Label-Imbalanced and Group-Sensitive Classification under Overparameterization

Anonymous Author(s)
Affiliation
Address
email

Abstract

The goal in label-imbalanced and group-sensitive classification is to optimize relevant metrics such as *balanced error* and *equal opportunity*. Classical methods, such as re-weighted cross-entropy, are known to fail when used in conjunction with the modern practice of training deep nets to the *terminal phase of training* (TPT), that is training beyond zero training error. This observation has motivated recent flurry of activity in developing heuristic alternatives following the intuitive mechanism of promoting larger margin for minorities. In contrast to previous heuristics, we follow a principled analysis explaining how different loss adjustments affect margins. First, we prove that for all linear classifiers trained in TPT, it is necessary to introduce *multiplicative*, rather than *additive*, logit adjustments so that the relative margins between classes change appropriately. To show this, we discover a connection of the multiplicative CE modification to the so-called cost-sensitive support-vector machines. Perhaps counterintuitively, we also find that, at the start of the training, the same multiplicative weights can actually harm the minority classes. Thus, while additive adjustments are ineffective deep in the TPT, we show numerically that they can speed up convergence by countering the initial negative effect of the multiplicative weights. Motivated by these findings, we formulate the *vector-scaling (VS) loss*, that captures existing techniques as special cases. Moreover, we introduce a natural extension of the VS-loss to group-sensitive classification, thus treating the two common types of imbalances (label/group) in a unifying way. Finally, for imbalanced Gaussian-mixtures data, we perform a generalization analysis, revealing tradeoffs between balanced / standard error and equal opportunity.

1 Introduction

1.1 Motivation

Equitable learning in the presence of data imbalances is a classical machine learning (ML) problem, but one with increasing importance as ML decisions are adapted in increasingly more complex applications that directly involve people [BS16]. Two common types of imbalances that have attracted particular attention are those present in *label-imbalanced* and *group-sensitive* classification. In the first type, examples from a target class are heavily outnumbered by examples from the rest of the classes. The standard metric of average misclassification error is insensitive to such imbalances and among several classical alternatives the *balanced error* is a widely used metric. In the second type, the broad goal is to ensure fairness with respect to a protected underrepresented group (e.g. gender, race). While acknowledging that there is no universal fairness metric [KMR16, FSV16], several suggestions have been made in the literature [CKP09, HPS16, ZVGRG17, WM19], such as the *Equal Opportunity* metric favoring same true positive rates across groups [HPS16].

Methods to cope with imbalances are broadly categorized into data-level and algorithm-level ones. In the latter category, belong *cost-sensitive methods* and specifically approaches that modify the *training*

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

loss function to account for varying class/group penalties. State-of-the-art (SOTA) research on such approaches is motivated by observations that classical methods, such as weighted cross-entropy (wCE) fail when used with current trends of training largely overparameterized deep nets without regularization and with train-loss minimization continuing well beyond zero train-error, in the so-called *terminal phase of training (TPT)* ([PHD20] and references therein). Intuitively, failure of wCE when trained in TPT is attributed to the failure to appropriately adjust the relative margins between different classes/groups in a way that properly accounts for imbalances. To overcome the challenge, recent works have proposed a so-called *logit-adjusted (LA) loss* that modifies the cross-entropy (CE) loss by including extra *additive* hyper-parameters acting on the logits [KHB+18, CWG+19, MJR+20]. Even more recently, [YCZC20] suggested yet another modification that introduces *multiplicative* hyper-parameters on the logits leading to a *class-dependent temperature (CDT) loss*. Empirically, both adjustments show performance improvements over wCE. However, it remains unclear: *Do both additive and multiplicative margins lead to favorable margin-adjustments (favoring minority classes)? If so, what are the individual mechanisms that lead to this behavior? How effective are different adjustments at each stage of training?*

This paper sheds light on these questions. We argue that multiplicative hyper-parameters are most effective for margin adjustments in TPT. However, additive weights can be useful in the beginning of training. Importantly, this intuition justifies our algorithmic contribution: we introduce the *vector-scaling (VS) loss* that combines both types of weights and attains improved performance on SOTA imbalanced datasets. Finally, using the same set of tools, we also study related questions and extend the VS-loss to instances of group-sensitive classification. We make multiple contributions as summarized below; see also Figure 1.

1.2 Contributions

• **Insights on existing techniques:** We show that when optimizing in the terminal phase of training (TPT) *multiplicative* logit adjustments are critical. Specifically, we prove for linear models that multiplicative adjustments find classifiers that are solutions to cost-sensitive support-vector-machines (CS-SVM), which by design create larger margins for minority classes. While effective in TPT, we also find that, at the start of the training, the same multiplicative weights can actually harm minorities. Instead, *additive* adjustments can speed up convergence by countering the initial negative effect of the multiplicative adjustments.

	Adjustment type		Inductive bias	
	Additive	Multiplicative	SVM	CS-SVM
LA-loss [MJR+20 +++]	✓		✓	
CDT-loss [YCZC20]		✓		✓
VS-loss [Ours]	✓	✓		✓

	Prior	Ours
Inductive bias	X	✓
Group imbalances	X	✓
Generalization + Tradeoffs	X	✓

Figure 1: Summary of contributions.

• **Algorithmic contributions:** Motivated by the above unique roles played by *both* multiplicative and additive weights, we propose the vector-scaling (VS) loss that captures existing techniques as special cases. We introduce a version of VS-loss tailored to group-sensitive settings, thus treating loss-adjustments for the two common types of imbalances (label/group) in a unifying way.

• **Generalization analysis and tradeoffs:** We present a sharp generalization analysis of the VS-loss on binary overparameterized Gaussian mixtures. Our formulae are explicit in terms of data geometry, class priors, parameterization ratio and the VS-loss hyperparameters leading to tradeoffs between balanced error, Equal Opportunity and standard error. We find that VS-loss can improve both balanced and standard error over CE. Interestingly, the optimal hyperparameters that minimize balanced error also optimize Equal Opportunity.

1.3 Connections to related literature

CE adjustments: The use of wCE for imbalanced data is rather old [XM89], but it becomes ineffective under overparameterization, e.g. [BL19]. This deficiency has led to the idea of *additive* label-based offset parameters ι_y on the logits (see (2)) [KHB+18, CWG+19, TWL+20, MJR+20, WCLL18]. Specifically, [MJR+20] proved that setting $\iota_y = \log(\pi_y)$ leads to a *Fisher consistent* loss (termed LA-loss) which outperformed other heuristics (e.g., focal loss [LGG+18]) on SOTA datasets. However, Fisher consistency is only relevant in the large sample size limit. Instead, our focus is on overparameterized models. In a recent work, [YCZC20] proposed the CDT-loss, which instead of additive uses *multiplicative* label-based parameters Δ_y on the logits (see (2)). The authors arrive at the CDT-loss as a heuristic means of compensating for the empirically observed phenomenon of *feature deviation of minorities* in deep nets, that is, the last-layer features deviating between training and test instances (see also [KK20]). Here, we arrive at the CDT-loss via a different viewpoint: we show that the multiplicative weights are necessary to move decision boundaries towards majority

classes when training overparameterized linear models in TPT. Moreover, we argue that while additive weights are not so effective in the TPT, they can help in the initial phase of training. Our analysis sheds light on the individual role of the two different modifications proposed in the literature and naturally motivates the VS-loss in (2). Compared to the above works we also demonstrate the successful use of VS-loss in overparamterized group-sensitive classification and show its competitive performance over popular alternatives based on distributionally robust optimization methods, e.g. [SKHL19, HNSS18, OSHL19].

Relation to vector-scaling calibration: Our naming of the VS-loss in Equation (2) is inspired by the vector scaling (VS) calibration [GPSW17], a *post-hoc procedure* that modifies the logits $\mathbf{v}$ of a neural net *after training* via $\mathbf{v} \to \boldsymbol{\Delta} \odot \mathbf{v} + \boldsymbol{\iota}$, where $\odot$ is the Hadamard product. Related to us, [ZCO20] shows that VS can improve calibration for imbalanced classes. But, in contrast to VS calibration, the multiplicative/additive scalings in our VS-loss are part of the loss and they directly affect training.

Blessings/curses of overparameterization: Overparameterization acts as a catalyst for state-of-the-art deep neural networks [NKB+19]. In terms of optimization, [SHN+18, OS19, JT18, AH18] show that gradient-based algorithms are *implicitly biased* towards certain min-norm solutions. Such solutions, are then analyzed in terms of generalization showing that they can in fact lead to benign overfitting [BLLT20, HMRT19, MMN18, MRSY19]. While implicit bias is key to benign overfitting it may also come with certain downsides. As an instance of this, we show that the parameters ω_y, ι_y in (2) can be ineffective in the interpolating regime in terms of balanced error/equal opportunity. Our argument essentially builds on characterizing the implicit bias of wCE/LA/CDT-losses. Related to this, [SRKL20] demonstrated the ineffectiveness of the weights ω_y in learning with groups.

Cost-sensitive SVM: [MSV10] arrived to CS-SVM by properly extending the SVM hinge-loss to guarantee Fisher consistency. Section 3.1 gives a new interpretation connecting it to our VS-loss.

2 Problem setup

Data model: Let training set $\{(\mathbf{x}_i, g_i, y_i)\}_{i=1}^n$ consisting of n i.i.d. samples from a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{G} \times \mathcal{Y}$. Here, $\mathcal{X} \subseteq \mathbb{R}^d$ is the input space, $\mathcal{Y} = [C] := \{1, \dots, C\}$ the set of C labels, and, $\mathcal{G} = [K]$ refers to group membership among $K \geq 1$ groups. The group-assignments are known for training data but are not known at test time. For concreteness, we focus here on the binary setting, that is $C = 2$ and $\mathcal{Y} = \{-1, +1\}$. Following standard convention, we assume throughout that $y = +1$ is the minority class. We discuss multiclass extensions in the Supplementary Material (SM).

Fairness metrics: Given a training set we build model $f_{\mathbf{w}} : \mathcal{X} \mapsto \mathcal{Y}$ parameterized by a parameter vector $\mathbf{w} \in \mathbb{R}^p$. For instance, linear models take the form $f_{\mathbf{w}} = \langle \mathbf{w}, h(\mathbf{x}) \rangle$ for some feature representation $h : \mathcal{X} \mapsto \mathbb{R}^p$. Given a new sample $\mathbf{x}$ we decide class membership $\hat{y} = \mathrm{sign}(f_{\mathbf{w}}(\mathbf{x}))$. The (standard) *risk* or **misclassification error** is $\mathcal{R} := \mathbb{P}\{\hat{y} \neq y\}$. Let $s = (y, g)$ define a subgroup for given values of y and g. We also define the **class-conditional risks** $\mathcal{R}_{\pm} = \mathbb{P}\{\hat{y} \neq y | y = \pm 1\}$, and, the **sub-group-conditional risks** $\mathcal{R}_{\pm,j} = \mathbb{P}\{\hat{y} \neq y | y = \pm 1, g = j\}, \quad j \in [K]$. The **balanced error** averages the conditional risks of the two classes: $\mathcal{R}_{\text{bal}} := (\mathcal{R}_+ + \mathcal{R}_-)/2$. Assuming $K = 2$ groups, the constraint of Equal Opportunity is satisfied if $\mathcal{R}_{+,1} = \mathcal{R}_{+,2}$ [HPS16]. More generally, we consider the (signed) **difference of equal opportunity (DEO)** $\mathcal{R}_{\text{deo}} := \mathcal{R}_{+,1} - \mathcal{R}_{+,2}$. In our experiments, we also measure the worst-case subgroup error $\max_{(y \in \pm 1, g \in [K])} \mathcal{R}_{y,g}$.

Terminal phase of training (TPT): Motivated by modern practice for training large models, we assume that $f_{\mathbf{w}}$ is overparameterized so that the training error $\mathcal{R}_{\text{train}} = \frac{1}{n} \sum_{i \in [n]} \mathbb{1}[\mathrm{sign}(\mathrm{f}_{\mathbf{w}}(\mathbf{x}_\mathrm{i})) \neq \mathrm{y_i}]$ can be driven to zero. Typically training overparameterized models continues well-beyond zero training error while the training loss is being pushed toward zero. For convenience, we follow the terminology introduced in [PHD20] and call this the *terminal phase of training* (TPT).

Training Algorithms

Cross-entropy adjustments: We introduce the **vector-scaling (VS) loss**, which combines *both* additive and multiplicative logit adjustments, previously suggested in the literature in isolation. The following is the **binary VS loss** for labels $y \in \{\pm 1\}$, weight parameters $\omega_\pm > 0$, additive logit parameters $\iota_\pm \in \mathbb{R}$, and multiplicative logit parameters $\Delta_\pm > 0$:

$$\ell_{\text{VS}}(y, f_{\mathbf{w}}(\mathbf{x})) = \omega_y \cdot \log\left(1 + e^{\iota_y} \cdot e^{-\Delta_y y f_{\mathbf{w}}(\mathbf{x})}\right). \tag{1}$$

For imbalanced C-class datasets, the **VS loss** takes the following form:

$$\ell_{\text{VS}}(y, \mathbf{f}_{\mathbf{w}}(\mathbf{x})) = -\omega_y \log\left(e^{\Delta_y \mathbf{f}_y(\mathbf{x}) + \iota_y} \Big/ \sum_{c \in [C]} e^{\Delta_c \mathbf{f}_c(\mathbf{x}) + \iota_c}\right). \tag{2}$$

Here $\mathbf{f_w} : \mathbb{R}^d \to \mathbb{R}^C$ and $\mathbf{f_w}(\mathbf{x}) = [\mathbf{f}_1(\mathbf{x}), \ldots, \mathbf{f}_C(\mathbf{x})]$ is the vector of logits. The VS-loss (Eqns. (1),(2)) captures existing techniques as special cases by tuning accordingly the additive/multiplicative hyperparameters; see also Figure 1. Specifically, we recover the following: (i) **weighted CE (wCE) loss** by $\Delta_y = 1, \iota_y = 0, \omega_y = \pi_y^{-1}$; (ii) **LA-loss** by $\Delta_y = 1$; (iii) **CDT-loss** by $\iota_y = 0$.
When the goal is to (additionally) ensure fairness with respect to a *sensitive group* within the classes, we naturally extend the VS-loss by introducing in (1) parameters $(\Delta_{y,g}, \iota_{y,g}, \omega_{y,g})$ that depend *both* on the class and group membership (specified by y and g, respectively). Our proposed binary **group-sensitive VS-loss** is formulated as follows (multiclass version can be defined accordingly):

$$\ell_{\text{Group-VS}}(y, g, f_{\mathbf{w}}(\mathbf{x})) = \omega_{y,g} \cdot \log\left(1 + e^{\iota_{y,g}} \cdot e^{-\Delta_{y,g} y f_{\mathbf{w}}(\mathbf{x})}\right). \tag{3}$$

CS-SVM: For linear classifiers with feature map $h : \mathcal{X} \to \mathbb{R}^p$, the cost-sensitive SVM [MSV10] learns model $f_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, h(\mathbf{x}) \rangle$ by solving:

$$\min_{\mathbf{w}} \|\mathbf{w}\|_2 \text{ sub. to} \begin{cases} \langle \mathbf{w}, h(\mathbf{x}_i) \rangle \geq \delta & , y_i = +1 \\ \langle \mathbf{w}, h(\mathbf{x}_i) \rangle \leq -1 & , y_i = -1 \end{cases}, i \in [n], \tag{4}$$

for hyper-parameter $\delta \in \mathbb{R}_+$ representing the ratio of margins between classes. Note that $\delta = 1$ corresponds to (standard) SVM. The more general formulation, naturally allows tuning $\delta > 1$ (resp. $\delta < 1$) to favor a larger margin $\delta / \|\hat{\mathbf{w}}_\delta\|_2$ for the minority class vs $1/\|\hat{\mathbf{w}}_\delta\|_2$ for the majority class. Thus, $\delta \to +\infty$ (resp. $\delta \to 0$) corresponds to the extreme scenario where the decision boundary starts right at the boundary of class $y = -1$ (resp. $y = +1$).
Group-sensitive SVM: Consider the following *group-sensitive* version of CS-SVM, which we refer to as GS-SVM: (For simplicity, we focus on $K = 2$ protected groups.) $\min_{\mathbf{w}} \|\mathbf{w}\|_2$ such that $\forall i \in [n]$: $y_i \langle \mathbf{w}, h(\mathbf{x}_i) \rangle \geq \delta$ if $g_i = 1$, and, $y_i \langle \mathbf{w}, h(\mathbf{x}_i) \rangle \geq 1$ if $g_i = 2$. By tuning $\delta > 1$, GS-SVM favors larger margin for the sensitive group $g = 1$. Refined versions are also possible when classes are also imbalanced themselves by modifying the constraints to $y_i \langle \mathbf{w}, \mathbf{h}(\mathbf{x}_i) \rangle \geq \delta_{y_i, g_i}$ for positive hyper-parameters $\delta_{y,g}, y \in \{\pm 1\}, g = 1, 2$ for each sub-group. Both CS-SVM and GS-SVM are feasible if and only if data are linearly separable (see SM). However, we caution that the role of the hyper-parameters in GS-SVM is in general harder to interpret as "margin-ratios" since its constraints may (not) be active depending on data geometry.

3 Insights on the VS-loss

In this section, we investigate the different roles of ω_y, ι_y and Δ_y in the VS-loss.

3.1 CDT-loss vs LA-loss: Why multiplicative weights?

We start by demonstrating the unique role played by the multiplicative weights Δ_y through a motivating experiment on synthetic data in Fig. 2. We generate a binary Gaussian-mixture dataset of $n = 100$ examples in $\mathbb{R}^{300}$ with data means sampled independently from the Gaussian distribution and normalized such that $\|\boldsymbol{\mu}_{+1}\|_2 = 2\|\boldsymbol{\mu}_{-1}\|_2 = 4$. We set prior $\pi_+ = 0.1$ for the minority class $+1$. For varying model size values $p \in [5 : 5 : 50, 75 : 25 : 300]$ we trained linear classifier $f_{\mathbf{w}}(x) = \langle \mathbf{w}, h(\mathbf{x}) \rangle$ using only the first p features, that is $h(\mathbf{x}) = \mathbf{x}(1 : p) \in \mathbb{R}^p$. This allows us to investigate performance with a varying parameterization ratio $\gamma = p/n$. [1] We train the model $\mathbf{w}$ using one of the following special cases of the VS-loss: (i) *CDT-loss* with $\Delta_+ = \delta_\star^{-1}, \Delta_- = 1$ ($\delta_\star > 0$ is set to the value shown in the inset plot; see SM for details). (ii) *LA-loss:* $\iota_+ = \pi^{-1/4}, \iota_- = (1-\pi)^{-1/4}$ (special case termed *LDAM-loss* in [CWG+19]). (iii) *LA loss:* $\iota_+ = \log\left(\frac{1-\pi}{\pi}\right), \iota_- = \log\left(\frac{\pi}{1-\pi}\right)$ (Fisher-consistent values proposed in [MJR+20]). We run gradient descent and average over 25 independent experiments. The *balanced error* is computed on a test set of size $10^4$and reported values are shown in red/blue/black markers. We also plot the training errors. Observe that for all losses the training error is zero for $\gamma \gtrsim 0.45$. The shaded region highlights the transition to the overparameterized regime where p is large enough to drive the training error to zero. In this regime, we continued training in the TPT (see SM for details). The depicted results reveal the following clear message: *The CDT loss has better balanced-error performance compared to the LA loss when both trained in TPT.* Moreover, the figure offers an intuitive explanation by revealing a connection to max-margin classifiers. We observe that *in the TPT: (a) LA-loss performs the same as SVM; (b) CDT-loss performs the same as CS-SVM.*

[1] Such simple models have been used in e.g. [HMRT19, BHX19, DKT19, KT20, CLOT20, DL20, SC19] for analytic studies of double descent [BMM18a, BRT19, MM19, NKB+19] in terms of misclassification error. Although not our focus here, we note that Fig. 2(a) reveals a double descent for the balanced error.

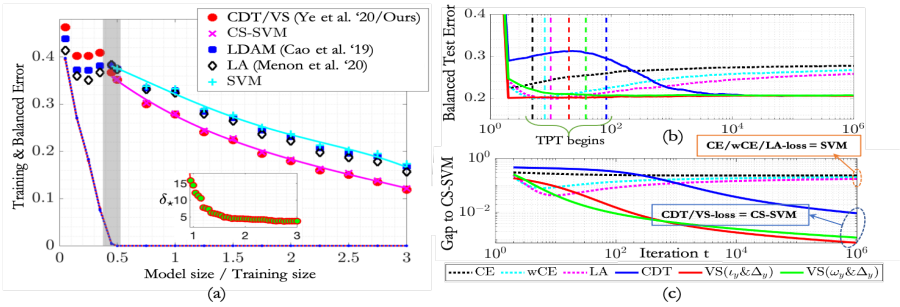


Figure 2: Insights on various cost-sensitive modifications of the CE-loss. **(a)** CDT has superior balanced-error performance over LA in the separable regime. Also, its performance matches that of CS-SVM, unlike LA matching SVM; Sec. 3.1 for more details. Solid lines follow theory of Sec. 4. **(b)** Although critical in TPT, multiplicative weights (aka CDT) can harm minority classes in initial phase of training by guiding the classifier in the wrong direction. Properly tuned additive weights (aka LA) can mitigate this effect and speed up convergence. This explains why VS can be superior compared to CDT (see Observation 1). Dashed lines show where TPT starts for each loss. **(c)** CDT and VS converge to CS-SVM, unlike LA and wCE. We prove this in Theorem 1.

We formalize the above empirical observations in the theorem below, for arbitrary linearly separable datasets (not necessarily Gaussian mixtures as in our experiment). Specifically, we consider a sequence of norm-constrained problems minimizing the VS-loss. As the norm constraint R gets big and we approach the original unconstrained VS-loss, we prove that the direction of the constrained minimizer $\mathbf{w}_R$ converges to the direction of the CS-SVM solution $\hat{\mathbf{w}}_\delta$ for $\delta = \Delta_-/\Delta_+$.

Theorem 1 (VS-loss=CS-SVM) *Fix a binary training set $\{\mathbf{x}_i, y_i\}_{i=1}^n$ with at least one example from each of the two classes. Assume feature map $h(\cdot)$ such that the data are linearly separable, that is $\exists \mathbf{w}: y_i \mathbf{w}^T h(\mathbf{x}_i) \geq 1, \forall i \in [n]$. Consider training a linear model $f_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, h(\mathbf{x}) \rangle$ by minimizing the VS-loss $\mathcal{L}_n(\mathbf{w}) := \sum_{i \in [n]} \ell_{\mathrm{VS}}(y_i, f_{\mathbf{w}}(\mathbf{x}_i))$ with ℓ_{VS} defined in (1) for positive parameters $\Delta_\pm, \omega_\pm \geq 0$ and arbitrary $\iota_\pm$. Define the norm-constrained optimal classifier $\mathbf{w}_R = \arg\min_{\|\mathbf{w}\|_2 \leq R} \mathcal{L}_n(\mathbf{w})$. Let $\hat{\mathbf{w}}_\delta$ be the CS-SVM solution of (4) with $\delta = \Delta_-/\Delta_+$. Then, $\lim_{R\to\infty} \mathbf{w}_R / \|\mathbf{w}_R\|_2 = \hat{\mathbf{w}}_\delta / \|\hat{\mathbf{w}}_\delta\|_2$.*

On the one hand, the theorem makes clear that $\omega_\pm$ and $\iota_\pm$ become ineffective in the TPT as they all result in the same SVM solutions. On the other hand, the multiplicative parameters $\Delta_\pm$ lead to the same classifier as that of CS-SVM, thus favoring solutions that move the classifier towards the majority class provided that $\Delta_- > \Delta_+ \Leftrightarrow \delta > 1$. The proof is given in the SM together with extensions for multiclass datasets. Finally, in the SM, we strengthen Theorem 1 by characterizing the *implicit bias* of gradient-flow iterations on VS-loss towards the CS-SVM solution with $\delta = \Delta_-/\Delta_+$. We also show that the group-sensitive VS-loss with $\Delta_{y,g} = \Delta_g$ converges to the corresponding GS-SVM.

Remark 1 *Thm 1 is reminiscent of Thm. 2.1 in [RZH03] who showed for a regularized ERM with CE-loss that when the regularization parameter vanishes, the normalized solution converges to the SVM classifier. Our result connects nicely to [RZH03] extending their theory to VS-loss / CS-SVM, as well as, to the group-case. In a similar way, our result on the inductive bias of gradient-flow on the VS-loss connects to more recent works [SHN+18, JT18] that pioneered corresponding results for CE-loss. Although related, our results on the properties of the VS-loss are not obtained as special cases of these existing works. As a final remark, in Fig. 2(b,c) we kept constant learning rate 0.1. Significantly faster convergence is observed with normalized GD schemes [NLG+19, JT21]; a detailed numerical study in SM.*

3.2 VS-loss: Best of two worlds

We have shown that multiplicative weights $\Delta_\pm$ are responsible for good balanced accuracy in the TPT. Here, we show that, at the initial phase of training, the same multiplicative weights can actually harm the minority classes. The following observation supports this claim.

Observation 1 *Assume $f_{\mathbf{w}}(x) = 0$ at initialization. Then, the gradients of CDT-loss with multiplicative logit factors Δ_y are identical to the gradients of wCE-loss with weights $\omega_y = \Delta_y$. Thus, we conclude the following where assume that $y = +1$ is minority. On the one hand, wCE-loss, which typically sets $\omega_+ > \omega_-$ (e.g., $\omega_y = 1/\pi_y$), helps minority examples by weighing down the loss over majority. On the other hand, the CDT-loss requires the reverse direction $\Delta_+ < \Delta_-$ as per Theorem 1, thus initially it guides the classifier in the wrong direction to penalize minorities.*

To see why the above is true note that for $f_{\mathbf{w}}(x) = \langle \mathbf{w}, h(\mathbf{x}) \rangle$ the gradient of VS-loss is $\nabla_{\mathbf{w}} \ell_{\mathrm{VS}}(y, f_{\mathbf{w}}(\mathbf{x})) = -\omega_y \Delta_y \, \sigma\big(-\Delta_y y f_{\mathbf{w}}(x) + \iota_y\big) \cdot y h(\mathbf{x})$ where $\sigma(t) = (1 + \exp(-t))^{-1}$ is the sigmoid function. It is then clear that at $f_{\mathbf{w}}(\mathbf{x}) = 0$, the logit factor Δ_y plays the same role as the weight ω_y. From Theorem 1, we know that pushing the margin towards majorities (which favors balancing the conditional errors) requires $\Delta_+ < \Delta_-$. Thus, gradient of minorities becomes smaller, initially pushing the optimization in the wrong direction. Now, we turn our focus at the impact of ι_y's at the start of training. Noting that $\sigma(\cdot)$ is increasing function, we see that setting $\iota_+ > \iota_-$ increases the gradient norm for minorities. This leads us to a second observation: *By properly tuning the additive logit adjustments ι_y we can counter the initial negative effect of the multiplicative adjustment, thus speeding up training.* The observations above naturally motivated us to formulate the VS-loss in Eqn. (2) bringing together the best of two worlds: the Δ_y's that play a critical role in the TPT and the ι_y's that compensate for the harmful effect of the Δ_y's in the beginning of training.

Figure 2(b,c) illustrate the discussion above. In a binary linear classification setting similar to that of Figure 2(a), we investigate the effect of the additive adjustments on the training dynamics. Specifically, we train using gradient descent: (i) *CE*; (ii) *wCE* with $\omega_y = 1/\pi_y$; (iii) *LA-loss* with $\iota_y = \log(1/\pi_y)$; (iv) *CDT-loss* with $\Delta_+ = \delta_\star^{-1}, \Delta_- = 1$; (v) *VS-loss* with $\Delta_+ = \delta_\star^{-1}, \Delta_- = 1$, $\iota_y = \log(1/\pi_y)$ and $\omega_y = 1$; (vi) *VS-loss* with same Δ, $\iota_y = 0$ and $\omega_y = 1/\pi_y$. Figures 2(b) and (c) plot balanced test error $\mathcal{R}_{\text{bal}}$ and angle-gap to CS-SVM solution as a function of gradient descent iterations (constant step size 0.1) for each algorithm. The vertical dashed lines mark the iteration after which training error stays zero and we enter the TPT. Observe in Fig. 2(c) that CDT/VS-losses, both converge to the CS-SVM solution as TPT progresses verifying Theorem 1. This also results in lowest test error in the TPT in Fig. 2(b). However, compared to CDT-loss, the VS-loss enters faster in the TPT and converges orders of magnitude faster to small values of $\mathcal{R}_{\text{bal}}$. Note in Fig. 2(c) that this behavior is correlated with the speed at which the two losses converge to CS-SVM. Following the discussion above, we attribute this favorable behavior at the initial phase of training to the inclusion of the ι_y's. This is also supported by Fig. 2(c) as we see that LA-loss (but also wCE) achieves significantly better values of $\mathcal{R}_{\text{bal}}$ at the first stage of training compared to CDT-loss.

4 Generalization analysis and tradeoffs

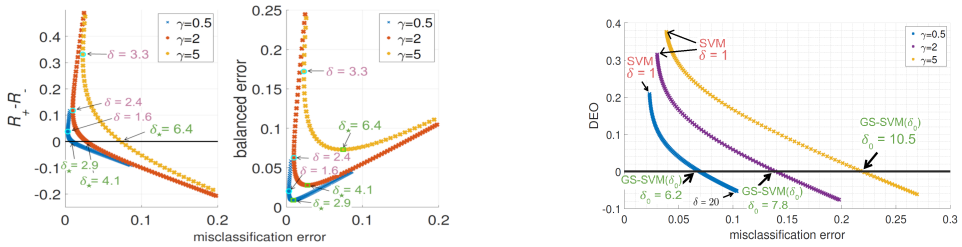
Our results in the previous section regarding VS-loss/CS-SVM hold for arbitrary linearly-separable training datasets. Here, under additional distributional assumptions, we establish a sharp asymptotic theory for VS-loss/CS-SVM and their group-sensitive counterparts.

Data model: We study a binary Gaussian-mixture generative model (GMM) for the data distribution $\mathcal{D}$. For the label $y \in \{\pm 1\}$ let $\pi := \mathbb{P}\{y = +1\}$. Group membership is decided conditionally on the label such that $\forall j \in [K] : \mathbb{P}\{g = j | y = \pm 1\} = p_{\pm,j}$, with $\sum_{j \in [K]} p_{+,j} = \sum_{j \in [K]} p_{-,j} = 1$. Finally, the feature conditional given label y and group g is a multivariate Gaussian of mean $\boldsymbol{\mu}_{y,g} \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma}$, that is, $\mathbf{x} \big| (y, g) \sim \mathcal{N}(\boldsymbol{\mu}_{y,g}, \boldsymbol{\Sigma})$. Specifically for *label-imbalances*, we let $K = 1$ and $\mathbf{x} \big| y \sim \mathcal{N}(\boldsymbol{\mu}_y, \mathbf{I}_d)$ (see SM for $\boldsymbol{\Sigma} \neq \mathbf{I}_d$). For *imbalances with respect to group-membership*, we focus on two groups $K = 2$ with $p_{+,1} = p_{-,1} = p < 1 - p = p_{+,2} = p_{-,2}$, $j = 1, 2$ and $\mathbf{x} \,|\, (y, g) \sim \mathcal{N}(y \boldsymbol{\mu}_g, \mathbf{I}_d)$. In both cases, let $\mathbf{M}$ denote the matrix of means (that is $\mathbf{M} = [\boldsymbol{\mu}_+ \quad \boldsymbol{\mu}_-]$ and $\mathbf{M} = [\boldsymbol{\mu}_1 \quad \boldsymbol{\mu}_2]$, respectively) and consider the eigen-decomposition of its Gramian: $\mathbf{M}^T \mathbf{M} = \mathbf{V} \mathbf{S}^2 \mathbf{V}^T, \quad \mathbf{S} \succ \mathbf{0}_{r \times r}, \mathbf{V} \in \mathbb{R}^{2 \times r}, r \in \{1, 2\}$, with $\mathbf{S}$ an $r \times r$ diagonal positive-definite matrix and $\mathbf{V}$ an orthonormal matrix obeying $\mathbf{V}^T \mathbf{V} = \mathbf{I}_r$. We study linear classifiers with $h(\mathbf{x}) = \mathbf{x}$.

Learning regime: We focus on a regime where the training data are linearly separable. For the data models above, it turns out that linear separability undergoes a sharp phase-transition as $d, n \to \infty$ at a proportional rate $\gamma = \frac{d}{n}$. That is, there exist thresholds $\gamma_\star := \gamma_\star(\mathbf{V}, \mathbf{S}, \pi) \leq 1/2$ and $\widetilde{\gamma}_\star := \widetilde{\gamma}_\star(\mathbf{V}, \mathbf{S}, \pi, p) \leq 1/2$ for label- and group- cases respectively, such that data are linearly separable with probability approaching one provided that $\gamma > \gamma_\star$ or $\gamma > \widetilde{\gamma}_\star$ [CS+20, MRSY19, DKT19, KA20, KT21]. We defer formal statements and explicit definitions of the thresholds to the SM. Onwards, we assume $\gamma > \gamma_\star$ or $\gamma > \widetilde{\gamma}_\star$, so that CS-SVM or GS-SVM are feasible with probability approaching 1.

Notation/definitions: We use $\xrightarrow{P}$ to denote convergence in probability and $Q(\cdot)$ the tail of the standard normal distribution. We let $(x)_- := \min\{x, 0\}$, denote $\mathbb{1}[\mathcal{E}]$ the indicator function of event $\mathcal{E}$, $\mathcal{B}_2^r$ the unit ball in $\mathbb{R}^r$, and $\mathbf{e}_1 = [1, 0]^T, \mathbf{e}_2 = [0, 1]^T$ the standard basis vectors in $\mathbb{R}^2$. We further need the following definitions before we state our results. Let random variables as follows:

$$\begin{cases} G \sim \mathcal{N}(0, 1), \quad Y \text{ symmetric Bernoulli with } \mathbb{P}\{Y = +1\} = \pi, \\ S \text{ takes values 1 or 2 with } \mathbb{P}\{Y = 1\} = p, \end{cases}$$



(a) **(b)**

Figure 3: Trade-offs between misclassification error and error imbalance / balanced error / DEO on GMM data achieved by **(a)** CS-SVM for class prior $\pi = 0.05$ and **(b)** GS-SVM for group prior $p = 0.05$, as a function of the margin-ratio hyperparameter $\delta \geq 1$ and for different values of γ. The plots are generated using predictions of Theorems 2 and 3. See text for interpretations.

$$\begin{cases} E_Y = \mathbf{e}_1 \mathbb{1}[Y=1] - \mathbf{e}_2 \mathbb{1}[Y=-1] & \text{and} \quad \widetilde{E}_S = \mathbf{e}_1 \mathbb{1}[S=1] + \mathbf{e}_2 \mathbb{1}[S=2] \\ \Delta_Y = \delta \cdot \mathbb{1}[Y=+1] + \mathbb{1}[Y=-1] & \text{and} \quad \widetilde{\Delta}_S = \delta \cdot \mathbb{1}[S=1] + 1 \cdot \mathbb{1}[S=2], \quad \text{for } \delta > 0. \end{cases}$$

With these define the following two key functions $\eta_\delta, \widetilde{\eta}_\delta : \mathbb{R}_{\geq 0} \times \mathcal{B}_2^r \times \mathbb{R} \to \mathbb{R}$ as $\eta_\delta(q, \boldsymbol{\rho}, b) := \mathbb{E}\Big[\big(G + E_Y^T \mathbf{V}\mathbf{S}\boldsymbol{\rho} + \frac{bY - \Delta_Y}{q}\big)_-^2\Big] - (1 - \|\boldsymbol{\rho}\|_2^2)\gamma$ and $\widetilde{\eta}_\delta(q, \boldsymbol{\rho}, b) := \mathbb{E}\Big[\big(G + \widetilde{E}_S^T \mathbf{V}\mathbf{S}\boldsymbol{\rho} + \frac{bY - \widetilde{\Delta}_S}{q}\big)_-^2\Big] - (1 - \|\boldsymbol{\rho}\|_2^2)\gamma$. Finally, define $(q_\delta, \boldsymbol{\rho}_\delta, b_\delta)$ as the *unique* triplet (see SM for uniqueness proof) satisfying $\eta_\delta(q_\delta, \boldsymbol{\rho}_\delta, b_\delta) = 0$ and $(\boldsymbol{\rho}_\delta, b_\delta) := \arg\min_{\|\boldsymbol{\rho}\|_2 \leq 1, b \in \mathbb{R}} \eta_\delta(q_\delta, \rho, b)$. Similarly $(\widetilde{q}_\delta, \widetilde{\boldsymbol{\rho}}_\delta, \widetilde{b}_\delta)$ is the unique triplet satisfying same equations, but with η_δ replaced with the function $\widetilde{\eta}_\delta$. Note that these triplets can be easily computed numerically for given values of γ, δ, π, p and means' Gramian $\mathbf{M}^T\mathbf{M} = \mathbf{V}\mathbf{S}^2\mathbf{V}^T$.

Analysis of CS/GS-SVM: We first present a precise analysis of the balanced error of CS-SVM.

Theorem 2 (Balanced error of CS-SVM) *Consider GMM data with label imbalances and learning regime as described above. Consider the CS-SVM classifier in (4) with $h(\mathbf{x}) = \mathbf{x}$, intercept b (i.e. constraints $\langle \mathbf{x}, \mathbf{w} \rangle + b \geq \delta$ or ≥ 1 in (4)) and fixed margin-ratio $\delta > 0$. Define $\overline{\mathcal{R}}_+ := Q\left(\mathbf{e}_1^T \mathbf{V}\mathbf{S}\boldsymbol{\rho}_\delta + b_\delta/q_\delta\right)$ and $\overline{\mathcal{R}}_- := Q\left(-\mathbf{e}_2^T \mathbf{V}\mathbf{S}\boldsymbol{\rho}_\delta - b_\delta/q_\delta\right)$. Then, in the limit of $n, d \to \infty$ with $d/n = \gamma > \gamma_\star$, it holds that $\mathcal{R}_+ \xrightarrow{P} \overline{\mathcal{R}}_+$ and $\mathcal{R}_- \xrightarrow{P} \overline{\mathcal{R}}_-$. In particular, $\mathcal{R}_{bal} \xrightarrow{P} \overline{\mathcal{R}}_{bal} := (\overline{\mathcal{R}}_+ + \overline{\mathcal{R}}_-)/2$.*

The theorem's proof further reveals that $(\|\hat{\mathbf{w}}_\delta\|_2, \frac{\hat{\mathbf{w}}_\delta^T \boldsymbol{\mu}_+}{\|\hat{\mathbf{w}}_\delta\|_2}, \frac{\hat{\mathbf{w}}_\delta^T \boldsymbol{\mu}_-}{\|\hat{\mathbf{w}}_\delta\|_2}, \hat{b}_\delta) \xrightarrow{P} (q_\delta, \mathbf{e}_1^T \mathbf{V}\mathbf{S}\boldsymbol{\rho}_\delta, \mathbf{e}_2^T \mathbf{V}\mathbf{S}\boldsymbol{\rho}_\delta, b_\delta)$. Thus, b_δ is the asymptotic value of the intercept, q_δ^{-1} is the asymptotic value of the classifier's margin $1/\|\mathbf{w}_\delta\|_2$ to the majority class, and $\boldsymbol{\rho}_\delta$ characterizes the asymptotic alignment of the classifier's hyperplane with the class means. The proof uses the convex Gaussian min-max theorem (CGMT) framework [Sto13, TOH15]; see SM for background, related works, the proof, as well as, (a) simpler expressions when the means are antipodal ($\pm\boldsymbol{\mu}$) and (b) extensions to general covariance model ($\boldsymbol{\Sigma} \neq \mathbf{I}$). The experiment (solid lines) in Figure 2(a) suggests the validity of the theorem's asymptotics predictions even at relatively small dimensions; see SM for additional experiments.

Next, we characterize the DEO of GS-SVM. Although similar in nature, the result differs to Thm. 2 since now each class itself is a Gaussian mixture.

Theorem 3 (DEO of GS-SVM) *Let group-sensitive GMM data and learning regime as described above. Consider the GS-SVM classifier with $h(\mathbf{x}) = \mathbf{x}$, intercept b and fixed margin-ratio $\delta > 0$ (corresponding to group VS-loss with $\Delta_{y,g} = \Delta_g, g = 1, 2$ such that $\delta = \Delta_2/\Delta_1$). Then, in the limit of $n, d \to \infty, d/n = \gamma > \widetilde{\gamma}_\star$ it holds $\mathcal{R}_{\pm,i} \xrightarrow{P} Q\big(\mathbf{e}_i^T \mathbf{V}\mathbf{S}\widetilde{\boldsymbol{\rho}}_\delta \pm \widetilde{b}_\delta/\widetilde{q}_\delta\big), i = 1, 2$. In particular, $\mathcal{R}_{\text{deo}} \xrightarrow{P} Q\big(\mathbf{e}_1^T \mathbf{V}\mathbf{S}\widetilde{\boldsymbol{\rho}}_\delta + \widetilde{b}_\delta/\widetilde{q}_\delta\big) - Q\big(\mathbf{e}_2^T \mathbf{V}\mathbf{S}\widetilde{\boldsymbol{\rho}}_\delta + \widetilde{b}_\delta/\widetilde{q}_\delta\big)$.*

Tradeoffs: The theorems above allow us to study tradeoffs between misclassification / balanced error / DEO in Fig. 3. Fig. 3(a) focuses on label imbalances. We make the following observations. (1) The optimal value $\delta_\star$ minimizing $\mathcal{R}_{\text{bal}}$ also achieves perfect balancing between the conditional errors of the two classes, that is $\mathcal{R}_+ = \mathcal{R}_- = Q(\frac{\ell_- + \ell_+}{2})$. In the SM, we prove this interesting property. The key to this is that, we are able to derive an explicit *closed-form* formula for $\delta_\star$ that only requires computing the triplet $(q_1, \boldsymbol{\rho}_1, b_1)$ for $\delta = 1$ corresponding to the standard SVM. We remark that this is rather unexpected given the seemingly involved nonlinear dependency of $\mathcal{R}_{\text{bal}}$ on δ in Thm. 2. In the SM, we also use this formula for the $\delta_\star$ to derive a theory-inspired heuristic for hyperparameter

tuning which shows good empirical performance on simple datasets such as imbalanced MNIST. (2) The value of δ minimizing standard error $\mathcal{R}$ (shown in magenta) is not equal to 1, hence CS-SVM also improves $\mathcal{R}$ (not only $\mathcal{R}_{\text{bal}}$). In Fig. 3(b), we investigate the effect of δ and the improvement of GS-SVM over SVM. The largest DEO and smallest misclassification error are achieved by the SVM ($\delta = 1$). But, with increasing δ, misclassification error is traded-off for reduction in absolute value of DEO. Interestingly, for some $\delta_0 = \delta_0(\gamma)$ (with value increasing with γ) GS-SVM guarantees Equal Opportunity (EO) $\mathcal{R}_{\text{deo}} = 0$ (without explicitly imposing such constraints as in [OA18, DOBD+18]).

5 Experiments

We present experimental results that further justify our theoretical findings in the previous sections. Additional experiments validating the insights of Section 3 are deferred to the SM.

5.1 Label-imbalanced data

Our first experiment on class-imbalanced CIFAR-10 aims to show that *non-trivial combinations* of additive/multiplicative logit adjustments in the VS-loss (2) can improve balanced error performance over *individual* additive/multiplicative adjustments in LA/CDT-losses.

Datasets: Table 1 evaluates LA/CDT/VS-losses on imbalanced instances of CIFAR-10 generated by [CWG+19]. Following [CWG+19], we consider: (1) *STEP* imbalance [BMM18b], which reduces the number of training instances of half of the classes to a fixed size. (2) Long-tailed (*LT*) imbalance [CJL+19], which exponentially decreases the number of training images across different classes. In both cases we set an imbalance ratio $N_{\max}/N_{\min} = 100$, where $N_{\max} = \max_y N_y$, $N_{\min} = \min_y N_y$ and N_y are the sample sizes of each class y. For consistency with [HZRS16, CWG+19, MJR+20, YCZC20] we keep the test set balanced and in addition to evaluating our models on it, we treat it as our validation set and use it to tune our hyperparameters. More sophisticated tuning strategies (perhaps using bi-level optimization) are deferred to future work. We use standard data-augmentation exactly as in [HZRS16, CWG+19, MJR+20, YCZC20]. Additional results are in the SM.

Model and Baselines: We compare performance of the following training losses: *(1) CE-loss* (standard cross-entropy). *(2) Re-Sampling*, where every training instance is included in the batch with probability ${\pi_y}^{-1}$. *(3) wCE*, weighted CE with weights $\omega_y = {\pi_y}^{-1}$. *(4) LDAM-loss* [CWG+19], a special case of LA-loss where $\iota_y = \frac{1}{2}(N_{\min}/N_y)^{1/4}$ is *subtracted* from the logits. *(5) LDAM-DRW* [CWG+19], combining LDAM with deferred re-weighting. *(6) LA-loss* [MJR+20], with the Fisher-consistent parametric choice $\iota_y = \tau \log(\pi_y)$ added to the logits. *(7) CDT-loss* [YCZC20], where the logits are multiplied by $\Delta_y = (N_y/N_{max})^{\gamma}$. *(8) VS-loss*, with combined hyperparameters $\iota_y = \tau \log(\pi_y)$ and $\Delta_y = (N_y/N_{\max})^{\gamma}$, parameterized by $\tau, \gamma > 0$ respectively. The three papers that first introduced (5)-(7) above, all trained CIFAR for a different number of epochs and with dissimilar regularization and learning rate schedules. For consistency, we follow the training setting by [CWG+19]. Thus, for LDAM we adapt results reported by [CWG+19], but for LA and CDT, we reproduce our own in that setting.

Table 1: Top-1 balanced accuracy results on balanced validation set (%); Average over 5 runs

Dataset	CIFAR 10	
Imbalance Type	**LT-100**	**STEP-100**
CE	71.1	65.3
Re-Sampling	71.2	65.0
wCE	72.6	67.3
LDAM [CWG+19].	73.35	66.58
LDAM-DRW [CWG+19]	77.03	76.92
LA ($\tau = 1$) [MJR+20]	77.15	73.67
CDT ($\gamma = \gamma^*$) [YCZC20]	79.22	73.26
VS ($\tau = \tau^*, \gamma = \gamma^*$)	**80.68**	**79.10**

Technical details: Following [CWG+19], we train a ResNet-32 [HZRS16], using batch size 128 and SGD with momentum 0.9 and weight decay 2×10^{-4}. For the first 5 epochs we use a linear warm up schedule until baseline learning rate of 0.1. We train for a total of 200 epochs, while decaying our learning rate by 0.1 at epochs 160 and 180. We use PyTorch [PGC+17] building on codes provided by [CWG+19, YCZC20]. Training is performed on 2 NVIDIA RTX-3080 GPUs. We remark that the values for LDAM (adapted from [CWG+19]) used learning rate decay 0.01 and last-layer feature/classifier normalization. We have found convergence difficult otherwise. For other losses, we do *not* use the above normalization of weights to isolate the impact of loss modifications.

Results: Table 1 shows Top-1 accuracy on the balanced validation set (averaged over 5 runs). We use a grid to pick the best γ for the CDT and the best pair (τ, γ) for VS on the validation set (see SM for tuning details). Since VS includes LA and CDT as special cases (corresponding to $\gamma = 0$ and $\tau = 0$ respectively), we expect that it is at least as good as the latter over our hyper-parameter grid search. We find that the optimal (τ^*, γ^*)-pairs correspond to non-trivial combinations of each individual parameter ($(1.25, 0.2)$ for LT-100 and $(1.75, 0.1)$ for STEP-100). Thus, VS-loss has better balanced error as shown in the table. Additional results, including histograms of the per class error rates and plots of training dynamics are included in the SM.

5.2 Group-sensitive data

The message of our second experiment on the Waterbirds dataset is three-fold. (1) We demonstrate experimentally the practical relevance of logit-adjusted CE modifications to settings with imbalances at the level of (sub)groups. (2) We show that such methods are competitive to alternative state-of-the-art; specifically, distributionally robust optimization (DRO) algorithms. (3) We demonstrate that logit-adjustments can be properly combined with DRO methods for even superior performance.

Dataset: We study a setting with spurious correlations —strong associations between label and background in image classification— which can be cast as a subgroup-sensitive classification problem [SKHL19]. We consider the Waterbirds dataset created by [SKHL19] from the original CUB dataset [WBW+11]. The goal is to classify images as either 'waterbirds' or 'landbirds', while their background —either 'water' or 'land'— can be spuriously correlated with the type of birds. Formally, each example has label $y \in \mathcal{Y} = \{\pm 1\} \equiv \{\text{waterbird}, \text{landbird}\}$ and belongs to a group $g \in \mathcal{G} = \{\pm 1\} \equiv \{\text{water}, \text{land}\}$. Let then $s = (y, g) \in \{\pm 1\} \times \{\pm 1\}$ be the four sub-groups with $(+1, -1)$, $(-1, +1)$ being minorities (specifically, $\hat{p}_{+1,+1} = 0.22, \hat{p}_{+1,-1} = 0.012, \hat{p}_{-1,+1} = 0.038$ and $\hat{p}_{-1,-1} = 0.73$.). Denote N_s the number of training examples belonging to sub-group s and $N_{\max} := \max_s N_s$. For notational consistency with Sec. 2, we note that the imbalance here is in subgroups; thus, Group-VS-loss in (3) consists of logit adjustments that depend on the subgroup $s = (y, g)$.

Model and Baselines: We use the code by [SKHL19] to train a ResNet50 [HZRS16] starting with pretrained weights on Imagenet (as per [SKHL19]). We use SGD with parameters exactly as in [SKHL19] (see SM for details) except we reduce batch size to 64 (from 128) due to memory constraints. Let $\beta_{s=(y,g)} = (N_{(y,g)}/N_{\max})$. We propose training with the group-sensitive VS-loss in (3) with $\Delta_{y,g} = \Delta_s = \beta_s^{\gamma}$ and $\iota_s = -\beta_s^{-\gamma}$ with $\gamma = 0.3$. We compare against CE and DRO used in [SKHL19]. We also implement a new training scheme that combines VS+DRO. In the SM, we show additional results for LA/CDT-losses (not previously used in group contexts). For fair comparison, we rerun the baseline experiments with CE and report our reproduced numbers. Since class $+1$ has no special meaning in this dataset, we use a symmetric measure of DEO Symm-DEO $= (|\mathcal{R}_{(+1,+1)} - \mathcal{R}_{(+1,-1)}| + |\mathcal{R}_{(-1,+1)} - \mathcal{R}_{(-1,-1)}|)/2$. We also report balanced and worst sub-group accuracies . We did not fine-tune γ as the heuristic choice already shows the benefit of VS-loss in this context. We expect further improvements tuning over validation examples.

Table 2: Symmetric DEO, balanced and worst-case sub-group accuracies on Waterbirds dataset; averages over 5 runs (SM for error bars)

Loss	Symm. DEO	Bal. acc.	Worst acc.
CE	25.1	84.9	67.6
VS	**17.8**	**88.2**	**77.1**
CE + DRO	16.3	88.6	75.7
VS + DRO	**11.8**	**90.3**	**78.8**

Results: Table 2 reports the test values obtained at the end of the last epoch (300 in total; see SM for details). We find that the VS-loss improves the performance with respect to all three metrics over CE, justifying the proposed use of logit-adjusted CE modifications in this group context. Moreover, VS-loss used in conjunction with DRO achieved the highest worst and balanced accuracies and the lowest Symm. DEO, as desired. Interestingly, VS improves the worst accuracy over CE+DRO, even though the latter is specifically designed to minimize that objective. A final observation is that Symm. DEO seems correlated with balanced accuracy, which we previously observed in Fig. 3(a).

6 Concluding remarks

We presented a theoretically-grounded study of recently introduced cost-sensitive modifications of the CE loss for imbalanced data. To optimize key fairness metrics, we formulated a new such modification subsuming previous techniques as special cases and provided theoretical justification, as well as, empirical evidence on its superior performance against existing methods. We suspect the VS-loss and our better understanding on the individual roles of different hyperparameters can benefit applications in NLP and computer vision with data imbalances; we expect future work to undertake this opportunity with additional experiments. When it comes to group-sensitive learning, it is of interest to extend our theory to other fairness metrics of interest. Ideally, our precise asymptotic theory could help contrast different definitions and assess their pros/cons. Our results are the first to theoretically justify the benefits/pitfalls of specific logit adjustments used in [KHB+18, CWG+19, MJR+20, YCZC20]. The current theory is limited to settings with fixed features (aka hand-crafted feature maps h). While this assumption is prevailing in most related theoretical works [JT18, NSS19, HMRT19, BLLT20, MNS+20], it is still far from deep-net practice where (last-layer) features are learnt jointly with the classifier. We expect very recent theoretical developments on that front [PHD20, MPP20, LS20] to be relevant in our setting when combined with ideas introduced here.

References

[AH18] Navid Azizan and Babak Hassibi. Stochastic gradient/mirror descent: Minimax optimality and implicit regularization. *arXiv preprint arXiv:1806.00952*, 2018.

[BHX19] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019.

[BL19] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881. PMLR, 2019.

[BLLT20] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[BMM18a] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549, 2018.

[BMM18b] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, Oct 2018.

[BRT19] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619, 2019.

[BS16] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[CJL+19] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples, 2019.

[CKP09] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

[CLOT20] Xiangyu Chang, Yingcong Li, Samet Oymak, and Christos Thrampoulidis. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks, 2020.

[CS+20] Emmanuel J Candès, Pragya Sur, et al. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42, 2020.

[CWG+19] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1567–1578, 2019.

[DKT19] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*, 2019.

[DL20] Oussama Dhifallah and Yue M. Lu. A precise performance analysis of learning with random features, 2020.

[DOBD+18] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *arXiv preprint arXiv:1802.08626*, 2018.

[FSV16] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

[GPSW17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[HMRT19] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[HNSS18] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2029–2037. PMLR, 10–15 Jul 2018.

[HPS16] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[JT18] Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

[JT21] Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

[KA20] Abla Kammoun and Mohamed-Slim Alouini. On the precise error analysis of support vector machines. *arXiv preprint arXiv:2003.12972*, 2020.

[KHB+18] Salman H. Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A. Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3573–3587, 2018.

[KK20] Byungju Kim and Junmo Kim. Adjusting decision boundary for class imbalanced learning. *IEEE Access*, 8:81674–81685, 2020.

[KMR16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[KT20] G. R. Kini and C. Thrampoulidis. Analytic study of double descent in binary classification: The impact of loss. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2527–2532, 2020.

[KT21] Ganesh Ramachandra Kini and Christos Thrampoulidis. Phase transitions for one-vs-one and one-vs-all linear separability in multiclass gaussian mixtures. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4020–4024, 2021.

[LGG+18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

[LS20] Jianfeng Lu and Stefan Steinerberger. Neural collapse with cross-entropy loss. *arXiv preprint arXiv:2012.08465*, 2020.

[MJR+20] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

[MM19] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[MNS+20] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *arXiv preprint arXiv:2005.08054*, 2020.

[MPP20] Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features, 2020.

[MRSY19] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[MSV10] Hamed Masnadi-Shirazi and Nuno Vasconcelos. Risk minimization, probability elicitation, and cost-sensitive svms. In *ICML*, pages 759–766. Citeseer, 2010.

[NKB+19] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

[NLG+19] Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

[NSS19] Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3051–3059, 2019.

[OA18] Mahbod Olfat and Anil Aswani. Spectral algorithms for computing fair support vector machines. In *International Conference on Artificial Intelligence and Statistics*, pages 1933–1942. PMLR, 2018.

[OS19] Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960. PMLR, 2019.

[OSHL19] Yonatan Oren, Shiori Sagawa, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust language modeling, 2019.

[PGC+17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[PHD20] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[RZH03] Saharon Rosset, Ji Zhu, and Trevor Hastie. Margin maximizing loss functions. In *NIPS*, pages 1237–1244, 2003.

[SC19] Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[SHN+18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[SKHL19] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[SRKL20] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

[Sto13] Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

[TOH15] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709, 2015.

[TWL+20] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11662–11671, 2020.

[WBW+11] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

[WCLL18] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, Jul 2018.

[WM19] Robert Williamson and Aditya Menon. Fairness risk measures. In *International Conference on Machine Learning*, pages 6786–6797. PMLR, 2019.

[XM89] Yu Xie and Charles F Manski. The logit model and response-based samples. *Sociological Methods & Research*, 17(3):283–302, 1989.

[YCZC20] Han-Jia Ye, Hong-You Chen, De-Chuan Zhan, and Wei-Lun Chao. Identifying and compensating for feature deviation in imbalanced deep learning, 2020.

[ZCO20] Yuan Zhao, Jiasi Chen, and Samet Oymak. On the role of dataset quality and heterogeneity in model confidence. *arXiv preprint arXiv:2002.09831*, 2020.

[ZVGRG17] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We summarize the contents and our contributions in the abstract and introduction

(b) Did you describe the limitations of your work? [Yes] We discuss some limitations of our current work and chalk out potential directions for improvement in Sec. 6

(c) Did you discuss any potential negative societal impacts of your work? [Yes] To the best of our knowledge, our contributions have the potential to benefit the society by being useful in practical problems including promoting algorithmic fairness, as discussed in Sec. 1

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes] In our theoretical results, we state our model and assumptions clearly

(b) Did you include complete proofs of all theoretical results? [Yes] Complete proofs are included in the supplementary material (due to space constraints)

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We will include in SM

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We include the details in Sec. 5 and further in SM

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report error bars in SM

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We have stated the compute resources that were used for our experiments in Sec. 5

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We cite the creators of code and data that we make use of

(b) Did you mention the license of the assets? [N/A] We cite the creators as said above, and we use publicly available assets

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We will include in SM

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We use publicly available data that has been employed in previous works

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]