Label-Imbalanced and Group-Sensitive Classification under Overparameterization

Ganesh Ramachandra Kini  
University of California, Santa Barbara  
kini@ucsb.edu

Orestis Paraskevas  
University of California, Santa Barbara  
orestis@ucsb.edu

Samet Oymak  
University of California, Riverside  
oymak@ece.ucr.edu

Christos Thrampoulidis  
University of British Columbia  
cthrampo@ece.ubc.ca

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 weighted cross-entropy, fail when 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 interclass margins change appropriately. To show this, we discover a connection of the multiplicative CE modification to the cost-sensitive support-vector machines. Perhaps counterintuitively, we also find that, at the start of training, the same multiplicative weights can actually harm the minority classes. Thus, while additive adjustments are ineffective in the TPT, we show 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. Importantly, our experiments on state-of-the-art datasets are fully consistent with our theoretical insights and confirm the superior performance of our algorithms. 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 and contributions

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 directly involving people [3]. Two common types of imbalances are those appearing 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.
Methods for imbalanced data are broadly categorized into data- and algorithm-level ones. In the latter category, belong cost-sensitive methods and, specifically, those that modify the training loss to account for varying class/group penalties. Corresponding state-of-the-art (SOTA) research is motivated by observations that classical methods, such as weighted cross-entropy (wCE) fail when training 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) ([43] 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 favors minorities. To overcome this 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 [24, 8, 32]. Even more recently, [54] 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 hyper-parameters lead to 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 answers the above questions. Specifically, we argue that multiplicative hyper-parameters are most effective for margin adjustments in TPT, while additive parameters can be useful in the initial phase of training. Importantly, this intuition justifies our algorithmic contribution: we introduce the vector-scaling (VS) loss that combines both types of adjustments and attains improved performance on SOTA imbalanced datasets. Finally, using the same set of tools, we extend the VS-loss to instances of group-sensitive classification. We make multiple contributions as summarized below; see also Figure 1.

- **Explaining the distinct roles of additive/multiplicative adjustments.** We show that when optimizing in 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 training, the same adjustments can actually harm minorities. Instead, additive adjustments can speed up convergence by counteracting the initial negative effect of the multiplicative ones. The analytical findings are consistent with our experiments.

- **An improved algorithm: VS-loss.** Motivated by the unique roles of the two different types of adjustments, we propose the vector-scaling (VS) loss that combines the best of both worlds and outperforms existing techniques on benchmark datasets.

- **Introducing logit-adjustments for group-imbalanced data.** We introduce a version of VS-loss tailored to group-imbalanced datasets, thus treating, for the first time, loss-adjustments for label and group imbalances in a unifying way. For the latter, we propose a new algorithm combining our VS-loss with the previously proposed DRO-method to achieve state-of-the-art performance in terms of both Equal Opportunity and worst-subgroup error.

- **Generalization analysis / fairness trade-offs.** We present a sharp generalization analysis of the VS-loss on binary overparameterized Gaussian mixtures. Our formulae are explicit in terms of data geometry, priors, parameterization ratio and hyperparameters; thus, leading to tradeoffs between standard error and fairness measures. 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.

<table>
<thead>
<tr>
<th>Adjustment type</th>
<th>Inductive bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive</td>
<td>Multiplicative</td>
</tr>
<tr>
<td>SVM</td>
<td>CS-SVM</td>
</tr>
<tr>
<td>LA-loss [MJR+20]</td>
<td>✓</td>
</tr>
<tr>
<td>CDT-loss [YCZC20]</td>
<td>✓</td>
</tr>
<tr>
<td>VS-loss [Ours]</td>
<td>✓</td>
</tr>
</tbody>
</table>

Figure 1: Summary of contributions.
1.2 Connections to related literature

CE adjustments. The use of wCE for imbalanced data is rather old [53], but it becomes ineffective under overparameterization, e.g., [6]. This deficiency has led to the idea of additive label-based parameters $w_y$ on the logits [24, 8, 50, 32, 52]. Specifically, [32] proved that setting $w_y = \log(\pi_y)$ ($\pi_y$ denotes the prior of class $y$) leads to a Fisher consistent loss, termed LA-loss, which outperformed other heuristics (e.g., focal loss [28]) on SOTA datasets. However, Fisher consistency is only relevant in the large sample size limit. Instead, we focus on overparameterized models. In a recent work, [54] proposed the CDT-loss, which instead uses multiplicative label-based parameters $\Delta_y$ on the logits. The authors arrive at the CDT-loss as a heuristic means of compensating for the empirically observed phenomenon of that the last-layer minority features deviate between training and test instances [25]. Instead, we arrive at the CDT-loss via a different viewpoint: we show that the multiplicative weights are necessary to move decision boundaries towards majorities 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 roles 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 group-imbalanced setting and show its competitive performance over alternatives in [45, 18, 40]. Beyond CE adjustments there is active research on alternative methods to improve fairness metrics, e.g. [23, 56, 29, 41]. These are orthogonal to CE adjustments and can potentially be used in conjunction.

Relation to vector-scaling calibration. Our naming of the VS-loss is inspired by the vector scaling (VS) calibration [14], a post-hoc procedure that modifies the logits $v$ after training via $v \rightarrow \Delta \odot v + \epsilon$, where $\odot$ is the Hadamard product. [55] 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 directly affect training.

Blessings/curses of overparameterization. Overparameterization acts as a catalyst for deep neural networks [38]. In terms of optimization, [47, 42, 20, 2] show that gradient-based algorithms are implicitly biased towards favorable min-norm solutions. Such solutions, are then analyzed in terms of generalization showing that they can in fact lead to benign overfitting e.g. [4, 16]. While implicit bias is key to benign overfitting it may come with certain downsides. As a matter of fact, we show here that certain hyper-parameters (e.g. additive ones) can be ineffective in the interpolating regime in promoting fairness. Our argument essentially builds on characterizing the implicit bias of wCE/LA/CDT-losses. Related to this, [46] demonstrated the ineffectiveness of $\omega_i$ in learning with groups.

2 Problem setup

Data. Let training set $\{(x_i, g_i, y_i)\}_{i=1}^n$ consisting of $n$ i.i.d. samples from a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{G} \times \mathcal{Y}$: $\mathcal{X} \subseteq \mathbb{R}^d$ is the input space, $\mathcal{Y} = [C] := \{1, \ldots, C\}$ the set of $C$ labels, and, $\mathcal{G} = [K]$ refers to group membership among $K \geq 1$ groups. Group-assignments are known for training data, but unknown at test time. For concreteness, we focus here on the binary setting, i.e. $C = 2$ and $\mathcal{Y} = \{-1, +1\}$; we present multiclass extensions in the Experiments and in the Supplementary Material (SM). We assume throughout that $y = +1$ is minority class.

Fairness metrics. Given a training set we learn $f_w : \mathcal{X} \rightarrow \mathcal{Y}$ parameterized by $w \in \mathbb{R}^p$. For instance, linear models take the form $f_w = \langle w, h(x) \rangle$ for some feature representation $h : \mathcal{X} \rightarrow \mathbb{R}^p$. Given a new sample $x$, we decide class membership $\hat{y} = \text{sign}(f_w(x))$. The (standard) risk or misclassification error is $R := \mathbb{P}\{\hat{y} \neq y\}$. Let $s = (y, g)$ define a subgroup for given values of $y$ and $g$. We also define the class-conditional risks $R_s := \mathbb{P}\{\hat{y} \neq y | g = \pm 1\}$, and, the sub-group-conditional risks $R_{g,j} := \mathbb{P}\{\hat{y} \neq y | g = \pm 1, j \in [K]\}$. The balanced error averages the conditional risks of the two classes: $R_{\text{bal}} := (R_{+1} + R_{-2})/2$. Assuming $K = 2$ groups, Equal Opportunity requires $R_{+1} = R_{-2}$ [15]. More generally, we consider the (signed) difference of equal opportunity (DEO) $R_{\text{deo}} := R_{+1} - R_{-2}$. In our experiments, we also measure the worst-case subgroup error $\max_{(y, g) \in [K]} \mathbb{E}_{x \sim \mathcal{D}}[R_{y, g}]$.

Terminal phase of training (TPT). Motivated by modern training practice, we assume overparameterized $f_w$ so that $R_{\text{train}} = \frac{1}{n} \sum_{i=1}^n I[\text{sign}(f_w(x_i)) \neq y_i]$ can be driven to zero. Typically, training such large models continues well-beyond zero training error as the training loss is being pushed toward zero. As in [43], we call this the terminal phase of training.
2.1 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 $\omega_+ > 0$, additive logit parameters $\iota_+ \in \mathbb{R}$, and multiplicative logit parameters $\Delta_y > 0$:

$$
\ell_{\text{VS}}(y, f_w(x)) = \omega_y \cdot \log \left( 1 + e^{\iota_y \cdot f_w(x)/y} \right).
$$

(1)

For imbalanced datasets with $C > 2$ classes, the VS-loss takes the following form:

$$
\ell_{\text{VS}}(y, f_w(x)) = -\omega_y \cdot \log \left( e^{\Delta_y \cdot f_w(x)/y} \sum_{c \in [C]} e^{-\iota_c \cdot f_w(x)/y} \right).
$$

(2)

Here $f_w : \mathbb{R}^d \to \mathbb{R}^C$ and $f_w(x) = [f_1(x), \ldots, f_C(x)]$ is the vector of logits. The VS-loss (Eqns. (1),(2)) captures existing techniques as special cases by tuning accordingly the additive/multiplicative hyperparameters. Specifically, we recover: (i) **weighted CE (wCE) loss** by $\Delta_y = 1, \iota_y = 0, \omega_y = \pi^{-1}$; (ii) **LA-loss** by $\Delta_y = 1$; (iii) **CDT-loss** by $\iota_y = 0$.

With the goal of (additionally) ensuring fairness with respect to sensitive groups, we extend the VS-loss by introducing parameters $(\Delta_{y,g}, \iota_{y,g}, \omega_{y,g})$ that depend both on class and group membership (specified by $y$ and $g$, respectively). Our proposed **group-sensitive VS-loss** is as follows (multiclass version can be defined accordingly):

$$
\ell_{\text{Group-VS}}(y, g, f_w(x)) = \omega_{y,g} \cdot \log \left( 1 + e^{\iota_{y,g} \cdot f_w(x)/y} \right).
$$

(3)

**CS-SVM.** For linear classifiers $f_w(x) = \langle w, h(x) \rangle$ with $h : \mathcal{X} \to \mathbb{R}^p$, CS-SVM [31] solves

$$
\min \quad |w|_2 \text{ sub. to } \begin{cases} 
\langle w, h(x_i) \rangle \geq \delta & , y_i = +1 \\
\langle w, h(x_i) \rangle \leq -1 & , y_i = -1 
\end{cases} , i \in [n],
$$

(4)

for hyper-parameter $\delta \in \mathbb{R}$, representing the ratio of margins between classes. $\delta = 1$ corresponds to (standard) SVM, while tuning $\delta > 1$ (resp. $\delta < 1$) favors a larger margin $\delta \cdot |\hat{w}|_2$ for the minority vs $1/|\hat{w}|_2$ for the majority classes. Thus, $\delta \to +\infty$ (resp. $\delta \to 0$) corresponds to the decision boundary starting right at the boundary of class $y = 1$ (resp. $y = +1$).

**Group-sensitive SVM.** The group-sensitive version of CS-SVM (GS-SVM), for $K = 2$ protected groups adjusts the constraints in (4) so that $y_i \langle w, h(x_i) \rangle \geq \delta$ (or $\geq 1$), if $y_i = 1$ (or $y_i = 2$). $\delta > 1$, GS-SVM favors larger margin for the sensitive group $g = 1$. Refined versions when classes are imbalanced modify the constraints to $y_i \langle w, h(x_i) \rangle \geq \delta_{y_i,g}$. Both CS-SVM and GS-SVM are feasible if data are linearly separable (see SM). However, we caution that the GS-SVM hyper-parameters are in general harder to interpret as “margin-ratios”.

3 Insights on the VS-loss

Here, we shed light on the distinct roles of the VS-loss hyper-parameters $\omega_y, \iota_y$ and $\Delta_y$.

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

We first demonstrate the unique role played by the multiplicative weights $\Delta_y$ through a motivating experiment on synthetic data in Fig. 2. We generated a binary Gaussian-mixture dataset of $n = 100$ examples in $\mathbb{R}^{100}$ with data means sampled independently from the Gaussian distribution and normalized such that $|\mu_{-1}|_2 = 2|\mu_{-1}|_2 = 4$. We set prior $\pi_+ = 0.1$ for the minority class $+1$. For varying model sizes $p \in \{5 : 5 : 50, 75 : 25 : 300\}$ we trained linear classifier $f_w(x) = \langle w, h(x) \rangle$ using only the first $p$ features, i.e. $h(x) = x(1 : p) \in \mathbb{R}^p$. This allows us to investigate performance versus the parameterization ratio $\gamma = p/n$. We train the model $w$ using the following special cases of the VS-loss (Eqn. (1)): (i) **CDT-loss** with $\Delta = \delta_{-1}^{-1}, \Delta = 1$ ($\delta_{-1} > 0$ is set to the value shown in the inset plot; see SM for details). (ii) **LDAM-loss**: $\iota_+ = \pi^{-1/4}, \iota_- = (1 - \pi)^{-1/4}$ (special case of LA-loss [8]). (iii) **LA-loss**: $\iota_+ = \log \left( \frac{1}{\pi_{-1}} \right), \iota_- = \log \left( \frac{\pi_{-1}}{\pi} \right)$ (Fisher-consistent values [32]). We ran gradient descent and averaged over 25 independent experiments. The balanced error was computed on a test set of size $10^4$ and reported values are shown in red/blue/black markers. We also plot the

---

1Such simple models have been used in e.g. [16, 10, 9, 11, 49] for analytic studies of double descent [5, 38] in terms of classification error. Fig. 2(a) reveals a double descent for the balanced error.
training errors, which are zero for $\gamma \geq 0.45$. The shaded region highlights the transition to the overparameterized / separable regime. In this regime, we continued training in the TPT. The plots reveal the following clear message: The CDT-loss has better balanced-error performance compared to the LA-loss when both trained in TPT. Moreover, they offer an intuitive explanation by uncovering a connection to max-margin classifiers: In the TPT, (a) LA-loss performs the same as SVM, and, (b) CDT-loss performs the same as CS-SVM.

We formalize those empirical observations in the theorem below, which holds for arbitrary linearly separable datasets (beyond Gaussian mixtures of the experiment). Specifically, for a sequence of norm-constrained minimizations of the VS-loss, we show that: As the norm constraint $R$ increases (thus, the problem approaches the original unconstrained loss), the direction of the constrained minimizer $w_R$ converges to that of the CS-SVM solution $\tilde{w}_{\Delta_-/\Delta_+}$.

**Theorem 1 (VS-loss=CS-SVM).** Fix a binary training set $\{(x_i, y_i)\}_{i=1}^n$ 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 $\exists w : y_i w^T h(x_i) \geq 1, \forall i \in [n]$. Consider training a linear model $f_w(x) = (w, h(x))$ by minimizing the VS-loss $\mathcal{L}_n(w) := \sum_{i \in [n]} \ell_{VS}(y_i, f_w(x_i))$ with $\ell_{VS}$ defined in (1) for positive parameters $\Delta_+, \omega_\pm \geq 0$ and arbitrary $\epsilon_\pm$. Define the norm-constrained optimal classifier $w_R = \arg\min_{\|w\|_2 \leq R} \mathcal{L}_n(w)$. Let $\tilde{w}_\delta$ be the CS-SVM solution of (4) with $\delta = \Delta_-/\Delta_+$. Then, $\lim_{R \to \infty} w_R / \|w_R\|_2 = \tilde{w}_\delta / \|\tilde{w}_\delta\|_2$.

On the one hand, the theorem makes clear that $\omega_\pm$ and $\epsilon_\pm$ 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 $\Delta_- > \Delta_+ \iff \delta > 1$. The proof is given in the SM together with extensions for multicase datasets. In the SM, we also strengthen Theorem 1 by characterizing the implicit bias of gradient-flow on VS-loss. Finally, we show that group-sensitive VS-loss with $\Delta_{\pm, g} = \Delta_g$ converges to the corresponding GS-SVM.

**Remark 1.** Thm 1 is reminiscent of Thm. 2.1 in [44] 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 [44] extending their theory to VS-loss / CS-SVM, as well as, to the group-case. In a similar way, our result on the implicit bias of gradient-flow on the VS-loss connects to more recent works [47, 20] 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. We remark that, when combined with a recent result by [19], our Theorem 1 also implies that gradient descent on the VS-loss with sufficiently small step size converges in direction to the solution of the CS-SVM. In other words, Theorem 1 characterizes the implicit bias of gradient descent on the VS-loss. As a final note, in Fig. 2(b,c) we kept constant learning rate 0.1. Significantly faster
convergence is observed with normalized GD schemes [36, 21]; see the SM for a detailed numerical study. We also note that Thm. 1 gives a modern interpretation to the CS-SVM via the lens of implicit bias theory.

3.2 VS-loss: Best of two worlds

We have shown that multiplicative weights \( \Delta_x \) 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 \( f_w(x) = 0 \) at initialization. Then, the gradients of CDT-loss with multiplicative logit factors \( \Delta_y \) are identical to the gradients of wCE-loss with weights \( \omega_y = \Delta_y \). Thus, we conclude the following where say \( y = +1 \) is minority. On the one hand, wCE, which typically sets \( \omega_y > \omega_\cdot \) (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 \( f_w(x) = \langle w, h(x) \rangle \) the gradient of VS-loss is \( \nabla_w \ell_{VS}(y, f_w(x)) = -\omega_y \Delta_y \sigma(- \Delta_y y f_w(x) + \epsilon_y) \cdot y h(x) \) where \( \sigma(t) = (1 + \exp(-t))^{-1} \) is the sigmoid function. It is then clear that at \( f_w(x) = 0 \), the logit factor \( \Delta_y \) plays the same role as the weight \( \omega_y \). 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 \( \epsilon_y \)'s at the start of training. Noting that \( \sigma(\cdot) \) is increasing function, we see that setting \( \epsilon_+ > \epsilon_- \) increases the gradient norm for minorities. This leads us to a second observation: By properly tuning the additive logit adjustments \( \epsilon_y \) 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 \( \Delta_y \)'s that play a critical role in the TPT and the \( \epsilon_y \)'s that compensate for the harmful effect of the \( \Delta_y \)'s in the beginning of training.

Figure 2(b,c) illustrate the discussion above. In the binary linear classification setting of Fig. 2(a), we investigate the effect of the additive adjustments on the training dynamics. Specifically, we trained using gradient descent: (i) CE; (ii) wCE with \( \omega_y = 1/\pi_y \); (iii) LA-loss with \( \epsilon_y = \log(1/\pi_y) \); (iv) CDT-loss with \( \Delta_+ = \delta_y^{-1}, \Delta_- = 1 \); (v) VS-loss with \( \Delta_+ = \delta_y^{-1}, \Delta_- = 1, \epsilon_y = \log(1/\pi_y) \) and \( \omega_y = 1 \); (vi) VS-loss with same \( \Delta \)'s, \( \epsilon_y = 0 \) and \( \omega_y = 1/\pi_y \). Figures 2(b) and (c) plot balanced test error \( R_{bal} \) and angle-gap to CS-SVM solution as a function of iteration number 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 \( R_{bal} \).

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 during the initial phase of training to the inclusion of the \( \epsilon_y \)'s. This is also supported by Fig. 2(c) as we see that LA-loss (but also wCE) achieves significantly better values of \( R_{bal} \) at the first stage of training compared to CDT-loss. In Sec. 5.1 we provide deep-net experiments on an imbalanced CIFAR-10 dataset that further support these findings.

4 Generalization analysis and fairness 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 binary Gaussian-mixture generative models (GMM) for the data distribution \( D \). For the label \( y \in \{ \pm 1 \} \), let \( \pi := \mathbb{P}(y = +1) \). Group membership is decided conditionally on the label such that \( \forall y \in [K] : \mathbb{P}(g = j | y = \pm 1) = p_{\pm,j} \), with \( \sum_{j \in [K]} p_{\pm,j} = 1 \). Finally, the feature conditional given label \( y \) and group \( g \) is a multivariate Gaussian of mean \( \mu_{y,g} \in \mathbb{R}^d \) and covariance \( \Sigma \), i.e. \( x(y,g) \sim \mathcal{N}(\mu_{y,g}, \Sigma) \). Specifically for label-imbalances, we let \( K = 1 \) and \( x(y) \sim \mathcal{N}(\mu_y, I_d) \) (see SM for \( \Sigma \neq I_d \)). For group-imbalances, we focus on two groups with \( p_{+,1} = p_{-,1} = p < 1 \) and \( p_{+,2} = p_{-,2} = 1 - p \), where \( j = 1, 2 \) and
Figure 3: Fairness tradeoffs between classification 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 various values of overparameterization $\gamma.$ Plots in (a) are generated using our sharp predictions in Theorem 2. Plots in (b) use corresponding result for GS-SVM given in the SM. See text for interpretations.

The theory 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 $\delta_*$ minimizing $R_{\text{bal}}$ also achieves perfect
5 Experiments

We show experimental results further justifying theoretical findings. (Code available in [1]).

5.1 Label-imbalanced data

Our first experiment (Table 1) shows that non-trivial combinations of additive/multiplicative adjustments can improve balanced accuracy over individual ones. Our second experiment (Fig. 4) validates the theory of Sec. 3 by examining how these adjustments affect training.

Datasets. Table 1 evaluates LA/CDT/VS-losses on imbalanced instances of CIFAR-10/100. Following [8], we consider: (1) CE-loss that includes each data point in the batch with probability \(\pi_y^{-1}\). (2) wCE with weights \(\omega_y = \pi_y^{-1}\). (3) LDAM-loss with deferred re-weighting. (4) LA-loss where \(\tau = \gamma = \gamma_0\) (with value increasing with \(\gamma\)) GS-SVM guarantees Equal Opportunity (EO) \(R_{\text{deo}} = 0\) (without explicitly imposing such constraints as in [39, 12]).

Model and Baselines. We compare the following: (1) CE-loss. (2) Re-Sampling that includes each data point in the batch with probability \(\pi_y^{-1}\). (3) wCE with weights \(\omega_y = \pi_y^{-1}\). (4) LDAM-loss [8], special case of LA-loss where \(\tau = \gamma = \gamma_0\) is subtracted from the logits. (5) LDAM-DRW [8], combining LDAM with deferred re-weighting. (6) LA-loss [32], with the Fisher-consistent parametric choice \(\tau \gamma / (N_{\max} / N_y)^{\gamma}\). (7) CDT-loss [54], with \(\Delta_y = (N_y / N_{\max})^\gamma\). (8) VS-loss, with combined hyperparameters \(\tau \gamma / (N_y / N_{\max})^\gamma\), parameterized by \(\tau, \gamma > 0\) respectively. The works introducing (5)-(7) above, all trained for a different number of epochs, with dissimilar regularization and learning rate schedules. For consistency, we follow the training setting in [8]. Thus, for LDAM we adapt results reported by [8], but for LA and CDT, we reproduce our own in that setting. Finally, for a fair comparison we ran LA-loss for optimized \(\tau = \tau^\ast\) (rather than \(\tau = 1\) in [32]).

Table 1: Top-1 accuracy results on balanced validation set (%).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CIFAR 10</th>
<th>CIFAR 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imbalance Prof</td>
<td>LT-100</td>
<td>STEP-100</td>
</tr>
<tr>
<td>CE</td>
<td>71.94 ± 0.58</td>
<td>62.69 ± 0.50</td>
</tr>
<tr>
<td>Re-Sampling</td>
<td>71.20</td>
<td>65.00</td>
</tr>
<tr>
<td>wCE</td>
<td>72.60</td>
<td>67.30</td>
</tr>
<tr>
<td>LDAM [8]</td>
<td>74.35</td>
<td>66.38</td>
</tr>
<tr>
<td>LDAM-DRW [8]</td>
<td>77.03</td>
<td>76.92</td>
</tr>
<tr>
<td>LA ((\tau = \tau^\ast)) [32]</td>
<td>80.81 ± 0.30</td>
<td>78.23 ± 0.52</td>
</tr>
<tr>
<td>CDT ((\gamma = \gamma^\ast)) [54]</td>
<td>79.55 ± 0.35</td>
<td>73.26 ± 0.29</td>
</tr>
<tr>
<td>VS ((\tau = \tau^\ast), (\gamma = \gamma^\ast))</td>
<td>80.82 ± 0.37</td>
<td>79.10 ± 0.06</td>
</tr>
</tbody>
</table>

VS-loss balanced accuracy. Table 1 shows Top-1 accuracy on balanced validation set (averaged over 5 runs). We use a grid to pick the best \(\tau / \gamma \) or \((\tau, \gamma)\)-pair for the LA /

---

2Here, the hyperparameter \(\gamma\) is used with some abuse of notation and is important to not be confused with the parameterization ratio in the linear models in Sec. 3 and 4. We have opted to use the same notation as in [54] to ease direct comparisons of experimental findings.
CDT / VS losses on the validation set. 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 $(\tau^*, \gamma^*)$-pairs correspond to non-trivial combinations of each individual parameter. Thus, VS-loss has better balanced accuracy as shown in the table. See SM for optimal hyperparameters choices.

(a) $\Delta_y$’s (parameterized by $\gamma$) can hurt training.

(b) LA trains easier than CDT.

(c) $\iota_y$’s mitigate the effect of $\Delta_y$’s (c1,c2), but $\Delta_y$’s dominate TPT performance (c3,c4).

**Figure 4:** Experiments on CIFAR10 with Long-tailed LT-100 imbalance demonstrating the effects of additive/multiplicative parameters at different phases of training. All results are averaged over 5 runs and shaded regions indicate the 95% confidence intervals. See text for details and interpretations.

**How hyperparameters affect training?** We perform three experiments. (a) Figure 4(a) shows that larger values of hyperparameter $\gamma$ (corresponding to more dispersed $\Delta_y$’s between classes) hurt training performance and delay entering to TPT. Complementary Figures 4(c1,c2) show that eventually, if we train longer, then, train accuracy approaches 100%. These findings are in line with Observation 1 in Sec. 3.2. (b) Figure 4(b) shows training accuracy of LA-loss for changing hyperparameter $\tau$ controlling additive adjustments. On the one hand, increasing values of $\tau$ delay training accuracy to reach 100%. On the other hand, when compared to the effect of $\Delta_y$’s in Fig. 4(a), we observe that the impact of additive adjustments on training is significantly milder than that of multiplicative adjustments. Thus, LA trains easier than CDT. (c) Figure 4(c) shows train and balanced accuracies for (i) CDT-loss in blue: $\tau = 0, \gamma = 0.15$, (ii) VS-loss in orange: $\tau = -0.5, \gamma = 0.15$. In Fig. 4(c1,c3) we trained for 200 epochs, while in Fig. 4(c2,c4) we trained for 300 epochs. For $\gamma = 0.15$, CDT-loss does not reach good training accuracy within 200 epochs (~93% at epoch 200 in Fig. 4(c1)), but the addition of $\iota_y$’s with $\tau = -0.5$ mitigates this effect achieving improved ~97% accuracy at 200 epochs. This also translates to balanced test accuracy: VS-loss has better accuracy at the end of training in Fig. 4(c3). Yet, CDT-loss has not yet entered the interpolating regime in this case. So, we ask: What changes if we train longer so that both CDT and VS loss get (closer) to interpolation. In Fig. 4(c2) train accuracy of both algorithms increases when training continues to 300 epochs. Again, thanks to the $\iota_y$’s VS-loss trains faster. However, note in Figure 4(c4) that the balanced accuracies of the two methods are now very close to each other. Thus, in the interpolating regime what dominates the performance are the multiplicative adjustments which are same for both losses. This is in agreement with the finding of Theorem 1 and the synthetic experiment in Fig. 2(b,c).

**5.2 Group-sensitive data**

The message of our experiments on group-imbalanced datasets is three-fold. (1) We demonstrate 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 propose combining logit-adjustments 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 [45]. We consider the Waterbirds dataset [45]. 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 $y \in \mathcal{Y} = \{\pm 1\} \equiv \{\text{waterbird, landbird}\}$ and belongs to a group $g \in \mathcal{G} = \{\pm 1\} \equiv \{\text{water, land}\}$. 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. As in [45], we train a ResNet50 starting with pretrained weights on Imagenet. 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 the DRO method of [45]. We also implement a new training scheme that combines Group-VS+DRO. We show additional results for Group-LA/CDT (not previously used in group contexts). For fair comparison, we reran the baseline experiments with CE and report our reproduced numbers. Since class $+1$ has no special meaning here, we use Symm-DEO $\equiv R_{(+1,+1)} - R_{(+1,-1)} + R_{(-1,+1)} - R_{(-1,-1)})/2$ and also report balanced and worst sub-group accuracies. We did not fine-tune $\gamma$ as the heuristic choice already shows the benefit of Group-VS-loss. We expect further improvements tuning over validation set.

Results. Table 2 reports test values obtained at last epoch (300 in total). Our Group-VS loss significantly improves performance (measured with all three fairness metrics) over CE, providing a cure for the poor CE performance under overparameterization reported in [46]. Group-CDT/VS have comparable performances, with or without DRO. Also, both outperform Group-LA that only uses additive adjustments. While these conclusions hold for the specific heuristic tuning of $\iota_y$'s, $\Delta_y$'s described above, they are in alignment with our Theorem 1. Interestingly, Group-VS improves by a small margin the worst accuracy over CE+DRO, despite the latter being specifically designed to minimize that objective. Our proposed Group-VS + DRO outperforms the CE+DRO algorithm used in [45] when training continues in TPT. Finally, Symm. DEO appears correlated with balanced accuracy, in alignment with our discussion in Sec. 4 (see Fig. 3(a)).

<table>
<thead>
<tr>
<th>Loss</th>
<th>Symm. DEO</th>
<th>Bal. acc.</th>
<th>Worst acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>25.3±0.66</td>
<td>89.9±0.29</td>
<td>68.1±2.2</td>
</tr>
<tr>
<td>Group LA</td>
<td>24.0±2.4</td>
<td>84.2±3.0</td>
<td>70.1±2.6</td>
</tr>
<tr>
<td>Group CDT</td>
<td>18.5±0.46</td>
<td>87.2±1.2</td>
<td>75.4±2.2</td>
</tr>
<tr>
<td>Group VS</td>
<td>18.1±0.65</td>
<td>88.1±0.38</td>
<td>76.7±2.3</td>
</tr>
<tr>
<td>CE + DRO</td>
<td>16.3±0.37</td>
<td>85.7±0.31</td>
<td>75.2±2.1</td>
</tr>
<tr>
<td>Group LA + DRO</td>
<td>16.3±0.82</td>
<td>88.7±0.40</td>
<td>74.3±2.5</td>
</tr>
<tr>
<td>Group CDT + DRO</td>
<td>11.7±0.15</td>
<td>90.3±0.2</td>
<td>79.9±1.5</td>
</tr>
<tr>
<td>Group VS + DRO</td>
<td>11.8±0.70</td>
<td>90.2±0.22</td>
<td>78.9±1.0</td>
</tr>
</tbody>
</table>

6 Concluding remarks

We presented a theoretically-grounded study of recently introduced cost-sensitive CE modifications 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 NLP and computer vision applications; 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 fairness definitions and assess their pros/cons. Our results are the first to theoretically justify the benefits/pitfalls of specific logit adjustments used in [24, 8, 32, 54]. The current theory is limited to settings with fixed features. While this assumption is prevailing in most related theoretical works [20, 37, 16, 4, 35], it is still far from deep-net practice where (last-layer) features are learnt jointly with the classifier. We expect recent theoretical developments on that front [43, 33, 30] to be relevant in our setting when combined with our ideas.
Acknowledgments

This work is supported by the National Science Foundation under grant Numbers CCF-2009030, by HDR-193464, by a CRG8 award from KAUST and by an NSERC Discovery Grant. C. Thrampoulidis would also like to acknowledge his affiliation with University of California, Santa Barbara. S. Oymak is partially supported by the NSF award CNS-1932254 and by the NSF CAREER award CCF-2046816.

References


