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

Tina Behnia[†§] **Ganesh Ramachandra Kini**[⋆§] **Vala Vakilian**[†§] **Christos Thrampoulidis**[†]
[†]University of British Columbia, Canada [⋆]University of California, Santa Barbara, USA

Abstract

Various logit-adjusted parameterizations of the cross-entropy (CE) loss have been proposed as alternatives to weighted CE for training large models on label-imbalanced data far beyond the zero train error regime. The driving force behind those designs has been the theory of *implicit bias*, which for linear(ized) models, explains why they successfully induce bias on the optimization path towards solutions that favor minorities. Aiming to extend this theory to non-linear models, we investigate the *implicit geometry* of classifiers and embeddings that are learned by different CE parameterizations. Our main result characterizes the global minimizers of a non-convex cost-sensitive SVM classifier for the unconstrained features model, which serves as an abstraction of deep-nets. We derive closed-form formulas for the angles and norms of classifiers and embeddings as a function of the number of classes, the imbalance and the minority ratios, and the loss hyperparameters. Using these, we show that logit-adjusted parameterizations can be appropriately tuned to learn symmetric geometries irrespective of the imbalance ratio. We complement our analysis with experiments and an empirical study of convergence accuracy in deep-nets.

1 INTRODUCTION

In the modern overparameterized regime, when 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 (Byrd and Lipton, 2019; Sagawa et al., 2020). In a growing literature, several alternatives have been proposed to guarantee equitable performance across majorities and minorities (e.g., Menon et al., 2020; Ye et al., 2020; Kini et al., 2021; Cao et al., 2019; Khan et al., 2017; Lin et al., 2018; Kim and Kim, 2020; Kang et al., 2020). Among these, the vector-scaling (VS) loss (Kini et al., 2021; Ye et al., 2020) introduces multiplicative hyperparameters 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 (Soudry et al., 2018; Ji and Telgarsky, 2018; Lyu and Li, 2019). Specifically for binary linear models, Kini et al. (2021) 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 learned models in non-linear settings where embeddings and classifiers are jointly learned. Unfortunately, implicit bias characterizations for non-linear models are more obscure compared to the linear case (Lyu and Li, 2019; Ji and Telgarsky, 2020). Particularly, it is unclear how to gain concrete insights from them on the way the learned models affect minorities.

This paper investigates the *implicit geometry* of classifiers and embeddings learned by CE parameterizations when trained on imbalanced data. The notion of implicit geometry, pioneered by Papyan et al. (2020) and further investigated by many others (e.g., Fang et al., 2021; Galanti et al., 2021; Graf et al., 2021; Han et al., 2021; Hui et al., 2022; Ji et al., 2021; Lu and Steinerberger, 2020; Mixon et al., 2020; Tirer and Bruna, 2022; Xie et al., 2022; Zhu et al., 2021; Zhou et al., 2022a; Thrampoulidis et al., 2022), is intimately related to

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

§: equal contribution
This work is supported by an NSERC Discovery Grant, NSF Grant CCF-2009030, and by a CRG8-KAUST award. The authors also acknowledge use of the Sockeye cluster by UBC Advanced Research Computing.
Code available at: https://github.com/valavakilian/Implicit_geometry

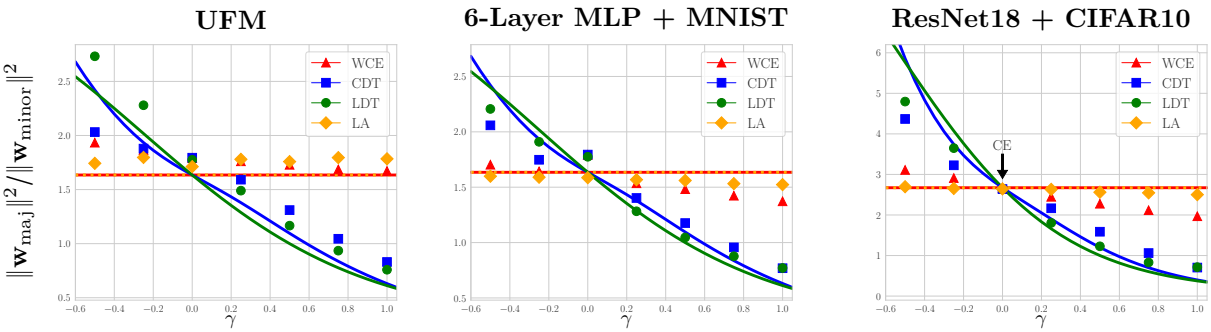

Figure 1: Ratio of classifier norms between majorities and minorities for ($R = 10$, ρ=1/2)-STEP imbalanced data (see Defn. 1) on (Left) UFM, (Middle) 6-layer MLP and MNIST, (Right) ResNet18 and CIFAR10. We train four different CE parameterizations with varying hyperparameter $\gamma \in [-0.5, 1]$: (1) weighted CE with $\omega_{\rm minor}/\omega_{\rm maj} := R^{\gamma}$; (2,3) CDT, LDT losses with $\Delta := \delta_{\rm maj}/\delta_{\rm minor} = R^{\gamma}$; (4) LA loss with $\iota_{\rm maj} - \iota_{\rm minor} = \gamma \log R$. $\gamma = 0$ corresponds to CE loss. Markers denote empirically measured quantities. Solid lines follow theoretical formulas (Eqn. (1)). See text for details.

that of implicit bias.[1] On the one hand, it is more restrictive as it focuses only on the classifiers and on the embeddings, rather than all the weights of the model. Also, it is insensitive to the specific architecture or dataset. On the other hand, it offers a more explicit characterization describing the entire geometry of the weights and promises to be "cross-situationally invariant" across architectures and datasets (Papyan et al., 2020).

Contributions. We study two parameterizations of the CE loss: (i) the class-dependent temperature (CDT) loss (Ye et al., 2020), which is a special case of the VS loss (Kini et al., 2021), and (ii) the label-dependent temperature (LDT) loss, which we introduce here as an alternative to the CDT loss. For both losses, we study the implicit geometry of learned features and classifiers when trained on label-imbalanced data without explicit regularization beyond zero training error. To do this, we rely on the unconstrained features model (UFM) (Mixon et al., 2020; Fang et al., 2021), which serves as a proxy for large overparameterized models and has been used recently to study the implicit geometry of the CE loss (see *Related work*). Relying on the implicit bias results, we relax the question of implicit geometry of the solutions found by stochastic gradient descent (SGD), to a question about the geometry of the global minimizers of a non-convex *Cost-Sensitive Support-Vector Machines* (CS-SVM) problem, which takes different forms for the CDT and LDT losses. Our main result characterizes the global minimizers of the CDT and LDT CS-SVM problems in terms of a new geometry, which we call the $(\boldsymbol{\delta}, R)$-*geometry* and is parameterized by a vector $\boldsymbol{\delta}$ of hyperparameters and the data imbalance ratio R. The new geometry has the following favorable properties: (i) It includes the previously discovered ETF (Papyan et al., 2020) and SELI (Thrampoulidis et al., 2022) geometries as special cases. Also, it captures both CDT and LDT. (ii) It admits an explicit characterization that involves closed-form formulas of the norms and angles in terms of the number of classes, the minority ratio, the imbalance ratio, and the vector of hyperparameters. (iii) It reveals appropriate tuning recipes for the hyperparameters to learn symmetric geometries with respect to minorities and majorities irrespective of the imbalance ratio. (iv) It shows that LDT and CDT can both mitigate minority collapse, i.e., the collapse of minority classifiers in the large imbalance-ratio limit. Beyond these, 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. This observation motivates further theoretical and algorithmic investigations towards faster training with CE parameterizations. As evidence of the utility of our geometry characterizations for the UFM, we present results on deep-net architectures and complex imbalanced datasets.

Example. Fig. 1 provides a graphical illustration of the impact of different CE parameterizations on the implicit geometry. Here, we focus on classifiers and specifically their norms.

In Fig. 1(Right), we train a ResNet18 on a (10,1/2)-STEP imbalanced CIFAR10 dataset (see Defn. 1). For the training we use four different parameterizations of

[1] Initially, Papyan et al. (2020) referred to their discovered geometry as "Neural Collapse" (NC). Later, to differentiate between the geometries learned by CE for balanced vs imbalanced data, Thrampoulidis et al. (2022) introduced the terms ETF and SELI geometries for the former and latter, respectively. We show here that different CE parameterizations result in yet different geometries, prompting us to adopt the more general term "implicit geometry".

the CE loss, namely the weighted CE (wCE), CDT (Ye et al., 2020) (Eqn. (3a)), LDT (Eqn. (3b)) and LA (Cao et al., 2019; Menon et al., 2020) losses. Each of these, comes with a set of corresponding hyperparameters, which we control by varying a single parameter $\gamma \in \mathbb{R}$ in the interval $[-0.5, 1]$. For $\gamma = 0$, all the losses reduce to standard CE loss. For each loss and for each value of γ, we compute the ratio of the classifier norms for each pair of majority-minority classes, and the markers report the average of these ratios. First, observe for $\gamma = 0$ (CE) that $\|\mathbf{w}_{\rm maj}\|_2 \approx 2.8\|\mathbf{w}_{\rm minor}\|_2$. This is different from the case of balanced classes where ETF geometry suggests $\|\mathbf{w}_{\rm maj}\|_2 \approx \|\mathbf{w}_{\rm minor}\|_2$ (Papyan et al., 2020). The fact that, under class imbalances, CE loss learns classifiers with larger norm for majorities compared to minorities has been empirically observed in the imbalanced deep-learning literature (Kim and Kim, 2020; Kang et al., 2020; Menon et al., 2020) and various heuristic methods have been proposed to mitigate this effect towards favoring minorities. One of these, the LA loss (Menon et al., 2020) is seen here to have minimal effect on changing the classifiers' imbalance ratio. The wCE loss has similar behavior as the ratio reduces only marginally with increasing γ. On the other hand, both CDT and LDT offer flexibility in tuning the ratio over a wide range by varying γ: as γ increases the norm of minorities increases relative to the majorities. Interestingly, for appropriate γ values the ratio can be made 1 (as in the balanced case).

Fig. 1(Middle) repeats the above experiment on a 6-layer MLP with imbalance MNIST data. The behavior is analogous: For CE the ratio is ≈ 2.8, while appropriately tuning LDT and CDT losses can tweak the classifiers' geometry and change the norm ratio.

Finally, Fig. 1(Left) repeats the experiment on the synthetic unconstrained features model (UFM) (see Sec. 3). Observe that the behavior is remarkably reflective of the trends seen previously on ResNet/MLP architectures and CIFAR10/MNIST data. Compared to the latter, the UFM is amenable to mathematical analysis. Specialized to classifiers' norms, our analysis yields the following explicit formulas for the CDT/LDT solutions of the UFM for hyperparameter $\Delta := R^{\gamma}$:

$$\text{CDT: } \frac{\|\mathbf{w}_{\rm maj}\|_2^2}{\|\mathbf{w}_{\rm minor}\|_2^2} = \frac{\frac{\sqrt{R}}{\Delta}(k-2)\big(1+\Delta^2\big)^{3/2} + 2\Delta^2\sqrt{R+1}}{(k-2)\big(1+\Delta^2\big)^{3/2} + 2\sqrt{R+1}},$$
$$\text{LDT: } \frac{\|\mathbf{w}_{\rm maj}\|_2^2}{\|\mathbf{w}_{\rm minor}\|_2^2} = \frac{(k-2)\sqrt{R} + \sqrt{(R+\Delta^2)/2}}{(k-2)\Delta + \sqrt{(R+\Delta^2)/2}}. \tag{1}$$

The solid blue (CDT) and green (LDT) curves graph those formulas for $k = 10$ classes and imbalance ratio $R = 10$. Note that the very same formulas capture the empirical trend for UFM Fig. 1(Left) and also for MLP and ResNet in Fig. 1(Middle,Right). For LDT simply setting $\Delta = \sqrt{R}$ ($\gamma = 1/2$) makes the norms of majorities and minorities equal. We will prove that the same choice also guarantees maximal angle separation and alignment of classifiers and embeddings. On the other hand, for CDT, the value of Δ (eqv. γ) making $\|\mathbf{w}_{\rm maj}\|_2 \approx \|\mathbf{w}_{\rm minor}\|_2$ depends on k and R in general.

Related works. In their inspiring work, Papyan et al. (2020) discover that the geometry of classifiers and embeddings that are learned by overparameterized models trained with CE far beyond zero-training error can be characterized in terms of a few simple properties. (i) *Neural Collapse (NC)*: the embeddings collapse to their class means. (ii) *Simplex Equiangular Tight-Frame (ETF) geometry*: the classifiers align with the embeddings of the corresponding class, they all have the same norm, and, they are maximally separated from each other. Notably, this characterization is shown to be cross-situationally invariant across different architectures and datasets. Important follow-up works (Mixon et al., 2020; Fang et al., 2021; Graf et al., 2021) introduce the Unconstrained Features Model (UFM), as a proxy model to complex deep-nets, and uses it (Zhu et al., 2021; Zhou et al., 2022a; Ji et al., 2021; Thrampoulidis et al., 2022; Zhou et al., 2022b) to give (partial) theoretical justification of the discovery made by Papyan et al. (2020). Extensions of the geometry characterization to mean-square loss and of the UFM to mean-square loss are also studied in Mixon et al. (2020); Zhou et al. (2022a); Tirer and Bruna (2022). A line of work also investigates potential connections to generalization (Hui et al., 2022; Han et al., 2021) and transfer-learning (Galanti et al., 2021, 2022), although this is an arguably less-understood topic. All these works assume that data are balanced. On the other hand, when data are imbalanced, Fang et al. (2021) shows a *minority collapse* phenomenon, i.e., the minority classifiers collapse to each other as the imbalance ratio R grows to infinity. The complete geometry of both classifiers and embeddings at finite imbalance ratios was only very recently characterized in Thrampoulidis et al. (2022) under the name: *Simplex-Encoded Label Interpolation (SELI)* geometry. The SELI geometry is parameterized by the imbalance ratio R, and it includes the ETF geometry as a special case. It also recovers the minority collapse when evaluating angles asymptotically in R. Extending this literature, we formulate a new and more general geometry (which includes SELI and ETF as special cases) and show that it describes the learned embeddings and classifiers of two CE parameterizations, the CDT and the LDT losses. Closely related are also the works Xie et al. (2022); Yang et al. (2022) which design loss functions for class-imbalanced learning in an attempt to enforce a geometry alike the ETF geometry for balanced data. However, they do not characterize the joint geometry of classifiers and embeddings as we do here. Besides, the loss functions

that they consider are different in nature from the CDT and LDT losses. The latter originate from Cao et al. (2019); Menon et al. (2020); Ye et al. (2020); Kini et al. (2021), which propose various logit-adjustments to the CE loss with the goal of mitigating label imbalances. Specifically, the CDT loss is proposed in Ye et al. (2020) and is a special case of the VS loss in Kini et al. (2021). Here, we also introduce a new loss, the LDT loss, and show that it forms a canonical extension of the binary VS loss of Kini et al. (2021). Unlike those prior works limiting their analytical studies to binary and linear models, our implicit geometry approach allows further investigating multiclass and feature-learning regimes.

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. $\mathbf{V}_{j:k} \in \mathbb{R}^{m\times(k-j+1)}$ chooses columns $j, j+1, \dots, k$ of $\mathbf{V}$, and $\mathbf{V}^T_{j:k} \in \mathbb{R}^{n\times(k-j+1)}$ does so on $\mathbf{V}^T$. We denote $\|\mathbf{V}\|_F$ the Frobenius norm of $\mathbf{V}$. We use $\mathbf{V} \propto \mathbf{X}$ whenever the two matrices are equal up to a scalar constant. For a vector $\mathbf{v} \in \mathbb{R}^k$, $\mathrm{diag}(\mathbf{v}) \in \mathbb{R}^{k\times k}$ is the diagonal matrix with $\mathbf{v}$ on its diagonal. $\otimes$ denotes Kronecker products. 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. Finally, we denote the set of positive rational numbers by $\mathbb{Q}_+$.

2 BACKGROUND

The Vector-Scaling (VS) loss is the following parameterization of the CE loss (Kini et al., 2021):

$$\mathcal{L}_{\mathrm{VS}}(\mathbf{W},\boldsymbol{\theta}) =: \sum_{i\in[n]} \log\Big(1+\sum_{c\neq y_i} e^{-(\delta_{y_i}\mathbf{w}_{y_i}-\delta_c\mathbf{w}_c)^T\mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_i)+\iota_{y_i}-\iota_c}\Big). \quad (2)$$

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 hidden layers of a neural network), $y_i \in [k], i\in[n]$ are labels, and $\mathbf{w}_c, c\in[k]$ are 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 $\delta_c = 1, \iota_c = 0$, recovers the CE loss. Setting $\delta_c = 1$ and only varying ι_c gives the LA loss (Menon et al., 2020), while setting $\iota_c = 0$ and only varying δ_c gives the CDT loss (Ye et al., 2020).

Prior art: Binary linear classification. In a binary setting with fixed feature map (non-trainable $\boldsymbol{\theta}$) Kini et al. (2021) studies the implicit bias of binary VS loss.

Proposition 1 (Kini et al. (2021))**.** *Consider a fixed feature map $\mathbf{h}_{\boldsymbol{\theta}}$, binary labels $\upsilon_i \in \{\pm 1\}$, $\mathbf{h}_i := \mathbf{h}_{\boldsymbol{\theta}}(x_i)$ for $i\in[n]$ and hyperparameters $\delta_{\pm 1}$. Then GD with sufficiently small learning rate on the binary VS loss $\mathcal{L}_{\mathrm{VS,bin}}(\mathbf{w}) := \sum_{i\in[n]} \log(1+e^{-\delta_{\upsilon_i}\upsilon_i\mathbf{w}^T\mathbf{h}_i+\iota_{\upsilon_i}})$ converges (asymptotically in the number of training steps) in direction to the Cost-Sensitive SVM (CS-SVM) classifier:* $\arg\min_{\mathbf{w}}\|\mathbf{w}\|_2$ subj. to $\upsilon_i\delta_{\upsilon_i}\mathbf{w}^T\mathbf{h}_i \geq 1, i\in[n]$.

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 $\upsilon = +1$ is minority, favors the minority by assigning larger margin to it. 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.

Open problem: Beyond linear models. Prop. 1 is limited to a setting with fixed features. While an extension of the loss to the learned-feature setting is easy to heuristically derive (see (2)), it is an open question to explicitly characterize the effect of the hyperparameters on the learned solution. For instance, how do they affect the relative margin between majorities and minorities or between minorities and 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 learned (asymptotically in the number of training steps) by GD. For this, we adopt the *unconstrained features model* (UFM) (Mixon et al., 2020; Fang et al., 2021). 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 classifiers 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 and $\sum_{c\in[k]} n_c = n$. Without loss of generality, we assume examples are ordered. Formally, defining $n_0 = 0$, examples $i = \sum_{c'=0}^{c-1} n_{c'} + 1, \dots, \sum_{c'=0}^{c} n_{c'}$ are in class c. In the UFM, features $\mathbf{h}_i, i\in[n]$ are trained *jointly* with the weights $\mathbf{w}_c, c\in[k]$ and are *unconstrained*, i.e. trained without abiding by an explicit parameterization by some weight vector $\boldsymbol{\theta}$ (as in (2)).

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

$$\mathcal{L}_{\mathrm{CDT}}(\mathbf{W}^T\mathbf{H};\boldsymbol{\delta}) := \sum_{i\in[n]} \log\Big(1+\sum_{c\neq y_i} e^{-(\delta_{y_i}\mathbf{w}_{y_i}-\delta_c\mathbf{w}_c)^T\mathbf{h}_i}\Big), \quad (3a)$$

$$\mathcal{L}_{\mathrm{LDT}}(\mathbf{W}^T\mathbf{H};\boldsymbol{\delta}) := \sum_{i\in[n]} \log\Big(1+\sum_{c\neq y_i} e^{-(\delta_{y_i}(\mathbf{w}_{y_i}-\mathbf{w}_c))^T\mathbf{h}_i}\Big). \quad (3b)$$

Both losses are parameterized by a positive vector $\boldsymbol{\delta} = \left[\delta_1, \delta_2, \ldots, \delta_k\right]^T \in \mathbb{R}_+^k$ of multiplicative hyperparameters. The CDT loss in (3a) was previously introduced by Ye et al. (2020); Kini et al. (2021) (which is a special case of (2) when ignoring the additive ι_c). Here, we also introduce the LDT loss in (3b) as an alternative parameterization.

CDT vs LDT. Observe the subtle distinction: CDT associates $\boldsymbol{\delta}$ with the class label of the classifiers $\mathbf{w}_c$, while LDT associates the same hyperparameters with the label of the feature vectors $\mathbf{h}_i$. Our initial motivation for introducing LDT is the following observation.

Lemma 3.1. *Assume binary linearly separable data and training of linear classifiers without regularization. The LDT classification rule coincides with the rule of the binary VS loss assuming same δ-tuning. On the other hand, minimizing CDT results in the same classification rule as CE, irrespective of the δ-tuning.*

In other words, for binary linear settings CDT does not improve over CE, while LDT does so by reducing to the binary VS loss of Prop. 1. While Lem. 3.1 motivates LDT, our results below show that the intuition gained from binary linear settings can be restrictive. Indeed, we show that both LDT and CDT losses induce rich behaviors in the multiclass learned-feature regime.

Unconstrained-features cost-sensitive SVM. We minimize the losses in (3) without explicit regularization. Note that in the UFM, minimization over the embedding map is not parameterized in terms of $\boldsymbol{\theta}$, as say in (2). 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 it can be shown following Lyu and Li (2019) 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: $\min_{\mathbf{W},\mathbf{H}} \ \|\mathbf{W}\|_F^2 + \|\mathbf{H}\|_F^2$

$$\text{subj. to} \quad (\delta_{y_i}\mathbf{w}_{y_i} - \delta_c\mathbf{w}_c)^T\mathbf{h}_i \geq 1, \ i \in [n], c \neq y_i, \quad \text{(4a)}$$

$$\text{subj. to} \quad \delta_{y_i}(\mathbf{w}_{y_i} - \mathbf{w}_c)^T\mathbf{h}_i \geq 1, \ i \in [n], c \neq y_i. \quad \text{(4b)}$$

Note the resemblence to the CS-SVM minimization of Prop. 1. But unlike that, the problems here are non-convex since minimization is also over $\mathbf{H}$. We refer to (4) as unconstrained CS-SVM or simply CS-SVM.

Remark 1. *It is straightforward to extend our results to a modified objective $\|\mathbf{W}\|_F^2 + \beta\|\mathbf{H}\|_F^2$, for some $\beta > 0$, as also suggested in Thrampoulidis et al. (2022). The global solutions of the two objectives have a one-to-one correspondence, differing only by a proper scaling.*

4 CS-SVM GEOMETRIES

In this section, we characterize the global minimizers $(\mathbf{W}^*, \mathbf{H}^*)$ of the non-convex programs in (4a) and (4b). 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 STEP-imbalanced setting. In this case, it is reasonable to assume (and we do so) that $\boldsymbol{\delta}$ also shares this STEP structure.

Definition 1 ((R,ρ)-STEP imbalance and STEP logit adjustment)**.** *In a setting with imbalance ratio $R \geq 1$ and minority fraction $\rho \in (0,1)$, an (R,ρ)-STEP imbalanced dataset has ρk minority classes with n_{min} samples each, and $\bar{\rho}k = (1-\rho)k$ majority classes with Rn_{min} samples. For STEP logit adjustment, the hyperparameter vector $\boldsymbol{\delta}$ shares this step structure: for majorities $\delta_c = \delta_{\text{maj}} > 0$ and for minorities $\delta_c = \delta_{\text{minor}} > 0$.*

Our results about CDT/LDT describe the geometry of the CS-SVM solutions in terms of an encoding matrix $\hat{\mathbf{Z}}$, which we call $(\boldsymbol{\delta}, R)$-SEL matrix and define below together with its SVD.

Definition 2 ($(\boldsymbol{\delta}, R)$-SEL matrix)**.** *For hyperparameters $\boldsymbol{\delta} \in \mathbb{R}_+^k$, minority fraction ρ ($\bar{\rho} := 1-\rho$), and k number of classes, define $\boldsymbol{\Xi} \in \mathbb{R}^{k\times k}$ such that $\forall c, j \in [k]$,*

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

Then, for a rational imbalance ratio $R \in \mathbb{Q}_+$,[2] the $(\boldsymbol{\delta}, R)$-Simplex-Encoding Label (SEL) matrix $\hat{\mathbf{Z}} \in \mathbb{R}^{k\times n}$ with $n := \alpha k(R\bar{\rho} + \rho)$ is defined as,

$$\hat{\mathbf{Z}} = \left[\boldsymbol{\Xi}_{1:\bar{\rho}k} \otimes \mathbb{1}_{\alpha R}^T \quad \boldsymbol{\Xi}_{(\bar{\rho}k+1):k} \otimes \mathbb{1}_{\alpha}^T\right], \quad (5)$$

where $\alpha \in \mathbb{N}$ is such that αR is an integer. Further let

$$\hat{\mathbf{Z}} = \mathbf{V}\boldsymbol{\Lambda}\left[\mathbf{U}_{1:\bar{\rho}k}^T \otimes \mathbb{1}_{\alpha R}^T \quad \mathbf{U}_{(\bar{\rho}k+1):k}^T \otimes \mathbb{1}_{\alpha}^T\right], \quad (6)$$

be the compact SVD of $\hat{\mathbf{Z}}$, where $\boldsymbol{\Lambda} \in \mathbb{R}^{(k-1)\times(k-1)}$ is a positive diagonal matrix and $\mathbf{U} \in \mathbb{R}^{k\times(k-1)}$, $\mathbf{V} \in \mathbb{R}^{k\times(k-1)}$ have orthonormal columns.

The pattern of the $(\boldsymbol{\delta}, R)$-SEL matrix $\hat{\mathbf{Z}}$ is clearly determined by the imbalance ratio R and the hyperparameters $\boldsymbol{\delta}$. However, it also depends on the number of classes k and the minority ratio ρ. We choose to drop the latter dependence from the name (δ, R)-SEL since our results focus on $R, \boldsymbol{\delta}$ and k, ρ are easily understood from context. 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 Thrampoulidis et al. (2022) calls the SEL matrix and shows that it characterizes the implicit geomtery

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

of the CE loss for imbalanced data. Our definition is strictly more general allowing us to describe the implicit geometry learned by CDT/LDT losses. We gather useful properties about the eigen-structure of $\hat{\mathbf{Z}}$ in Sec. B. Here, we note that $\hat{\mathbf{Z}}^T \operatorname{diag}(\boldsymbol{\delta})^{-1}\mathbb{1}_k = 0$. Thus, $\operatorname{rank}(\hat{\mathbf{Z}}) = k-1$. The $(\boldsymbol{\delta}, R)$-SEL matrix and its SVD induce a geometry, which is central to our results and we define it next.

Definition 3 ($(\boldsymbol{\delta}, R)$-SELI geometry)**.** *Consider a $(\boldsymbol{\delta}, R)$-SEL matrix $\hat{\mathbf{Z}}$, with SVD factors $\mathbf{U}$, $\boldsymbol{\Lambda}$ and $\mathbf{V}$ as defined in (6). 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:*

(i) $\mathbf{W}^T\mathbf{W} \propto \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$, **(ii)** $\mathbf{M}^T\mathbf{M} \propto \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, **(iii)** $\mathbf{W}^T\mathbf{M} \propto \mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^T = \boldsymbol{\Xi}$.

The first two statements characterize the relative norms and pair-wise angles of classifiers and mean embeddings, respectively. The third statement determines the relative margins between classes. The characterization is in terms of the SVD factors of an appropriate SEL-type encoding matrix. In Sec. D, we derive closed-form expressions for the norms, angles and margins as a function of $R, k, \boldsymbol{\delta}$ by explicitly computing the SVD factors of $\hat{\mathbf{Z}}$. Setting $(\boldsymbol{\delta} = \mathbb{1}_k, R)$ recovers the SELI geometry (Thrampoulidis et al., 2022), and $(\boldsymbol{\delta} = \mathbb{1}_k, R = 1)$ the ETF geometry (Papyan et al., 2020). We are now ready to state our main result. See Sec. C for proofs.

Theorem 1. *Suppose $d \geq k-1$ and (R, ρ)-STEP imbalance setting with STEP logit adjustments. Let $(\mathbf{W}^*, \mathbf{H}^*)$ be any minimizers of either (4a) and (4b), and $\mathbf{M}^*$ be the optimal class-wise mean-embeddings. Then, the following statements are true:*

(i) **[NC]** *All embeddings collapse to their class means, i.e., $\forall i \in [n]$ it holds that $\mathbf{h}_i^* = \boldsymbol{\mu}_{y_i}^*$.*

(ii) **[CDT (4a)]** *For CDT, $(\mathbf{W}^*, \mathbf{M}^*)$ follow the $(\boldsymbol{\delta}, R)$-SELI geometry.*

(iii) **[LDT (4b)]** *For LDT, $(\mathbf{W}^*, \mathbf{M}^* \operatorname{diag}(\boldsymbol{\delta}))$ follow the $(\mathbb{1}_k, \tilde{R})$-SELI geometry, where $\tilde{R} := R\big(\delta_{\text{minor}}/\delta_{\text{maj}}\big)^2$, provided $\tilde{R} \in \mathbb{Q}_+$.*[3]

Thm. 1 describes the geometry of both classifiers and embeddings that correspond to solutions of the non-convex CS-SVM for either CDT or LDT. Statement (i) shows that all optimal embeddings within the same class are equal. Thus, to analyze their geometry, it suffices to study their respective class means, which we arrange as columns of $\mathbf{M}^*$. Statements (ii) and (iii) describe the optimal classifiers and mean-embeddings in terms of the geometry in Defn. 3. Hence, we can find the angles and norms (up to a constant) of the classifiers/embeddings. It is also easy to see that the geometry only depends on the ratio $\Delta := \delta_{\text{maj}}/\delta_{\text{minor}}$ and not on the absolute magnitude of the hyperparameters.

When $\boldsymbol{\delta} = \mathbb{1}_k$, i.e., when the model is trained by CE loss, both statements (ii) and (iii) reduce to the SELI geometry of Thrampoulidis et al. (2022). Further assuming $R = 1$ (i.e., a balanced training set), recovers the ETF geometry (Papyan et al., 2020). 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. 2.

Angles and Norms. Expressing the geometry of the optimal solutions in terms of Defn. 3, enables us to derive explicit closed-form expressions for the angles between individual classifiers and embeddings, as well as, their norms. For example, the norm ratio for the classifiers is given by Eqn. (1). As an example for angle formulas, we can show for any R and Δ that:

$$\text{CDT: } \cos(\mathbf{w}_{\min}, \mathbf{w}'_{\min}) = \frac{-2 + 2\sqrt{R+1}\left(\sqrt{1+\Delta^2}\right)^{-3}}{k - 2 + 2\sqrt{R+1}\left(\sqrt{1+\Delta^2}\right)^{-3}},$$

$$\text{LDT: } \cos(\mathbf{w}_{\min}, \mathbf{w}'_{\min}) = \frac{-2\Delta + \sqrt{(R+\Delta^2)/2}}{(k-2)\Delta + \sqrt{(R+\Delta^2)/2}}. \tag{7}$$

See Sec. D for the complete list of closed-form formulas, all derived thanks to Thm. 1. Such explicit formulas allow studying optimal tunings and interesting asymptotics as R increases. We show these next.

Special tunings. We emphasize two notable special cases of geometries that arise respectively for LDT and CDT when setting $\delta_c = \sqrt{n_c} \Leftrightarrow \Delta = \sqrt{R}$.

Corollary 1.1 (Achieving alignment with CDT)**.** *In (4a), set $\Delta = \sqrt{R}$. Then, $\cos(\mathbf{w}_{y_i}^*, \mathbf{h}_i^*) = 1, \forall i \in [n]$, i.e., each feature embedding $\mathbf{h}_i^*$ perfectly aligns with its corresponding classifier $\mathbf{w}_{y_i}^*$.*

This results from the angle calculations detailed in Sec. D. While this simple tuning leads to perfect alignment of classifiers and mean-embeddings geometries, it does not guarantee equal norms or maximal separation. Thus, the geometry is in general still different from the ETF geometry for balanced data. In contrast, we show next that under the same tuning the implicit geometry of the LDT is an ETF modulo the scaling of the embeddings.

Corollary 1.2 (Achieving ETF with LDT)**.** *In (4b), set $\Delta = \sqrt{R}$. Then, $(\mathbf{W}^*, \mathbf{M}^* \operatorname{diag}(\boldsymbol{\delta}))$ follows the ETF geometry.*

Cor. 1.2 follows immediately from Thm. 1 by noting that $\delta_c = \sqrt{n_c}$ yields $\tilde{R} = 1$ and the $(\mathbb{1}_k, 1)$-SELI geometry coincides with the ETF geometry. This implies that classifiers and embeddings are perfectly aligned,

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

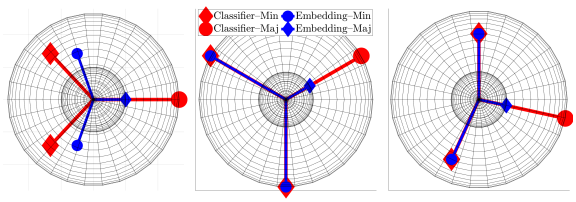

Figure 2: Geometries induced by CE, LDT and CDT for the respective CS-SVM minimizers. $k = 3$ with 2 minority and 1 majority classes, imbalance ratio $R = 10$ and hyperparameters ratio $\Delta = \delta_{\text{maj}}/\delta_{\text{minor}} = \sqrt{R}$. See Cors. 1.1 and 1.2.

but also all classifiers $\mathbf{w}_c^*, c \in [k]$ have equal norms, and both the classifiers and embeddings are maximally separated, i.e., $\cos(\mathbf{w}_c^*, \mathbf{w}_{c'}^*) = \cos(\mathbf{h}_c^*, \mathbf{h}_{c'}^*) = -1/(k-1)$. Notably, this holds irrespective of the imbalance ratio R. See Fig. 2 for the visualization.

Mitigating Minority Collapse. (Fang et al., 2021) discovered that when $R \to \infty$, the minority classifiers collapse, i.e., $\cos(\mathbf{w}_{\text{minor}}, \mathbf{w}'_{\text{minor}}) \to 1$ for any two minority classes. We show here that CDT and LDT losses can mitigate this effect when appropriately tuned. For this, we simply evaluate our closed-form formulas in (7) for $R \to \infty$. To obtain non-trivial results, we allow the hyperparameter Δ to scale with R, i.e., set $\Delta = R^\gamma$ for constant $\gamma \in \mathbb{R}$. This gives the following two results.

Corollary 1.3 (Mitigating classifier collapse with LDT)**.** *In (4b), set $\Delta = R^\gamma, \gamma \in \mathbb{R}$. Then, as $R \to \infty$ the minority/majority angles satisfy*

$\cos(\mathbf{w}_c, \mathbf{w}'_c)$	$\gamma < 0.5$	$\gamma = 0.5$	$\gamma > 0.5$
$c, c' \in$ minority	1	$-\frac{1}{k-1}$	$\frac{1-2\sqrt{2}}{1+\sqrt{2}(k-2)}$
$c, c' \in$ majority	$\frac{1-2\sqrt{2}}{1+\sqrt{2}(k-2)}$	$-\frac{1}{k-1}$	1

Corollary 1.4 (Mitigating classifier collapse with CDT)**.** *In (4a), set $\Delta = R^\gamma, \gamma \in \mathbb{R}$. Then, as $R \to \infty$ the minority/majority angles satisfy*

$\cos(\mathbf{w}_c, \mathbf{w}'_c)$	$\gamma < 1/6$	$\gamma = 1/6$	$\gamma > 1/6$
$c, c' \in$ minority	1	0	$-\frac{2}{k-2}$
$\cos(\mathbf{w}_c, \mathbf{w}'_c)$	$\gamma < 0$	$\gamma = 0$	$\gamma > 0$
$c, c' \in$ majority	$-\frac{2}{k-2}$	$\frac{1-2\sqrt{2}}{1+\sqrt{2}(k-2)}$	0

From Cor. 1.3, LDT with $\gamma \geq 0.5$ avoids the minority collapse. However, for $\gamma > 0.5$, majority classifiers collapse instead. Thus, we find that $\gamma = 0.5$ the only choice that keeps both majority and minority classifiers from collapsing. In fact, for this choice the angles of majorities and minorities are all equal, as expected by Cor. 1.2. On the other hand, from Cor. 1.4, CDT avoids minority collapse for any choice of $\gamma \geq 1/6$. Also, in this entire range the majority classifiers do not collapse either. Thus, for $R \to \infty$, CDT offers a wide tuning range for γ that avoids classifier collapse. Compare this to the single value of $\gamma = 0.5$ for LDT. Specifically for $\gamma = 0.5$, when classifiers and features are aligned in both CDT and LDT (see Cors. 1.1 and 1.2), the CDT minority angles are larger from the LDT angles since $-2/(k-2) < -1/(k-1)$; see also Fig. 2.

5 NUMERICAL RESULTS

For both CDT and LDT loss, we examine the convergence of the models trained by SGD to the implicit geometry proposed by Thm. 1. We train (i) UFM, (ii) MLP on MNIST, and (iii) ResNet18 on CIFAR10. All the models are trained in a $(R = 10, \rho = 1/2)$-STEP imbalanced setting. We further use STEP logit adjustment, and choose $\Delta = R^\gamma$ with $\gamma \in [-1.5, 1.5]$. We train the UFM by minimizing unregularized CDT/LDT, while for MLP and ResNet models, following the setup in Papyan et al. (2020), we use a small weight-decay (10^{-5}). We defer other experimental details to Sec. E.1.

Fig. 3 illustrates the empirical geometry discovered by SGD vs the prediction of Thm. 1. 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 for 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 closed-form expressions that result from Thm. 1 (see Sec. D). As reported in the figures, the empirical quantities follow the predicted theoretical trends. However, convergence becomes more challenging for the deep-net models, particularly for larger $|\gamma|$. Moreover, we encounter cases with non-zero training error for CDT loss for large $|\gamma|$ values. In addition to γ, the imbalance ratio R also affects the convergence to theory (see Sec. E.2 for details). 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 interesting value of $\gamma = 0.5$, with an exception of the majority classifier angles in the ResNet experiments. The mismatch is less severe for the 6-layer MLP. Also, as predicted by the theorem, for $\gamma = 0.5$ ($\Delta = \sqrt{R}$), the LDT geometry is the ETF, up to a scaling on the features: In Fig. 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.

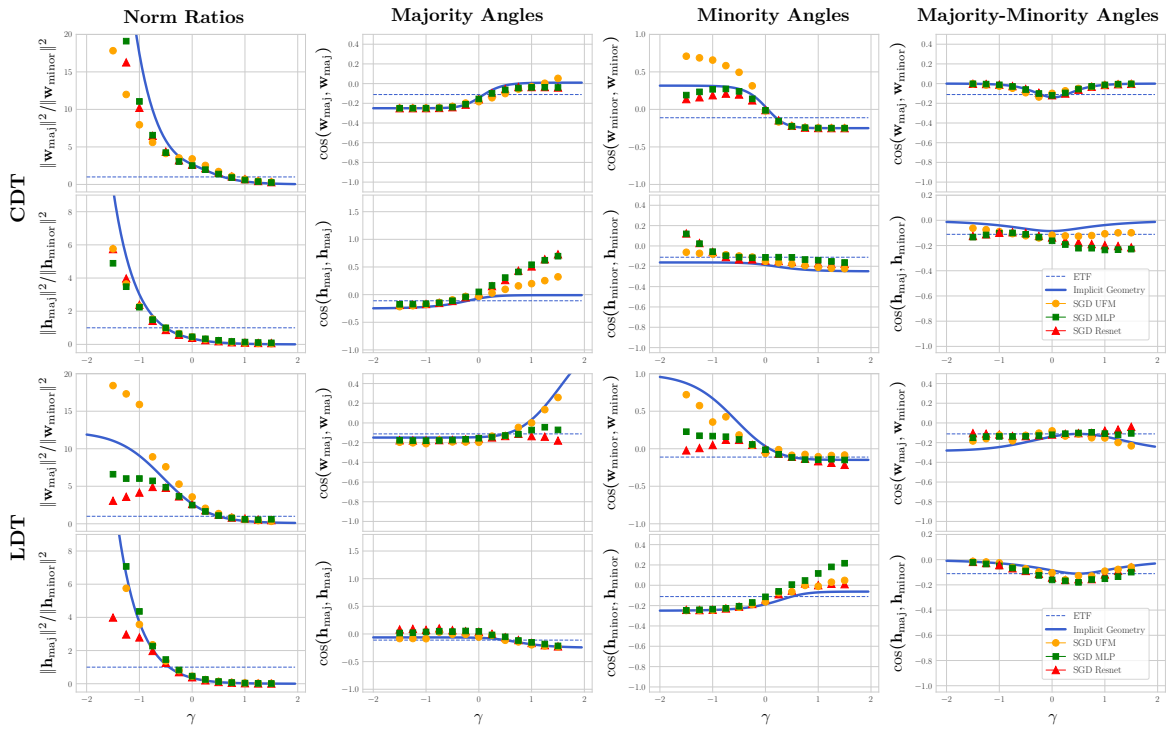

Figure 3: Comparison of models trained by SGD (markers) minimizing the CDT (3a)/LDT (3b) loss and the global minimizers of the CS-SVM in (4a), (4b) as given by Thm. 1 (solid line) in a (10, 1/2)-STEP imbalanced setting. The dashed line marks the ETF geometry (Papyan et al., 2020). See Sec. E.1 for details.

While the experiments in Fig. 3 correspond to a finite imbalance ratio of $R = 10$, there is resemblance to the asymptotic behavior of the classifier angles on LDT-trained UFM. Cor. 1.3 suggests $\gamma = 0.5$ is the only choice for $R \to \infty$ that avoids minority or majority classifiers collapsing. A similar trend is seen in Fig. 3, where the cosine of the minority classifiers goes towards 1 for $\gamma < 0.5$, while that of the majority classifiers approaches 1 for the complementary open interval of $\gamma > 0.5$. On the other hand, CDT does not attain equiangularity, but majority and minority angles are well controlled for a wider range of γ. This suggests that the CDT geometry is more robust to small changes in the hyperparameter γ compared to LDT geometry.

Remark 2. *In all our experiments with CDT and LDT, we center the embeddings before computing norms and angles. This is consistent with centering performed for experiments with balanced data in Papyan et al. (2020); Zhu et al. (2021); Thrampoulidis et al. (2022). In our case, the exact centering vector is different for each loss function (see Sec. D.1.3/D.2.3 for CDT/LDT loss).*

6 CONCLUDING REMARKS

Our paper is motivated by and contributes to two recent thrusts in the literature. The first seeks structural properties of the deep-nets trained far beyond the zero-train error regime (e.g., Papyan et al., 2020; Fang et al., 2021; Galanti et al., 2021; Graf et al., 2021; Han et al., 2021; Hui et al., 2022; Ji et al., 2021; Lu and Steinerberger, 2020; Mixon et al., 2020; Tirer and Bruna, 2022; Xie et al., 2022; Zhu et al., 2021; Zhou et al., 2022a; Thrampoulidis et al., 2022). The second investigates approaches to coping with class imbalances in training overparameterized models (e.g., Byrd and Lipton, 2019; Sagawa et al., 2019, 2020; Cao et al., 2019; Kang et al., 2020; Kim and Kim, 2020; Menon et al., 2020; Ye et al., 2020; Kini et al., 2021; Wang et al., 2021; Jitkrittum et al., 2022). We already discussed some of the most closely related works within each thrust as well as a few recent works (Fang et al., 2021; Xie et al., 2022; Yang et al., 2022) at the intersection (see *Related Work*). The goal of this section is to outline main take-aways of our work in the form of both contributions and limitations, together with some pointer for future directions.

Contributions. We extend the scope of the geometry characterizations of the embeddings and classifiers learned by deep-nets initiated by Papyan et al. (2020). To the best of our knowledge, all prior works study the geometries for either the CE or mean-square loss. Instead, we formulate a more general geometry that describes two alternative CE parameterizations and includes the previous geometries as special cases. Unlike previous works, our new geometry is parameterized in

terms of the loss hyperparameters, thus it involves rich structures (in terms of angles and norm-ratios) as these hyperparameters vary. Yet, like in previous works, the geometry is rather simple to describe, either implicitly in terms of a special encoding matrix or explicitly in terms of closed-form formulas for the angles and norms. We arrive at this new geometry by analyzing the simplified unconstrained features-model (specifically, its cost-sensitive version in Eqns. (4a),(4b)). Thus, we also extend the scope of the UFM model beyond the previously studied CE and square loss. Finally, we undertake an implicit-geometry view to loss modifications for imbalanced learning. Unlike the previously considered implicit-bias view in Byrd and Lipton (2019); Sagawa et al. (2020); Kini et al. (2021); Wang et al. (2021), which is limited to linear (thus, fixed-feature) models and/or binary settings, our approach applies to learned-feature models and multiclass settings.

Limitations. In the spirit of previous works (Papyan et al., 2020; Fang et al., 2021; Thrampoulidis et al., 2022) that our result builds upon, it also shares some of the same limitations. First, the characterizations of the involved geometries are asymptotic in the number of training epochs, and the convergence to the prescribed geometry can be (very) slow. The specific convergence behavior that we see for CDT/LDT losses is of similar nature to the CE loss in Papyan et al. (2020); Zhu et al. (2021); Thrampoulidis et al. (2022). For CDT/LDT losses, we also observe that convergence speed can vary significantly for different hyperparameter values. This issue appears already for the UFM itself and is consistent for deep-nets and complex data (see Sec. E.2). Second, the level of convergence reached in realistic training settings generally varies between architectures, data models and the loss function. For example, we find that CDT classifier geometry converges very well to its prescribed limit, but the same is not true for the CDT embeddings geometry or for the LDT classifiers geometry. Consistently, the experiments in Papyan et al. (2020) show different levels of convergence between different metrics (e.g., classifiers vs embeddings, norms vs angles) and different architectures/datasets. Third, there is no explicit known link between implicit geometry and generalization. This is one of the most pressing questions in the emerging literature and we expect more investigations to follow.[4] Finally, like Thrampoulidis et al. (2022), we also utilize findings of Lyu and Li (2019) on gradient flow convergence in homogeneous networks towards the KKT points of appropriate CS-SVM problems. Although we focus on the UFM, which falls under the homogeneous network category, exploring more intricate models (possibly using ideas from (Le and Jegelka, 2022; Jacot, 2022)) could provide insight into other aspects of deep-net training such as the inferior convergence of embeddings.

Outlook and future directions. While it is important to realize these shortcomings, it is equally important realizing that the quest for implicit geometries is by nature highly non-trivial: we seek geometry characterizations for classifiers and mean-embeddings that are learned by different complex deep architectures over different complex datasets. In our setting, we further have different losses (LDT vs CDT), different loss hyperparameters, and different imbalance ratios. Paraphrasing Papyan et al. (2020): one might anticipate that the classifier and embeddings being by-product of training in such complex environments display no underlying structure. In view of these, 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. For example, see first row of Fig. 3. Similarly, inspecting Fig. 1, why should one expect a priori that there is a simple formula parameterized by the loss hyperparameters that captures the norm-ratio of the classifiers learned by an MLP on MNIST and a ResNet on CIFAR10? In view of these, we deem our findings encouraging and supportive of the quest set by the emerging literature on such structural characterizations. At the same time, our findings are suggestive of several important research directions. First, while the UFM has proven powerful to be predictive of behaviors across different levels of imbalances and different losses, a major limitation remains that it does not capture the required centering needed for the embeddings (see Remark 2). This is a common theme also in previous works and is further highlighted here since in the new geometries the "correct" centering, done at a heuristic level in our experiments, is more intricate as it involves scaling with hyperparameter values. Second, while we characterize global minima of the CS-SVMs, it is not yet known whether SGD converges to those minima in our settings. Third, is it possible to speed up training for faster convergence to the asymptotic limits? Finally, more investigations are required both on theory and experiments to distill connections between geometries and generalization. We hope that some of our findings motivate further such investigations, which are otherwise beyond the scope of this paper.

[4] Initial findings and further discussion on generalization are included in the extended version of this paper, which can be accessed publicly on arXiv.

References

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

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. (2019). Learning imbalanced datasets with

label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1567–1578.

Fang, C., He, H., Long, Q., and Su, W. J. (2021). Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43).

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

Galanti, T., György, A., and Hutter, M. (2022). Generalization bounds for transfer learning with pretrained classifiers. *arXiv preprint arXiv:2212.12532*.

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

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

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

Jacot, A. (2022). Implicit bias of large depth networks: a notion of rank for nonlinear functions. *arXiv preprint arXiv:2209.15055*.

Ji, W., Lu, Y., Zhang, Y., Deng, Z., and Su, W. J. (2021). An unconstrained layer-peeled perspective on neural collapse. *arXiv preprint arXiv:2110.02796*.

Ji, Z. and Telgarsky, M. (2018). Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*.

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

Jitkrittum, W., Menon, A. K., Rawat, A. S., and Kumar, S. (2022). Elm: Embedding and logit margins for long-tail learning. *arXiv preprint arXiv:2204.13208*.

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

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., and Togneri, R. (2017). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587.

Kim, B. and Kim, J. (2020). Adjusting decision boundary for class imbalanced learning. *IEEE Access*, 8:81674–81685.

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

Le, T. and Jegelka, S. (2022). Training invariances and the low-rank phenomenon: beyond linear networks. *arXiv preprint arXiv:2201.11968*.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2018). Focal loss for dense object detection.

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

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

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

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

Papyan, V., Han, X., and Donoho, D. L. (2020). Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. (2020). An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. (2018). The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878.

Thrampoulidis, C., Kini, G. R., Vakilian, V., and Behnia, T. (2022). Imbalance trouble: Revisiting neural-collapse geometry. *arXiv preprint arXiv:2208.05512*.

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

Wang, K. A., Chatterji, N. S., Haque, S., and Hashimoto, T. (2021). Is importance weighting incompatible with interpolating classifiers? *arXiv preprint arXiv:2112.12986*.

Xie, L., Yang, Y., Cai, D., Tao, D., and He, X. (2022). Neural collapse inspired attraction-repulsion-balanced loss for imbalanced learning. *arXiv preprint arXiv:2204.08735*.

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

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

Zhou, J., Li, X., Ding, T., You, C., Qu, Q., and Zhu, Z. (2022a). On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. *arXiv preprint arXiv:2203.01238*.

Zhou, J., You, C., Li, X., Liu, K., Liu, S., Qu, Q., and Zhu, Z. (2022b). Are all losses created equal: A neural collapse perspective. *arXiv preprint arXiv:2210.02192*.

Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., and Qu, Q. (2021). A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34.

Contents

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. $\mathbf{V}_{j:k} \in \mathbb{R}^{m\times(k-j+1)}$ chooses columns $j, j+1, \dots, k$ of $\mathbf{V}$, and $\mathbf{V}^T_{j:k} \in \mathbb{R}^{n\times(k-j+1)}$ does so on $\mathbf{V}^T$. 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 $\mathbf{V}$. We use $\mathbf{V} \propto \mathbf{X}$ whenever the two matrices are equal up to a scalar constant. For a vector $\mathbf{v} \in \mathbb{R}^k$, $\mathrm{diag}(\mathbf{v}) \in \mathbb{R}^{k\times k}$ is the diagonal matrix with $\mathbf{v}$ on its diagonal. $\odot$ and $\otimes$ denote Hadammard and Kronecker products, respectively. 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. $\mathbf{e}_j$ is the j-th standard basis vector, a column vector with a single non-zero entry of 1 in the j-th entry. Finally, we denote the set of positive rational numbers by $\mathbb{Q}_+$.

A PROOF OF LEMMA 3.1

Lemma A.1 (Binary)**.** *Consider $k = 2$, linear model, separable data and minimizing un-regularized LDT/CDT/binary-CE/binary-VS losses. The LDT rule coincides with the classification rule of the binary VS loss assuming same δ-tuning. On the other hand, minimizing CDT results in the same classification rule as CE, irrespective of the δ-tuning.*

Proof. Let $\mathbf{W}^{\mathrm{CDT}}, \mathbf{W}^{\mathrm{LDT}} \in \mathbb{R}^{d\times 2}$ denote the CDT and LDT classifiers respectively. The corresponding classification rules is: $(\mathbf{w}_1 - \mathbf{w}_2)^T\mathbf{x} \underset{\hat{y}(\mathbf{W})=2}{\overset{\hat{y}(\mathbf{W})=1}{\gtrless}} 0$ for $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2]$ either $\mathbf{W}^{\mathrm{CDT}}$ or $\mathbf{W}^{\mathrm{LDT}}$, respectively. On the other hand, the CE or binary VS loss decision rule is $\mathbf{x}^T\mathbf{w}_* \underset{\hat{v}(\mathbf{w}_*)=-1}{\overset{\hat{v}(\mathbf{w}_*)=1}{\gtrless}} 0$, where $\mathbf{w}_*$ denotes a minimizer of either the CE or the binary VS loss. Here, we use $v \in \{\pm 1\}$ to denote the label encoding for binary CE loss, differentiating from the multiclass encoding $y \in \{1,2\}$ above. From the above two, we conclude that $\mathbf{w}_\star = \alpha(\mathbf{w}_1 - \mathbf{w}_2), \alpha > 0$ implies $\hat{y}(\mathbf{W}) = 1 \iff \hat{v}(\mathbf{w}_*) = 1$ (eqv. $\hat{y}(\mathbf{W}) = 2 \iff \hat{v}(\mathbf{w}_*) = -1$).

Since we minimize all losses without regularization and data are separable, it suffices by implicit bias (Soudry et al., 2018; Kini et al., 2021) to consider the solutions to the corresponding max-margin problems, i.e.,

$$\mathbf{W}^{\mathrm{CDT}} := \arg\min_{\mathbf{W}} \|\mathbf{W}\|_F^2 \quad \text{subj. to} \quad (\delta_{y_i}\mathbf{w}_{y_i} - \delta_c\mathbf{w}_c)^T\mathbf{h}_i \geq 1, c \neq y_i \tag{8a}$$

$$\mathbf{W}^{\mathrm{LDT}} := \arg\min_{\mathbf{W}} \|\mathbf{W}\|_F^2 \quad \text{subj. to} \quad \delta_{y_i}(\mathbf{w}_{y_i} - \mathbf{w}_c)^T\mathbf{h}_i \geq 1, c \neq y_i \tag{8b}$$

$$\mathbf{w}_\star^{\mathrm{CE,binary}} := \arg\min_{\mathbf{w}} \|\mathbf{w}\|_2^2 \quad \text{subj. to} \quad \upsilon_i\mathbf{w}_\star^T\mathbf{h}_i \geq 1, \tag{8c}$$

$$\mathbf{w}_\star^{\mathrm{VS,binary}} := \arg\min_{\mathbf{w}} \|\mathbf{w}\|_2^2 \quad \text{subj. to} \quad \upsilon_i\delta_{\upsilon_i}\mathbf{w}_\star^T\mathbf{h}_i \geq 1, \tag{8d}$$

First, we show $\hat{y}(\mathbf{W}^{\mathrm{LDT}}) = 1 \iff \hat{v}(\mathbf{w}_*^{\mathrm{VS,binary}}) = 1$ provided that the LDT and VS loss parameters are matching, i.e. $\delta_1^{\mathrm{LDT}} = \delta_1^{\mathrm{VS,binary}}$ and $\delta_2^{\mathrm{LDT}} = \delta_{-1}^{\mathrm{VS,binary}}$. This follows from the fact that $\mathbf{w}_1^{\mathrm{LDT}} + \mathbf{w}_2^{\mathrm{LDT}} = 0$ (see Lem. A.2). Thus, the minimization in (8b) does not change by adding the constraint $\mathbf{w}_2 = -\mathbf{w}_1$. But then, the solution set of (8b) is the same as the solution set of the minimization

$$\min_{\mathbf{w}_1} \|\mathbf{w}_1\|^2 \quad \text{subj. to} \quad \begin{cases} 2\delta_1\mathbf{w}_1^T\mathbf{h}_i \geq 1 & i : y_i = 1 \\ -2\delta_2\mathbf{w}_1^T\mathbf{h}_i \geq 1 & i : y_i = 2 \end{cases}.$$

Comparing this to (8d), it follows immediately that $\mathbf{w}_1^{\mathrm{LDT}} = \mathbf{w}_\star^{\mathrm{VS,binary}}/2$. Hence, $\mathbf{w}_1^{\mathrm{LDT}} - \mathbf{w}_2^{\mathrm{LDT}} = \mathbf{w}_\star^{\mathrm{VS,binary}}$, which proves the desired.

Second, we show that $\hat{y}(\mathbf{W}^{\mathrm{CDT}}) = \hat{y}(\mathbf{w}_*^{\mathrm{CE}})$. This is a consequence of the fact that $\delta_1^{-1}\mathbf{w}_1^{\mathrm{CDT}} + \delta_2^{-1}\mathbf{w}_2^{\mathrm{CDT}} = 0$ (see Lem. A.2). Indeed, we then have that the solution set of (8a) does not change by adding the constraint $\mathbf{w}_2 = -(\delta_2/\delta_1)\mathbf{w}_1$. But then, optimization is equivalent to:

$$\min_{\mathbf{w}_1} \|\mathbf{w}_1\|^2 \quad \text{subj. to} \quad \begin{cases} (\delta_1\mathbf{w}_1 - \delta_2\mathbf{w}_2)^T\mathbf{h}_i = (\delta_1 + \delta_2^2/\delta_1)\mathbf{w}_1^T\mathbf{h}_i \geq 1 & i : y_i = 1 \\ (\delta_2\mathbf{w}_2 - \delta_1\mathbf{w}_1)^T\mathbf{h}_i = -(\delta_1 + \delta_2^2/\delta_1)\mathbf{w}_1^T\mathbf{h}_i \geq 1 & i : y_i = 2 \end{cases}.$$

Comparing this to (8c), we find that $\mathbf{w}_1^{\mathrm{CDT}} = \frac{\delta_1}{\delta_1^2+\delta_2^2}\mathbf{w}_\star^{\mathrm{CE,binary}}$. Thus also, $\mathbf{w}_2^{\mathrm{CDT}} = -\frac{\delta_2}{\delta_1^2+\delta_2^2}\mathbf{w}_\star^{\mathrm{CE,binary}}$. In conclusion, $\mathbf{w}_1^{\mathrm{CDT}} - \mathbf{w}_2^{\mathrm{CDT}} = \frac{\delta_1+\delta_2}{\delta_1^2+\delta_2^2}\mathbf{w}_\star^{\mathrm{CE,binary}}$, from which the desired follows since $\delta_1, \delta_2 > 0$.

□

Lemma A.2. *For the CDT/LDT-SVM classifiers $\mathbf{W}^{CDT}, \mathbf{W}^{LDT}$ defined in (8b) and (8a), it holds that $\mathbf{w}_1^{LDT} + \mathbf{w}_2^{LDT} = 0$ and $\delta_1^{-1}\mathbf{w}_1^{CDT} + \delta_2^{-1}\mathbf{w}_1^{CDT} = 0$.*

Proof. We prove the claim for CDT. The proof for LDT is the same and is omitted for brevity. We use a symmetrization argument as follows. Set

$$\bar{\mathbf{w}} := (\delta_1^{-1}\mathbf{w}_1^{\mathrm{CDT}} + \delta_2^{-1}\mathbf{w}_2^{\mathrm{CDT}})/(\delta_1^{-2} + \delta_2^{-2}),$$

and assume for the sake of contradiction that $\bar{\mathbf{w}} \neq 0$. Consider a new classifier defined as $\widetilde{\mathbf{w}}_1 = \mathbf{w}_1^{\mathrm{CDT}} - \delta_1^{-1}\bar{\mathbf{w}}$ and $\widetilde{\mathbf{w}}_2 = \mathbf{w}_2^{\mathrm{CDT}} - \delta_2^{-1}\bar{\mathbf{w}}$. Clearly, it holds that $\delta_1\widetilde{\mathbf{w}}_1 - \delta_2\widetilde{\mathbf{w}}_2 = \delta_1\mathbf{w}_1^{\mathrm{CDT}} - \delta_2\mathbf{w}_2^{\mathrm{CDT}}$. Thus, $[\widetilde{\mathbf{w}}_1, \widetilde{\mathbf{w}}_2]$ is feasible in (8b). Moreover,

$$\|\widetilde{\mathbf{w}}_1\|_2^2 + \|\widetilde{\mathbf{w}}_2\|_2^2 = \|\mathbf{w}_1^{\mathrm{CDT}}\|_2^2 + \|\mathbf{w}_2^{\mathrm{CDT}}\|_2^2 - (\delta_1^{-2} + \delta_2^{-2})\|\bar{\mathbf{w}}\|^2 < \|\mathbf{w}_1^{\mathrm{CDT}}\|_2^2 + \|\mathbf{w}_2^{\mathrm{CDT}}\|_2^2.$$

But, these together contradict the optimality of $\mathbf{W}^{\mathrm{CDT}}$. □

B EIGEN-STRUCTURE OF THE (δ, R)-SEL MATRIX

In this section, we compute the eigen-structure of $(\boldsymbol{\delta}, R)$-SEL matrix $\hat{\mathbf{Z}}$ (Defn. 2) for a $(\boldsymbol{\delta}, R)$-STEP imbalanced setting with STEP logit adjustments. For simplicity, we let $\delta_{\mathrm{minor}} = 1$, $\delta_{\mathrm{maj}} = \Delta$ and $\alpha = 1$ (i.e. $R \in \mathbb{N}$).[5] 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 $\mathbf{S}_m(\sigma) := \mathbb{I}_m - \sigma\mathbb{1}_m\mathbb{1}_m^T \in \mathbb{R}^m$. Throughout the rest of the paper, we let $\mathbf{U}_\otimes = \begin{bmatrix}\mathbf{U}_{1:\bar{\rho}k}^T\otimes\mathbb{1}_{\alpha R}^T & \mathbf{U}_{(\bar{\rho}k+1):k}^T\otimes\mathbb{1}_\alpha^T\end{bmatrix}^T$.

Lemma B.1 ($(\boldsymbol{\delta}, R)$-SEL matrix SVD)**.** *Let $R \in \mathbb{N}$ and $\hat{\mathbf{Z}} \in \mathbb{R}^{k\times n}$ be the $(\boldsymbol{\delta}, R)$-SEL matrix described in Defn. 2, where recall that $n = k(R\bar{\rho} + \rho)$. Define the SVD of $\hat{\mathbf{Z}}$ as follows,*

$$\hat{\mathbf{Z}} = \mathbf{V}\boldsymbol{\Lambda}\begin{bmatrix}\mathbf{U}_{1:\bar{\rho}k}^T\otimes\mathbb{1}_R^T & \mathbf{U}_{(\bar{\rho}k+1):k}^T\end{bmatrix} =: \mathbf{V}\boldsymbol{\Lambda}\mathbf{U}_\otimes^T,$$

and further let $\mathbf{V} = [\mathbf{V}_{\mathrm{maj}}, \mathbf{v}, \mathbf{V}_{\mathrm{min}}]$ and $\mathbf{U} = [\mathbf{U}_{\mathrm{maj}}, \mathbf{u}, \mathbf{U}_{\mathrm{min}}]$. Then, the SVD factors are given by the following equations:

$$\boldsymbol{\Lambda} = \mathrm{diag}\left(\begin{bmatrix}\frac{\sqrt{R}}{\Delta}\mathbb{1}_{(\bar{\rho}k-1)}^T & \sqrt{\frac{\bar{\rho}+R\rho}{\bar{\rho}+\rho\Delta^2}} & \mathbb{1}_{(\rho k-1)}^T\end{bmatrix}\right), \tag{9}$$

$$\mathbf{V}_{maj} = \begin{bmatrix}\mathbb{P}_{\bar{\rho}k} \\ 0_{(\rho k)\times(\bar{\rho}k-1)}\end{bmatrix} \quad \mathbf{v} = \frac{1}{\sqrt{k(\bar{\rho}+\rho\Delta^2)}}\begin{bmatrix}-\Delta\sqrt{\frac{\rho}{\bar{\rho}}}\mathbb{1}_{\bar{\rho}k} \\ \sqrt{\frac{\bar{\rho}}{\rho}}\mathbb{1}_{\rho k}\end{bmatrix} \quad \mathbf{V}_{min} = \begin{bmatrix}0_{(\bar{\rho}k)\times(\rho k-1)} \\ \mathbb{P}_{\rho k}\end{bmatrix}, \tag{10}$$

$$\mathbf{U}_{maj} = \begin{bmatrix}\frac{1}{\sqrt{R}}\mathbb{P}_{\bar{\rho}k} \\ 0_{(\rho k)\times(\bar{\rho}k-1)}\end{bmatrix} \quad \mathbf{u} = \frac{1}{\sqrt{k(\bar{\rho}+R\rho)}}\begin{bmatrix}-\sqrt{\frac{\rho}{\bar{\rho}}}\mathbb{1}_{\bar{\rho}k} \\ \sqrt{\frac{\bar{\rho}}{\rho}}\mathbb{1}_{\rho k}\end{bmatrix} \quad \mathbf{U}_{min} = \begin{bmatrix}0_{(\bar{\rho}k)\times(\rho k-1)} \\ \mathbb{P}_{\rho k}\end{bmatrix}. \tag{11}$$

Proof. To prove the lemma, we only need to verify the correctness of the formulas. In particular: (1) $\mathbf{U}_\otimes$ and $\mathbf{V}$ are unitary matrices, and (2) $\mathbf{V}\boldsymbol{\Lambda}\mathbf{U}_\otimes^T = \hat{\mathbf{Z}}$. By recalling that $\mathbb{P}_m^T\mathbb{P}_m = \mathbb{I}_{m-1}$ and $\mathbb{P}_m^T\mathbb{1}_m = 0$ for $m \in \{\rho k, \bar{\rho}k\}$, it is easy to confirm $\mathbf{V}^T\mathbf{V} = \mathbb{I}_{k-1}$ and $\mathbf{U}_\otimes^T\mathbf{U}_\otimes = \mathbb{I}_{k-1}$. Since $\mathbf{U}_\otimes$ and $\hat{\mathbf{Z}}$ have the same pattern of repeated columns, proving $\mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^T = \boldsymbol{\Xi}$ verifies the decomposition. So, we start by expressing $\boldsymbol{\Xi}$ in block-form as follows:

$$\boldsymbol{\Xi} = \begin{bmatrix}\Delta^{-1}\mathbf{S}_{\bar{\rho}k}\left(\frac{1}{k(\bar{\rho}+\rho\Delta^2)}\right) & -\frac{\Delta}{k(\bar{\rho}+\rho\Delta^2)}\mathbb{1}_{\bar{\rho}k}\mathbb{1}_{\rho k}^T \\ -\frac{1}{k(\bar{\rho}+\rho\Delta^2)}\mathbb{1}_{\rho k}\mathbb{1}_{\bar{\rho}k}^T & \mathbf{S}_{\rho k}\left(\frac{\Delta^2}{k(\bar{\rho}+\rho\Delta^2)}\right)\end{bmatrix}. \tag{12}$$

[5] To relax these assumptions, we only need to change the scale of the eigen-factors. Particularly, singular values should be scaled by $\sqrt{\alpha}/\sqrt{\delta_{\mathrm{minor}}}$, and $\mathbf{U}$ by $1/\sqrt{\alpha}$. Thus, the results easily extend for general δ_{minor} and α, i.e., rational R.

Now, we can verify the equation by direct calculations:

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

□

With the eigen-structure of $\hat{\mathbf{Z}}$ at hand, we prove a useful property of the singular space in Lem. B.2. We will use this property later in Sec. C to characterize the solutions of the CS-SVM corresponding to CDT loss in (4a).

Lemma B.2. *Recall the setting of Lem. B.1 and the SVD* $\hat{\mathbf{Z}} = \mathbf{V}\mathbf{\Lambda}\mathbf{U}_{\otimes}{}^T$. *The matrix* $\mathbf{B}^* = \mathbf{U}_{\otimes}\mathbf{V}^T$ *satisfies the following element-wise* strict *inequalities:* $\mathbf{B}^* \odot \hat{\mathbf{Z}}^T > 0$.

Proof. We compute $\mathbf{B}^* := \begin{bmatrix} \mathbf{B}_{11}^* & \mathbf{B}_{12}^* \\ \mathbf{B}_{21}^* & \mathbf{B}_{22}^* \end{bmatrix}$ by plugging in the explicit SVD expressions in Lem. B.1.

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

Simplifying the expressions, we have

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

From (12), we can write $\hat{\mathbf{Z}}$ in block-form:

$$\hat{\mathbf{Z}}^T = \begin{bmatrix} \Delta^{-1}\mathbf{S}_{\overline{\rho}k}\big(\frac{1}{k(\overline{\rho}+\rho\Delta^2)}\big) \otimes \mathbb{1}_R & -\frac{1}{k(\overline{\rho}+\rho\Delta^2)}\mathbb{1}_{\overline{\rho}k}\mathbb{1}_{\rho k}^T \otimes \mathbb{1}_R \\ -\frac{\Delta}{k(\overline{\rho}+\rho\Delta^2)}\mathbb{1}_{\rho k}\mathbb{1}_{\overline{\rho}k}^T & \mathbf{S}_{\rho k}\big(\frac{\Delta^2}{k(\overline{\rho}+\rho\Delta^2)}\big) \end{bmatrix}.$$

The signs of the off-diagonal blocks of both $\hat{\mathbf{Z}}$ and $\mathbf{B}^*$ are negative. To inspect the sign agreement of the on-diagonal blocks, it is enough to see the following inequalities are always strictly satisfied,

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

□

C PROOF OF THEOREM 1

One of the paper's main contributions is introducing the $(\boldsymbol{\delta}, R)$-SELI geometry (Defn. 3) as the "correct" formalization that is able to capture the implicit geometries of *both* the CDT and LDT losses for all imbalance-ratio values R.[6] This property is captured by Thm. 1: thanks to the generality of Defn. 3, both CDT and LDT

[6]Since CE loss is a special case of CDT/LDT loss for $\boldsymbol{\delta} = \mathbb{1}_k$, the new geometry includes the previously introduced SELI (Thrampoulidis et al., 2022) and ETF (Papyan et al., 2020) geometries as special cases.

geometries, albeit different to each other, are formalized in terms of appropriate parameterizations of the same geometry. This unifying and concise formalization of the theorem is central to our work. For example, the eigenstructure properties of the (δ, R)-SEL matrix in Sec. B and the closed-form angles/norm-formulas in Sec. D apply immediately to both losses. Instead in this section, when proving Thm. 1, we find it more appropriate to treat the two losses separately: the proofs for CDT and LDT losses are included in Sec. C.1 and Sec. C.2, respectively.

Our proof in Sec. C.1 for CDT generalizes the proof of Thrampoulidis et al. (2022, Thm. 1), which only applies for the CE loss (a special case of CDT). At a high-level, the key innovations making this possible are: (i) formalizing the (δ, R)-SEL matrix (see Defn. 2) as the appropriate generalization of the SEL matrix in Thrampoulidis et al. (2022); (ii) expressing the dual of the CS-SVM corresponding to CDT (Eqn. (4a)) in a form that involves the (δ, R)-SEL and showing that it admits an explicit solution.

Our proof in Sec. C.2 for LDT relies on the following reduction idea: we prove that it is possible to re-parameterize the CS-SVM corresponding to LDT (Eqn. (4b)) such that it reduces to a weighted version of the standard unconstrained-features SVM (UF-SVM) for CE loss, CS-SVM with $\boldsymbol{\delta} = \mathbb{1}_k$, (see Prop. 2), albeit the new UF-SVM is over an artificial dataset with different imbalance ratio that is only introduced for the purpose of the proof. This reduction, together with the general formalization of the (δ, R)-SEL matrix, then allows us to leverage Thrampoulidis et al. (2022, Thm. 1).

C.1 CDT Loss: Theorem. 1 (ii)

Consider the CS-SVM of (4a):

$$\mathrm{p}_* = \min_{\mathbf{W},\mathbf{H}} \ \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{1}{2}\|\mathbf{H}\|_F^2 \qquad \text{sub. to} \quad (\delta_{y_i}\mathbf{w}_{y_i} - \delta_c \mathbf{w}_c)^T \mathbf{h}_i \geq 1, \ i \in [n], c \neq y_i, \tag{13}$$

and let the optimal parameters of the problem be $(\mathbf{W}^*, \mathbf{H}^*)$. We start by setting $\mathbf{X} = \begin{bmatrix} \mathbf{W}^T \\ \mathbf{H}^T \end{bmatrix} \begin{bmatrix} \mathbf{W} & \mathbf{H} \end{bmatrix} \in \mathbb{R}^{(k+n)\times(k+n)}$ and relaxing (13) as follows,

$$\begin{aligned} \mathrm{q}_* = \min_{\mathbf{X}\succeq 0} \quad & \frac{1}{2}\operatorname{tr}\left(\mathbf{X}\right) \\ \text{sub. to} \quad & \delta_{y_i}\mathbf{X}[y_i, k+i] - \delta_c \mathbf{X}[c, k+i] \geq 1, \ \forall i \in [n], c \neq y_i. \end{aligned} \tag{14}$$

Clearly, $\mathrm{p}_* \geq \mathrm{q}_*$. Our key insight in the analysis of (14) is writing its dual in a way that involves explicitly the $(\boldsymbol{\delta}, R)$-SEL matrix. Specifically, let $\hat{\mathbf{Z}}$ be the $(\boldsymbol{\delta}, R)$-SEL matrix of Defn. 2 with $\alpha = n_{\min}$. Then, we can formulate the dual of (14) as follows:

$$\mathrm{d}_* = \max_{\mathbf{B}\in\mathbb{R}^{n\times k}} \quad \operatorname{tr}(\hat{\mathbf{Z}}\mathbf{B}) \tag{15}$$

$$\begin{aligned} \text{sub. to} \quad & \begin{bmatrix} \mathbb{I}_d & -\mathbf{B}^T \\ -\mathbf{B} & \mathbb{I}_n \end{bmatrix} \succeq 1 \\ & \mathbf{B}\mathbf{D}^{-1}\mathbb{1}_k = 0 \\ & \mathbf{B} \odot \hat{\mathbf{Z}}^T \geq 0, \end{aligned} \tag{16}$$

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 (14) by satisfying Slater's condition. Thus, using the optimal solution of (15), we can characterize the optimizers (14).

To solve (15), we first relax the problem by ignoring constraint (16), and substituting the first constraint using Schur-complement argument:

$$\max_{\|\mathbf{B}\|_2\leq 1} \quad \operatorname{tr}(\hat{\mathbf{Z}}\mathbf{B}) \qquad \text{sub. to} \quad \mathbf{B}\mathbf{D}^{-1}\mathbb{1}_k = 0. \tag{17}$$

The optimal value of (17) is $\|\hat{\mathbf{Z}}\|_*$ and $\mathbf{B}^* = \mathbf{U}_\otimes \mathbf{V}^T$ is the unique solution (see Thrampoulidis et al. (2022, Lem. C.1))[7]. By Lem. B.2, $\mathbf{B}^*$ is strictly feasible in the relaxed condition (16). Therefore, the relaxation in (17)

[7] Thrampoulidis et al. (2022, Lem. C.1) holds for $(\mathbb{1}_k, R)$-SEL matrix $\hat{\mathbf{Z}}$, but inspecting the proof it remains unchanged for the general $(\boldsymbol{\delta}, R)$-SEL matrix.

is tight and $\mathbf{B}^*$ is in fact the dual optimal of (14). Since, strong duality holds for (14), we also have $q_* = \|\hat{\mathbf{Z}}\|_*$ and the optimizer $\mathbf{X}^*$ can be found by the complementary slackness conditions:

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

Let $\mathbf{X}^* = \begin{bmatrix} \mathbf{X}^*_{11} & \mathbf{X}^*_{12} \\ \mathbf{X}^*_{21} & \mathbf{X}^*_{22} \end{bmatrix}$, and recall that $\mathbf{B}^*$ satisfies (16) strictly. Then, the complementary slackness conditions imply:

$$\forall i \in [n], c \neq y_i: \quad 1 - \delta_{y_i}\mathbf{X}^*_{12}[y_i, i] + \delta_c \mathbf{X}^*_{12}[c, i] = 0$$
$$\mathbf{X}^*_{11} = \mathbf{B}^{*T}\mathbf{X}^{*T}_{12}, \quad \mathbf{X}^*_{22} = \mathbf{B}^*\mathbf{X}^*_{12}, \quad \mathbf{X}^*_{12} = \mathbf{B}^{*T}\mathbf{X}^*_{22}.$$

From the last condition, it is straightforward to see $\mathbf{X}^*_{22} = \mathbf{U}_\otimes \tilde{\mathbf{\Lambda}} \mathbf{V}^T$, $\mathbf{X}^*_{12} = \mathbf{V}\tilde{\mathbf{\Lambda}}\mathbf{U}_\otimes^T$ and $\mathbf{X}^*_{11} = \mathbf{V}\tilde{\mathbf{\Lambda}}\mathbf{V}^T$ for some $\tilde{\mathbf{\Lambda}} \in \mathbb{R}^{(k-1)\times(k-1)}$. Now, using the first condition, we have,

$$\begin{aligned}
&\delta_c^{-2}\delta_{y_i}\mathbf{X}^*_{12}[y_i,i] - \delta_c^{-1}\mathbf{X}^*_{12}[c,i] = \delta_c^{-2} \\
\xRightarrow{\sum_{c\neq y_i}} &\delta_{y_i}\mathbf{X}^*_{12}[y_i,i] \sum_{c\neq y_i}\delta_c^{-2} - \sum_{c\neq y_i}\delta_c^{-1}\mathbf{X}^*_{12}[c,i] = \sum_{c\neq y_i}\delta_c^{-2} \\
\Longrightarrow &\delta_{y_i}\mathbf{X}^*_{12}[y_i,i] \sum_{c\in[k]}\delta_c^{-2} - \sum_{c\in[k]}\delta_c^{-1}\mathbf{X}^*_{12}[c,i] = \sum_{c\neq y_i}\delta_c^{-2} \\
\xRightarrow{(i)} &\delta_{y_i}\mathbf{X}^*_{12}[y_i,i] \sum_{c\in[k]}\delta_c^{-2} = \sum_{c\neq y_i}\delta_c^{-2} \\
\Longrightarrow &\mathbf{X}^*_{12}[y_i,i] = \delta_{y_i}^{-1}\big(1 - \delta_{y_i}^{-2}/\sum_{c'\in[k]}\delta_{c'}^{-2}\big), \quad \mathbf{X}^*_{12}[c,i] = -\delta_c^{-1}\big(\delta_{y_i}^{-2}/\sum_{c'\in[k]}\delta_{c'}^{-2}\big) \\
\Longrightarrow &\mathbf{X}^*_{12} = \hat{\mathbf{Z}}.
\end{aligned} \tag{18}$$

In (i), we use the fact that $\mathbf{V}^T\mathbf{D}^{-1}\mathbb{1}_k = 0$ and thus $\mathbf{X}^{*T}_{12}\mathbf{D}^{-1}\mathbb{1}_k = 0$. By (18), and using $\mathbf{V}^T\mathbf{V} = \mathbf{U}_\otimes^T\mathbf{U}_\otimes = \mathbb{I}_{k-1}$ it is easy to show $\tilde{\mathbf{\Lambda}} = \mathbf{\Lambda}$ and thus,

$$\mathbf{X}^* = \begin{bmatrix} \mathbf{V} \\ \mathbf{U}_\otimes \end{bmatrix} \mathbf{\Lambda} \begin{bmatrix} \mathbf{V}^T & \mathbf{U}_\otimes^T \end{bmatrix}. \tag{19}$$

Now, it remains to show all the optimizers of (13) can be constructed by $\mathbf{X}^*$ and that the relaxation in (14) is tight. First, choose some partial orthonormal matrix $\mathbf{R} \in \mathbb{R}^{(k-1)\times d}$ with $\mathbf{R}\mathbf{R}^T = \mathbb{I}_{k-1}$, and construct $\mathbf{W}^* = \mathbf{R}^T\mathbf{\Lambda}^{1/2}\mathbf{V}^T$ and $\mathbf{H}^* = \mathbf{R}^T\mathbf{\Lambda}^{1/2}\mathbf{U}_R^T$. Then, $(\mathbf{W}^*, \mathbf{H}^*)$ is by construction feasible in (13) and,

$$q_* \leq p_* \leq \frac{1}{2}\|\mathbf{W}^*\|_F^2 + \frac{1}{2}\|\mathbf{H}^*\|_F^2 = \frac{1}{2}\operatorname{tr}(\mathbf{X}^*) = q_*.$$

Therefore, $q_* = p_*$ and indeed the relaxation is tight. On the other hand, if $(\tilde{\mathbf{W}}, \tilde{\mathbf{H}})$ is a minimizer of (13), $\tilde{\mathbf{X}} = \begin{bmatrix} \tilde{\mathbf{W}}^T \\ \tilde{\mathbf{H}}^T \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{W}} & \tilde{\mathbf{H}} \end{bmatrix}$ is feasible and optimal in (14) (since $q_* = p_*$), which implies $\tilde{\mathbf{X}}$ should satisfy (19). Hence, any minimizer of the CS-SVM (13) satisfies,

$$\begin{bmatrix} \mathbf{W}^{*T} \\ \mathbf{H}^{*T} \end{bmatrix} \begin{bmatrix} \mathbf{W}^* & \mathbf{H}^* \end{bmatrix} = \begin{bmatrix} \mathbf{V} \\ \mathbf{U}_\otimes \end{bmatrix} \mathbf{\Lambda} \begin{bmatrix} \mathbf{V}^T & \mathbf{U}_\otimes^T \end{bmatrix}. \tag{20}$$

The statement of the theorem is easy to see by (20). Specifically, by noting that $\mathbf{U}_\otimes$ has repeated columns, all the embeddings belonging to the same class are equal (NC occurs) and,

$$\mathbf{W}^{*T}\mathbf{W}^* = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T, \quad \mathbf{M}^{*T}\mathbf{M}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \quad \mathbf{W}^{*T}\mathbf{M}^* = \mathbf{\Xi}. \tag{21}$$

Remark 3. *For simplicity of exposition, we set $\alpha = n_{min}$ when using the $(\boldsymbol{\delta}, R)$-SEL matrix to formulate the dual problem. However, it is easy to see that by choosing some other α', the SVD factors would only change by a scaling factor. In particular, let $\tau = \sqrt{\alpha'/n_{min}}$, then $\mathbf{U}$ and $\boldsymbol{\Lambda}$ will be scaled by a factor of $1/\tau$ and τ respectively, and $\mathbf{V}$ remains unchanged. Hence, (21) changes as follows,*

$$\mathbf{W}^{*T}\mathbf{W}^* = \tau\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T, \quad \mathbf{M}^{*T}\mathbf{M}^* = \frac{1}{\tau}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T, \quad \mathbf{W}^{*T}\mathbf{M}^* = \boldsymbol{\Xi}.$$

C.2 LDT Loss: Theorem. 1 (iii)

We start the proof by restating a result from Thrampoulidis et al. (2022) regarding the optimal solutions of the unconstrained-features SVM.

Proposition 2 (Thrampoulidis et al. (2022, Sec. C.3))**.** *Consider the following k-class β-weighted unconstrained-features SVM (UF-SVM):*

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

in an (R, ρ)-STEP imbalanced setting. For any $\beta > 0$, the NC property holds, and the optimal solutions $(\hat{\mathbf{W}}_\beta, \hat{\mathbf{M}}_\beta)$ follow the $(\mathbb{1}_k, R)$-SELI geometry. Specifically,

$$\hat{\mathbf{W}}_\beta^T\hat{\mathbf{M}}_\beta = \boldsymbol{\Xi}, \quad \hat{\mathbf{M}}_\beta^T\hat{\mathbf{M}}_\beta = \frac{1}{\tau}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T, \quad \hat{\mathbf{W}}_\beta^T\hat{\mathbf{W}}_\beta = \tau\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T,$$

where $\mathbf{V}$, $\boldsymbol{\Lambda}$, $\mathbf{U}$ are the SVD factors of the $(\mathbb{1}_k, R)$-SEL matrix as described in Defn. 2, and τ is a positive scalar depending on β and n_{minor}.

Consider the k-class CS-SVM problem of (4b), restated below for convenience:

$$(\mathbf{W}^*, \mathbf{H}^*) \in \arg\min_{\mathbf{W},\mathbf{H}} \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{1}{2}\|\mathbf{H}\|_F^2 \qquad \text{sub. to} \quad \delta_{y_i}(\mathbf{w}_{y_i} - \mathbf{w}_c)^T\mathbf{h}_i \geq 1,\ i \in [n], c \neq y_i. \tag{22}$$

Also recall that $n_c, c \in [k]$ is the number of examples in class c. We will relate the above optimization problem to an equivalent UF-SVM, whose solution can be found by Prop. 2. The resulting solution will be used to state the minimizers of (22).

First, it is easy to verify the NC property: for a fixed $\mathbf{W}$, the optimization in (22) is separable in $\mathbf{h}_i$, and for all the samples in the same class, the separable problems are identical and strongly-convex. Thus, for all $i : y_i = c$ there is a unique minimzer for the fixed $\mathbf{W}$. So, at the optimal solution, all the embeddings within a class are equal to their means, i.e. $\forall i \in [n] : y_i = c, \mathbf{h}_i = \boldsymbol{\mu}_c$. Defining $\mathbf{M} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k]$, we can re-formulate (22) as follows,

$$(\mathbf{W}^*, \mathbf{M}^*) \in \arg\min_{\mathbf{W},\mathbf{M}} \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{1}{2}\sum_{c\in[k]} n_c\|\boldsymbol{\mu}_c\|_2^2 \qquad \text{sub. to} \quad \delta_c(\mathbf{w}_c - \mathbf{w}_{c'})^T\boldsymbol{\mu}_c \geq 1,\ c, c' \in [k], \tag{23}$$

and by the NC property, there is a one-to-one correspondence between the optimal solutions of (22) and (23).

Now, let $\mathbf{D} = \text{diag}(\boldsymbol{\delta})$ and $\widetilde{\mathbf{M}} = \mathbf{M}\mathbf{D}$, i.e. $\tilde{\boldsymbol{\mu}}_c = \delta_c\boldsymbol{\mu}_c$. Applying this reparametrization to (23), we have,

$$(\mathbf{W}^*, \widetilde{\mathbf{M}}^*) \in \arg\min_{\mathbf{W},\tilde{\mathbf{M}}} \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{1}{2}\sum_{c\in[k]} \frac{n_c}{\delta_c^2}\|\tilde{\boldsymbol{\mu}}_c\|_2^2 \qquad \text{sub. to} \quad (\mathbf{w}_c - \mathbf{w}_{c'})^T\tilde{\boldsymbol{\mu}}_c \geq 1,\ c, c' \in [k]. \tag{24}$$

Define $\tilde{R} = R\big(\delta_{\text{minor}}/\delta_{\text{maj}}\big)^2$, which is rational by assumption. Thus, there exists $\alpha \in \mathbb{N}$ such that $\alpha\tilde{R}$ is an integer. Now, set $\beta^2 = n_{\min}/(\alpha\delta_{\text{minor}}^2)$, and re-write (24) as follows:

$$(\mathbf{W}^*, \widetilde{\mathbf{M}}^*) \in \arg\min_{\mathbf{W},\tilde{\mathbf{M}}} \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{\beta^2}{2}\sum_{c\in[k]} \tilde{n}_c\|\tilde{\boldsymbol{\mu}}_c\|_2^2 \qquad \text{sub. to} \quad (\mathbf{w}_c - \mathbf{w}_{c'})^T\tilde{\boldsymbol{\mu}}_c \geq 1,\ c, c' \in [k], \tag{25}$$

where

$$\tilde{n}_c = \begin{cases} \alpha\tilde{R}, & \text{if } c \in \{1, \ldots, \bar{\rho}k\}, \\ \alpha, & \text{if } c \in \{\bar{\rho}k + 1, \ldots, k\} \end{cases}$$

By a similar argument that led to the equivalence of (22) and (23), it is easy to see $(\mathbf{W}^*, \widetilde{\mathbf{M}}^*)$ is the optimal parameters of a β-weighted UF-SVM trained on an imbalanced dataset with imbalance ratio $\tilde{R}$ and $\tilde{n}_c$ samples per class for $c \in [k]$. Thus, $(\mathbf{W}^*, \widetilde{\mathbf{M}}^*)$ follows the $(\mathbb{1}_k, \tilde{R})$-SELI geometry as in Prop. 2. The proof is complete by noting that $(\mathbf{W}^*, \widetilde{\mathbf{M}}^*) = (\mathbf{W}^*, \mathbf{M}^*\mathbf{D})$.

D CLOSED-FORM FORMULAS FOR THE $(\boldsymbol{\delta}, R)$-SELI GEOMETRY

As stated in the Thm. 1, the optimal parameters of the CS-SVM 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 parameters' geometry as a function of R, ρ, k and of the hyper-parameters $\boldsymbol{\delta}$.

Similar to Sec. B, throughout this section, we assume the data is STEP imbalanced and STEP logit adjustment is adopted. For simplicity, we consider the case $\rho = 1/2$, $\delta_{\text{minor}} = 1$ and $\delta_{\text{maj}} = \Delta$. This choice is without loss of generality since the geometry only depends on the ratio $\delta_{\text{maj}}/\delta_{\text{minor}}$. We use the closed-form SVD in Sec. B derived by assuming $\alpha = 1$. It is easy to see that a general α only introduces an appropriate scaling to the SVD factors. (See Remark 3). Thus, using the closed-form expressions in Lemma B.1 for the corresponding $\mathbf{V}$, $\boldsymbol{\Lambda}$, and $\mathbf{U}$, the optimal parameters satisfy:

$$\mathbf{W}^{*T}\mathbf{W}^* = \tau \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T, \quad \mathbf{M}^{*T}\mathbf{M}^* = \frac{1}{\tau}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T, \quad \mathbf{W}^{*T}\mathbf{M}^* = \boldsymbol{\Xi}, \tag{26}$$

for some positive scalar τ (that depends on $n_{\min}$ and α). Since, τ onlys affects the scale of the geometry, in the lemmas we assume $\tau = 1$ for brevity.

In Sec. D.1, we describe the geometric and asymptotic properties of the solutions of (4a), the CS-SVM under CDT loss. In Sec. D.2 we characterize the same properties for problem (4b) corresponding to the LDT loss. In the following lemmas, we use $\mathbf{w}_{\text{maj}}$ when referring to any majority classifier $\mathbf{w}_c, c \in \{1, \dots, k/2\}$, and $\mathbf{w}_{\text{minor}}$ for any minority classifier $\mathbf{w}_c, c \in \{k/2+1, \dots, k\}$. Similarly, $\mathbf{h}_{\text{maj}}$ denotes any $\mathbf{h}_j$ with $j \in \{i \in [n] : y_i = 1, \dots, k/2\}$ and $\mathbf{h}_{\text{minor}}$ denotes any $\mathbf{h}_j$ with $j \in \{i \in [n] : y_i = k/2+1, \dots, k\}$.

D.1 CDT Loss

D.1.1 Norms and Angles

Lemma D.1 (CDT classifiers)**.** *Let $\mathbf{V}$, $\boldsymbol{\Lambda}$, $\mathbf{U}$ be the eigen-factors of the $(\boldsymbol{\delta}, R)$-SEL matrix. For the optimal classifier $\mathbf{W}$ of the CS-SVM (4a):*

(a) (Norms) *All the majority/minority classes have equal norms,*

$$\|\mathbf{w}_{\text{maj}}\|_2^2 = \frac{\sqrt{R}}{\Delta}(1-2/k) + \frac{2\Delta^2\sqrt{R+1}}{k\left(\sqrt{1+\Delta^2}\right)^3}, \quad \|\mathbf{w}_{\text{minor}}\|_2^2 = (1-2/k) + \frac{2\sqrt{R+1}}{k\left(\sqrt{1+\Delta^2}\right)^3}, \tag{27}$$

and the majority-minority norm-ratio is,

$$\frac{\|\mathbf{w}_{\text{maj}}\|_2^2}{\|\mathbf{w}_{\text{minor}}\|_2^2} = \frac{\frac{\sqrt{R}}{\Delta}(k-2)\left(1+\Delta^2\right)^{3/2} + 2\Delta^2\sqrt{R+1}}{(k-2)\left(1+\Delta^2\right)^{3/2} + 2\sqrt{R+1}}.$$

(b) (Angles) *For each pair of majority/minority classifiers the angles are equal and,*

$$\cos(\mathbf{w}_{\rm maj}, \mathbf{w}'_{\rm maj}) = \frac{-2\sqrt{R}+2\sqrt{R+1}\left(\sqrt{1+\Delta^{-2}}\right)^{-3}}{(k-2)\sqrt{R}+2\sqrt{R+1}\left(\sqrt{1+\Delta^{-2}}\right)^{-3}}$$
$$\cos(\mathbf{w}_{\rm minor}, \mathbf{w}'_{\rm minor}) = \frac{-2+2\sqrt{R+1}\left(\sqrt{1+\Delta^{2}}\right)^{-3}}{k-2+2\sqrt{R+1}\left(\sqrt{1+\Delta^{2}}\right)^{-3}}$$
$$\cos(\mathbf{w}_{\rm maj}, \mathbf{w}_{\rm minor}) = -\frac{2\Delta\sqrt{R+1}}{k\left(\sqrt{1+\Delta^{2}}\right)^{3}\|\mathbf{w}_{\rm maj}\|_2\|\mathbf{w}_{\rm minor}\|_2}.$$

Proof. Let $m = k/2$. From Thm. 1, $\mathbf{W}^T\mathbf{W} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$. Using Lem. B.1 we have,

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

Inspecting the diagonal entries proves the norm equations. To prove part **(b)**, we use the off-diagonals entries that specify the inner-product of each pair of classifiers. Particularly,

$$\mathbf{w}_{\rm maj}^T\mathbf{w}'_{\rm maj} = \frac{-2\sqrt{R}+2\sqrt{R+1}\left(\sqrt{\Delta^{-2}+1}\right)^{-3}}{k\Delta},$$
$$\mathbf{w}_{\rm minor}^T\mathbf{w}'_{\rm minor} = \frac{-2+2\sqrt{R+1}\left(\sqrt{1+\Delta^{2}}\right)^{-3}}{k},$$
$$\mathbf{w}_{\rm minor}^T\mathbf{w}_{\rm maj} = \frac{-2\Delta\sqrt{R+1}}{k\left(\sqrt{\Delta^2+1}\right)^3}.$$

These equations together with (27) complete the proof. □

Lemma D.2 (CDT embeddings)**.** *Let* $\mathbf{V}$*,* $\mathbf{\Lambda}$*,* $\mathbf{U}$ *be the eigen-factors of the* $(\boldsymbol{\delta}, R)$*-SEL matrix. For the optimal embeddings* $\mathbf{H}$ *of the CS-SVM* (4a)*:*

(a) (Norms) *All the embeddings in the majority/minority classes have equal norms,*

$$\|\mathbf{h}_{\rm maj}\|_2^2 = \frac{(1-2/k)}{\Delta\sqrt{R}} + \frac{2}{k\sqrt{R+1}\sqrt{1+\Delta^2}}, \quad \|\mathbf{h}_{\rm minor}\|_2^2 = (1-2/k) + \frac{2}{k\sqrt{R+1}\sqrt{1+\Delta^2}},$$

and the majority-minority norm-ratio is as follows,

$$\frac{\|\mathbf{h}_{\rm maj}\|_2^2}{\|\mathbf{h}_{\rm minor}\|_2^2} = \frac{\frac{1}{\Delta\sqrt{R}}(k-2)\sqrt{R+1}\sqrt{1+\Delta^2}+2}{(k-2)\sqrt{R+1}\sqrt{1+\Delta^2}+2}.$$

(b) (Angles) *For each pair of majority/minority embeddings the angles are equal, and,*

$$\cos(\mathbf{h}_{\rm maj}, \mathbf{h}'_{\rm maj}) = \frac{-2\sqrt{\Delta^{-2}+1}\sqrt{R+1}+2\sqrt{R}}{(k-2)\sqrt{\Delta^{-2}+1}\sqrt{R+1}+2\sqrt{R}}$$
$$\cos(\mathbf{h}_{\rm minor}, \mathbf{h}'_{\rm minor}) = \frac{-2\sqrt{\Delta^{2}+1}\sqrt{R+1}+2}{(k-2)\sqrt{\Delta^{2}+1}\sqrt{R+1}+2}$$
$$\cos(\mathbf{h}_{\rm maj}, \mathbf{h}_{\rm minor}) = \frac{-2}{k\sqrt{\Delta^2+1}\sqrt{R+1}\|\mathbf{h}_{\rm maj}\|_2\|\mathbf{h}_{\rm minor}\|_2}.$$

Proof. By the NC property, to find the norms and angles of the embeddings, it suffices to analyze the mean-embeddings $\mathbf{M}$, for which, following Thm. 1, we have $\mathbf{M}^T\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$. By Lemma B.1,

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

The diagonal entries determine the norm of the embeddings as in part **(a)** and the off-diagonals entries specify the inner-product of each pair of the embeddings. Particularly,

$$\mathbf{h}_{\text{maj}}^T\mathbf{h}'_{\text{maj}} = -\frac{2}{k}\left(\frac{1}{\Delta\sqrt{R}} - \frac{1}{\sqrt{R+1}\sqrt{\Delta^2+1}}\right)$$
$$\mathbf{h}_{\text{minor}}^T\mathbf{h}'_{\text{minor}} = -\frac{2}{k}\left(1 - \frac{1}{\sqrt{R+1}\sqrt{\Delta^2+1}}\right)$$
$$\mathbf{h}_{\text{minor}}^T\mathbf{h}_{\text{maj}} = -\frac{2}{k\sqrt{R+1}\sqrt{\Delta^2+1}}.$$

Combining these with the norm calculations of part **(a)** 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 $\cos(\mathbf{w}_c, \mathbf{h}_i)$ for $c \in [k], i : y_i = c$, which can be thought of the degree of alignment between classifiers and embeddings.

Lemma D.3 (CDT: Alignment of classifiers and embeddings)**.** *The angles between majority/minority embeddings and the their corresponding classifiers are all equal:*

$$\cos(\mathbf{w}_{\text{maj}}, \mathbf{h}_{\text{maj}}) = \frac{k\Delta^2 + (k-2)}{k\,\Delta\,(\Delta^2+1)\,\|\mathbf{w}_{\text{maj}}\|_2\|\mathbf{h}_{\text{maj}}\|_2},$$
$$\cos(\mathbf{w}_{\text{minor}}, \mathbf{h}_{\text{minor}}) = \frac{k\Delta^{-2} + (k-2)}{k\,(\Delta^{-2}+1)\,\|\mathbf{w}_{\text{minor}}\|_2\|\mathbf{h}_{\text{minor}}\|_2}.$$

Proof. Recalling $\mathbf{W}^T\mathbf{H} = \hat{\mathbf{Z}}$, 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. □

D.1.2 Asymptotics

We present the limiting values of the norm-ratios and angles in the asymptotic regime $\Delta = R^\gamma, \gamma \in \mathbb{R}$ and $R \to \infty$. This parameterization is interesting because it can guide us on how to maintain finite angles between classifiers and embeddings as the imbalance ratio grows large. Specifically, the angles are as shown in Table 1.

$\cos(\mathbf{w}_c, \mathbf{w}'_c)$	$\gamma < 1/6$	$\gamma = 1/6$	$\gamma > 1/6$
$c, c' \in$ minority	1	0	$-\frac{2}{k-2}$
$\cos(\mathbf{w}_c, \mathbf{w}'_c)$	$\gamma < 0$	$\gamma = 0$	$\gamma > 0$
$c, c' \in$ majority	$-\frac{2}{k-2}$	$\frac{1-2\sqrt{2}}{1+\sqrt{2}(k-2)}$	0
$\cos(\mathbf{h}_c, \mathbf{h}'_c)$	$\gamma < 0$	$\gamma = 0$	$\gamma > 0$
$c, c' \in$ minority	0	$-\frac{2}{k-2}$	$-\frac{2}{k-2}$
$c, c' \in$ majority	0	$\frac{2-2\sqrt{2}}{2+(k-2)\sqrt{2}}$	$-\frac{2}{k-2}$

Table 1: Asymptotic values of angles for CDT with $\Delta = R^\gamma, \gamma \in \mathbb{R}$ and $R \to \infty$.

D.1.3 Centering

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

$$\sum_{i\in[n]} \frac{1}{n_{y_i}} \mathbf{h}_i^* = 0. \tag{28}$$

Note that this reduces to $\sum_{i\in[n]} \mathbf{h}_i^*$ for balanced data, and remains unchanged for any choice of the hyperparameters $\boldsymbol{\delta}$. Eqn. (28) 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.

D.2 LDT Loss

D.2.1 Norms and Angles

From Thm. 1, solutions $(\mathbf{W}^*, \mathbf{M}^*\mathbf{D})$ of the CS-SVM under LDT loss in (4b), follow the SELI geometry (Thrampoulidis et al., 2022), with imbalance ratio $\tilde{R} = R(\delta_{\min}/\delta_{\text{maj}})^2$. Thus, the corresponding norms and angles can be found by analyzing the $(\mathbb{1}_k, \tilde{R})$-SELI structure (up to a norm scaling by $\mathbf{D}$ for the mean embeddings $\mathbf{M}^*$). We refer the reader to Thrampoulidis et al. (2022, Sec. B.1) for closed form expressions of the $(\mathbb{1}_k, \tilde{R})$-SELI. We repeat some key formulas below for showing explicit dependence on Δ.

Corollary 1.5 (LDT: Norm ratios and classifier angles)**.** *For the optimal solution* $(\mathbf{W}, \mathbf{H})$ *of the CS-SVM* (4b)*:*

$$\frac{\|\mathbf{w}_{\text{maj}}\|_2^2}{\|\mathbf{w}_{\text{minor}}\|_2^2} = \frac{(k-2)\sqrt{R} + \sqrt{(R+\Delta^2)/2}}{(k-2)\Delta + \sqrt{(R+\Delta^2)/2}},$$

$$\frac{\|\mathbf{h}_{\text{maj}}\|_2^2}{\|\mathbf{h}_{\text{minor}}\|_2^2} = \frac{\frac{1}{\sqrt{R}}(k-2) + \frac{1}{\sqrt{(R+\Delta^2)/2}}}{(k-2)\Delta + \frac{\Delta^2}{\sqrt{(R+\Delta^2)/2}}},$$

$$\cos(\mathbf{w}_{\text{maj}}, \mathbf{w}'_{\text{maj}}) = \frac{-2\sqrt{R} + \sqrt{(R+\Delta^2)/2}}{(k-2)\sqrt{R} + \sqrt{(R+\Delta^2)/2}},$$

$$\cos(\mathbf{w}_{\text{min}}, \mathbf{w}'_{\text{min}}) = \frac{-2\Delta + \sqrt{(R+\Delta^2)/2}}{(k-2)\Delta + \sqrt{(R+\Delta^2)/2}}.$$

D.2.2 Asymptotics

Similar to the calculations for the CDT case, we present the limiting values of the norm-ratios and angles in the asymptotic regime $\Delta = R^\gamma, \gamma \in \mathbb{R}$ and $R \to \infty$. Then, the angles are as given in Table 2:

$\cos(\mathbf{w}_c, \mathbf{w}'_c)$	$\gamma < 1/2$	$\gamma = 1/2$	$\gamma > 1/2$
$c, c' \in$ minority	1	$-\frac{1}{k-1}$	$\frac{1-2\sqrt{2}}{1+\sqrt{2}(k-2)}$
$c, c' \in$ majority	$\frac{1-2\sqrt{2}}{1+\sqrt{2}(k-2)}$	$-\frac{1}{k-1}$	1
$\cos(\mathbf{h}_c, \mathbf{h}'_c)$	$\gamma < 1/2$	$\gamma = 1/2$	$\gamma > 1/2$
$c, c' \in$ minority	$-\frac{2}{k-2}$	$-\frac{1}{k-1}$	$\frac{2-2\sqrt{2}}{2+\sqrt{2}(k-2)}$
$c, c' \in$ majority	$-\frac{\sqrt{2}}{-\sqrt{2}+k(1+\sqrt{2})}$	$-\frac{1}{k-1}$	$-\frac{2}{k-2}$

Table 2: Asymptotic values of angles for LDT with $\Delta = R^\gamma, \gamma \in \mathbb{R}$ and $R \to \infty$.

D.2.3 Centering

The optimal classifiers and features $(\mathbf{W}^*, \mathbf{M}^*\mathbf{D})$ follow the $(\mathbb{1}_k, \tilde{R})$-SELI structure. Thus (see Thrampoulidis et al. (2022, 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,

$$\sum_{i\in[n]} \frac{\delta_{y_i}}{n_{y_i}} \mathbf{h}_i^* = 0. \tag{29}$$

E NUMERICAL RESULTS

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

E.1 Additional Experimental Details

As mentioned in Sec. 5, we investigate the convergence of SGD steps for CDT/LDT loss in (3a)/(3b) to the implicit geometries of Thm. 1. Here, we describe the experimental setup in more details.

UFM experiments. 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 $\mathbb{R}^n$. The labels for each vector are chosen such that the dataset is $(R = 10, \rho = 1/2)$-STEP imbalanced, with $n_{\min} = 5$ and a batch size of 5. We further use STEP logit adjustment, and choose $\Delta = R^\gamma$ with $\gamma \in [-1.5, 1.5]$. We train all models with the same constant learning rate for 6000 epochs. We normalize $\boldsymbol{\delta}$ so that $\mathbb{1}_k^T \boldsymbol{\delta} = k$, since we empirically observe that for a fixed ratio Δ the convergence speed depends on the magnitude of $\boldsymbol{\delta}$.

Deep-net experiments. We train (i) ResNet18 on CIFAR10, and, (ii) a 6-layered fully connected MLP with batch-norm and ReLU activations on MNIST, both under a $R = 10$ imbalance ratio. The MLP model consists of 6 fully-connected layers of width 2048, each followed by batch-norm and ReLU activations. We train the models 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 Papyan et al. (2020); Thrampoulidis et al. (2022), we set momentum and weight decay to 0.9 and 10^{-5} respectively. We also normalize $\boldsymbol{\delta}$ to sum to k, similar to UFM. We perform the experiments in Fig. 3 without any data augmentation, remaining consistent with previous works on neural collapse (e.g., Papyan et al., 2020)).

We conducted additional experiments with imbalance ratio $R = 2, 5, 20$. We have not included those results due to their similarity to $R = 10$; however, we will discuss the impacts of higher imbalance ratio R and hyperparameter γ in the following section.

E.2 Speed of Convergence

We empirically observe that the UFM parameters converge more slowly to the global optimizers in Thm. 1 as the imbalance ratio R and hyperparameter $|\gamma|$ increase. A similar observation for large values of R is also reported in Thrampoulidis et al. (2022). To illustrate the speed of convergence, we measure the distance of the SGD steps to the predicted implicit geometry during training. In particular, at each step $(\mathbf{W}_t, \mathbf{M}_t)$, we compute $\| \frac{\mathbf{W}_t^T \mathbf{W}_t}{\|\mathbf{W}_t^T \mathbf{W}_t\|} - \frac{\mathbf{W}^{*T} \mathbf{W}^*}{\|\mathbf{W}^{*T} \mathbf{W}^*\|} \|_F$ for the classifiers and $\| \frac{\mathbf{M}_t^T \mathbf{M}_t}{\|\mathbf{M}_t^T \mathbf{M}_t\|} - \frac{\mathbf{M}^{*T} \mathbf{M}^*}{\|\mathbf{M}^{*T} \mathbf{M}^*\|} \|_F$ for the centered mean-embeddings, where $(\mathbf{W}^*, \mathbf{M}^*)$ are as described by Thm. 1. Fig. 4 illustrates the convergence behaviour of the parameters for UFM and ResNet18. While as training progresses, the classifiers/embeddings get closer to the predicted geometry, imbalance ratio and hyperparameter values can significantly slow down the convergence. This behaviour appear for both UFM and deep-net experiments.

In addition to the worse convergence, it becomes more challenging to achieve zero training error as $|\gamma|$ increases. We illustrate this in Fig. 5, where we report the training accuracy of ResNet model trained on imbalanced CIFAR10 at different epochs. We empirically observe that it is in general easier to enter the zero-error regime by LDT loss. On the other hand, we do not achieve 100% training accuracy for large values of γ on CDT loss. This is consistent with similar observation on CDT training in Kini et al. (2021).

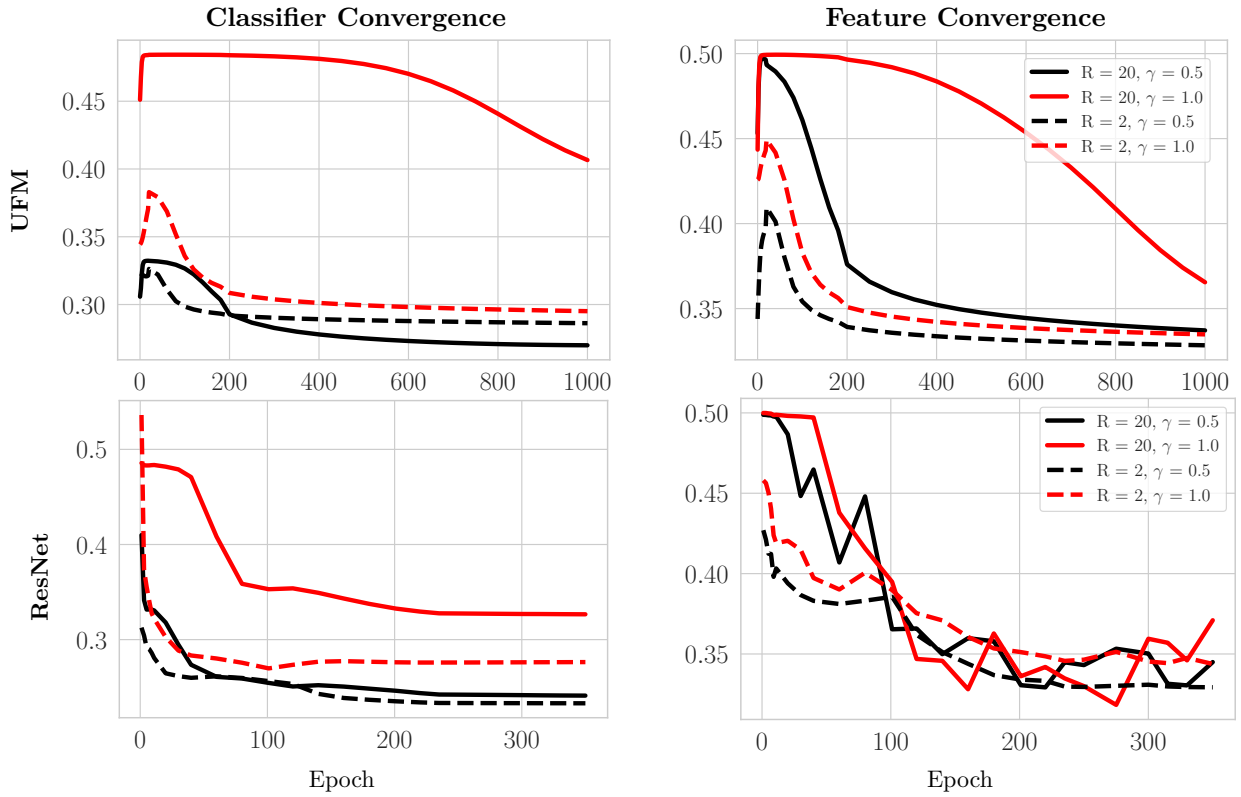

Figure 4: Convergence of classifiers and mean-embeddings to the implicit geometry in Thm. 1: (first row) UFM, (second row) ResNet18 trained on CIFAR10. The models are trained by SGD on CDT loss. Larger R and γ lead to slower convergence to the expected structure.

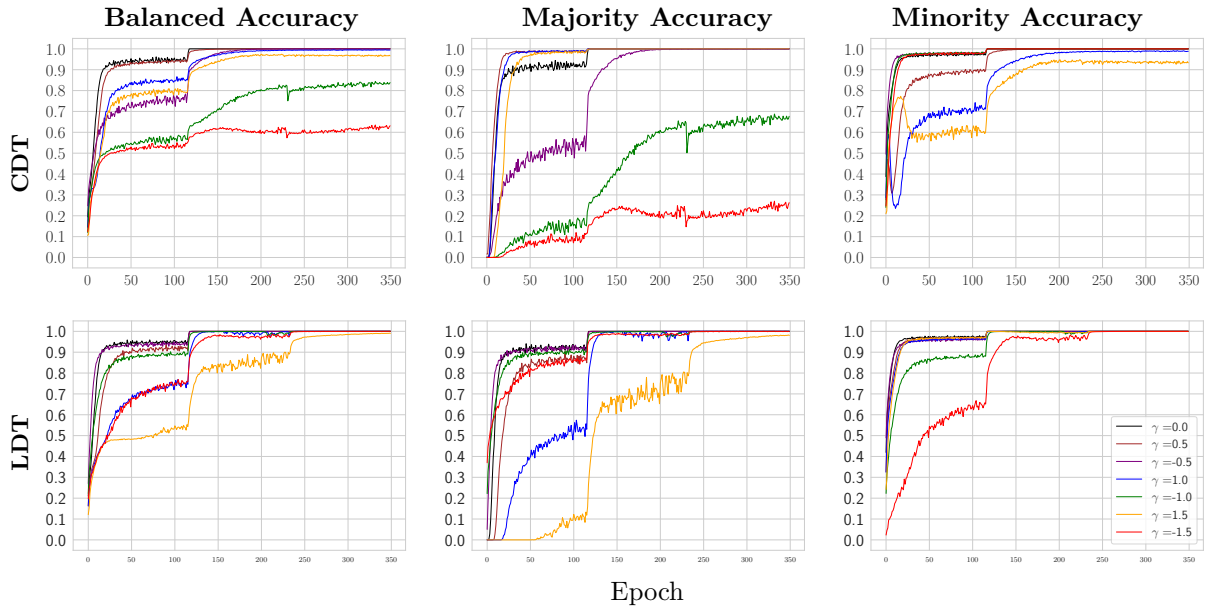

Figure 5: Training accuracy across epochs of ResNet18 model trained on ($R = 10$, ρ=1/2)-STEP imbalanced CIFAR10 dataset with CDT/LDT loss and different values of γ. It becomes harder to enter zero training error regime for larger $|\gamma|$ with the impact being more noticable on CDT.