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

author names withheld

Under Review for OPT 2022

Abstract

It has been empirically observed that training large models with weighted cross-entropy (CE) beyond the zero-training-error regime is not a satisfactory remedy for label-imbalanced data. Instead, researchers have proposed the vector-scaling (VS) loss, as a parameterization of the CE loss that is 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*. Specifically for linear(ized) models, this theory allows to explain why weighted CE fails and how the VS-loss biases the optimization path towards solutions that favor 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 learnt classifiers and embeddings. Our main result characterizes the global minimizers of a non-convex cost-sensitive SVM classifier for the so-called unconstrained features model, which serves as an abstraction of deep models. We also study empirically the convergence of SGD to this global minimizer observing slow-downs with increasing imbalance ratios and scalings of the loss hyperparameters.

1. Introduction

It has been shown that in the modern overparameterized regime, where training continues beyond zero-training error, traditional techniques, such as oversampling minorities or minimizing a weighted cross-entropy (CE) loss can be ineffective in mitigating label-imbalances [1, 21]. In a growing literature, several alternatives have been proposed to guarantee equitable performance across majorities and minorities, e.g. [2, 11–15, 18, 26]. Among these, the vector-scaling (VS)-loss [14, 26] introduces multiplicative hyper-parameters on the logits of the CE loss. The idea behind this parameterization is rooted in the theory of *implicit bias*, which seeks characterizing the bias introduced by gradient-based algorithms during training [9, 17, 22]. 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]. In particular, it is unclear how to gain from them 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, 23–25, 27, 28], 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 [26], which is a special case of the VS-loss [14]. Second, the label-dependent temperature (LDT)-loss, which we introduce here 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], which serves as a proxy for large overparameterized models and has been used recently to study the implicit geometry of the CE loss [8, 23, 27, 28]. Relying on implicit-bias results, we relax the question of implicit geometry of the solutions found by SGD, 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 ratio, the imbalance ratio, and the hyperparameters of the losses. We also show numerically that SGD training on the UFM converges to the uncovered geometries. However, we observe that convergence slows down for increasing imbalance ratios and increasing values of the hyperparameters. We believe this motivates further theoretical and algorithmic investigations towards faster training. Finally, as preliminary evidence of the utility of our geometry characterizations for the UFM, we present preliminary results on deep-learning architectures and complex imbalanced datasets.

2. Background

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

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

Here $\mathbf{x}_i, i \in [n]$ are n examples, $\mathbf{h}_{\boldsymbol{\theta}}(.)$ is the feature map parameterized by trainable parameters $\boldsymbol{\theta}$ (e.g. weights of the hidden layers of a neural network), $y_i \in [k], i \in [n]$ are the labels, and $\mathbf{w}_c, c \in [k]$ are the classifier vectors (e.g. 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 $\mathbf{h}_{\boldsymbol{\theta}}$, $y_i \in \{\pm 1\}$, and $\mathbf{h}_i := \mathbf{h}_{\boldsymbol{\theta}}(i), i \in [n]$, GD with sufficiently small learning rate on the binary VS-loss $\mathcal{L}_{\mathrm{VS,binary}}(\mathbf{w}) := \sum_{i\in[n]} \log(1 +$*

[1] In the literature this is known as *neural collapse* (NC) [20]. However, [23] showed that training CE loss on imbalanced data learns a geometry different from that in [20] for balanced data. To draw distinction, 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".

$e^{-\delta_{y_i} y_i \mathbf{w}^T \mathbf{h}_i + \iota_{y_i}}$), *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 $y_i \delta_{y_i} \mathbf{w}^T \mathbf{h}_i \geq 1, i \in [n]$.

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 $y = +1$ is minority, favors the minority by assigning larger margin to it [14]. Note, the additive hyperparameters ι_c do *not* have any 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 learnt-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. For instance, 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 propose studying their implicit geometry, i.e. the geometry of classifiers and embeddings learnt (asymptotically in the number of training steps) by GD. For this, we adopt the *unconstrained feature model* (UFM) [3, 19]. To describe the model, 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. Formally, defining $n_0 = 0$, examples $i = \sum_{c'=0}^{c-1} n_{c'} + 1, \ldots, \sum_{c'=0}^{c} n_{c'}$ have labels $y_i = c$. 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 study the implicit geometry of CE loss on balanced (e.g. [3, 19, 27, 28]), and more recently, on imbalanced data [23]. Instead, we study different parameterizations of the CE loss.

CDT and LDT losses on the UFM. Consider training on the UFM with the following two parameterization of the CE loss:

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

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

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, i.e. $\boldsymbol{\delta} > 0$. The CDT loss in (2a) was previously introduced by [14, 26] (note is a special case of (1) when ignoring the additive ι_c). Here, we also introduce the LDT loss in (2b) as an alternative parameterization. Observe the following subtle distinction: CDT associates the hyperparameters $\boldsymbol{\delta}$ with the class label of the classifier vectors $\mathbf{w}_c$, while LDT associates the same hyperparameters $\boldsymbol{\delta}$ with the label of the example 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 [23]) 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 unconstrained CS-SVM. Our main result characterizes their global solutions.

4. Global structure of the unconstrained CS-SVM

In this section, we characterize the global minimizers of the non-convex programs in (3a) and (3b). Our result is be in terms of the following matrix.

Definition 1 ($\boldsymbol{\delta}$-SEL matrix) *The* $\boldsymbol{\delta}$-parameterized simplex-encoding label ($\boldsymbol{\delta}$-SEL) matrix $\hat{\mathbf{Z}}_{\boldsymbol{\delta}} \in \mathbb{R}^{k \times n}$ *is such that*

$$\forall c \in [k], i \in [n]: \quad \hat{\mathbf{Z}}_{\boldsymbol{\delta}}[c,i] = \begin{cases} \delta_{y_i}^{-1}\left(1 - \delta_{y_i}^{-2}/\sum_{c' \in [k]} \delta_{c'}^{-2}\right) & , c = y_i \\ -\delta_c^{-1}\left(\delta_{y_i}^{-2}/\sum_{c' \in [k]} \delta_{c'}^{-2}\right) & , c \neq y_i \end{cases}. \tag{4}$$

Let $\hat{\mathbf{Z}}_{\boldsymbol{\delta}}^T = \mathbf{U}_{\boldsymbol{\delta}} \boldsymbol{\Lambda}_{\boldsymbol{\delta}} \mathbf{V}_{\boldsymbol{\delta}}^T$ *be the compact SVD of* $\hat{\mathbf{Z}}_{\boldsymbol{\delta}}^T$*. Specifically,* $\boldsymbol{\Lambda}_{\boldsymbol{\delta}} \in \mathbb{R}^{(k-1)\times(k-1)}$ *is a positive diagonal matrix and* $\mathbf{U}_{\boldsymbol{\delta}} \in \mathbb{R}^{n \times (k-1)}$*,* $\mathbf{V}_{\boldsymbol{\delta}} \in \mathbb{R}^{k \times (k-1)}$ *have orthonormal columns. For the special case* $\boldsymbol{\delta} = \mathbb{1}_k$*, we denote* $\hat{\mathbf{Z}} := \hat{\mathbf{Z}}_{\mathbb{1}}$ *and refer to this simply as the SEL matrix.*

Rows of $\hat{\mathbf{Z}}_{\boldsymbol{\delta}}$ correspond to classes and columns to examples. The matrix is explicitly parameterized by $\boldsymbol{\delta}$, but it also depends (implicitly) on the labels $y_i, i \in [n]$ of the examples. Also note that the matrix $\hat{\mathbf{Z}}$ is a centering of the one-hot encoding matrix $\mathbf{Y}$, i.e. $\hat{\mathbf{Z}} = \mathbf{Y} - \mathbb{1}_k \mathbb{1}_k^T/k$. Previous work [14] shows that $\hat{\mathbf{Z}}$ characterizes the implicit geometry of the CE loss. We show how its extension $\hat{\mathbf{Z}}_{\boldsymbol{\delta}}$ plays a crucial role for CDT/LDT losses. We gather useful properties about the eigenstructure of $\hat{\mathbf{Z}}_{\boldsymbol{\delta}}$ in Sec. A. Here, we note that $\hat{\mathbf{Z}}_{\boldsymbol{\delta}}^T \operatorname{diag}(\boldsymbol{\delta})^{-1} \mathbb{1}_k = 0$. Thus, $\operatorname{rank}(\hat{\mathbf{Z}}_{\boldsymbol{\delta}}) = k - 1$.

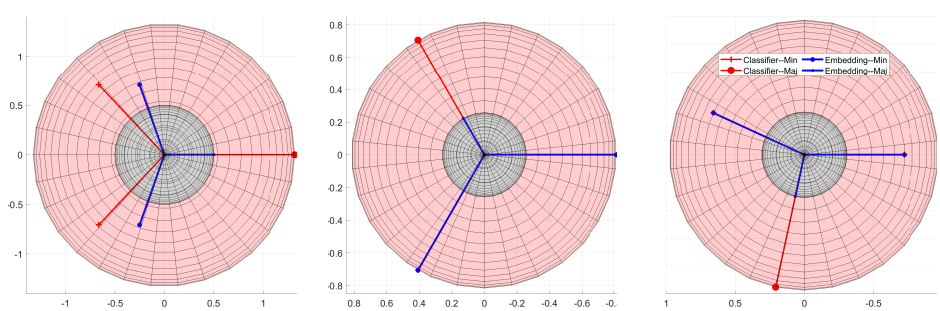
For simplicity, we restrict attention to a STEP-imbalanced setting. In this case, it is reasonable to assume (and we do so) that $\boldsymbol{\delta}$ shares this STEP structure.

Definition 2 ((R,ρ)-STEP imbalance and STEP logit adjustment) *All minority classes have the same sample size* n_{min} *and all majority classes have same sample size* Rn_{min} *for imbalance ratio* $R \geq 1$*. There are* $(1-\rho)k$ *majority and* ρk *minority classes for* $\rho \in (0,1)$*. Without loss of generality, we assume classes* $\{1, \ldots, (1-\rho)k\}$ *are majorities. Finally, the hyperparameters' vector* $\boldsymbol{\delta}$ *has also a step structure:* $\delta_c = \delta_{\min} > 0$ *for all minorities* $c \in \{1, \ldots, (1-\rho)k\}$*, and* $\delta_c = \delta_{\text{maj}} > 0$ *for all majorities* $c \in \{(1-\rho)k+1, \ldots, k\}$*.*

Theorem 3 *Suppose* $d \geq k-1$ *and* (R,ρ)*-STEP imbalance setting with STEP logit adjustments. Let* $(\hat{\mathbf{W}}, \hat{\mathbf{H}})$ *be any minimizers of either* (3a) *and* (3b)*. The following statements are true*

(i) [Neural collapse] *$\forall i, j \in [n]$ such that $y_i = y_j$ it holds that $\hat{\mathbf{h}}_i = \hat{\mathbf{h}}_j$.*
(ii) [Logits] *For (3a), $\hat{\mathbf{W}}^T\hat{\mathbf{H}} = \hat{\mathbf{Z}}_{\boldsymbol{\delta}}$. For (3b), $\hat{\mathbf{W}}^T\hat{\mathbf{H}} = \hat{\mathbf{Z}}\,\mathrm{diag}(\mathbf{Y}^T\boldsymbol{\delta})$, where recall $\mathbf{Y} \in \mathbb{R}^{k\times n}$ is the one-hot encoding matrix.*
(iii) [Gram matrices of (3a)] *For (3a), $\hat{\mathbf{H}}^T\hat{\mathbf{H}} = \mathbf{U}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}_{\boldsymbol{\delta}}\mathbf{U}_{\boldsymbol{\delta}}^T$ and $\hat{\mathbf{W}}^T\hat{\mathbf{W}} = \mathbf{V}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}_{\boldsymbol{\delta}}\mathbf{V}_{\boldsymbol{\delta}}^T$.*
(iv) [Gram matrices of (3b)] *Suppose $\delta_c, c \in [k]$ are rationals. Let $\alpha > 0$ such that $n_c^{\mathit{eff}} := \alpha n_c/\delta_c^2$ are integers. Construct $\hat{\mathbf{Z}}^{\mathit{eff}} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^T$ the SEL-matrix for a dataset with n_c^{eff} number of examples per class. Then, for (3b), $\hat{\mathbf{W}}^T\hat{\mathbf{W}} = \alpha\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$ and $\hat{\mathbf{H}}^T\hat{\mathbf{H}}$ is also determined explicitly by the singular factors $\mathbf{U}$ and $\boldsymbol{\Lambda}$; see App. B.2 for details.*

The theorem characterizes the entire geometry. Statements (iii,iv) characterize all angles and norms (up to global scaling) of both classifiers and embeddings. When combined with statement (ii) this determines the relative margin between all classes. The characterization is in terms of SVD factors of appropriate SEL-type encoding matrices. In the Appendix, we show that it is possible to compute those SVD factors explicitly. This is important because it leads to closed-form expressions for the norms, angles and margins as a function of $R, k, \boldsymbol{\delta}$. Thus, we can study how these parameters affect the geometry (see Fig. 1 for visualization). For $\boldsymbol{\delta} = 1$, we can show that the formulas of Thm 3 recover the SELI geometry of [23] for the CE loss as a special case. Further assuming R=1, recovers the ETF geometry of [20]. For general 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. A notable special case for LDT loss arises when setting $\delta_c = \sqrt{n_c}$. Then, $n_c^{\mathrm{eff}} = 1$ meaning that the geometry of the LDT loss is exactly the same (for all values of R) as the perfectly symmetric ETF geometry modulo a scaling on the norms of the embeddings (minority norms are larger than majorities by factor of $\sqrt{R}$).



SELI LDT CDT

Figure 1: Example geometries of SELI, LDT and CDT for the respective unconstrained features SVM minimizers. Here $k = 3, R = 10, n_{\min} = 1, \delta_c = \sqrt{n_c}$ with 2 minority and 1 majority classes.

5. Numerical results on GD path

In this section, we investigate convergence of GD trajectory for unregularized CDT/LDT-loss in 2a and 2b to the implicit geometries in Thm. 3. We measure the properties (i.e norms and angles) of the trained geometry and compare them with the statement in Thm. 3 (See App.

C). For the simulations, we model the UFM as a two-layer neural network (without bias) with n inputs, $d = 20$ hidden units and $k = 10$ classes, trained on the set of unit basis vectors in $\mathbb{R}^n$. The labels for each vector are chosen such that the dataset is $(R = 10, \rho = 1/2)$-STEP imbalanced, with $n_{\min} = 5$. We further use STEP logit adjustment, and choose $\Delta := \frac{\delta_{\text{maj}}}{\delta_{\min}} = R^{\gamma}$ with γ ranging in $[-1.5, 1.5]$. Models are trained with constant learning rate for 6000 epochs. We normalize $\boldsymbol{\delta}$ to sum to k, since we empirically observe that for a fixed Δ, the convergence speed depends crucially on the magnitude of $\boldsymbol{\delta}$.

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 in Fig. 2. We note that when trained with GD on CDT/LDT losses, the classifiers and mean embeddings show a general convergence towards respective CDT/LDT implicit geometries described by Thm. 3. Note that the convergence is relatively better for the classifier geometry compared to the feature geometry. This behavior is also been observed for models trained with CE [23] (note that this CE is a special case of our results for $\gamma = 0$). As $\delta_{\text{maj}}/\delta_{min}$ gets far from 1, it becomes more challenging for the model to reach the optimal parameters. Furthermore, some models fail to achieve zero-training-error in this setting: We differentiate these cases with the purple (rather than green) markers. For more details and results, including experiments on deep-net architectures, see Sec. D.

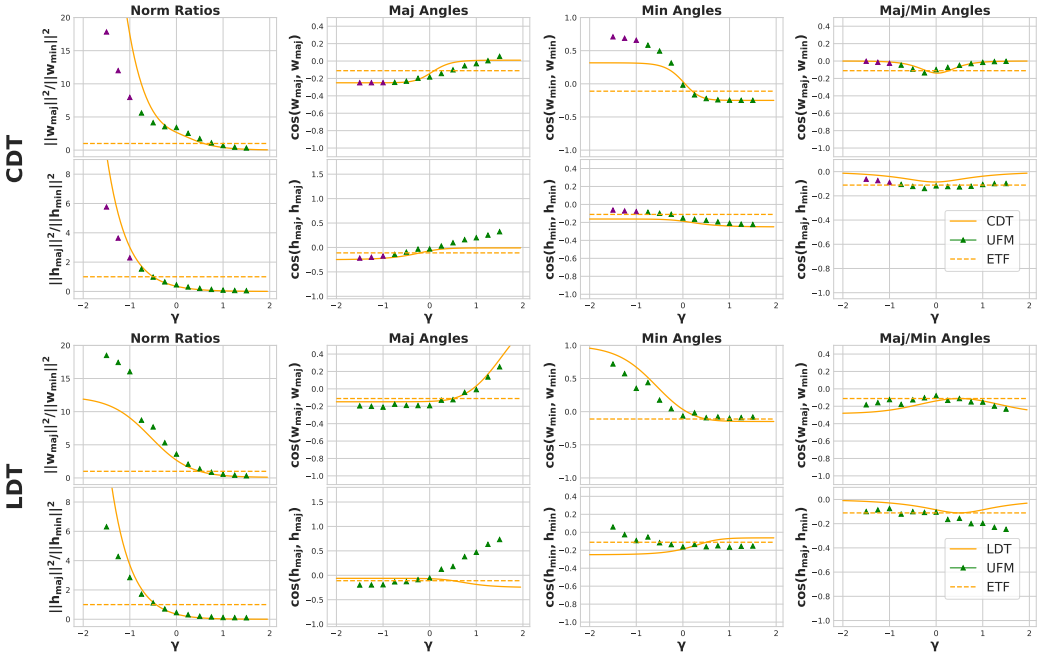


Figure 2: Comparing the solution found by GD (markers) on the CDT/LDT losses in 2a, 2b to the global minimizers of the unconstrained CS-SVM in 3a, 3b as given by Thm. 3 (solid line). The dashed line marks the perfectly symmetric ETF-geometry of balanced data [20]. Purple markers distinguish the cases where the model did not enter a zero-training-error regime. Refer to Sec. C and D for more details.

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] 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.

[22] 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.

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

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

[25] 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.

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

[27] 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.

[28] 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.

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

Appendix A. Eigen-structure of the δ-SEL matrix

In this section, we assume the data is (R,ρ)-STEP imbalanced, and we utilize STEP logit adjustment. With these assumptions, we compute the eigen-structure of the $\boldsymbol{\delta}$-SEL matrix. For simplicity, we assume $n_{\min} = 1$, $\delta_{\min} = 1$ and $\delta_{\mathrm{maj}} = \Delta$[2]. Also, for $m \in [k]$, define $\mathbb{P}_m \in \mathbb{R}^{m\times(m-1)}$ as 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 $\mathbb{P}_m^T\mathbb{P}_m = \mathbb{I}_{m-1}$, and $\mathbf{S}_m(\sigma) := \mathbb{I}_m - \sigma\mathbb{1}_m\mathbb{1}_m^T \in \mathbb{R}^m$.

Lemma 4 (δ-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 $n = (\rho + R\overline{\rho})k$. Let $\hat{\mathbf{Z}}_{\boldsymbol{\delta}} = \mathbf{V}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}_{\boldsymbol{\delta}}\mathbf{U}_{\boldsymbol{\delta}}^T$ be the SVD decomposition of $\hat{\mathbf{Z}}_{\boldsymbol{\delta}}$. Also, let $\mathbf{V}_{\boldsymbol{\delta}} = [\mathbf{V}_{maj}, \mathbf{v}, \mathbf{V}_{min}]$ and $\mathbf{U}_{\boldsymbol{\delta}} = [\mathbf{U}_{maj}, \mathbf{u}, \mathbf{U}_{min}]$. Then, the SVD factors are given as follows:*

$$\boldsymbol{\Lambda}_{\boldsymbol{\delta}} = \mathrm{diag}\left(\begin{bmatrix} \frac{\sqrt{R}}{\Delta}\mathbb{1}_{(\overline{\rho}k-1)}^T & \sqrt{\frac{\overline{\rho}+R\rho}{\overline{\rho}+\rho\Delta^2}} & \mathbb{1}_{(\rho k-1)}^T \end{bmatrix}\right), \tag{5}$$

$$\mathbf{V}_{maj} = \begin{bmatrix} \mathbb{P}_{\overline{\rho}k} \\ 0_{(\rho k)\times(\overline{\rho}k-1)} \end{bmatrix} \quad \mathbf{v} = \frac{1}{\sqrt{k(\overline{\rho}+\rho\Delta^2)}}\begin{bmatrix} -\Delta\sqrt{\frac{\rho}{\overline{\rho}}}\mathbb{1}_{\overline{\rho}k} \\ \sqrt{\frac{\overline{\rho}}{\rho}}\mathbb{1}_{\rho k} \end{bmatrix} \quad \mathbf{V}_{min} = \begin{bmatrix} 0_{(\overline{\rho}k)\times(\rho k-1)} \\ \mathbb{P}_{\rho k} \end{bmatrix}, \tag{6}$$

$$\mathbf{U}_{maj} = \begin{bmatrix} \frac{1}{\sqrt{R}}\mathbb{P}_{\overline{\rho}k}\otimes\mathbb{1}_R \\ 0_{(\rho k)\times(\overline{\rho}k-1)} \end{bmatrix} \quad \mathbf{u} = \frac{1}{\sqrt{k(\overline{\rho}+R\rho)}}\begin{bmatrix} -\sqrt{\frac{\rho}{\overline{\rho}}}\mathbb{1}_{R\overline{\rho}k} \\ \sqrt{\frac{\overline{\rho}}{\rho}}\mathbb{1}_{\rho k} \end{bmatrix} \quad \mathbf{U}_{min} = \begin{bmatrix} 0_{(R\overline{\rho}k)\times(\rho k-1)} \\ \mathbb{P}_{\rho k} \end{bmatrix}. \tag{7}$$

Proof To prove the lemma, we only need to verify the correctness of the formulas by direct calculations. By recalling that $\mathbb{P}_m^T\mathbb{P}_m = \mathbb{I}_{m-1}$ and $\mathbb{P}_m^T\mathbb{1}_m = 0$ for $m \in \{\rho k, \overline{\rho}k\}$, we can confirm that $\mathbf{V}_{\boldsymbol{\delta}}^T\mathbf{V}_{\boldsymbol{\delta}} = \mathbb{I}_{k-1}$ and $\mathbf{U}_{\boldsymbol{\delta}}^T\mathbf{U}_{\boldsymbol{\delta}} = \mathbb{I}_{k-1}$. To show $\mathbf{V}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}_{\boldsymbol{\delta}}\mathbf{U}_{\boldsymbol{\delta}}^T = \hat{\mathbf{Z}}_{\boldsymbol{\delta}}$, consider the block-form of $\hat{\mathbf{Z}}_{\boldsymbol{\delta}}$ as follows

$$\hat{\mathbf{Z}}_{\boldsymbol{\delta}} = \begin{bmatrix} \hat{\mathbf{Z}}_{11} & \hat{\mathbf{Z}}_{12} \\ \hat{\mathbf{Z}}_{21} & \hat{\mathbf{Z}}_{22} \end{bmatrix} = \begin{bmatrix} \Delta^{-1}\mathbf{S}_{\overline{\rho}k}\left(\frac{1}{k(\overline{\rho}+\rho\Delta^2)}\right)\otimes\mathbb{1}_R^T & -\frac{\Delta}{k(\overline{\rho}+\rho\Delta^2)}\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\rho k}^T \\ -\frac{1}{k(\overline{\rho}+\rho\Delta^2)}\mathbb{1}_{\rho k}\mathbb{1}_{R\overline{\rho}k}^T & \mathbf{S}_{\rho k}\left(\frac{\Delta^2}{k(\overline{\rho}+\rho\Delta^2)}\right) \end{bmatrix}. \tag{8}$$

[2]Without these assumptions, the structure of the matrix would only change up to a scaling. In particular, singular values will be scaled by $\frac{\sqrt{n_{\min}}}{\delta_{\min}}$, $\mathbf{V}$ remains unchanged and $\mathbf{U}_{\mathrm{maj}}$ and $\mathbf{U}_{\min}$ blocks are multiplied by $1/\sqrt{n_{\min}}$.

Now, we can verify the equation by direct calculations:

$$
\begin{aligned}
\mathbf{V}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}_{\boldsymbol{\delta}}\mathbf{U}_{\boldsymbol{\delta}}^T &= \frac{\sqrt{R}}{\Delta}\mathbf{V}_{\text{maj}}\mathbf{U}_{\text{maj}}^T + \sqrt{\frac{\overline{\rho}+R\rho}{\overline{\rho}+\rho\Delta^2}}\mathbf{v}\mathbf{u}^T + \mathbf{V}_{\text{min}}\mathbf{U}_{\text{min}}^T \\
&= \begin{bmatrix} \Delta^{-1}\mathbb{P}_{\overline{\rho}k}\mathbb{P}_{\overline{\rho}k}^T \otimes \mathbb{1}_R^T & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{k(\overline{\rho}+\rho\Delta^2)}\begin{bmatrix} \Delta\frac{\rho}{\overline{\rho}}\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{R\overline{\rho}k}^T & -\Delta\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\rho k}^T \\ -\mathbb{1}_{\rho k}\mathbb{1}_{R\overline{\rho}k}^T & \frac{\overline{\rho}}{\rho}\mathbb{1}_{\rho k}\mathbb{1}_{\rho k}^T \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbb{P}_{\rho k}\mathbb{P}_{\rho k}^T \end{bmatrix} \\
&= \begin{bmatrix} \Delta^{-1}\left(\mathbb{I}_{\overline{\rho}k} - \frac{1}{k(\overline{\rho}+\rho\Delta^2)}\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\overline{\rho}k}^T\right) \otimes \mathbb{1}_R^T & -\frac{\Delta}{k(\overline{\rho}+\rho\Delta^2)}\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\rho k}^T \\ -\frac{1}{k(\overline{\rho}+\rho\Delta^2)}\mathbb{1}_{\rho k}\mathbb{1}_{R\overline{\rho}k}^T & \mathbb{I}_{\rho k} - \frac{\Delta^2}{k(\overline{\rho}+\rho\Delta^2)}\mathbb{1}_{\rho k}\mathbb{1}_{\rho k}^T \end{bmatrix} \\
&= \hat{\mathbf{Z}}_{\boldsymbol{\delta}}.
\end{aligned}
$$

■

With the eigen-structure of $\hat{\mathbf{Z}}_{\boldsymbol{\delta}}$ at hand, we prove a useful property of the singular space in Lemma 5. We will use this property later in Sec. B to characterize the solutions of the CDT-SVM (3a).

Lemma 5 *Recall the setting of Lemma 4 and the SVD* $\hat{\mathbf{Z}}_{\boldsymbol{\delta}} = \mathbf{V}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}_{\boldsymbol{\delta}}\mathbf{U}_{\boldsymbol{\delta}}^T$. *The matrix* $\hat{\mathbf{B}} = \mathbf{U}_{\boldsymbol{\delta}}\mathbf{V}_{\boldsymbol{\delta}}^T$ *satisfies the following element-wise* strict *inequalities:* $\hat{\mathbf{B}} \odot \hat{\mathbf{Z}}_{\boldsymbol{\delta}}^T > 0$.

Proof We can compute $\hat{\mathbf{B}} := \begin{bmatrix} \hat{\mathbf{B}}_{11} & \hat{\mathbf{B}}_{12} \\ \hat{\mathbf{B}}_{21} & \hat{\mathbf{B}}_{22} \end{bmatrix}$ by plugging in the explicit SVD expressions in Lemma 4.

$$
\mathbf{U}_{\boldsymbol{\delta}}\mathbf{V}_{\boldsymbol{\delta}}^T = \begin{bmatrix} \frac{1}{\sqrt{R}}\mathbb{P}_{\overline{\rho}k}\mathbb{P}_{\overline{\rho}k}^T \otimes \mathbb{1}_R & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{k\sqrt{\overline{\rho}+R\rho}\sqrt{\overline{\rho}+\rho\Delta^2}}\begin{bmatrix} \Delta\frac{\rho}{\overline{\rho}}\mathbb{1}_{R\overline{\rho}k}\mathbb{1}_{\overline{\rho}k}^T & -\mathbb{1}_{R\overline{\rho}k}\mathbb{1}_{\rho k}^T \\ -\Delta\mathbb{1}_{\rho k}\mathbb{1}_{\overline{\rho}k}^T & \frac{\overline{\rho}}{\rho}\mathbb{1}_{\rho k}\mathbb{1}_{\rho k}^T \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbb{P}_{\rho k}\mathbb{P}_{\rho k}^T \end{bmatrix}.
$$

Simplifying the expressions, we have

$$
\begin{aligned}
\hat{\mathbf{B}}_{11} &= \frac{1}{\sqrt{R}}\left(\mathbb{I}_{\overline{\rho}k} - \frac{1}{\overline{\rho}k}\left(1 - \Delta\sqrt{\frac{R\rho}{(R+\overline{\rho}/\rho)(\overline{\rho}+\rho\Delta^2)}}\right)\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\overline{\rho}k}\right) \otimes \mathbb{1}_R\,, \\
\hat{\mathbf{B}}_{12} &= -\frac{1}{k\sqrt{\overline{\rho}+R\rho}\sqrt{\rho+\overline{\rho}\Delta^2}}\mathbb{1}_{R\overline{\rho}k}\mathbb{1}_{\rho k}^T\,, \\
\hat{\mathbf{B}}_{21} &= -\frac{\Delta}{k\sqrt{\overline{\rho}+R\rho}\sqrt{\rho+\overline{\rho}\Delta^2}}\mathbb{1}_{\rho k}\mathbb{1}_{\overline{\rho}k}^T\,, \\
\hat{\mathbf{B}}_{22} &= \mathbb{I}_{\rho k} - \frac{1}{\rho k}\left(1 - \sqrt{\frac{\overline{\rho}}{(1+R(\rho/\overline{\rho}))(\overline{\rho}+\rho\Delta^2)}}\right)\mathbb{1}_{\rho k}\mathbb{1}_{\rho k}^T.
\end{aligned} \tag{9}
$$

Recall the block-form of $\hat{\mathbf{Z}}_{\boldsymbol{\delta}}$ in (8). The signs of the off-diagonal blocks of both $\hat{\mathbf{Z}}_{\boldsymbol{\delta}}$ and $\hat{\mathbf{B}}$ are negative. To inspect the sign agreement of the on-diagonal blocks, it is enough to see the following inequalities are always strictly satisfied

$$
1 > 1 - \Delta\sqrt{\frac{R\rho}{(R+\overline{\rho}/\rho)(\overline{\rho}+\rho\Delta^2)}} > 0 \qquad \text{and} \qquad 1 > 1 - \sqrt{\frac{\overline{\rho}}{(1+R(\rho/\overline{\rho}))(\overline{\rho}+\rho\Delta^2)}} > 0.
$$

■

Appendix B. Proof of Theorem 3

B.1. CDT

First, by setting $\mathbf{X} = \begin{bmatrix} \mathbf{W}^T \\ \mathbf{H}^T \end{bmatrix} \begin{bmatrix} \mathbf{W} & \mathbf{H} \end{bmatrix} \in \mathbb{R}^{(k+n)\times(k+n)}$, we define the relaxation of the unconstrained CS-SVM problem (3a) as follows,

$$\begin{aligned} \mathrm{q}_* = \min_{\mathbf{X}\succeq 0} \quad & \frac{1}{2}\operatorname{tr}\left(\mathbf{X}\right) \\ \text{sub. to} \quad & \delta_{y_i}\mathbf{X}[y_i, k+i] - \delta_c \mathbf{X}[c, k+i] \geq 1, \ \ \forall i \in [n], c \neq y_i. \end{aligned} \tag{10}$$

Our key insight in the analysis of (10) is writing its dual in a way that involves explicitly the $\boldsymbol{\delta}$-SEL matrix. Specifically, the dual of (10) can be formulated as in follows:

$$\mathrm{d}_* = \max_{\mathbf{B}\in\mathbb{R}^{n\times k}} \quad \operatorname{tr}(\hat{\mathbf{Z}}_{\boldsymbol{\delta}}\mathbf{B}) \tag{11}$$

$$\begin{aligned} \text{sub. to} \quad & \|\mathbf{B}\|_2 \geq 1 \\ & \mathbf{B}\mathbf{D}^{-1}\mathbb{1}_k = 0 \\ & \mathbf{B} \odot \hat{\mathbf{Z}}_{\boldsymbol{\delta}}^T \geq 0\,, \end{aligned} \tag{12}$$

where $\mathbf{B}$ contains the dual variables and $\mathbf{D} = \operatorname{diag}(\boldsymbol{\delta}) \in \mathbb{R}^{k\times k}$. It is easy to see that strong duality holds for the convex problem (10) by satisfying Slater's condition. Thus, using the optimal solution of the (11), we can characterize the optimal set of (10).

To solve (11), we first relax the problem by ignoring the constraint (12):

$$\max_{\|\mathbf{B}\|_2\leq 1} \quad \operatorname{tr}(\hat{\mathbf{Z}}_{\boldsymbol{\delta}}\mathbf{B}) \qquad \text{sub. to} \ \ \mathbf{B}\mathbf{D}^{-1}\mathbb{1}_k = 0. \tag{13}$$

The optimal solution of this problem is

$$\hat{\mathbf{B}} = \mathbf{U}_\delta \mathbf{V}_\delta^T\,, \tag{14}$$

where $\mathbf{U}_\delta$ and $\mathbf{V}_\delta$ are the SVD factors of $\hat{\mathbf{Z}}_{\boldsymbol{\delta}}$. The rest of the proof follows similar to [23, Section C.1]: It can be shown that $\hat{\mathbf{B}}$ is the unique solution of (13), and according to Lemma 5, it also satisfies (12). Thus, the relaxation in (13) is tight and $\hat{\mathbf{B}}$ is also the unique maximizer of (11). Next, by strong duality, we can construct the solution to the primal relaxation (10). Namely, together with complementary slackness conditions and strict feasibility in (12), the optimal minmizer of the primal is

$$\hat{\mathbf{X}} = \begin{bmatrix} \mathbf{V}_\delta \\ \mathbf{U}_\delta \end{bmatrix} \boldsymbol{\Lambda}_\delta \begin{bmatrix} \mathbf{V}_\delta^T & \mathbf{U}_\delta^T \end{bmatrix}. \tag{15}$$

The proof completes by showing that any solution of the original problem in (3a) is of the following form,

$$\hat{\mathbf{W}} = \mathbf{R}^T \boldsymbol{\Lambda}_\delta^{1/2} \mathbf{V}_\delta^T \qquad \text{and} \qquad \hat{\mathbf{H}} = \mathbf{R}^T \boldsymbol{\Lambda}_\delta^{1/2} \mathbf{U}_\delta^T, \tag{16}$$

for some $\mathbf{R} \in \mathbb{R}^{(k-1)\times d}$ (recall that $d \geq k-1$) with $\mathbf{R}\mathbf{R}^T = \mathbb{I}_{k-1}$.

B.2. LDT

Consider the k-class unconstrained features CS-SVM problem of (3b):

$$(\hat{\mathbf{W}}, \hat{\mathbf{H}}) \in \arg\min_{\mathbf{W},\mathbf{H}} \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{1}{2}\|\mathbf{H}\|_F^2 \qquad \text{sub. to} \quad \delta_{y_i}(\mathbf{w}_{y_i} - \mathbf{w}_c)^T \mathbf{h}_i \geq 1,\ i \in [n], c \neq y_i. \tag{17}$$

Let n_c denote the number of examples in the class c. We will relate the above optimization problem to an equivalent UF-SVM problem, whose solution is known to satisfy a certain SELI geometry [23]. The resulting solution will be used to state the minimizers of (3b). First, let us reparameterize the problem as follows:

$$\tilde{\mathbf{h}}_i = \delta_{y_i}\mathbf{h}_i, \forall\ i \in [n], \text{ or in matrix notation, } \tilde{\mathbf{H}} = \mathbf{H}\mathbf{D}_n$$

Letting $\overline{\mathbf{H}}$ denote an optimal reparameterized feature matrix, the minimization then becomes:

$$\begin{aligned}(\hat{\mathbf{W}}, \overline{\mathbf{H}}) &\in \arg\min_{\mathbf{W},\tilde{\mathbf{H}}} \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{1}{2}\|\tilde{\mathbf{H}}\mathbf{D}_n^{-1}\|_F^2 \qquad \text{sub. to} \quad (\mathbf{w}_{y_i} - \mathbf{w}_c)^T \tilde{\mathbf{h}}_i \geq 1,\ i \in [n], c \neq y_i. \\ &= \arg\min_{\mathbf{W},\tilde{\mathbf{H}}} \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{1}{2}\sum_{i\in[n]} \frac{1}{\delta_{y_i}^2}\|\tilde{\mathbf{h}}_i\|^2 \qquad \text{sub. to} \quad (\mathbf{w}_{y_i} - \mathbf{w}_c)^T \tilde{\mathbf{h}}_i \geq 1,\ i \in [n], c \neq y_i.\end{aligned}$$

It can be checked that NC property holds on the global solution, i.e. $\forall i, j$ s.t. $y_i = y_j := c, \tilde{\mathbf{h}}_i = \tilde{\mathbf{h}}_j := \tilde{\mathbf{h}}_c$. With this we have

$$\arg\min_{\mathbf{W},\{\tilde{\mathbf{h}}_c\}} \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{1}{2}\sum_{c\in[k]} \frac{n_c}{\delta_c^2}\|\tilde{\mathbf{h}}_c\|^2 \qquad \text{sub. to} \quad (\mathbf{w}_c - \mathbf{w}_{c'})^T \tilde{\mathbf{h}}_c \geq 1,\ c \in [k], c' \neq c.$$

Now let $n_c/\delta_c^2 := \beta_c$. Suppose $\beta_c \in \mathbf{Z}_+, c \in [k]$, the above minimization is equivalent to a UF-SVM with $n_c^{\text{eff}} = \beta_c$ number of examples in class c shown below:

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

where $n^{\text{eff}} = \sum_{c=1}^k n_c^{\text{eff}}$. In showing the equivalence, we again used the NC property on the global minimizer of the effective UF-SVM problem. Note that the effective problem has a new effective imbalance ratio given by $R^{\text{eff}} := \frac{\beta_{\text{maj}}}{\beta_{\text{min}}} = \frac{n_{\text{maj}}}{n_{\text{min}}} \cdot \frac{\delta_{\text{min}}^2}{\delta_{\text{maj}}^2} = R \cdot \frac{\delta_{\text{min}}^2}{\delta_{\text{maj}}^2}$. Let $\hat{\mathbf{Z}}^{\text{eff}} \in \mathbb{R}^{k \times n^{\text{eff}}}$ be the SEL matrix of the effective UF-SVM problem with n_c^{eff} examples in class c. Let $\hat{\mathbf{Z}}^{\text{eff}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ be its compact SVD. Then, invoking [23], the global solution of the above problem satisfies the following properties:

- $\hat{\mathbf{W}}^T \overline{\mathbf{H}}^{\text{eff}} = \hat{\mathbf{Z}}^{\text{eff}}$
- $\overline{\mathbf{H}}^{\text{eff}T} \overline{\mathbf{H}}^{\text{eff}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, $\hat{\mathbf{W}}^T \hat{\mathbf{W}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$

In order to reconcile the dimensions, we will state the structure of the embeddings in terms of the Gram matrices of mean-embeddings. Let $\overline{\mathbf{M}}^{\text{eff}} \in \mathbb{R}^{k \times k}$ be the matrix whose columns are class-wise mean of optimal embeddings. Let $\mathbf{n} = \begin{bmatrix} n_1, & n_2, & \dots & n_k \end{bmatrix}^T$. Also let

$\mathbf{m} = \text{cumsum}(\mathbf{n}) - n_1$. We define the following operator that operates parameterized by $\mathbf{n}$ on any square matrix $\mathbf{A}$ of size $n = \sum_c \mathbf{n}$:

$$\mathcal{T}(\mathbf{A};\mathbf{n}) : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}^{k\times k}, \text{ such that}$$
$$T[c,c'] = \mathcal{T}(\mathbf{A};\mathbf{n})[c,c'] = \mathbf{A}[\mathbf{m}[c],\mathbf{m}[c']]$$

With this notation the the Gram matrix of the mean optimal embeddings becomes

$$(\overline{\mathbf{M}}^{\text{eff}})^T\overline{\mathbf{M}}^{\text{eff}} = \mathcal{T}((\overline{\mathbf{H}}^{\text{eff}})^T\overline{\mathbf{H}}^{\text{eff}};\mathbf{n}^{\text{eff}}) = \mathcal{T}(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T;\mathbf{n}^{\text{eff}})$$

Due to NC, by definition of the mean-embeddings of $\overline{\mathbf{H}}$ and $\overline{\mathbf{H}}^{\text{eff}}$, we have that $\overline{\mathbf{M}} = \overline{\mathbf{M}}^{\text{eff}}$, therefore $\overline{\mathbf{M}}^T\overline{\mathbf{M}} = \overline{\mathbf{M}}^{\text{eff}T}\overline{\mathbf{M}}^{\text{eff}}$. Now, we connect the above solution of the effective UF-SVM to the original LDT-SVM as follows. Since, $\mathbf{h}_i = \tilde{\mathbf{h}}_i/\delta_{y_i}$, it can be checked that the Gram matrix of mean-embeddings of the original LDT-SVM is:

$$\begin{aligned}\hat{\mathbf{M}}^T\hat{\mathbf{M}} &= (\overline{\mathbf{M}}\mathbf{D}^{-1})^T(\overline{\mathbf{M}}\mathbf{D}^{-1})\\ &= \mathbf{D}^{-1}\overline{\mathbf{M}}^T\overline{\mathbf{M}}\mathbf{D}^{-1}\\ &= \mathbf{D}^{-1}\overline{\mathbf{M}}^{\text{eff}T}\overline{\mathbf{M}}^{\text{eff}}\mathbf{D}^{-1}\\ &= \mathbf{D}^{-1}\mathcal{T}(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T;\mathbf{n}^{\text{eff}})\mathbf{D}^{-1},\end{aligned}$$

where $\mathbf{D} = \text{diag}(\boldsymbol{\delta})$. Now suppose that β_c is not an integer for some c. Then, letting δ_c to be a positive rational, β_c is a positive rational for every c. Say $\beta_c = \beta_{c,p}/\beta_{c,q}$,where $\beta_{c,p}, \beta_{c,q} \in \mathbb{Z}_+$ is its canonical representation, then we can define $n_c^{\text{eff}} := \alpha^2\beta_c$, where $\alpha^2 := \text{LCM}_c(\beta_{c,q})$. Now n_c^{eff} becomes a positive integer. From the extended results of [23], it can be seen that this additional scaling affects the geometry only up to a scaling factor α. The final geometry of global minimizers of LDT-SVM can then be described as shown below:

- Classifiers: $\hat{\mathbf{W}}^T\hat{\mathbf{W}} = \alpha\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$
- Mean-embeddings: $\hat{\mathbf{M}}^T\hat{\mathbf{M}} = \frac{1}{\alpha}\mathbf{D}^{-1}\mathcal{T}(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T;\mathbf{n}^{\text{eff}})\mathbf{D}^{-1}$

Appendix C. The implicit geometry

As stated in the Thm. 3, the optimal parameters of the UFM under the CDT/LDT loss have a unique description in terms of the SVD factors of a corresponding label-encoding matrix. In this section, we use this characterization to derive explicit expressions for the network's geometry as a function of R, ρ, k and of the hyper-parameters $\boldsymbol{\delta}$.

Similar to Sec. A, throughout this section, we assume the data is STEP imbalanced and STEP logit adjustment is adopted. For simplicity, we consider the case $\rho = 1/2$, $\delta_{\text{maj}} = \Delta$, and $\delta_{\text{min}} = 1$. The geometry in more general settings can be derived similarly. Assuming that the classifiers $\mathbf{w}_c$ and embeddings $\mathbf{h}_i$ for $c \in [k], i \in [n]$ follow the geometry in Thm. 3, we present closed-form expressions for the norms and pairwise angles of the parameters in the following lemmas.

C.1. CDT Loss

C.1.1. Norms

Lemma 6 (Norms of classifiers) *It holds that the classifier norms across all majority / minority classes are the same, i.e.,*

$$\forall c = 1, \dots, k/2\,, \quad \|\mathbf{w}_c\|_2^2 =: \|\mathbf{w}_{\text{maj}}\|_2^2 = \frac{\sqrt{R}}{\Delta}(1-2/k) + \frac{2\Delta^2\sqrt{R+1}}{k\left(\sqrt{1+\Delta^2}\right)^3}, \tag{18}$$

$$\forall c = k/2+1, \dots, k\,, \quad \|\mathbf{w}_c\|_2^2 =: \|\mathbf{w}_{\text{min}}\|_2^2 = (1-2/k) + \frac{2\sqrt{R+1}}{k\left(\sqrt{1+\Delta^2}\right)^3}. \tag{19}$$

where $\|\mathbf{w}_{\text{maj}}\|_2$ / $\|\mathbf{w}_{\text{min}}\|_2$ denotes the classifier norm of a majority / minority class.

Proof Let $m = k/2$. From Theorem 3 $\mathbf{W}^T\mathbf{W} = \mathbf{V}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}_{\boldsymbol{\delta}}\mathbf{V}_{\boldsymbol{\delta}}^T$. Using the SVD factors derived in Sec. A we have,

$$\begin{aligned}\mathbf{V}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}\mathbf{V}_{\boldsymbol{\delta}}^T &= \frac{\sqrt{R}}{\Delta}\begin{bmatrix}\mathbb{P}_m\mathbb{P}_m^T & 0\\ 0 & 0\end{bmatrix} + \frac{2\sqrt{R+1}}{k(\sqrt{1+\Delta^2})^3}\begin{bmatrix}\Delta^2\mathbb{1}_m\mathbb{1}_m^T & -\Delta\mathbb{1}_m\mathbb{1}_m^T\\ -\Delta\mathbb{1}_m\mathbb{1}_m^T & \mathbb{1}_m\mathbb{1}_m^T\end{bmatrix} + \begin{bmatrix}0 & 0\\ 0 & \mathbb{P}_m\mathbb{P}_m^T\end{bmatrix}\\
&= \begin{bmatrix}\frac{\sqrt{R}}{\Delta}\mathbb{I}_{k/2} - \frac{2}{k}\left(\frac{\sqrt{R}}{\Delta} - \frac{\Delta^2\sqrt{R+1}}{(\sqrt{1+\Delta^2})^3}\right)\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T & -\frac{2\Delta\sqrt{R+1}}{k(\sqrt{1+\Delta^2})^3}\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T\\ -\frac{2\Delta\sqrt{R+1}}{k(\sqrt{1+\Delta^2})^3}\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T & \mathbb{I}_{k/2} - \frac{2}{k}\left(1 - \frac{\sqrt{R+1}}{(\sqrt{1+\Delta^2})^3}\right)\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T\end{bmatrix}.\end{aligned} \tag{20}$$

Inspecting the diagonal entries of the Gram matrix above, completes the proof of the lemma. ■

Lemma 7 (Norms of embeddings) *It holds that the embedding norms across all majority / minority classes have equal norms, i.e.,*

$$\forall j \in \{i \in [n] : y_i = 1, \dots, k/2\}\,, \quad \|\mathbf{h}_j\|_2^2 =: \|\mathbf{h}_{\text{maj}}\|_2^2 = \frac{1}{\Delta\sqrt{R}}(1-2/k) + \frac{2}{k\sqrt{R+1}\sqrt{1+\Delta^2}}, \tag{21}$$

$$\forall j \in \{i \in [n] : y_i = k/2+1, \dots, k\}\,, \quad \|\mathbf{h}_j\|_2^2 =: \|\mathbf{h}_{\text{min}}\|_2^2 = (1-2/k) + \frac{2}{k\sqrt{R+1}\sqrt{1+\Delta^2}}. \tag{22}$$

where $\|\mathbf{h}_{\text{maj}}\|_2$ / $\|\mathbf{h}_{\text{min}}\|_2$ denotes the embeddings norm of a sample in a majority / minority class.

Proof Similar to the previous lemma, following Thm. 3, $\mathbf{H}^T\mathbf{H} = \mathbf{U}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}_{\boldsymbol{\delta}}\mathbf{U}_{\boldsymbol{\delta}}^T$. Recalling (7), the diagonal enteries of $\mathbf{U}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}_{\boldsymbol{\delta}}\mathbf{U}_{\boldsymbol{\delta}}^T$ can be found as follows.

$$\begin{aligned}\mathbf{U}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}_{\boldsymbol{\delta}}\mathbf{U}_{\boldsymbol{\delta}}^T &= \frac{\sqrt{R}}{\Delta}\begin{bmatrix}\frac{1}{R}\mathbb{P}_m\mathbb{P}_m^T \otimes \mathbb{1}_R\mathbb{1}_R^T & 0\\ 0 & 0\end{bmatrix} + \frac{2}{k\sqrt{R+1}\sqrt{\Delta^2+1}}\begin{bmatrix}\mathbb{1}_{Rm}\mathbb{1}_{Rm}^T & -\mathbb{1}_{Rm}\mathbb{1}_m^T\\ -\mathbb{1}_m\mathbb{1}_{Rm}^T & \mathbb{1}_m\mathbb{1}_m^T\end{bmatrix} + \begin{bmatrix}0 & 0\\ 0 & \mathbb{P}_m\mathbb{P}_m^T\end{bmatrix}\\
&= \begin{bmatrix}\left(\frac{1}{\Delta\sqrt{R}}\mathbb{I}_{k/2} - \frac{2}{k}\left(\frac{1}{\Delta\sqrt{R}} - \frac{1}{\sqrt{R+1}\sqrt{\Delta^2+1}}\right)\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T\right) \otimes \mathbb{1}_R\mathbb{1}_R^T & -\frac{2}{k\sqrt{R+1}\sqrt{\Delta^2+1}}\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T\\ -\frac{2}{k\sqrt{R+1}\sqrt{\Delta^2+1}}\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T & \mathbb{I}_{k/2} - \frac{2}{k}\left(1 - \frac{1}{\sqrt{R+1}\sqrt{\Delta^2+1}}\right)\mathbb{1}_{k/2}\mathbb{1}_{k/2}^T\end{bmatrix}.\end{aligned} \tag{23}$$

■

C.1.2. Angles

Lemma 8 (Angles of classifiers) *For each pair of classifiers belonging to majority / minority classes, the angles are the same. That is*

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

And the angles can be found by the following equations,

$$\begin{aligned}
\mathrm{Cos}(\mathbf{w}_{\mathrm{maj}}, \mathbf{w}'_{\mathrm{maj}}) &= \frac{-2\sqrt{R} + 2\sqrt{R+1}\left(\sqrt{1+\Delta^{-2}}\right)^{-3}}{(k-2)\sqrt{R} + 2\sqrt{R+1}\left(\sqrt{1+\Delta^{-2}}\right)^{-3}} \\
\mathrm{Cos}(\mathbf{w}_{\mathrm{min}}, \mathbf{w}'_{\mathrm{min}}) &= \frac{-2 + 2\sqrt{R+1}\left(\sqrt{1+\Delta^{2}}\right)^{-3}}{k - 2 + 2\sqrt{R+1}\left(\sqrt{1+\Delta^{2}}\right)^{-3}} \\
\mathrm{Cos}(\mathbf{w}_{\mathrm{maj}}, \mathbf{w}_{\mathrm{min}}) &= -\frac{2\Delta\sqrt{R+1}}{k\left(\sqrt{1+\Delta^{2}}\right)^{3} \|\mathbf{w}_{\mathrm{maj}}\|_2 \|\mathbf{w}_{\mathrm{min}}\|_2}\,.
\end{aligned}$$

Proof Recalling the geometry $\mathbf{W}^T\mathbf{W} = \mathbf{V}_{\boldsymbol{\delta}}\Lambda\mathbf{V}_{\boldsymbol{\delta}}^T$, the off-diagonals entries in (20) specifies the inner-product of each pair of classifiers. Particularly,

$$\begin{aligned}
\mathbf{w}_{\mathrm{maj}}^T\mathbf{w}'_{\mathrm{maj}} &= \frac{-2\sqrt{R} + 2\sqrt{R+1}\left(\sqrt{\Delta^{-2}+1}\right)^{-3}}{k\Delta}, \\
\mathbf{w}_{\mathrm{min}}^T\mathbf{w}'_{\mathrm{min}} &= \frac{-2 + 2\sqrt{R+1}\left(\sqrt{1+\Delta^{2}}\right)^{-3}}{k}, \\
\mathbf{w}_{\mathrm{min}}^T\mathbf{w}_{\mathrm{maj}} &= \frac{-2\Delta\sqrt{R+1}}{k\left(\sqrt{\Delta^{2}+1}\right)^{3}}\,.
\end{aligned}$$

These equations together with Lemma 6 completes the proof. ■

Lemma 9 (Angles of classifiers) *For each pair of embeddings belonging to majority / minority classes, the angles are the same. That is*

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

And the angles can be found using the following equations,

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

Proof Recalling the geometry $\mathbf{H}^T\mathbf{H} = \mathbf{U}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}_{\boldsymbol{\delta}}\mathbf{U}_{\boldsymbol{\delta}}^T$, the off-diagonals entries in (23) specifies the inner-product of each pair of the mebeddings. Particularly,

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

Combine these with the norm calculations in Lemma 7 completes the proof. ■

In the next lemma, we calculate the angles between an embedding and its corresponding classifier. Particularly, we give closed-form expression for $\mathrm{Cos}(\mathbf{w}_c, \mathbf{h}_i)$ for $c \in [k], i : y_i = c$, which can be thought of the degree of alignment between classifiers and embeddings.

Lemma 10 (Alignment of classifiers and embeddings) *For the angles between majority / minority embeddings and their corresponding classifiers are equal, i.e.,*

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

It also holds that,

$$\mathrm{Cos}(\mathbf{w}_{\mathrm{maj}}, \mathbf{h}_{\mathrm{maj}}) = \frac{k\Delta^2 + (k-2)}{k\,\Delta\,(\Delta^2+1)\,\|\mathbf{w}_{\mathrm{maj}}\|_2\|\mathbf{h}_{\mathrm{maj}}\|_2}, \tag{24}$$

$$\mathrm{Cos}(\mathbf{w}_{\mathrm{min}}, \mathbf{h}_{\mathrm{min}}) = \frac{k\Delta^{-2} + (k-2)}{k\,(\Delta^{-2}+1)\,\|\mathbf{w}_{\mathrm{min}}\|_2\|\mathbf{h}_{\mathrm{min}}\|_2}. \tag{25}$$

Proof Recalling $\mathbf{W}^T\mathbf{H} = \hat{\mathbf{Z}}_{\boldsymbol{\delta}}$, for all $c \in [k]$ and $i : y_i = c$ it holds that $\mathbf{w}_c^T\mathbf{h}_i = \Delta^{-1}\left(1 - \frac{2}{k(\Delta^2+1)}\right)$ if c is a majority class, and $\mathbf{w}_c^T\mathbf{h}_i = \left(1 - \frac{2}{k(\Delta^{-2}+1)}\right)$ otherwise. ■

C.1.3. Centering

Assuming that the classifiers follow the geometry in Thm. 3, $\mathbf{w}_c, c \in [k]$ are centered around zero after some re-weighting, i.e. $\sum_{c\in[k]} \mathbf{w}_c/\delta_c = 0$. This is immediate from $\hat{\mathbf{W}} = \mathbf{R}^T \mathbf{\Lambda}_{\boldsymbol{\delta}}^{1/2} \mathbf{V}_{\boldsymbol{\delta}}^T$ and $\mathbf{V}^T \mathbf{D}^{-1} \mathbb{1}_k = 0$.
The embeddings $\mathbf{h}_i, i \in [n]$ are also not centered around zero in general. Instead, it holds that

$$\sum_{i\in[n]} \frac{1}{n_{y_i}} \mathbf{h}_i = 0. \tag{26}$$

Note that this reduces to $\sum_{i\in[n]} \mathbf{h}_i$ for balanced data, however remains unchanged for any choice of the hyper-parameters $\boldsymbol{\delta}$ for imbalanced data.

C.2. LDT Loss

From Thm. 3, we can see that if $(\hat{\mathbf{W}}, \hat{\mathbf{H}})$ are the solutions of the unconstrained CS-SVM of LDT-loss in (3b), $(\hat{\mathbf{W}}, \hat{\mathbf{M}}\mathbf{D})$ follows the SELI geometry [23], with imbalance ratio $R^{\text{eff}} = R \cdot \delta_{\text{min}}^2/\delta_{\text{maj}}^2$. Thus, the corresponding norms and angles can be found by analyzing the R^{eff}-SELI structure (up to a norm scaling by $\mathbf{D}$ for the mean embeddings $\hat{\mathbf{M}}$). For the closed form expressions of the implicit geometry under the LDT-loss refer to [23, Sec. B.1].

Appendix D. Additional Experimental Results

D.1. UFM Results

In order to confirm our experimental results on other imbalance ratios we have conducted experiments under setting $(R = 5, \rho = 1/2)$-STEP Imbalance with $n_{\text{min}} = 9$ and $(R = 2, \rho = 1/2)$-STEP Imbalance with $n_{\text{min}} = 18$. We use these n_{min} values in to ensure that the total number of examples across different experimental settings are the same as the total number of data points used in experiments with $R = 10$. In all cases, we train the model with zero weight dacay and momentum. We use constant learning rate of 0.1 and batch size of 8.

To inspect the geometries, we measure the properties of the implicit geometry defined in Sec. C. In particular, for the geometry of the final trained classifier , we compute the norm ratio $\sigma_{\mathbf{w}}^{\text{norm}}(c, c') = \|\mathbf{w}_c\|/\|\mathbf{w}_{c'}\|$ and angles $\sigma_{\mathbf{w}}^{\text{angle}}(c, c') = \cos(\mathbf{w}_c, \mathbf{w}_{c'})$ for c and c' a majority/minority class. We similarly calculate these values for the embeddings, and compare the average of the norm ratio and angles (markers in Fig. 2-4) with the corresponding theoretical values in Sec. C (solid lines).

D.2. Results on Deep Networks

In addition to the UFM results, we conduct experiments with ResNet18 model trained on CIFAR-10 under a $R = 10$ imbalance ratio. 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. Following the same setting as in [20, 23], momentum and weight decay were set to 0.9 and $1e-5$ respectively and the batchsize was set to 128. We use CDT/LDT-loss with $\boldsymbol{\delta}$ normalized to sum to $k = 10$, similar to the UFM cases. Based on the results in Fig. 5, the classifiers show a

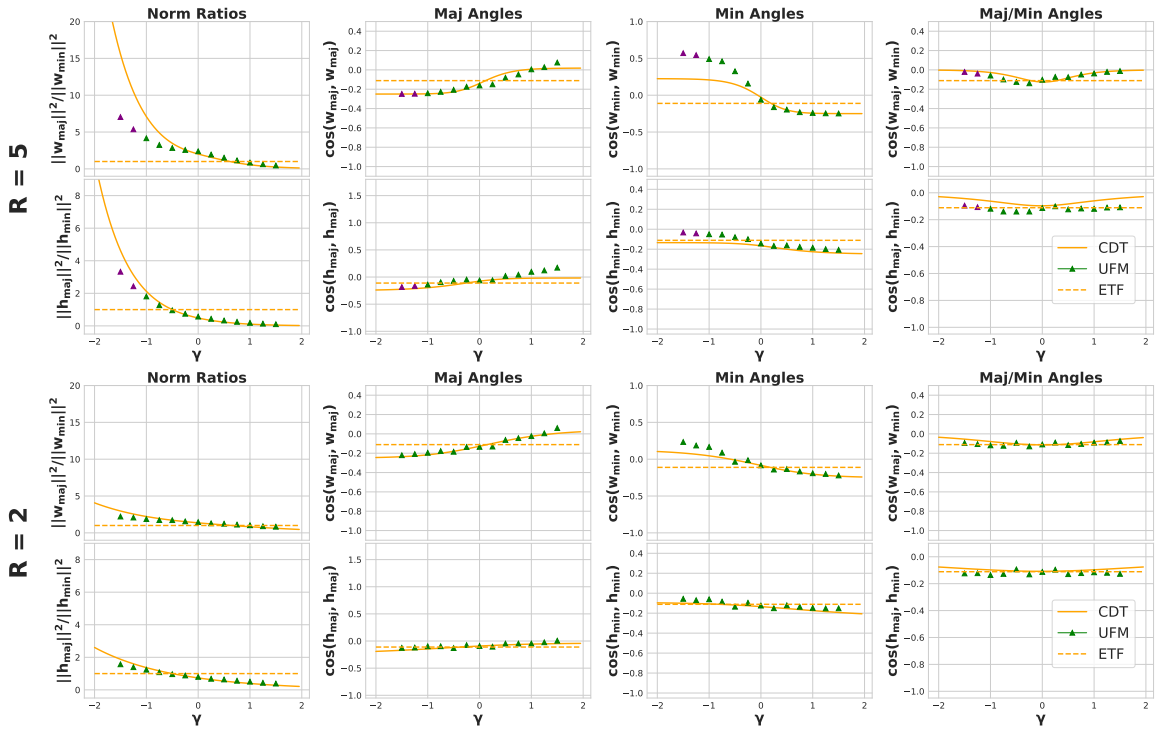


Figure 3: The geometric properties of the UFM (green/purple markers) trained by CDT-loss vs implicit geometry in Thm. 3 (solid line) and ETF (dashed line), for imabalance ratio $R = 2, 5$. Purple markers are used to distinguish the cases the model did not enter the zero-training-error regime.

faster convergence to the proposed geometry compared to features, indicating that the loss function affects the classifier geometry more visibly. Similar to the UFM case, achieving zero training error tends to be easier with the LDT loss even with larger γ.

D.3. Optimization Concerns

Across our experiments we have attempted to use the same training regiment regardless of imbalance ratios or $\boldsymbol{\delta}$ values. However, Fig. 6 indicates that higher imbalance ratios and larger γs tend to converge slower. This suggests that we might need to consider different hyper-parameters and methods depending on the experimental setup for better convergence.

We observe slower convergence to the theoretical model with higher imbalance ratios (compare Figs. 2-4). These figures further demonstrate that it becomes harder to achieve zero-training-error as Δ gets far from 1, i.e $\delta_{\text{maj}} = \delta_{\text{min}}$. On top of the convergence issue, when considering the deepnet experiments it becomes significantly harder to achieve zero training error for larger γ values. Fig. 7 demonstrates the effect of γ on the training accuracy of the minority classes. We also observe empirically that entering this zero-error regime is easier for the LDT model, given that both models are trained under the same setting. The experimental results seem to suggest that positive γ values help achieve zero train error for minority classes for the CDT case. We further discovered that this trend translates into higher test minority and balanced test accuracy for the CDT case as indicated. Fig. 8 shows

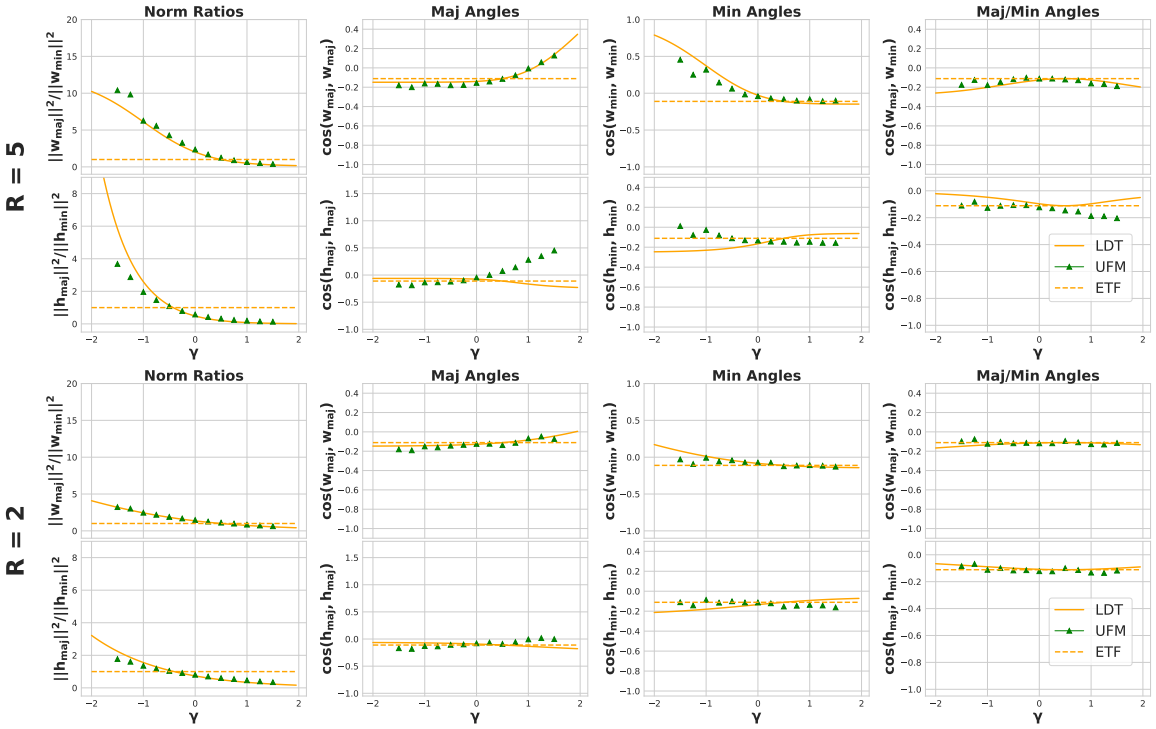


Figure 4: Same setting as Fig. 3, with LDT-loss.

how the gamma value has a positive effect on the balanced test accuracy. We will continue to study the effects of LDT and CDT to discover new methods for increasing the model accuracies.

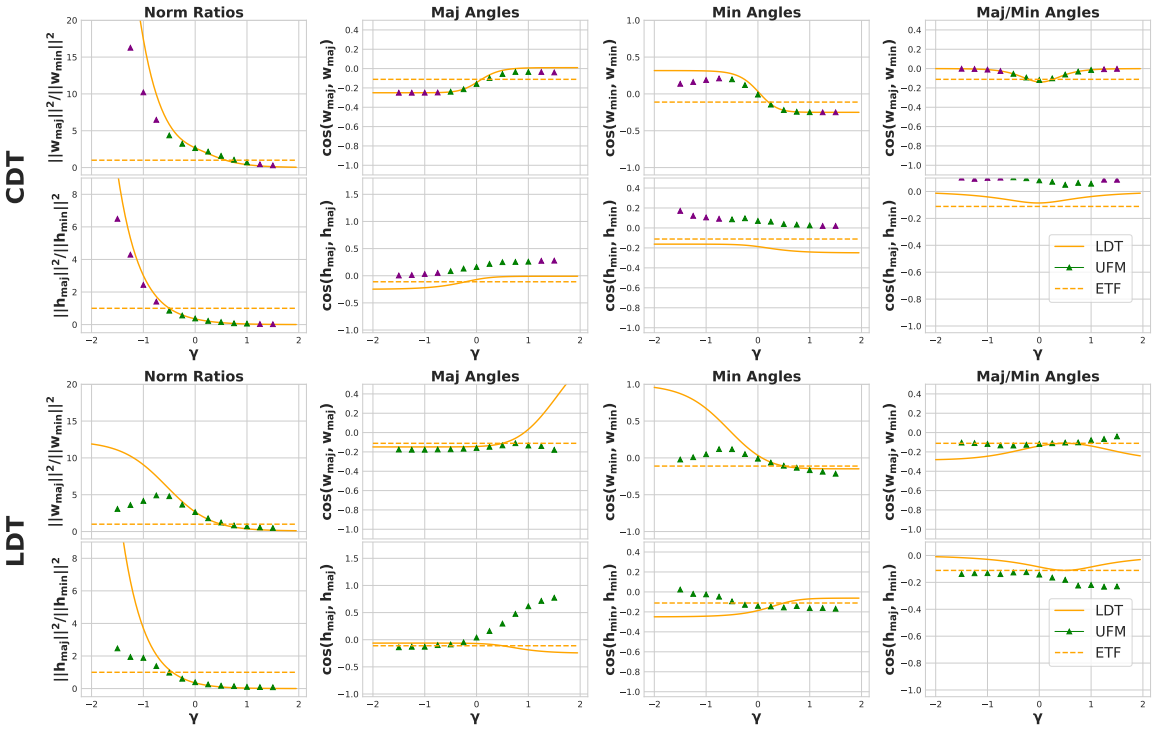


Figure 5: The geometry of the classifiers and embeddings of a ResNet18 network trained on ($R = 10$, ρ=1/2)-STEP imbalanced CIFAR-10 (green/purple markers) trained by CDT/LDT-loss vs implicit geometry in Thm. 3 (solid line) and ETF (dashed line). Purple markers are used to distinguish the cases the model did not enter the zero-training-error regime.

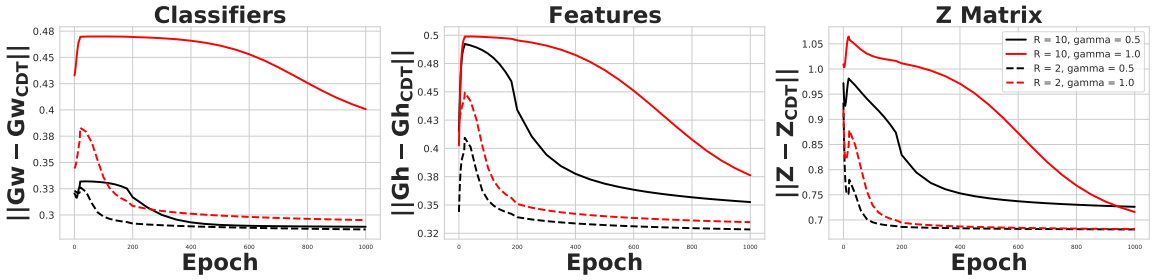


Figure 6: This plot compares the convergence to CDT geometry trained UFM model. The plots above demonstrate how higher ratios of imbalance and larger γ values correlate to slower speed of convergence for classifier, feature and logits to the CDT structure. This plot is only for the first 1000 epochs of training.

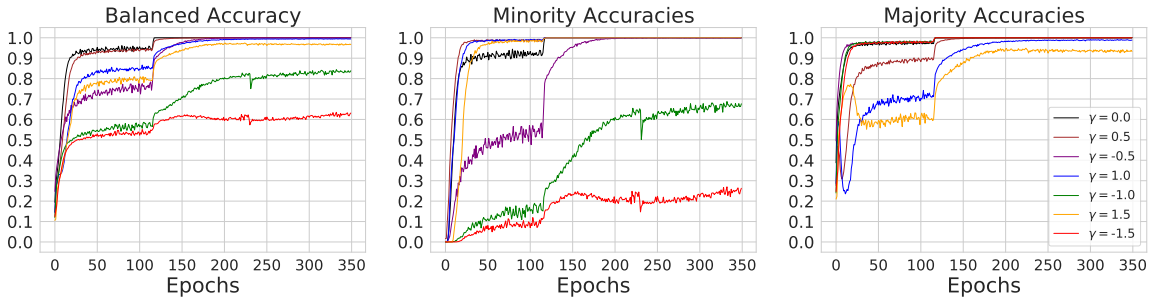


Figure 7: Train accuracy for ResNet18 network trained on (R = 10, ρ=1/2)-STEP imbalanced CIFAR-10 trained with CDT with a range of different γ values. ($\frac{\delta_{\delta_{\text{maj}}}}{\delta_{\delta_{\text{min}}}} = R^{\gamma}$).

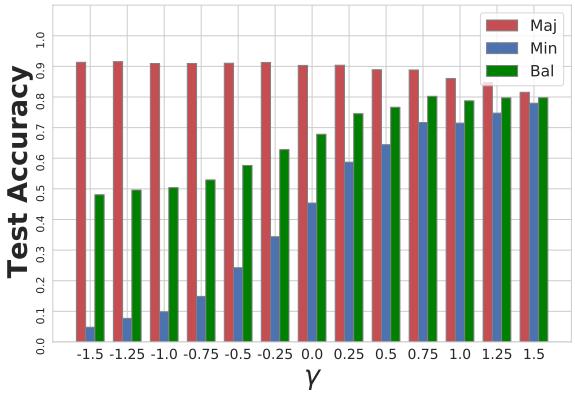


Figure 8: Test accuracy for ResNet18 network trained on (R = 10, ρ=1/2)-STEP imbalanced CIFAR-10 trained with CDT with a range of different γ values ($\frac{\delta_{\delta_{\text{maj}}}}{\delta_{\delta_{\text{min}}}} = R^{\gamma}$). The plot indicates that value of γ has a positive correlation with the minority accuracy.