OPT2022: 14th Annual Workshop on Optimization for Machine Learning

On the Implicit Geometry of Cross-Entropy Parameterizations for Label-Imbalanced Data

Tina Behnia[⋆] TINA.BEHNIA@ECE.UBC.CA
Ganesh Ramachandra Kini[⋆] KINI@UCSB.EDU
Vala Vakilian[⋆] VAALAA@ECE.UBC.CA
Christos Thrampoulidis CTHRAMPO@ECE.UBC.CA

Abstract

It has been empirically observed that training large models with weighted cross-entropy (CE) beyond zero training error is not a satisfactory remedy for label-imbalanced data. Instead, the recently introduced vector-scaling (VS) loss parameterizes the CE loss in a way tailored to this modern training regime. The driving force to understand the impact of such parameterizations on the gradient-descent path has been the theory of *implicit bias*. For linear(ized) models, this theory allows to explain why weighted CE fails and how the VS-loss biases the optimization path towards solutions favoring minorities. However, beyond linear models the description of implicit bias is more obscure. In order to gain insights on the impact of different CE-parameterizations in non-linear models, we investigate their *implicit geometry* of learned classifiers and embeddings. Our main result characterizes the global minimizers of a non-convex cost-sensitive SVM classifier for the unconstrained features model (UFM), which serves as a deep-net abstraction. We also study empirically, both for the UFM and for deep-nets, the convergence of SGD to this global minimizer observing slow-downs with increasing imbalance ratios and scalings of the loss hyperparameters.

1. Introduction

In the overparameterized regime, when training continues beyond zero-training error, traditional techniques, such as oversampling minorities or minimizing weighted cross-entropy (CE) can be ineffective in mitigating label-imbalances [1, 22]. A growing literature has proposed several alternatives to guarantee equitable performance across majorities and minorities [2, 11–15, 18, 27]. Among these, the vector-scaling (VS)-loss [14, 27] introduces multiplicative hyperparameters on the CE logits. The idea is rooted in the theory of *implicit bias*, which seeks characterizing the bias introduced by gradient-based algorithms during training [9, 17, 23]. Specifically for binary linear models, [14] uncovers a favorable bias of the VS-loss towards classifiers with larger margin for the minority. However, this leaves open the question how the VS-loss changes the learnt model in non-linear settings where embeddings and classifiers are jointly learnt. Unfortunately, implicit bias characterizations for non-linear models are more obscure compared to the linear case [10, 17] making it unclear how to gain concrete insights on the way the learnt models affect minorities.

This paper undertakes the investigation of the *implicit geometry* of classifiers and embeddings learnt by CE-parameterizations when training on imbalanced data. The notion of implicit geometry,[1] pioneered by [20] and further investigated by [3–8, 16, 19, 24–26, 28, 29],

§: equal contribution. The long version of this paper is available on arXiv.

[1] In the literature this is known as *neural collapse* (NC) [20]. However, [24] showed that training CE loss on imbalanced data learns a geometry different from that in [20] for balanced data. To draw distinction,

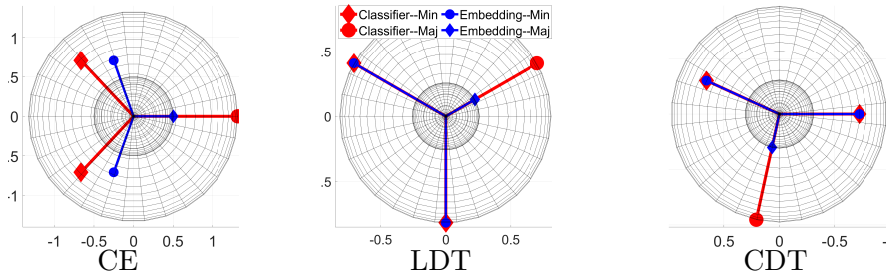

Figure 1: Geometries induced by CE and its two parameterizations: LDT and CDT. $k = 3$ classes (2 minorities); imbalance ratio $R = 10$; hyperparameters $\delta_{\text{maj}}/\delta_{\text{min}} = \sqrt{R}$. See Cor. 3 for details.

is intimately related to that of implicit bias. On the one hand, it is more restrictive as it focuses only on the classifiers and on the embeddings, rather than the weights of the entire model. Also, it is insensitive to the specific architecture or dataset. On the other hand, it offers a more explicit characterization that describes the involved geometry of the weights and promises to be "cross-situationally invariant" across architectures and datasets [20].

Contributions. We study two parameterizations of the CE loss. First, the class-dependent temperature (CDT)-loss [27], which is a special case of the VS-loss [14]. Second, the label-dependent temperature (LDT)-loss, which we introduce as an alternative to the CDT-loss. For both losses, we investigate the implicit geometry of learnt features and classifiers when training on label-imbalanced data without explicit regularization beyond zero training error. To do this, we rely on the unconstrained features model (UFM) [3, 19], a proxy for large models recently used to study the implicit geometry of the CE loss [8, 24, 28, 29]. Relying on implicit-bias results, we relax the question of implicit geometry of SGD solutions, to a question about the geometry of the global minimizers of a non-convex cost-sensitive support-vector machines (SVM) problem. Our main result characterizes the global minimizers of this problem (which takes different form for the CDT and LDT losses). The characterization is explicit in terms of the number of classes, the minority and imbalance ratios, and the loss-hyperparameters. It also leads to closed-form expressions for the norms and pairwise angles of the learned geometries. This show explicitly how to best tune the hyperparameters to induce symmetry in the geometry (see Fig. 1). We also show numerically that SGD training on the UFM converges to the uncovered geometries. However, we observe that convergence slows down for increasing values of the imbalance ratio and of the hyperparameters. This motivates further theoretical and algorithmic investigations towards faster training. Finally, as evidence of the utility of our geometry characterizations for the UFM, we present preliminary results on deep-nets and complex imbalanced datasets.

2. Background

The VS-loss is the following parameterization of the CE-loss [14]:

$$\mathcal{L}_{\text{VS}}(\mathbf{W};\boldsymbol{\theta}) := \sum_{i\in[n]} \log\big(1 + \sum_{c\in[k]:\ c\neq y_i} e^{-(\delta_{y_i}\mathbf{w}_{y_i} - \delta_c \mathbf{w}_c)^T \mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_i) + \iota_{y_i} - \iota_c}\big). \tag{1}$$

Here $\mathbf{x}_i, i \in [n]$ are n examples, $\mathbf{h}_{\boldsymbol{\theta}}(.)$ is the feature map parameterized by trainable parameters $\boldsymbol{\theta}$ (weights of the hidden layers of a neural network), $y_i \in [k]$ are the labels,

they refer to the former and latter as SELI and ETF geometries, respectively. We show that different CE parameterizations result in yet different geometries. This motivates the term "implicit geometry".

and $\mathbf{w}_c, c \in [k]$ are the classifier vectors (head of the network) in a k-class classification setting. The parameters δ_c, and $\iota_c, c \in [k]$ are multiplicative and additive hyperparameters, respectively. Setting $\delta_c = 1, \iota_c = 0, \forall c \in [k]$ recovers the CE loss.

Prior art: Binary linear classification. In a binary setting with fixed feature map (i.e. non-trainable $\boldsymbol{\theta}$) [14] studies the implicit bias of gradient descent (GD) on binary VS loss.

Proposition 1 ([14]) *For fixed feature map $\mathbf{h}_{\boldsymbol{\theta}}$, $v_i \in \{\pm 1\}$, and $\mathbf{h}_i := \mathbf{h}_{\boldsymbol{\theta}}(i), i \in [n]$, GD with sufficiently small learning rate on the binary VS-loss $\mathcal{L}_{\text{VS,binary}}(\mathbf{w}) := \sum_{i\in[n]} \log(1 + e^{-\delta_{v_i} v_i \mathbf{w}^T \mathbf{h}_i + \iota_{v_i}})$, converges (asymptotically in the number of training steps) in direction to the Cost-Sensitive SVM (CS-SVM) classifiers:* $\arg\min_{\mathbf{w}} \|\mathbf{w}\|_2^2$ subj. to $v_i \delta_{v_i} \mathbf{w}^T \mathbf{h}_i \geq 1, i \in [n]$.

Prop. 1 explicitly describes how the hyperparameters affect training asymptotically: the GD path is implicitly biased towards a classifier that assigns margins to the two classes with relative ratio δ_{-1}/δ_{+1}. Thus, tuning $\delta_{-1} > \delta_{+1}$ if class $v = +1$ is minority, favors the minority by assigning larger margin to it. Note, the additive hyperparameters ι_c have *no* have effect on the implicit bias asymptotically. Our focus here is on the asymptotic training regime, hence onwards we restrict attention to the multiplicative hyperparameters δ_y.

Open problem: Beyond linear models. Prop. 1 is limited to a setting with fixed features. While an extension of the loss itself to the learned-feature setting is easy to heuristically derive (see Eqn. (1)), it is an open question to explicitly characterize the effect of the hyperparameters on the learnt solution: How do they affect the relative margin between majorities and minorities or between minorities vs minorities?

3. An implicit-geometry view

To better understand the impact of different CE modifications, we study their implicit geometry, i.e. the geometry of classifiers and embeddings learned (asymptotically in the number of training steps) by GD. For this, we adopt the *unconstrained features model* (UFM) [3, 19]. Let $\mathbf{W}_{d\times k} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_k]$ and $\mathbf{H}_{d\times n} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n]$ be the matrix of k classifier weights and n feature embeddings corresponding to each example in the training set. Here, $d \geq k-1$ is the feature dimension. We assume each class $c \in [k]$ has $n_c \geq 1$ examples so that $\sum_{c\in[k]} n_c = n$. Without loss of generality, we assume examples are ordered, i.e. samples $i = 1, \ldots, n_1$ have label $y_i = 1$ and so on. In the UFM the features $\mathbf{h}_i, i \in [n]$ are trained *jointly* with the weights $\mathbf{w}_c, c \in [k]$ and are trained *unconstrained*, i.e. without abiding by an explicit parameterization by some weight vector $\boldsymbol{\theta}$ (as in (1)). Previous works use the UFM to derive the implicit geometry of CE loss on balanced (e.g. [3, 19, 28, 29]) and on imbalanced data [24]. Instead, we study different parameterizations of the CE loss.

CDT/LDT losses on the UFM. Consider training the UFM by minimizing the following two parameterization of the CE loss:

$$\mathcal{L}_{\text{CDT}}(\mathbf{W}^T\mathbf{H}; \boldsymbol{\delta}) := \sum_{i\in[n]} \log\Big(1 + \sum_{c\neq y_i} e^{-(\delta_{y_i}\mathbf{w}_{y_i} - \delta_c \mathbf{w}_c)^T \mathbf{h}_i}\Big), \tag{2a}$$

$$\mathcal{L}_{\text{LDT}}(\mathbf{W}^T\mathbf{H}; \boldsymbol{\delta}) := \sum_{i\in[n]} \log\Big(1 + \sum_{c\neq y_i} e^{-(\delta_{y_i}(\mathbf{w}_{y_i} - \mathbf{w}_c))^T \mathbf{h}_i}\Big). \tag{2b}$$

Both losses are parameterized by a vector $\boldsymbol{\delta} = \big[\delta_1, \delta_2, \ldots, \delta_k\big]^T \in \mathbb{R}^k$ of positive multiplicative hyperparameters. The CDT loss (2a) was previously introduced by [14, 27] (note is a special case of (1) when ignoring ι_c). Here, we also introduce the LDT loss (2b) as an alternative

parameterization: CDT associates the hyperparameters $\boldsymbol{\delta}$ with the class label of the classifier vectors $\mathbf{w}_c$, while LDT assigns the same $\boldsymbol{\delta}$ to the label of the feature vector $\mathbf{h}_i$.

Unconstrained-features cost-sensitive SVM. We minimize the losses in (2) without explicit regularization. Note that in the UFM, minimization over the embedding map is not parameterized in terms of $\boldsymbol{\theta}$, as say in (1). Thus, the minimization is (joint) over classifiers $\mathbf{W}$ and embeddings $\mathbf{H}$. Specifically, consider performing this minimization using gradient flow (i.e. GD with infinitesimal step-size.) Then, by interpreting the UFM as a two-layer linear model (e.g. see [24]) it can be shown following [17] that gradient flow will converge (asymptotically in time) in direction to a KKT point of the following two non-convex minimizations for CDT and LDT losses respectively:

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2 \quad \text{subj. to} \quad (\delta_{y_i}\mathbf{w}_{y_i} - \delta_c \mathbf{w}_c)^T \mathbf{h}_i \geq 1,\ i \in [n], c \neq y_i, \tag{3a}$$

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2 \quad \text{subj. to} \quad \delta_{y_i}(\mathbf{w}_{y_i} - \mathbf{w}_c)^T \mathbf{h}_i \geq 1,\ i \in [n], c \neq y_i. \tag{3b}$$

Note the resemblence of the above to the CS-SVM minimization of Prop. 1. But unlike in Prop. 1, the problems here are non-convex since minimization is also over $\mathbf{H}$. We refer to (3) as CS-SVM. Our main result characterizes their global solutions.

4. Global structure of the CS-SVM

In this section, we use $\mathbf{M}_{d\times k} = [\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k]$ to denote the mean-embeddings of $\mathbf{H}$, i.e. $\boldsymbol{\mu}_c = (1/n_c)\sum_{i:y_i=c} \mathbf{h}_i, \forall c \in [k]$. For simplicity, we focus on a (R,ρ)-STEP-imbalanced setting, where we have ρk minority classes with $n_{\min}$ samples each, and $\overline{\rho}k = (1-\rho)k$ majority classes with $Rn_{\min}$ samples. In this case, it is reasonable to assume that $\boldsymbol{\delta}$ also shares this STEP structure, i.e. for majorities $\delta_c = \delta_{\text{maj}} > 0$ and for minorities $\delta_c = \delta_{\min} > 0$. We describe the geometry of the unconstrained CS-SVM solutions $(\mathbf{W}^*, \mathbf{H}^*)$ of (3a) and (3b) in terms of an encoding matrix and a geometry induced by the SVD factors of this matrix.

Definition 1 *For hyperparameters $\boldsymbol{\delta} \in \mathbb{R}_+^k$, minority fraction ρ ($\overline{\rho} := 1-\rho$), k classes, and a rational imbalance ratio R [2], the $(\boldsymbol{\delta}, R)$-Simplex-Encoding Label (SEL) matrix and the corresponding Simplex-Encoded-Labels Interpolation (SELI) geometry are defined as follows:*

(a) $(\boldsymbol{\delta}, R)$-SEL matrix. *Define $\boldsymbol{\Xi}_{k\times k}$ such that $\forall c, j \in [k]$,*

$$\boldsymbol{\Xi}[c,j] = \begin{cases} \delta_c^{-1}\left(1 - \delta_c^{-2}/\sum_{c'\in[k]}\delta_{c'}^{-2}\right) & , c = j \\ -\delta_c^{-1}\left(\delta_j^{-2}/\sum_{c'\in[k]}\delta_{c'}^{-2}\right) & , c \neq j \end{cases}.$$

Choose $\alpha \in \mathbb{N}$ such that $\alpha R \in \mathbb{N}$, and let $n := \alpha k(R\overline{\rho} + \rho)$. Then, the $(\boldsymbol{\delta}, R)$-SEL matrix, along with its SVD, is defined as follows,

$$\hat{\mathbf{Z}}_{k\times n} = \left[\boldsymbol{\Xi}_{1:\overline{\rho}k} \otimes \mathbb{1}_{\alpha R}^T \quad \boldsymbol{\Xi}_{(\overline{\rho}k+1):k} \otimes \mathbb{1}_{\alpha}^T\right] = \mathbf{V}\boldsymbol{\Lambda}\left[\mathbf{U}_{1:\overline{\rho}k}^T \otimes \mathbb{1}_{\alpha R}^T \quad \mathbf{U}_{(\overline{\rho}k+1):k}^T \otimes \mathbb{1}_{\alpha}^T\right], \tag{4}$$

where $\boldsymbol{\Lambda} \in \mathbb{R}_+^{(k-1)\times(k-1)}$ is a diagonal matrix and $\mathbf{U}$, $\mathbf{V} \in \mathbb{R}_{k\times(k-1)}$ have orthonormal columns.

(b) $(\boldsymbol{\delta}, R)$-SELI geometry. *The classifier and mean-embeddings matrices $\mathbf{W}, \mathbf{M} \in \mathbb{R}^{d\times k}$ follow the $(\boldsymbol{\delta}, R)$-SELI geometry if the following conditions are satisfied:*

$$\mathbf{W}^T\mathbf{W} \propto \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T, \quad \mathbf{M}^T\mathbf{M} \propto \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T, \text{ and } \quad \mathbf{W}^T\mathbf{M} \propto \mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^T = \boldsymbol{\Xi}. \tag{5}$$

[2] This assumption is not restrictive since under STEP imbalance $R := n_{\text{maj}}/n_{\text{minor}}$ for integers $n_{\text{maj}}, n_{\text{minor}}$.

The $(\boldsymbol{\delta}, R)$-SELI geometry is specified by the SVD factors of the corresponding SEL matrix. The first two properties describe the relative norms and pair-wise angles of classifiers/mean embeddings, and the third one determines the relative margins between classes. The pattern of the $(\boldsymbol{\delta}, R)$-SEL matrix is determined by R and $\boldsymbol{\delta}$, as well as ρ and k. We drop the latter dependence in the name, since they are understood from context and our results focus on the role of R and $\boldsymbol{\delta}$. When $\boldsymbol{\delta} = \mathbb{1}_k$, $\hat{\mathbf{Z}}$ takes a special form: it reduces to a matrix with entries $1 - 1/k$ and $-1/k$, which [24] calls the SEL matrix and shows that it characterize the implicit geometry of the CE loss for imbalanced data (aka SELI geometry) [24]. Further, considering the balanced case $R = 1$, recovers the ETF geometry. Our definition is strictly more general allowing us to describe the implicit geometry learned by CDT/LDT losses. We note that $\hat{\mathbf{Z}}^T \operatorname{diag}(\boldsymbol{\delta})^{-1} \mathbb{1}_k = 0$. Thus, $\operatorname{rank}(\hat{\mathbf{Z}}) = k - 1$.

With these we are ready to state our main result.

Theorem 2 *Suppose $d \geq k - 1$ in an (R, ρ)-STEP imbalance setting. Let $(\mathbf{W}^*, \mathbf{H}^*)$ be any minimizers of either (3a) or (3b), and $\mathbf{M}^*$ be the optimal class-wise mean-embeddings. Then, all embeddings collapse to their class means, i.e., $\forall i \in [n]$, $\mathbf{h}_i^* = \boldsymbol{\mu}_{y_i}^*$. Also,*

(i) **[CDT (3a)]** *$(\mathbf{W}^*, \mathbf{M}^*)$ follow the $(\boldsymbol{\delta}, R)$-SELI geometry.*

(ii) **[LDT (3b)]** *$(\mathbf{W}^*, \mathbf{M}^* \operatorname{diag}(\boldsymbol{\delta}))$ follow $(\mathbb{1}_k, \tilde{R})$-SELI geometry, where $\tilde{R} := R\big(\delta_{\min}/\delta_{\mathrm{maj}}\big)^2$, provided $\tilde{R}$ is rational.*[3]

Thm. 2 describes the geometry of both classifiers and embeddings that solve the non-convex CS-SVM for either CDT or LDT. From statements (i) and (ii) we can find the optimal classifiers and mean-embeddings, their norms (up to a constant) and pair-wise angles in terms of the geometry in Defn. 1. It is also easy to see that the geometry only depends on the ratio $\Delta := \delta_{\mathrm{maj}}/\delta_{\min}$. The special case of $\boldsymbol{\delta} = \mathbb{1}_k$ and $R = 1$ recovers the SELI [24] and ETF [20] geometries. For general R and tuning of $\boldsymbol{\delta}$, the LDT/CDT geometries are different than both the SELI and ETF geometries. We visualize changes in the geometry in Fig. 1.

Closed-form geometry formulas. Explicit computation of the SVD factors of the (δ, R)-SEL matrix yields closed-form expressions for the norms and pair-wise angles of the mean-embeddings and classifiers. These formulas can then be evaluated in the limit $R \to \infty$ giving further insights for mitigating collapse of minority classifiers; see long version for details. (The collapse of minorities as $R \to \infty$ is studied in [3]).

Special tunings. Two notable special cases arise when setting $\delta_c = \sqrt{n_c}$, as shown below.

Corollary 3 *Set $\Delta := \delta_{\mathrm{maj}}/\delta_{\min} = \sqrt{R}$. Then, the followings hold: **(i)** For CDT loss, each feature embedding $\mathbf{h}_i^*$ perfectly aligns with its corresponding classifier $\mathbf{w}_{y_i}^*$, i.e. $\cos(\mathbf{w}_{y_i}^*, \mathbf{h}_i^*) = 1, \forall i \in [n]$. **(ii)** For LDT loss, $(\mathbf{W}^*, \mathbf{M}^* \operatorname{diag}(\boldsymbol{\delta}))$ follows the ETF geometry.*

Note the LDT geometry becomes exactly the same (for all R) as the perfectly symmetric ETF geometry modulo a scaling on the norms of the embeddings (minority norms are larger than majorities by a factor of $\sqrt{R}$). For the same tuning, the CDT geometry aligns perfectly the embeddings of each class with the corresponding classifier. See Fig. 1 for an illustration.

[3] This a technical requirement. In our experiments we apply the same formulas even when $\tilde{R}$ is not rational.

Numerical experiments. We investigate convergence of SGD trajectory for unregularized CDT/LDT-losses in (2a)/(2b) to the implicit geometries in Thm. 2 for both UFM and deep-nets with standard datasets. In both cases, we parameterize $\Delta = R^\gamma$ with $\gamma \in [-1.5, 1.5]$. For each choice of γ and loss function, we compute the average norm ratios and angles between each pair of majority and minority classifiers/embeddings and compare it to CDT/LDT geometries. Fig. 2 shows the results for $(R = 10, \rho = 1/2)$-STEP imbalanced UFM, while Fig. 3 in the SM shows the results for ResNet-18 model on imbalanced CIFAR-10 dataset. The details of the experiments are given in the SM. Additional theoretical and experimental results are available in the long version of the paper.

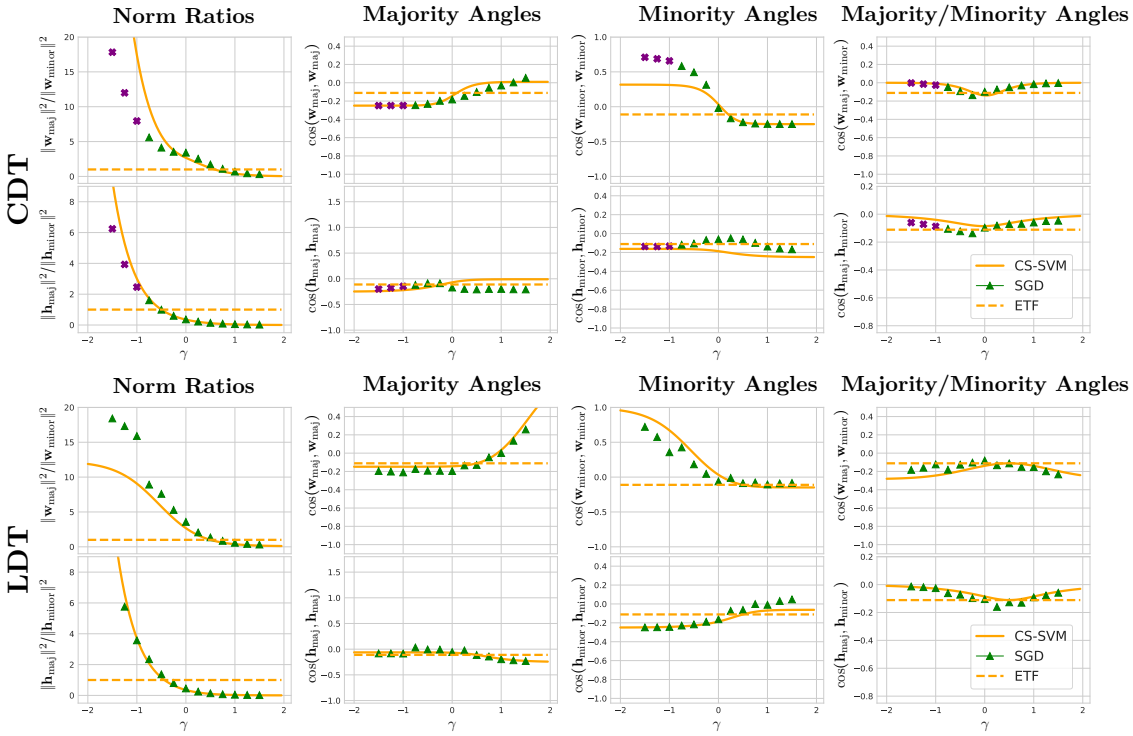

Figure 2: Properties of the solutions found by SGD (markers) trained on the CDT/LDT losses in (2a)/(2b) versus the global minimizers of the CS-SVM in (3a)/(3b) as given by Thm. 2 (solid line). The dashed line marks ETF [20]. Purple markers distinguish the cases where the model did not enter a zero-training-error regime. Refer to SM for more details.

Concluding remarks on experiments. In our empirical studies, we observe the following: the level of convergence accuracy that can be reached in practical SGD training generally varies between architectures, data models and the loss that being optimized. For example, for CDT loss, we find that the classifier geometry converges very well to its prescribed limit, but the same is not true for the embeddings geometry for the same loss or for the classifiers geometry for the LDT. Consistently, the experiments in [20, 24] show different levels of convergence between different metrics (e.g. classifiers vs embeddings, norms vs angles) and different architectures/datasets. Similar to observations in [24] of slow convergence under high imbalance ratios, our results indicate worse convergence under "strong" parameterizations of CE as well. Despite these limitations, our results indicate that our geometry characterizations

are remarkably well-predictive of the behavior in complex deep-nets. Specifically accounting for the fact that classifiers and embeddings are by-products of training in very complex environments, we find the level of agreement of the empirically measured angles/norms to the respective (closed-form) $(\boldsymbol{\delta}, R)$-SELI geometry values rather striking.

References

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

[2] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1567–1578, 2019.

[3] Cong Fang, Hangfeng He, Qi Long, and Weijie 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] Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. *arXiv preprint arXiv:2112.15121*, 2021.

[5] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021.

[6] XY Han, Vardan Papyan, and David L Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021.

[7] Like Hui, Mikhail Belkin, and Preetum Nakkiran. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022.

[8] Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, and Weijie J Su. An unconstrained layer-peeled perspective on neural collapse. *arXiv preprint arXiv:2110.02796*, 2021.

[9] Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

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

[11] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition, 2020.

[12] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017.

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

[14] Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34:18970–18983, 2021.

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

[16] Jianfeng Lu and Stefan Steinerberger. Neural collapse with cross-entropy loss. *arXiv preprint arXiv:2012.08465*, 2020.

[17] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

[18] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

[19] Dustin G Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *arXiv preprint arXiv:2011.11619*, 2020.

[20] Vardan Papyan, XY Han, and David 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.

[21] Vardan Papyan, XY Han, and David 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.

[22] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

[23] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[24] Christos Thrampoulidis, Ganesh R Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *arXiv preprint arXiv:2208.05512*, 2022.

[25] Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. *arXiv preprint arXiv:2202.08087*, 2022.

[26] Liang Xie, Yibo Yang, Deng Cai, Dacheng Tao, and Xiaofei He. Neural collapse inspired attraction-repulsion-balanced loss for imbalanced learning. *arXiv preprint arXiv:2204.08735*, 2022.

[27] Han-Jia Ye, Hong-You Chen, De-Chuan Zhan, and Wei-Lun Chao. Identifying and compensating for feature deviation in imbalanced deep learning, 2020.

[28] Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. *arXiv preprint arXiv:2203.01238*, 2022.

[29] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34, 2021.

Appendix A. Experiment details

In this section, we provide additional details and discussions on our experiments.

A.1. UFM experimental setting

We investigate convergence of SGD trajectory for unregularized CDT/LDT loss in (2a)/(2b) to the implicit geometries in Thm. 2. We train the UFM as a two-layer network (no biases) with $n = 275$ inputs, $d = 20$ hidden units and $k = 10$ classes, trained on the basis vectors in $\mathbb{R}^n$. We measure the properties (i.e norms and angles) of the trained geometry and compare them with the statement in Thm. 2. The labels for each vector are chosen such that the dataset is $(R = 10, \rho = 1/2)$-STEP imbalanced, with $n_{\min} = 5$, and choose $\Delta = \frac{\delta_{\text{maj}}}{\delta_{\min}} = R^{\gamma}$ with $\gamma \in [-1.5, 1.5]$. Models are trained with constant learning rate for 6000 epochs. We normalize $\boldsymbol{\delta}$ so that $\mathbb{1}_k^T \boldsymbol{\delta} = k$, since we empirically observe for a fixed ratio Δ that the convergence speed depends on the magnitude of $\boldsymbol{\delta}$.

For the trained classifiers and embeddings, we compute: (1) squared ratios of majority-minority norms, (2) cosine of angles between pairs of majority-majority, minority-minority, majority-minority classifiers and mean-embeddings. For each choice of γ and loss function, we compute each metric on all the respective pairs, and compare their average to the values predicted by Thm. 2. As Δ gets far from 1, some models fail to achieve zero-training-error in this setting: We differentiate these cases with the purple (rather than green) markers.

A.2. Deep-net experimental setting

We train a ResNet-18 on imbalanced CIFAR-10 with $R = 10$. We train the model for 350 epochs with an initial learning rate of 0.1 reduced at epochs 116 and 232 by a factor of 10, with a batch size of 128. Following the same setting as in [20, 24], momentum and weight decay are set to 0.9 and 10^{-5} respectively. Similar to UFM, we normalize $\boldsymbol{\delta}$ to sum to k. At the end on the training, we compute the same metrics described in Sec. A.1.

A.3. Discussion on experimental results

Figs. 2 and 3 illustrate the empirical geometry vs the prediction of Thm. 2. The convergence to the theoretical values is more challenging for the deep-net models, particularly for large $|\gamma|$. Further, the theory gives a more accurate prediction of the mean-embeddings' geometry in case of the LDT, and of the classifiers' in case of the CDT loss. This is consistent for both UFM and deep-net models. For LDT, the prediction is well followed by UFM and ResNet empirics around the special case of $\gamma = 0.5$, with an exception of the majority classifier angles in the ResNet experiments. Also, as predicted by the theorem, for $\gamma = 0.5$ ($\Delta = \sqrt{R}$), the LDT geometry is the Simplex ETF, up to a scaling on the features: In Figs. 2 and 3 the LDT cosine plots intersect with the ETF angles, i.e. $-1/(k-1)$, thus achieving equiangularity and maximal angular separation. The classifier norm ratios also attain the value 1, which along with the equiangularity describe an ETF structure for classifiers.

Remark 4 *In our experiments with CDT and LDT, we center the embeddings before computing the geometrical quantities like norms and angles. This is consistent with centering performed for experiments with balanced data by [21, 24, 29], although the exact centering*

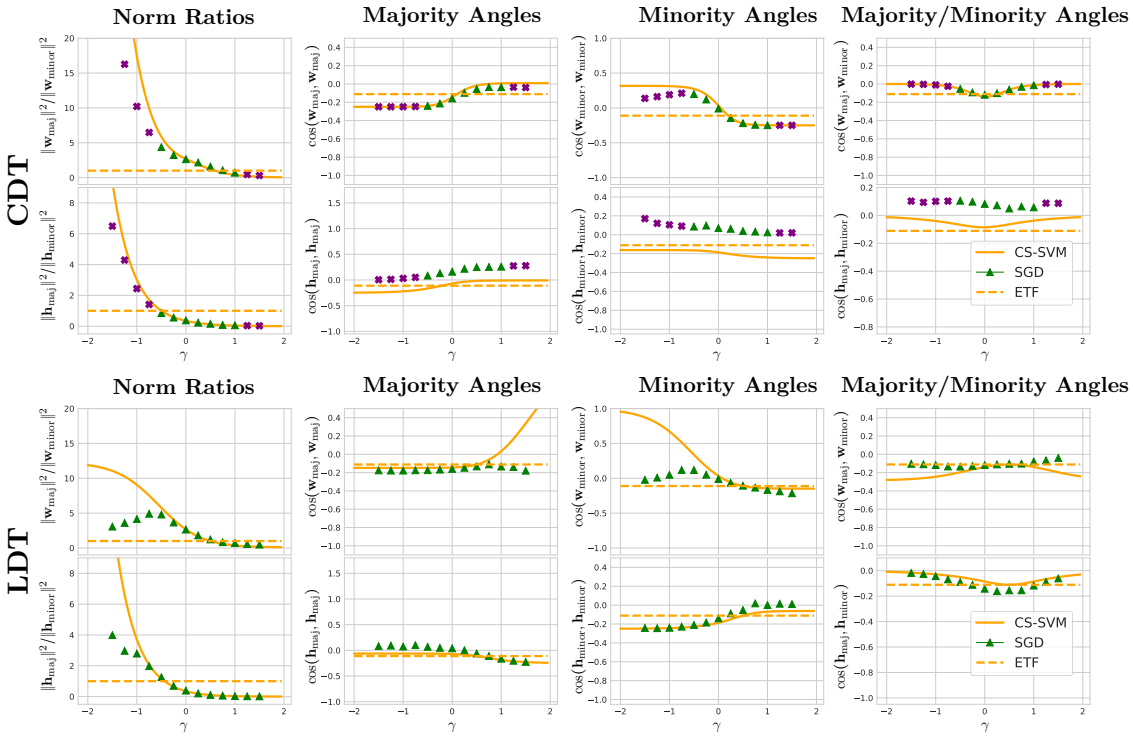

Figure 3: Same as Fig. 2, instead with ResNet-18 trained on (10, 1/2)-STEP imbalanced CIFAR-10.

vector is now as discussed in the section below. Further, even in the UFM experiments with LDT, we observe that the centered mean embeddings follow the predicted implicit geometry more closely. Thus, we employ the centering as shown in (7).

CDT centering: Classifiers that follow the geometry in Thm. 2, $\mathbf{w}_c^*,\ c \in [k]$ are centered around zero after some re-weighting, i.e. $\sum_{c\in[k]} \mathbf{w}_c^*/\delta_c = 0$. The embeddings $\mathbf{h}_i, i \in [n]$ are also not centered around zero in general. Instead, it holds that

$$\sum_{i\in[n]} \frac{1}{n_{y_i}} \mathbf{h}_i^* = 0. \tag{6}$$

Note that this reduces to $\sum_{i\in[n]} \mathbf{h}_i^*$ for balanced data, and remains unchanged for any choice of the hyper-parameters $\boldsymbol{\delta}$. Eqn. (6) is also equivalent to $\sum_{c\in[k]} \boldsymbol{\mu}_c^* = 0$, with $\boldsymbol{\mu}_c^*,\ c \in [k]$ the mean embeddings of each class.

LDT centering: The optimal classifiers and features $(\mathbf{W}^*, \mathbf{M}^*\mathbf{D})$ follow the $(\mathbb{1}_k, \tilde{R})$-SELI structure. Thus (see [24, Sec. B.1.4]), the classifiers $\mathbf{w}_c^*, c \in [k]$ are centered around zero. However the embeddings are centered around zero after a reweighting that depends both on δ_c and $n_c,\ c \in [k]$. Specifically, $\sum_{c\in[k]} \delta_c \boldsymbol{\mu}_c^* = 0$, or equivalently,

$$\sum_{i\in[n]} \frac{\delta_{y_i}}{n_{y_i}} \mathbf{h}_i^* = 0. \tag{7}$$

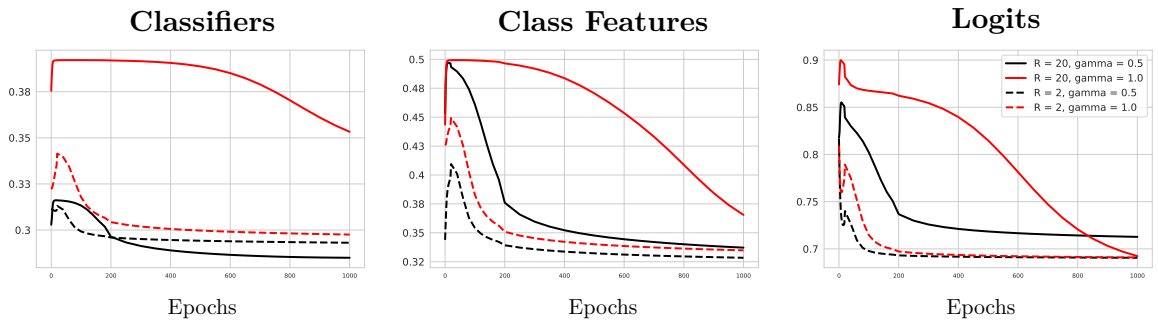

Figure 4: Convergence of UFM parameters to the optimal solution in Thm. 2, trained by SGD on CDT loss. Larger imbalance ratio R and γ can lead to slower convergence to the expected structure.

A.4. Optimization Concerns

In our experiments, we observe that large R and Δ affect the convergence negatively. This observation is also made in [24]. We compare the convergence speed for different values of R and γ in Fig. 4. To measure the distance of the SGD solution to the predicted implicit geometry at each epoch, we compute $\| \frac{\mathbf{W}^T\mathbf{W}}{\|\mathbf{W}^T\mathbf{W}\|} - \frac{\mathbf{W}^{*T}\mathbf{W}^*}{\|\mathbf{W}^{*T}\mathbf{W}^*\|} \|_F$ for the classifier weights, $\| \frac{\mathbf{M}^T\mathbf{M}}{\|\mathbf{M}^T\mathbf{M}\|} - \frac{\mathbf{M}^{*T}\mathbf{M}^*}{\|\mathbf{M}^{*T}\mathbf{M}^*\|} \|_F$ for centered mean embeddings and $\| \frac{\mathbf{W}^T\mathbf{M}}{\|\mathbf{W}^T\mathbf{M}\|} - \frac{\mathbf{W}^{*T}\mathbf{M}^*}{\|\mathbf{W}^{*T}\mathbf{M}^*\|} \|_F$ for logits, where $(\mathbf{W}^*, \mathbf{M}^*)$ are as described by Thm. 2.

In addition to the worse convergence, as suggested by Fig. 3, larger $|\gamma|$ values make it more challenging to reach models with zero training error by SGD. In particular, for a ResNet-18 model trained with CDT on imbalanced CIFAR-10, Fig. 5 illustrates how larger γ values significantly reduce the training accuracy of minority classes, further emphasizing the difficulties with optimization in this setting.

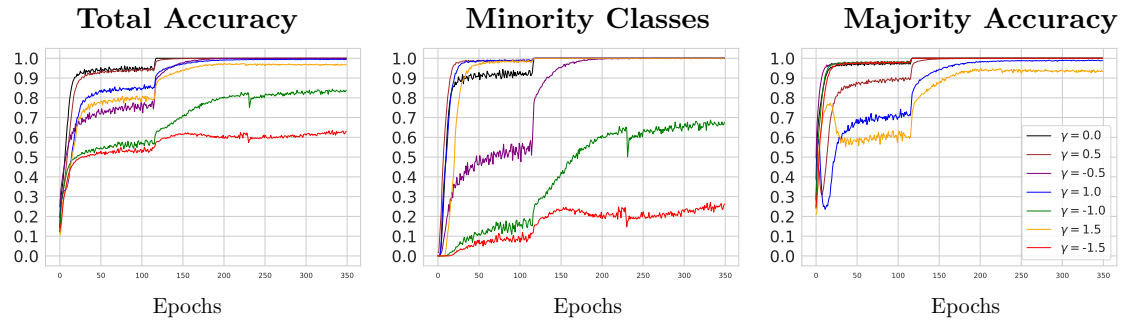

Figure 5: Balanced training accuracy during training of ResNet-18 models on ($R = 10$, ρ=1/2)-STEP imbalance CIFAR10 dataset with CDT loss using a range of γ values.