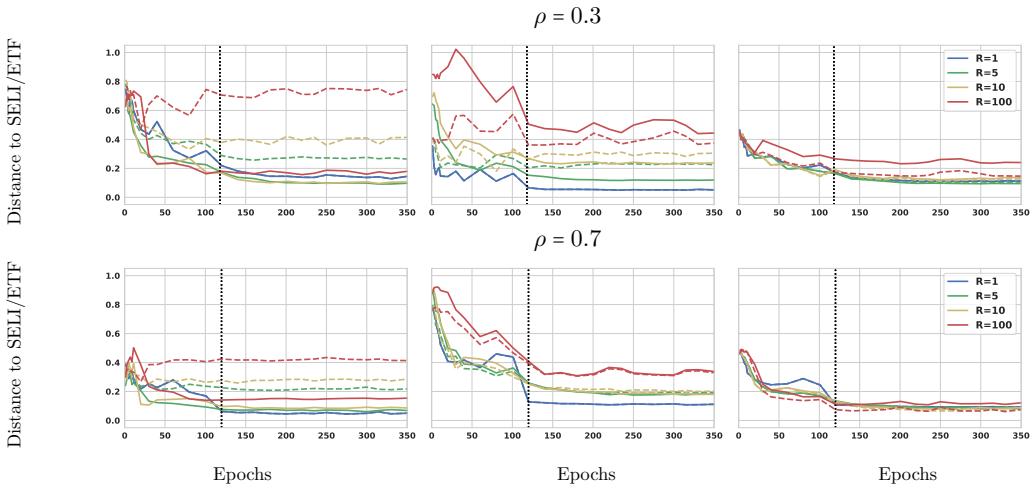
FHLS(2) They also find numerically that the solution to the same constrained UFM gives $\mathrm{Cos}(\mathbf{w}_{\text{minor}}, \mathbf{w}'_{\text{minor}}) = 1$ for any $R > R_0$ for some *finite* threshold R_0.

FHLS(3) Their experiments, specifically [3, Fig. 3], suggest that for fixed number of classes k, the value of the threshold R_0 increases as the constraint parameter gets relaxed and also as the ratio ρ of minority classes increases.

FHLS(4) Finally, they validate the minority collapse phenomenon on real imbalanced datasets trained with deep-nets. Their real-data experiments (e.g. [3, Fig. 2,4]) suggest the following.

(a) Consistently, as R increases, the cosine similarity between minority classes increases until it reaches one.

(b) The value of R after which the cosine becomes one (i.e. minority collapse is reached) depends critically on the weight-decay. For small weight decay



(a) $\underline{\mathbf{G_M}}^{\text{maj−maj}}$ (b) $\underline{\mathbf{G_M}}^{\text{minor−minor}}$ (c) $\underline{\mathbf{G_M}}^{\text{maj−minor}}$

Figure 21: Impact of ρ on the convergence of majority/minority embeddings of ResNet model trained on CIFAR10.

($\sim 5e-4$), it takes $R > 1000$ to reach minority collapse. It is only for larger weight decay ($\sim 5e-3$) that minority collapse occurs for $R \sim 100$.

Our novelties. Before discussing implications of our results for minority collapse, we highlight the following key features of our study.

- **Entire geometry:** We describe the *entire geometry of classifiers and embeddings for both majority and minority classes* (not only the geometry of minority classes.)
- **Finite imbalance levels:** Our geometric characterizations (aka SELI geometry) hold for *all finite values of the imbalance ratio* R (not only asymptotically.)
- **Vanishing regularization:** We focus on CE training with vanishing regularization. (As such, our geometry characterizations result from analyzing the UF-SVM.)

Contact points: What do our results say about minority collapse?

- **For zero regularization (aka UF-SVM), there is no minority collapse for any finite value of R.** Specifically, we show in Lemma B.3 that for $(R, 1/2)$-STEP imbalance $\text{Cos}(\mathbf{w}_{\text{minor}}, \mathbf{w}'_{\text{minor}}) = {}^{(R-7)}/\big(R-7+2k(2+\sqrt{(R+1)/2})\big) < 1$. This does not contradict Finding **FHLS(2)** since here we consider zero regularization; all their numerical evaluations are with finite regularization.
- **Minority collapse occurs (only) *asymptotically* in R for the UF-SVM.** Specifically, we show in Corollary 2.4 that $\lim_{R\to\infty} \text{Cos}(\mathbf{w}_{\text{minor}}, \mathbf{w}'_{\text{minor}}) = 1$; see also Fig. 6c. This can be seen as a different manifestation of the Finding **FHLS(1)**, this time for the UF-SVM, instead of constrained CE minimization.
- **There is no minority collapse for small regularization strength when training the UFM with regularized CE minimization.** Specifically, we show in Lemma D.4 that for any finite imbalance values R, ρ, there is no minority collapse in the solutions of Eqn. (1) when $\lambda < 1/2$. Instead, the CE solution perfectly separates the training data. Our result theoretical justifies the numerical Findings **FHLS(3)** and **FHLS(4)**(b). To establish the connection to the setting of Fang et al. [3], we need to normalize the CE loss in (1) by a factor of $1/n$ (see [3, Eqn. (15)]). For this normalized minimization, Lemma D.4 ensures *no* minority collapse provided that

$$2\lambda < \frac{1}{n} = \frac{1}{(\rho + R\overline{\rho})k} = \frac{1}{(R - \rho(R-1))k}\,. \tag{54}$$

Put in terms of imbalance ratio, we prove that

$$\text{Minority collapse requires} \quad R > f(\lambda, \rho) := \frac{\frac{1}{2k\lambda} - \rho}{1 - \rho}. \tag{55}$$

It is straightforward to check that $f(\lambda, \rho)$ is increasing in ρ and decreasing in λ. Thus, we show that the minority collapse threshold R_0 (see Finding **FHLS(3)**) increases with the minority ratio ρ and with the inverse regularization parameter $1/\lambda$. This finding explains the behavior reported empirically in [3, Fig. 3] for the UFM and in [3, Fig. 2,4] for real data. [9]

- **The angle between minority classifiers decreases monotonically with the imbalance ratio R for the UF-SVM.** Specifically, it can be checked easily by direct differentiation that the formula in Lemma B.3 giving $\mathrm{Cos}(\mathbf{w}_{\mathrm{minor}}, \mathbf{w}'_{\mathrm{minor}})$ is increasing in R. This can be seen as a theoretical justification of the empirical Finding **FHLS(4)**(a).

I Additional related work

As discussed in Sec. 1.1, our work is inspired and is most closely related to the recent literature on Neural Collapse. In Sec. 1.1 we reviewed the most closely related of these works. A few others are referenced in Sec. I.1 below. Beyond Neural collapse, our results and analysis tools are also related to the literatures on implicit bias, matrix factorization and imbalanced deep-learning. We elaborate on these connections below.

I.1 Additional works on Neural Collapse

Beyond CE minimization, a series of recent works study and analyze the neural collapse phenomenon when training with square loss. Interestingly, Graf et al. [5], Fang et al. [3] discover and analyze neural collapse for similarity-type losses, such as the self-supervised contrastive loss, which trains only for embeddings. To the best of our knowledge, all these works restrict attention to balanced classification. Our work shows that it is possible to obtain explicit geometric characterizations in class-imbalanced settings when training with CE. Hence, it also opens the way to extending the analysis to square-loss minimization.

Potential connections of neural collapse to generalization and transferability is a less well-understood topic. Some initial investigations appear in the recent works [11, 4, 9]. Our results are not immediately conclusive about generalization. However, as mentioned in Sec. 6 our results have the potential to 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, 2]).

Remark I.1 (Last-layer peeled model (LPM))**.** *Around the same time that the UFM was proposed by Mixon et al. [24] (see also Lu and Steinerberger [21], Graf et al. [5]), the same model was independently formulated and analyzed by Fang et al. [3] under the alternative name of "last-layer peeled model (LPM)".*

Remark I.2 (**(NC)** and **(ETF)** properties in [26])**.** *The formalization of Neural collapse by Papyan et al. [26] involved four NC properties. The first property concerns the collapse of class embeddings to their corresponding means. Properties two and three concern the geometry of class-means and classifier-weights, specifically their alignment and convergence to a simplex ETF geometry. Property four is a consequence of the other three properties, hence is less important in the formalization and we do not discuss it further. Motivated by our findings, we propose and use here a regrouping/renaming of the aforementioned three properties. We refer to the first property as the **(NC)** property, and, to the second and third properties as the **(ETF)** property. We argue that this distinction is important towards a formulation that is invariant across class-imbalances, by showing that the ETF property is* not *invariant and replacing it with the **(SELI)** property. For balanced data, the latter simplifies to the ETF property.*

[9] When referring to [3, Fig. 3] keep in mind that they simulate constrained, rather than regularized, CE minimization. Hence, larger constraint parameters mean larger *inverse* regularization parameter $1/\lambda$.

I.2 Implicit bias

Neural collapse is intimately related to the recent literature on implicit bias, which started from a series of influential works [30, 13, 7, 14, 25]. (This connection is already recognized by the seminal work of Papyan et al. [26].) For example, [7, Theorem 7, Remark 2] concerns a bilinear non-convex SVM-type minimization that bears similarities to the UF-SVM and establishes a connection to a convex nuclear-norm minimization problem. However, the two factors in their bilinear formulation are the same (unlike the UF-SVM) and also they restrict attention to binary classification. Another very closely related work is that by Lyu and Li [22] who show that gradient descent on deep homogeneous networks converges (in direction) to a KKT point of a corresponding non-convex max-margin classifier. While their focus is again on binary classification, they briefly discuss extension to multiclass settings in their appendix. Recently, Ji et al. [12] leveraged their results and formally showed that gradient descent on (1) with zero regularization converges to a KKT point of the UF-SVM. The max-margin classifiers corresponding to multi-layer linear networks (with the UF-SVM being a special case) are non-convex. Hence there is no guarantee that the KKT point where gradient descent converges to is a global optimum. Whether this is the case or not is investigated for various settings by Vardi et al. [33]. Specifically for linear fully-connected networks, which are of interest to us, they show that, when trained on binary data, the point of convergence is always a global optimum [33, Theorem 3.1]. Their proof uses another nice result on implicit bias by Ji and Telgarsky [15]. In fact, their results are more intimately connected to neural collapse as it can be checked that [15, Proposition 4.4] provides a direct proof that gradient descent on unregularized CE for the UFM and binary data finds embeddings and a classifier that satisfy the NC and ETF properties (irrespective of imbalance). Note here that all these works focus almost exclusively on binary settings. A salient message of our results is that rich and possibly complicated behaviors can occur in multiclass ($k > 2$) settings. There are several findings supporting this. For example, we show that regularization in the UFM only matters when data are multiclass and imbalanced. Similarly, it is only then that the model found by the two-layer UFM can differ from what a one-layer convex network would find. We also show empirically for the UFM that convergence rates in the absence of regularization can be heavily impacted by imbalances. Related to this, we highlight a missing piece in our analysis: we characterize the global optimum of the UF-SVM, but we do not prove, or are aware of a proof, that gradient descent converges to this global optimum. Proving or disproving this can be of great interest in its own way. On the one hand, if the conjecture holds, then our results warn that imbalance levels can severely impact convergence rates. On the other hand, if the conjecture is refuted, then this would be the simplest model to have been discovered where convergence to global optimum fails.

I.3 Matrix factorization and low-rank recovery

The connection between the study of the UFM for the purpose of neural collapse analysis and the literatures on matrix factorization and low-rank matrix recovery (see for example [36]) is uncovered and first exploited by Zhu et al. [41] and Fang et al. [3]. Thus, we refer the interested reader to those papers for a list of references and detailed discussion (specifically see Zhu et al. [41, Sec. 3.2]). Specializing this discussion to the UF-SVM that is of main interest to us, we note its close ties to the formulation of the hard-margin matrix factorization problem as studied by Srebro et al. [31]. The author formulated the problem of fitting a binary target matrix $\mathbf{Y}$ (ie. with entries ± 1) with a low-rank matrix $\mathbf{W}^T\mathbf{H}$ as the minimization

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{W}^T\mathbf{H}\|_* \quad \text{sub. to} \quad \Big(\mathbf{Y} \odot \mathbf{W}^T\mathbf{H}\Big)[c,i] \geq 1, \ \forall (c,i) \in \mathcal{S}, \tag{56}$$

where $\mathcal{S}$ is a given subset of observed entries of $\mathbf{Y}$. Despite being non-convex, they derived a convex reformulation based on duality and a corresponding procedure for finding a solution to (56) via essentially solving an SDP and an appropriate system of linear equations corresponding to the active constraints. The non-convex max-margin problem (2) that we investigate bears similarities to (56). Specifically, for an one-hot encoding and fully observed $\mathbf{Y}$, (56) is essentially the binary analogue of (2). Importantly in our setting, we are able to calculate the solution to (2) in closed form, that is without requiring numerically solving an SDP. Finally, as mentioned in Sec. E.4 a byproduct of Theorem 1 is that the

nuclear-norm max-margin minimization (5) has the same solution as the vanilla max-margin with Frobenius norm. Although different, somewhat related settings were frobenius-norm penalized problems give same solutions as nuclear-norm penalized ones have been studied in the low-rank representation literature, e.g. [34, 27].

I.4 Class-imbalanced deep learning

The past few years have seen a surge of research activity towards substituting the vanilla CE empirical risk minimization, which leads to poor accuracy for minorities, with better alternatives that are particularly suited for training large models, e.g. [17, 18, 2, 23, 39]. Among the many solutions suggested in the recent literature, most closely related to the topic of neural collapse are [17, 18]. Interestingly (as it happened chronologically before the conception of Neural collapse by Papyan et al. [26]), Kang et al. [17] and Kim and Kim [18] observed that the classifier weights found by deep-nets when trained with CE on class-imbalanced data yield larger norms for majority rather than minority classes. This empirical observation led them to propose post-hoc schemes that normalize the logits before deciding on the correct class, thus leading to better performance on the minorities. Our Lemma B.1, not only proves this behavior for the unconstrained feature model, but it also precisely quantifies the norm-ratio between minorities and majorities. Interestingly, our deep-net experiments in Fig. 13 confirm the predicted behavior. Evidently then, Lemma 3.1 offers a plausible theoretical justification of the empirical findings by Kang et al. [17], Kim and Kim [18] and also quantifies the norm ratio. It is conceivable, that this characterization in terms of a simple formula that only involves the imbalance ratio and number of classes, can be used to turn some of the heuristic post-hoc normalizations of Kang et al. [17], Kim and Kim [18] to principled methods. Beyond that, our results are conclusive not only about classifiers and norms, but also about embeddings and angles. It is exciting to investigate leveraging these findings to design better techniques for class-imbalanced deep learning (see also discussion by the end of Sec. 6.)

Related ideas appeared very recently in the contemporaneous works [38, 37], where the authors design loss functions for class-imbalanced learning in an attempt to enforce a geometry that is alike the ETF geometry for balanced data. However, they do not characterize the joint geometry of classifiers and embeddings under class imbalances as we do here.

I.5 On the Simplex-encoding interpolation

Our analysis of the UF-SVM uncovers the unique role played by the SEL matrix. Specifically, Theorem 1 shows that the global minimizers $(\hat{\mathbf{W}}, \hat{\mathbf{H}})$ of the non-convex UF-SVM are such that the resulting logit matrix $\hat{\mathbf{W}}^T\hat{\mathbf{H}}$ satisfies all constraints with equality and also it equals the SEL matrix $\hat{\mathbf{Z}}$. Our finding is related to (in fact, can be seen as an extension of) a recent result by Wang et al. [35]. In a different context and with different research objective, Wang et al. [35] derived a deterministic condition under which the solution to the (convex) vanilla SVM (like the one in (52) but with general inputs, not necessarily the basis vectors $\mathbf{e}_i$ resulting in the UFM) finds logits that interpolate a simplex encoding of the labels. Theorem 1 goes far beyond: it studies a non-convex SVM and since it applies directly to the UFM, it does not involve deterministic input conditions for the result to hold. That said, it might be interesting future work to derive similar conditions on the inputs of the non-convex max-margin problem such that it gives optimal logits interpolating the simplex encoding of the labels.