The role of class encoding in neural collapse

Bastien Massion
ICTEAM, UCLouvain
Louvain-la-Neuve, Belgium
bastien.massion@uclouvain.be

Roy Makhlouf
ICTEAM, UCLouvain
Louvain-la-Neuve, Belgium
roy.makhlouf@uclouvain.be

Estelle Massart
ICTEAM, UCLouvain
Louvain-la-Neuve, Belgium
estelle.massart@uclouvain.be

***Abstract*—Neural collapse is a structural property of the last-hidden-layer activations in neural network classification models, when trained beyond a zero classification error. In this work, we explore the role of label encoding in neural collapse by relying on the unrestricted feature model with mean squared error training loss. We demonstrate that, for one-hot encoded labels and balanced data, the uncentered mean features associated with each class transition from a simplex equiangular tight frame to an orthogonal frame when increasing the bias regularization coefficient associated with the final classifier. These structures are reminiscent of the orthogonal frame structure of one-hot encoded labels. For any arbitrary encoding, we also show that the final classifier's bias aims at centering the labels, compensating for the discrepancy between the global mean of the labels and the origin. We further discuss the role of the encoding in other neural collapse properties.**

***Index Terms*—Deep learning, Neural Collapse, Class encoding**

I. Introduction

Neural collapse (NC) refers to a phenomenon occurring during the terminal stage of learning deep neural network classifiers. By terminal stage of learning, we refer to the last epochs, where the training loss is further reduced while the training error is zero (i.e., all inputs are correctly classified). It was noticed in [17] that some structure appears in the activations of the last hidden layer; namely, the activations corresponding to inputs of a same class concentrate around a mean value, leading to the *neural collapse* terminology, and these mean vectors, after centering, localize at the vertices of a simplex equiangular tight frame (ETF). This nice geometric structure was exploited in the design of out-of-distribution detection algorithms [8] and transfer learning methods [5] among others; see [12] for a review on NC.

Understanding the emergence of NC has been the focus of many works over the past years. The original work [17] numerically observed NC on classifiers trained with the cross-entropy loss across various datasets and architectures. Other works further explored the occurrence of NC for different losses [9], [28], and aimed at generalizing NC to intermediate layers [18]–[20], to regression models [1] or to account for imbalanced data [4], [14], [22]. Additionally, motivated by large language models, the works [11], [25] investigated NC in the regime where the number of classes outnumbers the last-hidden-layer dimension.

Arguably the most common approach towards a theoretical characterization of NC relies on the *unconstrained features model* (UFM) [16], or the related *layer-peeled model* [4]. These models formulate the training problem as the joint optimization of last-hidden-layer features and a final linear classifier, i.e., the last-hidden-layer features are not restricted by the expressivity of the previous layers. These simplified models, which can be seen as (overparametrized) matrix factorization problems, were used to prove that global minimizers satisfy the NC properties for various loss functions [3], [4], [6], [9], [15], [16], [23], [28]. Subsequent works provide global landscape analyses, showing that the UFM/layer-peeled model has a benign loss landscape and, consequently, that the training algorithm will converge to a point satisfying the NC properties as soon as it is able to avoid saddle points with a strictly negative curvature in some directions [26], [27], [29].

In this paper, we focus on the mean squared error (MSE) loss function. The use of this loss function gained attraction over recent years for classification problems [2], [10]. In [23], the authors showed that the geometric structure of the optimal solutions of the UFM for the MSE loss with one-hot encoding differs depending on the use of a bias term in the final classifier: the (uncentered) mean activations are organized as a simplex ETF when some (unregularized) bias is used, while they form an orthogonal frame (OF) in the absence of bias. In this work, we build a bridge between these two structures, by showing that OFs and simplex ETFs are merely shifted versions of one another, see Figure 1. The transition between these two structures is due to the final classifier's bias, which aims at compensating for the possible non-centering of the labels and whose magnitude depends on the strength of its regularization. Note that these structures are reminiscent of the OF structure of one-hot encoded labels. We explore further the role of the encoding on NC, and demonstrate that even if variability collapse does not depend on the encoding, this is not the case of other NC properties.

The structure of the paper is as follows. We recall fundamental definitions in Section II, show the equivalence between simplex ETFs and OFs, and discuss the role of the bias of the final classifier for general encodings in Section III, and address other NC properties in Section IV.

II. Preliminaries

Let us consider a set of input data $x_1, \dots, x_N \in \mathbb{R}^{d_x}$, to be classified into K classes. Let us write $\bar{y}_k \in \mathbb{R}^K$, the label of class $k \in \{1, \dots, K\}$ and define $\bar{Y} := [\bar{y}_1 \cdots \bar{y}_K] \in \mathbb{R}^{K \times K}$. Note that we consider the dimension of the encoding

to be equal to the number of classes K, which includes one-hot encoding and label smoothing, among others. Let $y_i \in \{\bar{y}_1, \dots, \bar{y}_K\}$ be the target label vector associated with input x_i and define $Y := [y_1 \cdots y_N] \in \mathbb{R}^{K \times N}$. For notational convenience, we assume that the samples are ordered by class:

$$Y = \bar{Y}C, \tag{1}$$

where $C := \begin{bmatrix} e_1 1_{n_1}^\top & \cdots & e_K 1_{n_K}^\top \end{bmatrix} \in \mathbb{R}^{K \times N}$ is a class assignment matrix, where e_k denotes the k^{th} canonical vector, n_k the number of occurrences of the target label $\bar{y}_k$ in the dataset, and 1_{n_k} the vector of ones of length n_k. This notation generalizes the Kronecker product formulation widely used in the literature for balanced classes [23], [24].

NC refers to four properties that occur simultaneously.

- $\mathcal{NC}1$ (Variability collapse): the last-hidden-layer features concentrate around their class mean.
- $\mathcal{NC}2$ (Convergence to a simplex ETF): the class means, after centering by their global mean, converge to a simplex ETF (see Definition II.1).
- $\mathcal{NC}3$ (Convergence to self-duality): the last layer's classifier weights converge to the class means, up to rescaling.
- $\mathcal{NC}4$ (Simplification to nearest-class-center classification): the linear classifier's decision rule reduces to choosing the class whose mean is the nearest to the last-hidden-layer activations.

NC is tightly related to the notion of simplex ETF [17].

Definition II.1 (Simplex ETF)**.** A simplex equiangular tight frame [17] is a collection of K vectors in $\mathbb{R}^d$ (with $d \geqslant K$) specified by the columns of

$$M = \alpha P \left(I_K - \frac{1_K 1_K^\top}{K} \right),$$

where $P \in \mathbb{R}^{d \times K}$ is semi-orthogonal (i.e., has orthonormal columns) and $\alpha > 0$. Note that, equivalently, M satisfies

$$M^\top M = \alpha^2 \left(I_K - \frac{1_K 1_K^\top}{K} \right).$$

Definition II.2 (OF)**.** By orthogonal frame, we refer to a collection of K vectors in $\mathbb{R}^d$ (with $d \geqslant K$) specified by the columns of

$$M = \alpha P,$$

where $P \in \mathbb{R}^{d \times K}$ is semi-orthogonal and $\alpha > 0$.

In this work, we analyze NC through the lens of the Unconstrained Features Model (UFM) proposed in [16], expressed here for the MSE loss:

$$\min_{W,H,b} \frac{1}{2N} \|WH + b1_N^\top - Y\|_{\text{F}}^2 + \frac{\lambda_W}{2} \|W\|_{\text{F}}^2 + \frac{\lambda_H}{2} \|H\|_{\text{F}}^2 + \frac{\lambda_b}{2} \|b\|_2^2, \tag{2}$$

where $H \in \mathbb{R}^{d \times N}$ with $d \geqslant K$ are unconstrained features, which model here the last-hidden-layer activations, omitting expressivity restrictions due to the network architecture, $W \in \mathbb{R}^{K \times d}$ and $b \in \mathbb{R}^K$ are respectively the final classifier's weights and bias, and $\lambda_W > 0$, $\lambda_H > 0$ and $\lambda_b \geqslant 0$ are regularization parameters. Note that the linearity assumption on the classifier is standard [16].

III. From OFs to simplex ETFs: the role of bias

A. One-hot encoded labels and balanced data

We address here the UFM problem (2) with one-hot encoded labels and balanced data, in line with [23], [27].

Theorem III.1 (Theorem 3.1 in [27], shortened)**.** *Assume that the labels are one-hot encoded i.e., $\bar{Y} = I_K$, that the data are balanced, i.e., $n_1 = \cdots = n_K =: n = \frac{N}{K}$, and define $c := K\sqrt{n\lambda_W \lambda_H}$. Then, if $c < 1$, any global minimizers (W^*, H^*, b^*) of* (2) *satisfy $\mathcal{NC}1$ and $\mathcal{NC}3$:*

$$H^* = \bar{H}^* C \quad \textit{and} \quad W^{*\top} = \sqrt{\frac{n\lambda_H}{\lambda_W}} \bar{H}^*, \tag{3}$$

where $\bar{H}^* \in \mathbb{R}^{d \times K}$ is such that

$$\bar{H}^{*\top} \bar{H}^* = \begin{cases} \alpha_1 \left(I_K - \frac{1_K 1_K^\top}{K} \right) & \text{if } \lambda_b \leqslant \frac{c}{1-c}, \\ \alpha_2 \left(I_K - \frac{c}{\lambda_b (1-c)} \frac{1_K 1_K^\top}{K} \right) & \text{otherwise,} \end{cases} \tag{4}$$

for some $\alpha_1 > 0$ and $\alpha_2 > 0$ that depend on λ_W, λ_H and λ_b. Moreover, the optimal bias b^* satisfies

$$b^* = \begin{cases} \frac{1}{1+\lambda_b} \frac{1_K}{K} & \text{if } \lambda_b \leqslant \frac{c}{1-c}, \\ \frac{c}{\lambda_b} \frac{1_K}{K} & \text{otherwise.} \end{cases} \tag{5}$$

Theorem III.1 reveals a transition in the structure of $\bar{H}^*$ as λ_b decreases, shifting from an OF (for $\lambda_b \to \infty$) to a simplex ETF (for $\lambda_b \leq c/(1-c)$). This observation suggests a link between OFs and simplex ETFs. Indeed, centering any orthogonal frame $M = \alpha P$ gives

$$M_c = \alpha P - (\alpha P) \frac{1_K 1_K^\top}{K},$$

which, according to Definition II.1, is a simplex ETF; see Figure 1 for an illustration. This led us to propose the following definition.

Definition III.1. (Shifted simplex ETF) A shifted simplex equiangular tight frame (SSETF) is a collection of K vectors in $\mathbb{R}^d$ (with $d \geqslant K$) specified by the columns of

$$M = \alpha P \left(I_K - \beta \frac{1_K 1_K^\top}{K} \right),$$

where $P \in \mathbb{R}^{d \times K}$ is semi-orthogonal, $\alpha > 0$ and $\beta \in \mathbb{R}$.

This proposed structure generalizes both OFs ($\beta = 0$) and simplex ETFs ($\beta = 1$). Note that centering any SSETF leads to a simplex ETF.

Proposition III.2. *If $M \in \mathbb{R}^{d \times K}$ is an SSETF, then $M_c = M \left(I_K - \frac{1_K 1_K^\top}{K} \right)$ is a simplex ETF.*

Proof. By Definition III.1, $M = \alpha P (I_K - \beta \frac{1_K 1_K^\top}{K})$ for some $\alpha, \beta \in \mathbb{R}$. By definition, $M_c = M(I_K - \frac{1_K 1_K^\top}{K})$, which

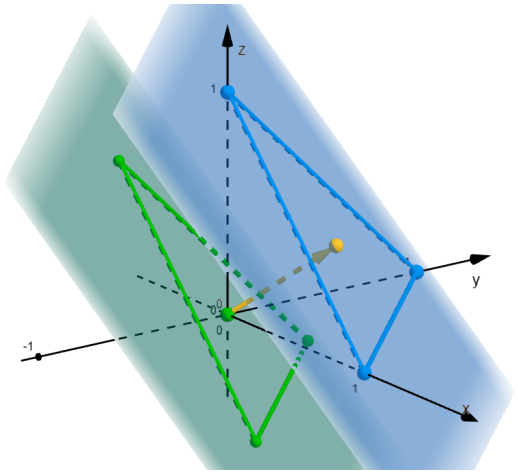

Fig. 1. Centering an OF (blue) always results in a simplex ETF (green). For this figure, the parameters in Definitions II.1 and II.2 are chosen as $d = K = 3$, $P = I_K$ and $\alpha = 1$.

simplifies to $M_c = \alpha P(I_K - \frac{1_K 1_K^\top}{K})$ for any β, i.e., M_c is a simplex ETF. □

Notice in (5) how the optimal bias b^* evolves as λ_b decreases: it transitions continuously from 0_K in the bias-free UFM ($\lambda_b \to \infty$) to $\frac{1}{K}1_K$ in the unregularized-bias UFM ($\lambda_b = 0$), namely, the mean of the labels. In other words, for one-hot encoded labels, the bias in the unregularized-bias UFM compensates for the non-centering of the labels.

B. Arbitrary labels

We generalize here the results of previous section to arbitrary label encoding $\bar{Y} \in \mathbb{R}^{K\times K}$ and imbalanced data. In this setting, the authors of [27] showed that, for the bias-free UFM, the optimal unconstrained feature matrix H^* is closely related to the right singular vectors of Y, i.e., depends on the encoding. To address the UFM with bias (2), we first derive optimality conditions in Lemma III.3, from which Theorem III.4 is proven.

Lemma III.3. *Let $\bar{Y} \in \mathbb{R}^{K\times K}$ be an arbitrary encoding matrix and let $Y = \bar{Y}C \in \mathbb{R}^{K\times N}$ for some class assignment matrix C, see* (1)*. Then, any global minimizer (W^*, H^*, b^*) of* (2) *satisfies the first-order optimality conditions:*

$$Y M_\gamma H^{*\top} = W^* H^* M_\gamma H^{*\top} + \lambda_W N W^*, \tag{6}$$

$$W^{*\top} Y M_\gamma = W^{*\top} W^* H^* M_\gamma + \lambda_H N H^*, \tag{7}$$

$$b^* = \frac{\gamma}{N}(Y - W^* H^*)1_N, \tag{8}$$

where $\gamma = \frac{1}{1+\lambda_b}$ and $M_\gamma = I_N - \gamma \frac{1_N 1_N^\top}{N}$.

Proof. Since (2) is strictly convex and coercive in b, setting the gradient w.r.t. b of (2) to 0 allows us to find b^*, leading to (8). Then, requiring the gradients w.r.t. W and H to be 0 give the stationarity conditions (6) and (7), after plugging in (8). As a result, any global minimizer (W^*, H^*, b^*) of (2) must satisfy (6)-(8). □

Theorem III.4. *Let $\bar{Y} \in \mathbb{R}^{K\times K}$ be an arbitrary encoding matrix and let $Y = \bar{Y}C \in \mathbb{R}^{K\times N}$ for some class assignment matrix C, see* (1)*.*

- *The UFM* (2) *with $\lambda_b = 0$ has optimal bias $b^* = Y\frac{1_N}{N}$. It reduces to the UFM with no bias for centered labels $Y - Y\frac{1_N 1_N^\top}{N}$.*
- *If Y is centered, i.e., $Y1_N = 0_K$, then for all $\lambda_b \geqslant 0$, the optimal bias of the UFM* (2) *is $b^* = 0_K$.*

Proof. In Lemma III.3, M_γ satisfies $M_\gamma 1_N = (1-\gamma)1_N$. Multiplying (7) on the right by 1_N thus gives

$$(1-\gamma)W^{*\top} Y 1_N = ((1-\gamma)W^{*\top} W^* + \lambda_H N I_d)(H^* 1_N).$$

If $\lambda_b = 0$ (i.e., $\gamma = 1$) or Y is centered, the left-hand side of the above equation vanishes. Moreover, since $(1-\gamma)W^{*\top}W^* + \lambda_H N I_d$ is always invertible for $\lambda_H > 0$ and $\lambda_b \geqslant 0$, it follows that $H^* 1_N = 0_K$ in both cases. Hence, by (8), $b^* = \frac{\gamma}{N} Y 1_N$. On the one hand, if $\lambda_b = 0$, the bias simplifies to $b^* = \frac{1}{N} Y 1_N$ and the objective in (2) becomes

$$\frac{1}{2N}\left\| WH + Y\frac{1_N 1_N^\top}{N} - Y \right\|_{\mathrm{F}}^2 + \frac{\lambda_W}{2}\|W\|_{\mathrm{F}}^2 + \frac{\lambda_H}{2}\|H\|_{\mathrm{F}}^2.$$

Thus, solving the UFM with $\lambda_b = 0$ and labels Y reduces to solving the bias-free UFM with centered labels $Y - Y\frac{1_N 1_N^\top}{N}$. On the other hand, if Y is centered, i.e., $Y1_N = 0_K$, then $b^* = 0_K$, regardless of the value of $\lambda_b \geqslant 0$. □

Theorem III.4 confirms that the bias in the UFM aims at centering the labels beyond one-hot encoding and balanced data.

IV. Impact of encoding on NC properties

We explore in this section the impact of the encoding on the NC properties, from the viewpoint of the UFM with MSE loss. First, we show that variability collapse $\mathcal{NC}1$ holds for any encoding. Then, for balanced scenarios, we relate convergence of the centered features to a simplex ETF (i.e., $\mathcal{NC}2$) to SSETF encodings and demonstrate that a weaker version of self-duality, denoted $\mathcal{NC}3'$, is preserved regardless of the encoding.

A. Robustness of variability collapse ($\mathcal{NC}1$)

$\mathcal{NC}1$ has been shown to hold for multiple instances of the UFM under MSE loss, such as in the bias-free case with one-hot encoded labels and imbalanced data [13] or for $\lambda_b \geqslant 0$, with one-hot labels and balanced data [27]. To our knowledge, our next result is the first to ensure that $\mathcal{NC}1$ holds more generally for any bias regularization $\lambda_b \geqslant 0$ and for any encoding as well as for imbalanced scenarios. The independence of variability collapse on the encoding and on the value of λ_b can be explained intuitively: assuming that the model has found an optimal feature representation for a sample of a given class, then all samples of the same class should be represented identically, because of the separability of the problem regarding the samples.

Theorem IV.1. *Let $\bar{Y} \in \mathbb{R}^{K\times K}$ be an arbitrary encoding matrix and let $Y = \bar{Y}C \in \mathbb{R}^{K\times N}$ for some class assignment*

matrix C, see (1)*. Assume that classes are balanced, and let (W^*, H^*, b^*) be an optimal solution of* (2)*. Then, for any $\lambda_H, \lambda_W > 0$ and $\lambda_b \geqslant 0$, H^* satisfies $\mathcal{NC}1$, i.e.,*

$$H^* = \bar{H}^* C \tag{9}$$

for some $\bar{H}^* \in \mathbb{R}^{d \times K}$.

Proof. Let $h_{k,i}$ be the feature vector corresponding to the i^{th} sample of class k, with $1 \leqslant i \leqslant n_k$ and $1 \leqslant k \leqslant K$. We decompose the two terms involving H in (2) as

$$\frac{1}{2N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|W h_{k,i} + b - \bar{y}_k\|_2^2 + \frac{\lambda_H}{2} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|h_{k,i}\|_2^2.$$

For the k^{th} sum of the second term, we get

$$\sum_{i=1}^{n_k} \|h_{k,i}\|_2^2 \geqslant \frac{1}{n_k} \left\| \sum_{i=1}^{n_k} h_{k,i} \right\|_2^2 = n_k \|\bar{h}_k\|_2^2, \tag{10}$$

where we first used Jensen's inequality applied to the convex function $f(x) = \|x\|_2^2$ for $x \in \mathbb{R}^d$, and where we defined $\bar{h}_k := (\sum_{i=1}^{n_k} h_{k,i})/n_k$. Note that, for each class k, equality can always be reached in (10) and holds if and only if $h_{k,i} = \bar{h}_k$ for all $1 \leqslant i \leqslant n_k$. Then, we apply the same reasoning for the k^{th} sum of the first term:

$$\sum_{i=1}^{n_k} \|W h_{k,i} + b - \bar{y}_k\|_2^2 \geqslant n_k \|W \bar{h}_k + b - \bar{y}_k\|_2^2, \tag{11}$$

with equality in (11) if and only if $W h_{k,i} + b - \bar{y}_k = W \bar{h}_k + b - \bar{y}_k$ for all $1 \leqslant i \leqslant n_k$, or equivalently $W h_{k,i} = W \bar{h}_k$. These conditions are weaker than the previously derived conditions $h_{k,i} = \bar{h}_k$, and consequently do not impose any additional constraints on H. Then, any optimal H^* in (2) can be written as $H^* = \bar{H}^* C$ with $\bar{H}^* = [\bar{h}_1 \cdots \bar{h}_K]$. □

B. Convergence to a simplex ETF ($\mathcal{NC}2$) for SSETF encodings

Currently, the complete determination of which encodings result in $\mathcal{NC}2$ for the UFM with MSE loss remains an open question, both in the balanced and imbalanced cases. We next show that a sufficient condition for $\mathcal{NC}2$ is the fact that the optimal mean activations $\bar{H}^*$ form an SSETF; this directly follows from Proposition III.2.

Corollary IV.2. *Let $\bar{Y} \in \mathbb{R}^{K \times K}$ be an arbitrary class encoding matrix. If, for any $\lambda_W, \lambda_H > 0$ and $\lambda_b \geqslant 0$, the optimal mean features $\bar{H}^*$ of* (2) *defined in* (9) *form an SSETF, then their centered counterpart $\bar{H}_c^* = \bar{H}^*(I_K - \frac{1_K 1_K^\top}{K})$ form a simplex ETF, in agreement with $\mathcal{NC}2$.*

This corollary shows that the optimal solutions of the UFM with MSE loss for balanced one-hot encoded data, provided in Theorem III.1, satisfy $\mathcal{NC}2$, as empirically observed by the authors of [23]. Future work will aim at deriving necessary and sufficient conditions on the encoding matrix $\bar{Y} \in \mathbb{R}^{K \times K}$ to result in mean features $\bar{H}^*$ with an SSETF structure, and consequently, to lead in $\mathcal{NC}2$. We conjecture that, in balanced scenarios, $\bar{H}^* \in \mathbb{R}^{d \times K}$ is an SSETF if and only if the encoding $\bar{Y} \in \mathbb{R}^{K \times K}$ is an SSETF. The family of SSETFs encompasses two widely used encodings: one-hot encoding (for $\beta = 0$, $\alpha = 1$, $d = K$ and $P = I_K$ in Definition III.1), and label smoothing, a "smoothed" variant of one-hot encoding. Indeed, smoothed labels can be written as $\bar{Y}^\delta = (1-\delta) I_K + \frac{\delta}{K} 1_K 1_K^\top$ for some smoothing parameter $0 < \delta < 1$ [7], [21], and form an SSETF; simply take $\alpha = 1 - \delta$, $\beta = \frac{\delta}{1-\delta}$, $d = K$ and $P = I_K$ in Definition III.1. While we showed earlier that $\mathcal{NC}2$ holds for one-hot encoded labels (see Corollary IV.2), this is still unproven for label smoothing, though empirical results supporting IV.2 were obtained using the cross-entropy loss [7].

C. Failure of self-duality ($\mathcal{NC}3$) and success of a weaker form of self-duality ($\mathcal{NC}3'$)

For balanced data and one-hot encoded labels, Theorem III.1 showed that ${W^*}^\top$ and $\bar{H}^*$ are necessarily aligned, i.e., the global minimizers of the UFM satisfy $\mathcal{NC}3$. We show here that $\mathcal{NC}3$ may fail, depending on the encoding. In particular, applying an orthogonal transformation Q to the labels destroys the alignmnent between ${W^*}^\top$ and $\bar{H}^*$.

Proposition IV.3. *Let $\bar{Y} \in \mathbb{R}^{K \times K}$ be arbitrary and let $Y = \bar{Y} C \in \mathbb{R}^{K \times N}$ for some class assignment matrix C, see* (1)*. Let $Q \in \mathbb{R}^{K \times K}$ be an orthogonal matrix and let $Y_Q = QY$. Then, the global minimizers of the UFM with labels Y_Q are of the form (QW^*, H^*, Qb^*), where (W^*, H^*, b^*) is any global minimizer of the UFM with labels Y.*

Proof. Replacing (W, H, b) in the UFM (2) with modified labels Y_Q by (QW', H, Qb') and using the invariance of the Frobenius norm under multiplication with an orthogonal matrix yields the result. Since Q is invertible, this change of variables is reversible, providing a bijection between the set of global minimizers of the original problem and those of the problem with new labels. □

The optimal classification weights W^* (and the bias b^*) directly follow the encoding Y undergoing any left orthogonal transformation, indicating that the left singular vectors of W^* are strongly related to the ones of $\bar{Y}$. This observation matches the result from [27] in the bias-free case. Meanwhile, the mean class features $\bar{H}^*$ stay unchanged, so the self-duality between ${W^*}^\top$ and $\bar{H}^*$ is broken. However, inverting the isometry reconstructs their alignment, indicating that the property is not denied at a more fundamental level. Based on this remark, we propose the following weaker variant of $\mathcal{NC}3$, that we will show to hold in a broader setting.

- $\mathcal{NC}3'$ (Variant of $\mathcal{NC}3$): the last layer's classifier weights converge to the class means, up to rescaling and rotation/reflections.

Mathematically, $\mathcal{NC}3'$ is satisfied if and only if there exists an orthogonal matrix $Q \in \mathbb{R}^{K \times K}$ and a scalar $\kappa > 0$ such that

$${W^*}^\top = \kappa \bar{H}^* Q^\top. \tag{12}$$

This equality can be equivalently stated in terms of the Gram matrices of ${W^*}^\top$ and $\bar{H}^*$:

$${W^*}^\top W^* = \kappa^2 \bar{H}^* {\bar{H}^*}^\top.$$

As stated in Theorem IV.4, $\mathcal{NC}3'$ is always verified in the UFM with MSE loss and balanced classes. Besides, the theorem implies that there always exists an orthogonal matrix $Q \in \mathbb{R}^{K\times K}$ such that it reverts to the original self-duality property $\mathcal{NC}3$ when considering the labels $Q^\top Y$.

Theorem IV.4. *Let $\bar{Y} \in \mathbb{R}^{K\times K}$ be an arbitrary class encoding matrix, and assume that the data are balanced. Then, for any $\lambda_W, \lambda_H > 0$ and $\lambda_b \geqslant 0$, the optimal mean features $\bar{H}^*$ of (2) defined in (9) satisfy $\mathcal{NC}3'$, with scaling coefficient $\kappa = \sqrt{\frac{\lambda_H n}{\lambda_W}}$ in (12).*

Proof. Multiply the optimality conditions (6) and (7) respectively by ${W^*}^\top$ on the right and by ${\bar{H}^*}^\top$ on the left. Subtracting these two equations gives $\lambda_W N {W^*}^\top W^* = \lambda_H N H^* {H^*}^\top$. By Theorem IV.1, we can substitute $H^* = \bar{H}^* C$, where C is the class assignment matrix. Finally, leveraging the equality $CC^\top = nI_K$ for balanced classes yields ${W^*}^\top W^* = \frac{\lambda_H n}{\lambda_W} \bar{H}^* {\bar{H}^*}^\top$, where we identify $\kappa^2 = \frac{\lambda_H n}{\lambda_W}$. □

V. Conclusion

This work studies the role of class encoding in NC. It demonstrates that within the UFM under MSE loss, the final classifier's bias aims at compensating for the non-centering of the labels. In the case of balanced, one-hot encoded labels, increasing the bias regularization parameter yields a continuous transition of the unconstrained mean features from an OF to an ETF. We finally discuss the dependency of other NC properties on class encoding.

Acknowledgment

R. M. is a FRIA grantee of the Fonds de la Recherche Scientifique - FNRS. E. M. work is partly funded by the FRS-FNRS Research Project NTTN (grant number TW02223) and the Concerted Research Action (ARC) "Gravit-AI".

References

[1] G. Andriopoulos, Z. Dong, L. Guo, Z. Zhao, and K. Ross. The prevalence of neural collapse in neural multivariate regression. In *Proceedings of the 38th Conference on Neural Information Processing Systems*, 2024.

[2] A. Demirkaya, J. Chen, and S. Oymak. Exploring the role of loss functions in multiclass classification. In *Proceedings of the 54th Annual Conference on Information Sciences and Systems*, 2020.

[3] W. E and S. Wojtowytsch. On the emergence of simplex symmetry in the final and penultimate layers of neural network classifiers. In *Proceedings of 2nd Annual Conference on Mathematical and Scientific Machine Learning*, 2021.

[4] 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.

[5] T. Galanti, A. György, and M. Hutter. On the role of neural collapse in transfer learning. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[6] F. Graf, C. D. Hofer, M. Niethammer, and R. Kwitt. Dissecting supervised constrastive learning. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[7] L. Guo, G. Andriopoulos, Z. Zhao, S. Ling, Z. Dong, and K. Ross. Cross entropy versus label smoothing: A neural collapse perspective. arXiv:2402.03979, 2025.

[8] J. Haas, W. Yolland, and B. T. Rabus. Linking neural collapse and l2 normalization with improved out-of-distribution detection in deep neural networks. *Transactions on Machine Learning Research*, 2022.

[9] X. Y. Han, V. Papyan, and D. L. Donoho. Neural collapse under MSE loss: proximity to and dynamics on the central path. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[10] L. Hui and M. Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[11] J. Jiang, J. Zhou, P. Wang, Q. Qu, D. G. Mixon, C. You, and Z. Zhu. Generalized neural collapse for a large number of classes. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[12] V. Kothapalli. Neural collapse: A review on modelling principles and generalization. *Transactions on Machine Learning Research*, 2023.

[13] H. Liu. The exploration of neural collapse under imbalanced data. arXiv:2411.17278, 2024.

[14] X. Liu, J. Zhang, T. Hu, H. Cao, L. Pan, and Y. Yao. Inducing neural collapse in deep long-tailed learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 2023.

[15] J. Lu and S. Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022.

[16] D. G. Mixon, H. Parshall, and J. Pi. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis*, 20(11), 2020.

[17] V. Papyan, X. Y. 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.

[18] A. Rangamani, M. Lindegaard, T. Galanti, and T. Poggio. Feature learning in deep classifiers through intermediate neural collapse. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[19] P. Sukenik, C. H. Lampert, and M. Mondelli. Neural collapse versus low-rank bias: Is deep neural collapse really optimal? In *Proceedings of the 38th Conference on Neural Information Processing Systems*, 2024.

[20] P. Sukenik, M. Mondelli, and C. H. Lampert. Deep neural collapse is provably optimal for the deep unconstrained features model. In *Proceedings of the 37th Conference on Neural Information Processing Systems*, 2023.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[22] C. Thrampoulidis, G. R. Kini, V. Vakilian, and T. Behnia. Imbalance trouble: Revisiting neural-collapse geometry. In *Proceedings of the 36th Conference on Neural Information Processing Systems*, 2022.

[23] T. Tirer and J. Bruna. Extended unconstrained features model for exploring deep neural collapse. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[24] T. Tirer, H. Huang, and J. Niles-Weed. Perturbation analysis of neural collapse. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[25] R. Wu and V. Papyan. Linguistic collapse: Neural collapse in (large) language models. In *Proceedings of the 38th Conference on Neural Information Processing Systems*, 2024.

[26] C. Yaras, P. Wang, Z. Zhu, L. Balzano, and Q. Qu. Neural collapse with normalized features: A geometric analysis over the Riemannian manifold. In *Proceedings of the 36th Conference on Neural Information Processing Systems*, 2022.

[27] 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. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[28] J. Zhou, C. You, X. Li, K. Liu, S. Liu, Q. Qu, and Z. Zhu. Are all losses created equal: A neural collapse perspective. In *Proceedings of the 36th Conference on Neural Information Processing Systems*, 2022.

[29] Z. Zhu, T. Ding, J. Zhou, X. Li, C. You, J. Sulam, and Q. Qu. A geometric analysis of neural collapse with unconstrained features. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, 2021.