Supplementary Material for "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.

Organization of the supplementary material

The supplementary material (SM) is organized as follows.

1. In Section A we provide additional technical information on the label-imbalanced experiments of Sec. 5.1. We also show experiments of imbalanced MNIST dataset.

2. In Section B we provide missing details and additional results on the group-imbalanced experiments of Sec. 5.2.

3. In Section C we present synthetic experiments on both label-imbalanced and group-sensitive datasets further supporting our theoretical findings in Sections 3 and 4.

4. In Section D we present and prove a more general version of Theorem 1 (specifically, see Theorem 3) on the connection of overparameterized VS-loss and to CS-SVM. We also discuss multiclass extensions (see Theorem 4) and implicit bias of gradient flow (see Theorem 5).

5. In Section E, we present theoretical results on optimal tuning of CS-SVM. First, we state and prove Lemma 2 which establishes a structural connection between the solution of CS-SVM to the solution of the standard SVM, allowing to view the former as a post-hoc adjustment to the latter. Then, we use this property together with the sharp characterizations of Theorem 2 to derive an explicit formula for the optimal margin ratio under Gaussian mixture data.

6. In Section F we prove Theorem 2 on generalization of CS-SVM. We also discuss related works on sharp high-dimensional asymptotics and provide necessary background on the convex Gaussian min-max theorem. Finally, we include formulas for the phase-transition threshold of CS-SVM.

7. Finally, in Section G we state and prove Theorem 7 characterizing the DEO of GS-SVM as mentioned in Section 4.

To ease readability and accessibility, we also opted to keep the main manuscript. The SM starts at page 17.

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 [6]. 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 (e.g. gender, race). While acknowledging that there is no universal fairness metric [37, 20], several suggestions have been made in the literature including *Equal Opportunity* favoring same true positive rates across groups [23].

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)* ([58] 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 [32, 13, 46]. Even more recently, [78] 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 countering 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.

	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.

• **Generalization analysis / fairness 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, 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.

1.2 Connections to related literature

CE adjustments. The use of wCE for imbalanced data is rather old [77], but it becomes ineffective under overparameterization, e.g. [11]. This deficiency has led to the idea of *additive* label-based parameters ι_y on the logits [32, 13, 71, 46, 75]. Specifically, [46] proved that setting $\iota_y = \log(\pi_y)$ (π_y denotes the prior of class y) leads to a *Fisher consistent* loss, termed LA-loss, which outperformed other heuristics (e.g., focal loss [40]) 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, [78] proposed the CDT-loss, which instead uses *multiplicative* label-based parameters Δ_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 [33]. 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 [62, 26, 55]. Beyond CE adjustments there is active research on alternative methods to improve fairness metrics, e.g. [31, 80, 41, 56]. 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 [22], a *post-hoc procedure* that modifies the logits **v**

after training via $\mathbf{v} \to \boldsymbol{\Delta} \odot \mathbf{v} + \boldsymbol{\iota}$, where $\odot$ is the Hadamard product. [79] 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 [53]. In terms of optimization, [66, 57, 28, 5] 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. [7, 24]. 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, [63] demonstrated the ineffectiveness of ω_y in learning with groups.

2 Problem setup

Data. 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 $\mathcal{X} \times \mathcal{G} \times \mathcal{Y}$; $\mathcal{X} \subseteq \mathbb{R}^d$ 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 $f_\mathbf{w} : \mathcal{X} \mapsto \mathcal{Y}$ parameterized by $\mathbf{w} \in \mathbb{R}^p$. 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 $\hat{y} = \text{sign}(f_\mathbf{w}(\mathbf{x}))$. 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, Equal Opportunity requires $\mathcal{R}_{+,1} = \mathcal{R}_{+,2}$ [23]. 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 training practice, we assume overparameterized $f_\mathbf{w}$ so that $\mathcal{R}_{\text{train}} = \frac{1}{n} \sum_{i \in [n]} \mathbb{1}[\text{sign}(f_\mathbf{w}(\mathbf{x}_i)) \neq y_i]$ 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 [58], 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 $\omega_\pm > 0$, 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 datasets with $C > 2$ classes, 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 $\mathbf{f}_\mathbf{w} : \mathbb{R}^d \to \mathbb{R}^C$ and $\mathbf{f}_\mathbf{w}(\mathbf{x}) = [\mathbf{f}_1(\mathbf{x}), \ldots, \mathbf{f}_C(\mathbf{x})]$ is the vector of logits. The VS-loss (Eqns. (1),(2)) captures existing techniques as special cases by tuning accordingly the additive/multiplicative hyperparameters. Specifically, we recover: (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 $\iota_y = 0$.

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):

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

4

CS-SVM. For linear classifiers $f_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, h(\mathbf{x}) \rangle$ with $h : \mathcal{X} \to \mathbb{R}^p$, CS-SVM [44] solves

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

for hyper-parameter $\delta \in \mathbb{R}_+$ representing the ratio of margins between classes. $\delta = 1$ corresponds to (standard) SVM, while tuning $\delta > 1$ (resp. $\delta < 1$) favors a larger margin $\delta / \|\hat{\mathbf{w}}_\delta\|_2$ for the minority vs $1/\|\hat{\mathbf{w}}_\delta\|_2$ 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 $y_i \langle \mathbf{w}, h(\mathbf{x}_i) \rangle \geq \delta$ (or ≥ 1), if $g_i = 1$ (or $g_i = 2$.) $\delta > 1$, GS-SVM favors larger margin for the sensitive group $g = 1$. Refined versions when classes are also imbalanced modify the constraints to $y_i \langle \mathbf{w}, \mathbf{h}(\mathbf{x}_i) \rangle \geq \delta_{y_i, g_i}$. Both CS-SVM and GS-SVM are feasible iff 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 ω_y, ι_y and Δ_y.

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

We first demonstrate the unique role played by the multiplicative weights Δ_y through a motivating experiment on synthetic data in Fig. 2. We generated a binary Gaussian-mixture dataset of $n = 100$ examples in $\mathbb{R}^{300}$ 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, i.e. $h(\mathbf{x}) = \mathbf{x}(1 : p) \in \mathbb{R}^p$. This allows us to investigate performance versus the parameterization ratio $\gamma = p/n$. [1] We train the model $\mathbf{w}$ using the following special cases of the VS-loss (Eqn. (1)): (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) *LDAM-loss:* $\iota_+ = \pi^{-1/4}, \iota_- = (1 - \pi)^{-1/4}$ (special case of *LA-loss* [13]). (iii) *LA loss:* $\iota_+ = \log\left(\frac{1-\pi}{\pi}\right), \iota_- = \log\left(\frac{\pi}{1-\pi}\right)$ (Fisher-consistent values [46]). 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 training errors, which are zero for $\gamma \gtrsim 0.45$. 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 $\mathbf{w}_R$ converges to that of the CS-SVM solution $\hat{\mathbf{w}}_{\Delta_-/\Delta_+}$.

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 $\mathcal{L}_n(\mathbf{w}) := \sum_{i \in [n]} \ell_{VS}(y_i, f_{\mathbf{w}}(\mathbf{x}_i))$ 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

---

[1] Such simple models have been used in e.g. [24, 15, 14, 16, 68] for analytic studies of double descent [10, 53] in terms ofclassification error. Fig. 2(a) reveals a double descent for the balanced error.
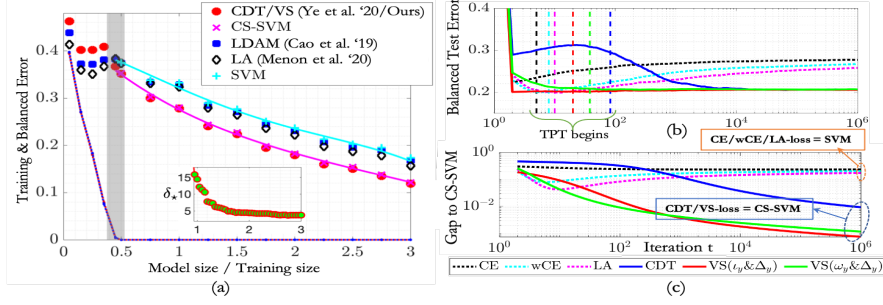
Figure 2: Insights on various cost-sensitive modifications of the CE-loss. **(a)** CDT has superior balanced-error performance over LA in the separable regime. Also, its performance matches that of CS-SVM, unlike LA matching SVM; Sec. 3.1 for more details. Solid lines follow theory of Sec. 4. **(b)** 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.

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. 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 $\Delta_{y,g} = \Delta_g$ converges to the corresponding GS-SVM.

Remark 1. *Thm 1 is reminiscent of Thm. 2.1 in [61] 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 [61] 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 [66, 28] 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 [27], 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 [51, 29]; 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 $\Delta_{\pm}$ 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 <u>identical</u> 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 $\omega_+ > \omega_-$ (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 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 $\iota_y = \log(1/\pi_y)$; (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's, $\iota_y = 0$ 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 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 $\mathcal{R}_{\text{bal}}$. 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 during 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. 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 $\mathcal{D}$. 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 $\boldsymbol{\Sigma}$, i.e. $\mathbf{x} | (y, g) \sim \mathcal{N}(\boldsymbol{\mu}_{y,g}, \boldsymbol{\Sigma})$. Specifically for *label-imbalances*, we let $K = 1$ and $\mathbf{x} | y \sim \mathcal{N}(\boldsymbol{\mu}_y, \mathbf{I}_d)$ (see SM for $\boldsymbol{\Sigma} \neq \mathbf{I}_d$). For *group-imbalances*, we focus on two groups 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, $\mathbf{M}$ denotes the matrix of means, i.e. $\mathbf{M} = [\boldsymbol{\mu}_+ \quad \boldsymbol{\mu}_-]$ and $\mathbf{M} = [\boldsymbol{\mu}_1 \quad \boldsymbol{\mu}_2]$, respectively. Also, consider the eigen-decomposition: $\mathbf{M}^T \mathbf{M} = \mathbf{V}\mathbf{S}^2\mathbf{V}^T$, $\mathbf{S} > \mathbf{0}_{r \times r}, \mathbf{V} \in \mathbb{R}^{2 \times r}, r \in \{1, 2\}$, 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 the separable regime. For the models above, linear separability undergoes a sharp phase-transition as $d, n \to \infty$ at a proportional rate $\gamma = \frac{d}{n}$. That is, there exists threshold $\gamma_\star := \gamma_\star(\mathbf{V}, \mathbf{S}, \pi) \leq 1/2$ for the label-case, such that data are linearly separable with probability approaching one provided that $\gamma > \gamma_\star$ (accordingly for the group-case) [12, 49, 15, 30, 36]. See SM for formal statements and explicit definitions.

Analysis of CS/GS-SVM. We use $\xrightarrow{P}$ to denote convergence in probability and $Q(\cdot)$ the standard normal tail. We let $(x)_- := \min\{x, 0\}$; $\mathbb{1}[\mathcal{E}]$ the indicator function of event $\mathcal{E}$; $\mathcal{B}_2^r$ the unit ball in $\mathbb{R}^r$; and, $\mathbf{e}_1 = [1, 0]^T, \mathbf{e}_2 = [0, 1]^T$ standard basis vectors in $\mathbb{R}^2$. We further need the following definitions. Let random variables as follows: $G \sim \mathcal{N}(0, 1)$, Y symmetric Bernoulli with $\mathbb{P}\{Y = +1\} = \pi$, $E_Y = \mathbf{e}_1 \mathbb{1}[Y = 1] - \mathbf{e}_2 \mathbb{1}[Y = -1]$ and $\Delta_Y = \delta \cdot \mathbb{1}[Y = +1] + \mathbb{1}[Y = -1]$, for $\delta > 0$. With these define key function $\eta_\delta : \mathbb{R}_{\geq 0} \times \mathcal{B}_2^r \times \mathbb{R} \to \mathbb{R}$ as $\eta_\delta(q, \boldsymbol{\rho}, b) := \mathbb{E}\left[\left(G + E_Y^T \mathbf{V}\mathbf{S}\boldsymbol{\rho} + \frac{bY - \Delta_Y}{q}\right)_-^2\right] - (1 - \|\boldsymbol{\rho}\|_2^2)\gamma$. Finally, define $(q_\delta, \boldsymbol{\rho}_\delta, b_\delta)$ as the *unique* triplet (see SM for proof) satisfying $\eta_\delta(q_\delta, \boldsymbol{\rho}_\delta, b_\delta) = 0$ and $(\boldsymbol{\rho}_\delta, b_\delta) := \arg\min_{\|\boldsymbol{\rho}\|_2 \leq 1, b \in \mathbb{R}} \eta_\delta(q_\delta, \rho, b)$. Note that these triplets can be easily computed numerically for given values of γ, δ, π, p and means' Gramian $\mathbf{M}^T \mathbf{M} = \mathbf{V}\mathbf{S}^2\mathbf{V}^T$.
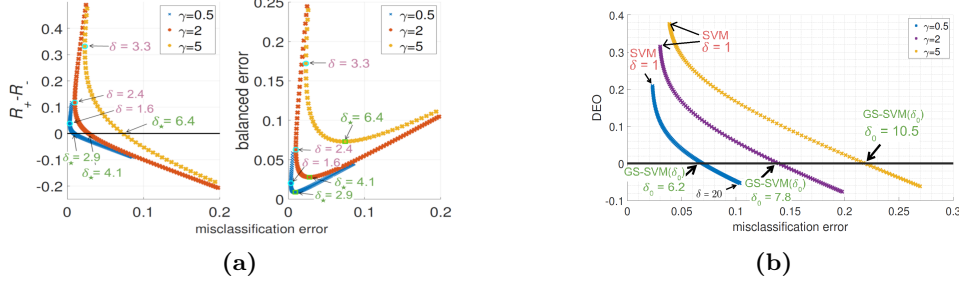
Figure 3: Fairness tradeoffs between classification error and error-imbalance/balanced-error/DEO on GMM data achieved by **(a)** CS-SVM for class prior $\pi = 0.05$ and **(b)** GS-SVM for group prior $p = 0.05$, as a function of the margin-ratio hyperparameter $\delta \geq 1$ and for various values of overparameterization γ. 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.

Theorem 2 (Balanced error of CS-SVM). *Let 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 $\langle \mathbf{x}, \mathbf{w} \rangle + b \geq \{\delta \text{ or } 1\}$ in* (4)*) and fixed margin-ratio $\delta > 0$. Define $\overline{\mathcal{R}}_+ := Q\left(\mathbf{e}_1^T \mathbf{VS} \boldsymbol{\rho}_\delta + b_\delta/q_\delta\right)$ and $\overline{\mathcal{R}}_- := Q\left(-\mathbf{e}_2^T \mathbf{VS}\boldsymbol{\rho}_\delta - b_\delta/q_\delta\right)$. Then, as $n, d \to \infty$ 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 further shows $(\|\hat{\mathbf{w}}_\delta\|_2, \frac{\hat{\mathbf{w}}_\delta^T \boldsymbol{\mu}_+}{\|\hat{\mathbf{w}}_\delta\|_2}, \frac{\hat{\mathbf{w}}_\delta^T \boldsymbol{\mu}_-}{\|\hat{\mathbf{w}}_\delta\|_2}, \hat{b}_\delta) \xrightarrow{P} (q_\delta, \mathbf{e}_1^T \mathbf{VS}\boldsymbol{\rho}_\delta, \mathbf{e}_2^T \mathbf{VS}\boldsymbol{\rho}_\delta, b_\delta)$. Thus, b_δ is the asymptotic the intercept, q_δ^{-1} is the asymptotic classifier's margin $1/\|\hat{\mathbf{w}}_\delta\|_2$ to the majority, and $\boldsymbol{\rho}_\delta$ determines the asymptotic alignment of the classifier with the class mean. The proof uses the convex Gaussian min-max theorem (CGMT) framework [67, 73]; see SM for background, 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) validates the theorem's predictions. Also, in the SM, we characterize the DEO of GS-SVM for GMM data. Although similar in nature, that characterization differs to Thm. 2 since each class is now itself a Gaussian mixture as described in the model above.

Fairness tradeoffs. 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 $\delta_\star$ minimizing $\mathcal{R}_{bal}$ also achieves perfect balancing between the conditional errors of the two classes, that is $\mathcal{R}_+ = \mathcal{R}_- = Q(\frac{\ell_- + \ell_+}{2})$. We prove this interesting property in the SM by deriving an explicit 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. Such closed-form formula is rather unexpected given the seemingly involved nonlinear dependency of $\mathcal{R}_{bal}$ on δ in Thm. 2. In the SM, we also use this formula to formulate 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 $\mathcal{R}_{bal}$). 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) $\mathcal{R}_{deo} = 0$ (without explicitly imposing such constraints as in [54, 17]).

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 [13], we consider: (1) *STEP* imbalance, reducing the sample size of half of the classes to a fixed number. (2) Long-tailed (*LT*) imbalance, which exponentially decreases the number of training images across different classes. 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 sample sizes of class y. For consistency with [25, 13, 46, 78] we keep a balanced test set 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 data-augmentation exactly as in [25, 13, 46, 78]. See SM for more implementation details.

Model and Baselines. We compare the following: *(1) CE-loss. (2) Re-Sampling* that includes each data point in the batch with probability π_y^{-1}. *(3) wCE* with weights $\omega_y = \pi_y^{-1}$. *(4) LDAM-loss* [13], 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* [13], combining LDAM with deferred re-weighting. *(6) LA-loss* [46], with the Fisher-consistent parametric choice $\iota_y = \tau \log(\pi_y)$. *(7) CDT-loss* [78], with $\Delta_y = (N_y/N_{max})^\gamma$. *(8) VS-loss*, with combined hyperparameters $\iota_y = \tau \log(\pi_y)$ and $\Delta_y = (N_y/N_{\max})^\gamma$, pa-

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

Dataset	CIFAR 10		CIFAR 100	
Imbalance Profile	LT-100	STEP-100	LT-100	STEP-100
CE	71.94 ± 0.38	62.69 ± 0.50	38.82 ± 0.69	39.49 ± 0.16
Re-Sampling	71.2	65.0	34.7	38.4
wCE	72.6	67.3	40.5	40.1
LDAM [13].	73.35	66.58	39.60	39.58
LDAM-DRW [13]	77.03	76.92	42.04	45.36
LA ($\tau = \tau^*$) [46]	80.81 ± 0.30	78.23 ± 0.52	42.87 ± 0.32	45.69 ± 0.27
CDT ($\gamma = \gamma^*$) [78]	79.55 ± 0.35	73.26 ± 0.29	42.57 ± 0.32	44.12 ± 0.17
VS ($\tau = \tau^*, \gamma = \gamma^*$)	**80.82** ± 0.37	**79.10** ± 0.66	**43.52** ± 0.46	**46.53** ± 0.17

rameterized by $\tau, \gamma > 0$ respectively [2]. 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 [13]. Thus, for LDAM we adapt results reported by [13], but for LA and CDT, we reproduce our own in that setting. Finally, for a fair comparison we ran LA-loss for optimized $\tau = \tau^*$ (rather than $\tau = 1$ in [46]).

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 τ / γ / (τ, γ)-pair for the LA / 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 (τ^*, γ^*)-pairs correspond to non-trivial combinations of each individual parameter. Thus, VS-loss has better balanced accucy as shown in the table. See SM for optimal hyperparameters choices.

How hyperparameters affect training? We perform three experiments. **(a)** Figure 4(a) shows that larger values of hyperparameter γ (corresponding to more dispersed Δ_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 τ controlling additive adjustments. On the one hand, increasing values of τ delay training accuracy to reach 100%. On the other hand, when compared to the effect of Δ_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 ι_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

---

[2]Here, the hyperparameter γ 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 [78] to ease direct comparisons of experimental findings.
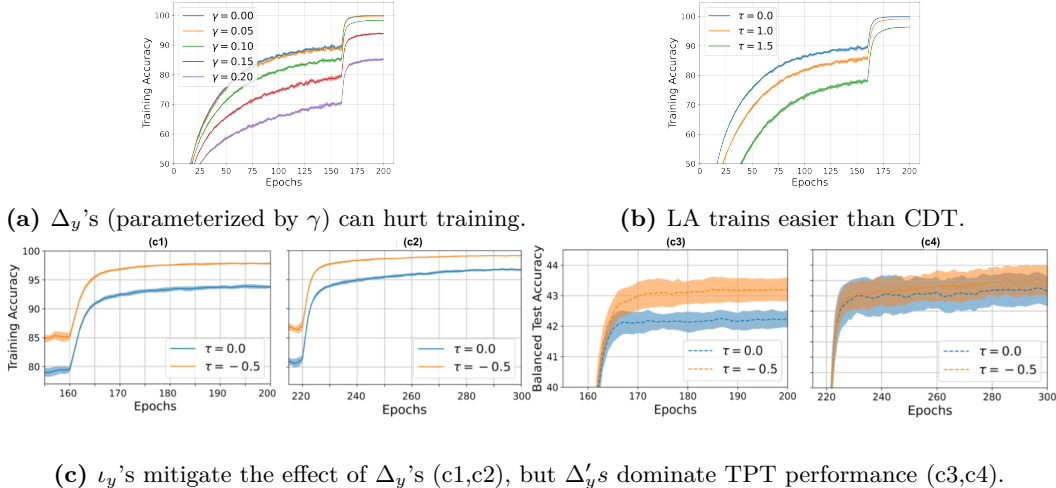
(a) Δ_y's (parameterized by γ) can hurt training.

(b) LA trains easier than CDT.

(c) ι_y's mitigate the effect of Δ_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.

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 ι_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 [62]. We consider the Waterbirds dataset [62]. 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. As in [62], 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 [62]. 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 $= (|\mathcal{R}_{(+1,+1)} - \mathcal{R}_{(+1,-1)}| + |\mathcal{R}_{(-1,+1)} - \mathcal{R}_{(-1,-1)}|)/2$ and also report balanced and worst sub-group accuracies. We did not fine-tune γ as the heuristic choice already shows the benefit of Group-VS-loss. We expect further improvements tuning over validation set.

10

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 [63]. 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 ι_y's, Δ_y's described above, they are in alignment with our Theorem 1. Interestingly, Group-VS improves by a small margin the worst

Table 2: Symmetric DEO, balanced and worst-case subgroup accuracies on Waterbirds dataset; averages over 10 runs, along with standard deviations.

Loss	Symm. DEO	Bal. acc.	Worst acc.
CE	25.3±0.66	84.9±0.29	68.1±2.2
Group LA	24.0±2.4	84.2±3.0	70.1±2.6
Group CDT	18.5±0.46	87.2±1.2	75.4±2.2
Group VS	**18.1±0.65**	**88.1±0.38**	**76.7±2.3**
CE + DRO	16.3±0.37	88.7±0.31	75.2±2.1
Group LA + DRO	16.3±0.82	88.7±0.40	74.3±2.5
Group CDT + DRO	**11.7±0.15**	**90.3±0.2**	**79.9±1.5**
Group VS + DRO	11.8±0.70	90.2±0.22	78.9±1.0

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 [62] 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)).

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 [32, 13, 46, 78]. The current theory is limited to settings with fixed features. While this assumption is prevailing in most related theoretical works [28, 52, 24, 7, 50], 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 [58, 48, 43] to be relevant in our setting when combined with our ideas.

Acknowledgments

References

[1] Code for paper: Label-imbalanced and group-sensitive classification under overparameterization. https://github.com/orparask/VS-Loss.

[2] Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1370–1378. PMLR, 2019.

[3] Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

[4] Benjamin Aubin, Florent Krzakala, Yue M Lu, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. *arXiv preprint arXiv:2006.06560*, 2020.

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

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

[7] 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.

[8] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *Information Theory, IEEE Transactions on*, 57(2):764–785, 2011.

[9] Derek Bean, Peter J Bickel, Noureddine El Karoui, and Bin Yu. Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568, 2013.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

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

[17] 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.

[18] David Donoho and Andrea Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 2016.

[19] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

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

[21] Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

[22] 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.

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

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

[25] 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.

[26] 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.

[27] Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136. PMLR, 2020.

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

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

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

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

[32] 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.

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

[34] 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.

[35] Ganesh Kini and Christos Thrampoulidis. Phase transitions for one-vs-one and one-vs-all linear separability in multiclass gaussian mixtures. *International Conference on Acoustics, Speech, and Signal Processing*, 2021.

[36] 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.

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

[38] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

[39] Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and min-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*, 2020.

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

[41] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world, 2019.

[42] Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pages 1540–1552. PMLR, 2020.

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

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

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

[46] 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.

[47] Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. The role of regularization in classification of high-dimensional noisy gaussian mixture. In *International Conference on Machine Learning*, pages 6874–6883. PMLR, 2020.

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

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

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

[56] Wanli Ouyang, Xiaogang Wang, Cong Zhang, and Xiaokang Yang. Factors in finetuning deep model for object detection, 2016.

[57] 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.

[58] 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.

[59] 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.

[60] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

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

[62] 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.

[63] 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.

[64] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. *arXiv preprint arXiv:1906.03761*, 2019.

[65] Grigoris Karakoulas John Shawe-Taylor and Grigoris Karakoulas. Optimizing classifiers for imbalanced training sets. *Advances in neural information processing systems*, 11(11):253, 1999.

[66] 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.

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

[68] 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.

[69] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Fundamental limits of ridge-regularized empirical risk minimization in high dimensions. *arXiv preprint arXiv:2006.08917*, 2020.

[70] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Sharp asymptotics and optimal performance for inference in binary models. In *International Conference on Artificial Intelligence and Statistics*, pages 3739–3749. PMLR, 2020.

[71] 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.

[72] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[73] 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.

[74] Christos Thrampoulidis, Weiyu Xu, and Babak Hassibi. Symbol error rate performance of box-relaxation decoders in massive mimo. *IEEE Transactions on Signal Processing*, 66(13):3377–3392, 2018.

[75] 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.

[76] Gang Wu and Edward Y Chang. Class-boundary alignment for imbalanced dataset learning. In *ICML 2003 workshop on learning from imbalanced data sets II, Washington, DC*, pages 49–56, 2003.

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

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

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

[80] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition, 2020.

A   Additional Experiments on Label-Imbalanced Datasets

In this section, we provide omitted information (due to space limits) on the results of Section 5.1, as well, as additional experiments.

A.1   Deep-net experiments

Here we provide additional implementation details and a more extensive discussion on the results presented in Table 1 in Section 5.1 of the main text.

Technical details: Following [13], we train a ResNet-32 [25], 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. For STEP-100 imbalance we trained for 300 rather than 200 epochs and adjusted the learning rate accordingly as we found this type of imbalance more difficult to learn. We remark that the values for LDAM (adapted from [13]) 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.

Implementation details. A seed is used for each of the 5 runs and the weights of the network are initialized with the same values for all the losses that we train. We only show 95% confidence intervals for CE, LA, CDT and VS losses which we implemented. For the remaining algorithms (e.g., LDAM), we report averages over 5 realizations as given in [13]. For LA, CDT and VS losses, we have tuned the hyper-parameters (τ, γ) over the validation set as described in Section 5.1 (see Remark 2 and Table 3). More sophisticated tuning strategies over the validation set (e.g., based on bilevel optimization or Hyperband [42, 38]) and the corresponding performance assessment on test set are left to future work. Same as in [25, 13, 46, 78] before training we augment the data by padding the images to size 40×40, flipping them horizontally at random and then random cropping them to their original size. We use PyTorch [59] building on codes provided by [13, 78]. Training is performed on 2 NVIDIA RTX-3080 GPUs.

Remark 2 (On the (τ, γ) parameterization of ι_y's & Δ_y's)**.** *As mentioned in Section 5.1, our deep-net experiments with VS-loss for label-imbalances, use the following parameterization for the additive and multiplicative logit factors in terms of two hyperparameters τ and γ:*

$$\iota_y = \tau \log(N_y/N_{\text{tot}}) \qquad and \qquad \Delta_y = (N_y/N_{\max})^{\gamma}, \tag{5}$$

where N_y is the train-sample size of class y, $N_{\max} = \max_y N_y$ and $N_{\text{tot}} = \sum_y N_y$. This parameterizations follow [46] and [78], respectively. A convenient feature is that setting $\tau = 0$ recovers the CDT-loss, and setting $\gamma = 0$ recovers the LA-loss.

Results and discussion. Table 1 shows that our VS-loss performs favorably over the other methods across all experiments. The margins of improvement depend on the dataset / imbalance-type. Also, observe that in most cases LA-loss performs better than CDT-loss. This is likely because the CDT loss enters the TPT slower for the shown amount of training. Interestingly, VS-loss, even though it resembles the CDT-loss in the fact that it also adjusts the logits multiplicatively, does *not* seem to suffer from the same problem. In Section 3.1, we presented experiments showing that: (i) If given enough time to train, CDT-loss can achieve similar or better results than LA-loss. (ii) The addition of the ι_y's in the VS-loss can mitigate the effect of Δ_y on the speed of convergence. In that sense, VS-loss fulfills the theoretical intuition in Section 3.1, as the method that combines additive and multiplicative adjustments for high accuracy and fast convergence.

Tuning results. To promote reproducibility of our results and to give some insight on the range of τ and γ, in Table 3 we present the values of the hyperparameters that we determined through tuning and used to generate Table 1. As we discussed in Sec. 3.1, large values of τ and γ can hinder training. Thus, when training with the VS loss, which adjusts the logits both in an additive and in a multiplicative way, it seems beneficial to use smaller values of these parameters, than when training with the LA or the CDT losses. Additionally, note that if searching over a grid, it is possible that the best values found for the VS-loss, will be

Table 3: Hyperparameter tuning results for each dataset, imbalance profile and loss function.

Dataset	CIFAR 10		CIFAR 100	
Imbalance Profile	**LT-100**	**STEP-100**	**LT-100**	**STEP-100**
LA ($\tau = \tau^*$) [46]	2.25	2.25	1.375	0.875
CDT ($\gamma = \gamma^*$) [78]	0.4	0.3	0.1	0.1
VS ($\tau = \tau^*, \gamma = \gamma^*$)	$(1.25, 0.15)$	$(1.5, 0.2)$	$(0.75, 0.05)$	$(0.5, 0.05)$
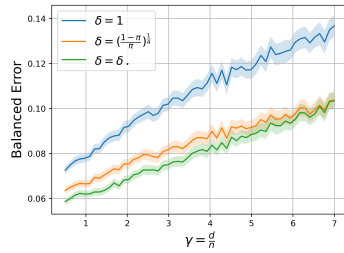
(a) Linear classifier
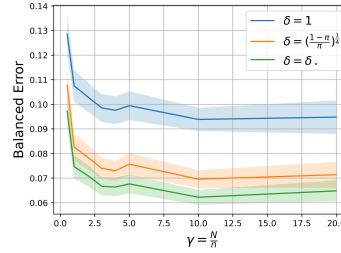
(b) Random-features classifier

Figure 5: A comparison of CS-SVM balanced error against the overparameterization ratio γ, for the standard hard margin SVM ($\delta = 1$), for a heuristic $\delta = \left(\frac{1-\pi}{\pi}\right)^{\frac{1}{4}}$ and for our *approximation* of the optimal δ ($\delta = \tilde{\delta}_\star$) obtained by the data-dependent heuristic in Section E.1.1. The experiment is performed on the MNIST dataset in a one-vs-rest classification task where the goal is to separate the minority class containing images of the digit 7 from the majority class containing images of all other digits. See text for details.

the same as those of the LA or CDT losses (but never worse than them). Searching over a fine enough grid though should yield parameter values for which VS-loss outperforms both of them. Finally, note that the (τ, γ)-parameterization of the ι_y, Δ_y's is itself restrictive and other alternatives might yield further improvements when combining both types of adjustments as observed in the other cases.

A.2 Experiments on the MNIST dataset

Here, we present additional results on imbalanced MNIST data trained with linear and random-feature models. These results complement the synthetic experiment of Figure 2(a).

Specifically, we designed an experiment where we perform binary one-vs-rest classification on the MNIST dataset to classify digit 7 from the rest. Specifically, we split the dataset in two classes, the minority class containing images of the digit 7 and the majority class containing images of all other digits. To be consistent with our notation we assign the label +1 to the minority class and the label −1 to the majority class. Here, $d = 784$ and $\pi = 0.1$ is the prior for the minority class. All test-error evaluations were performed on a test set of 1000 samples. The results of the experiments were averaged over 200 realizations and the 90% confidence intervals for the mean are shown in Figure 5 as shaded regions.

We ran two experiments. In the first one depicted in Figure 5(a), we trained linear classifiers using the standard SVM (blue), the CS-SVM with a heuristic value $\delta = \left(\frac{1-\pi}{\pi}\right)^{\frac{1}{4}}$ (orange), and the CS-SVM with our heuristic data-dependent estimate of the optimal $\tilde{\delta}_\star$ (green). We compute such an estimate based on a recipe inspired by our exact expression in (33) for the GMM; see Section E.1.1 for details. We compute the three classifiers on training sets of varying sizes $n = d/\gamma$ for a range of values of γ and report their balanced error. We observe that CS-SVM always outperforms SVM (aka $\delta = 1$) and the heuristic optimal tuning of CS-SVM consistently outperforms the choice $\delta = \left(\frac{1-\pi}{\pi}\right)^{\frac{1}{4}}$.
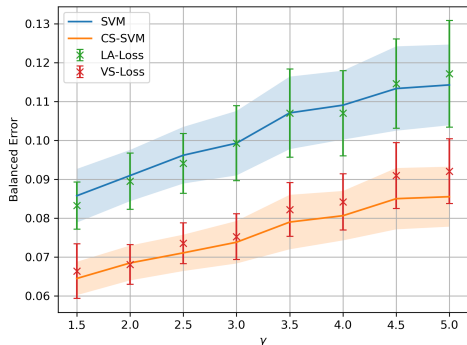
Figure 6: In the overparameterized regime, our VS loss converges to the CS-SVM classifier, while the LA-loss converges to the inferior —in terms of balanced-error performance— SVM. The experiment was performed on the MNIST dataset in a one-vs-rest classification task where the goal is to separate the minority class containing images of the digit 7 from the majority class containing images of all other digits. See text for details.

Next, in Figure 5(b) for the same dataset we trained a Random-features classifier. Specifically, for each one of the $n = 300$ training samples $\mathbf{x}_i \in \mathbb{R}^{d=784}$ we generate random features $\widetilde{\mathbf{x}}_i = \text{ReLU}(\mathbf{A}\mathbf{x}_i)$ for a matrix $\mathbf{A} \in \mathbb{R}^{N \times d}$ which we sample once such that it has entries IID standard normal and is then standardized such that each column becomes unit norm. In this case we control γ by varying the number $N = \gamma n$ of rows of that matrix $\mathbf{A}$. Observe here that the balanced error decreases as γ increases (an instance of benign overfitting, e.g. [24, 7, 45] and that again the estimated optimal $\delta_\star$ results in tuning of CS-SVM that outperforms the other depicted choices.

In Figure 6 we repeat the experiment of Figure 5(a) only this time additionally to training CS-SVM for $\delta = 1$ and for $\delta = \tilde{\delta}_\star$ we also train using the LA-loss and our VS-loss. For the VS loss we use (1) with the following choice of parameters: $\omega_\pm = 1$, $\iota_\pm = 0$ and $\Delta_y = \tilde{\delta}_\star^{-1} \mathbb{1}[y = +1] + \mathbb{1}[y = -1]$ (see Section E.1.1 for $\tilde{\delta}_\star$). In a similar manner, LA-loss is defined using the same formula (1), but with parameters $\Delta_\pm = 1$, $\omega_\pm = 1$ and $\iota_+ = \pi^{-1/4}, \iota_- = (1 - \pi)^{-1/4}$ (as suggested in [13]).

The figure confirms our theoretical expectations: training with gradient descent on the LA and VS losses asymptotically (in the number of iterations) converge to the SVM and CS-SVM solutions respectively.

The training is performed over 200 epochs and for computing the gradient we iterate through the dataset in batches of size 64. The results are averaged over 200 realizations and the 90% confidence intervals are plotted as shaded regions for the CS-SVM model and as errorbars for the VS loss.

B Further details and additional experiments on group-imbalances

B.1 Deep-net experiments

In this section, we elaborate on our proposed method of combining our group logit-adjusted losses with the DRO method. In all experiments, we chose $\Delta_s = (N_s/N_{\max})^\gamma$, $\iota_s = -(N_s/N_{\max})^{-\gamma}$ with $\gamma = 0.3$. For example, Group-LA has $\iota_s = -(N_s/N_{\max})^{-0.3}$ and $\Delta_s = 0$.

Group-VS+DRO algorithm. For completeness, we elaborate on our proposed method of combining DRO with our Group VS-loss (see bottom half of Table 2). We recall from [62] that their proposed CE+DRO algorithm seeks a model that minimizes the worst subgroup empirical risk by instead minimizing the worst subgroup CE-loss: $\max_{s \in \mathcal{S}} \mathbb{E}_{(\mathbf{x},y) \sim \hat{P}_s}[\ell_{\text{CE}}(y, f_\mathbf{w}(\mathbf{x}))]$,

where $\hat{P}_s$ is the empirical distribution on training samples from subgroup s. Instead, our

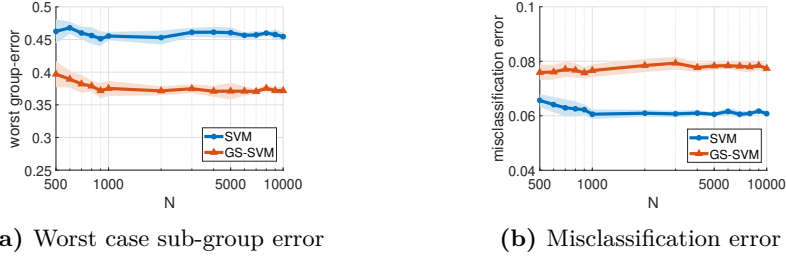**(a)** Worst case sub-group error **(b)** Misclassification error

Figure 7: The benefit of GS-SVM (corr. Group-VS loss) compared to SVM (corr. wCE) in achieving smaller *worst case sub-group error* without significant loss on the misclassification error in the Waterbirds dataset. Training a linear model with N-dimensional Random-feature map over pretrained ResNet-18 features as in [63].

Group-VS+DRO method attempts to solve the following distributionally robust optimization problem:

$$\min_{\mathbf{w}} \max_{s \in \mathcal{S}} \mathbb{E}_{(\mathbf{x},y) \sim \hat{P}_s} [\ell_{\text{Group-VS}}(y, s, f_{\mathbf{w}}(\mathbf{x}))],$$

with $\ell_{\text{Group-VS}}(y, s, f_{\mathbf{w}}(\mathbf{x})) = \omega_s \cdot \log\left(1 + e^{\iota_s} \cdot e^{-\Delta_s y f_{\mathbf{w}}(\mathbf{x})}\right)$ (see Equation (3)). To solve the above non-convex non-differentiable minimization, we employ the same online optimization algorithm given in [62, Algorithm 1], but changing the CE loss to the Group-VS.

B.2 GS-SVM experiments

Section 5.2 demonstrated, for a deep-net model trained on the Waterbird dataset, the efficacy of the Group-VS loss compared to the CE and DRO algorithms used in [62]. Here, we follow [63] who, similar to us, focused in overparameterized training in the TPT. Specifically, [63] showed that wCE trained on a Random-feature model applied on top of a pretrained ResNet results in large *worst-group error* when trained in TPT. In their analysis, they observed that this is because weighted logistic loss in the separable regime behaves like SVM, which is insensitive to groups. Here, we repeat their experiment only this time we use the Group-VS loss. In line with our results thus far, Group-VS loss shows improved performance in this setting as well.

Algorithm. Concretely, since we are training linear models (on random feature maps), we know from Theorem 1 that Group-VS loss converges to GS-SVM. Thus, for simplicity, we directly trained the following instance of GS-SVM and compared it against SVM:

$$\min_{\mathbf{w}} \quad \|\mathbf{w}\|_2 \qquad \text{sub. to} \quad y_i(h(\mathbf{x}_i)^T \mathbf{w} + b) \geq \delta_{s_i}, \; i \in [n]. \tag{6}$$

Above, $\delta_{s_i} = \delta_{(y_i, g_i)} = \left(\frac{1}{\hat{p}_{(y,g)}}\right)^4$, $h : \mathcal{X} \to \mathbb{R}^N$ is the random-feature map (see Section A.2), and $\mathbf{x}_i, i \in [n]$ are d-dimensional pretrained ResNet18 features (same as those used in [63]). Here, $n = 4795$, N took a range of values from 500 to 10000 and $d = 512$. For those values of N the data are separable, thus SVM/GS-SVM are feasible.

Experiment #1: GS-SVM vs SVM (or, Group-VS vs wCE). Figure 7 shows worst-group and missclassification errors of GS-SVM and SVM as a function of the feature dimension N. The curves show averages over 10 realizations of the random projection matrix along with standard deviations depicted using shaded error-bars. We confirm that:

- GS-SVM consistently outperforms standard SVM in the overparameterized regime in terms of worst-group error
- This gain comes without significant losses on the misclassification error.

Experiment #2: GS-SVM vs Sub-sampling. As a means of improving over wCE, [63] proposed instead the use of *CE with subsampling,* for better *worst case sub-group error.* In Figure 8 we compare the performance of three algorithms: (i) SVM, (ii) GS-SVM, and (iii) SVM with subsampling (corresponding to CE with subsampling). For the latter, we chose
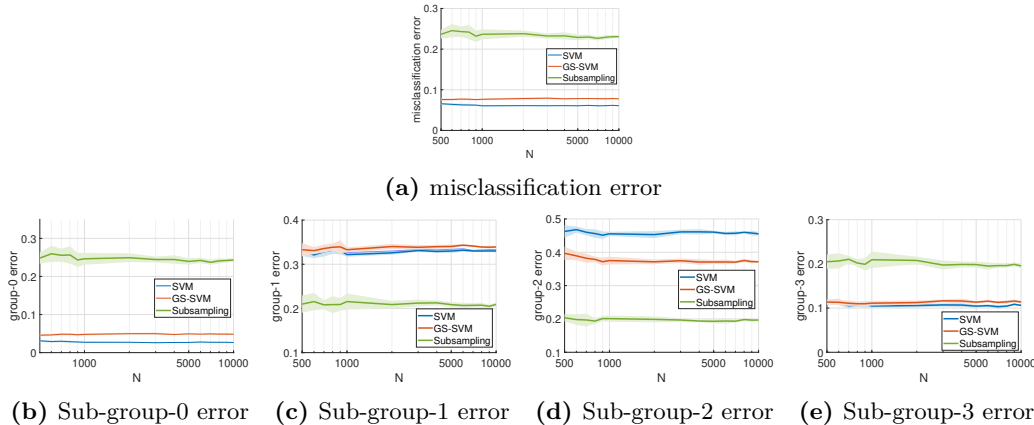
(a) misclassification error

(b) Sub-group-0 error   **(c)** Sub-group-1 error   **(d)** Sub-group-2 error   **(e)** Sub-group-3 error

Figure 8: Misclassification and conditional sub-group errors of SVM (blue), GS-SVM with heuristic tuning $\delta_{(y,g)} = p_{(y,g)}^{-4}$ (red), and, SVM with subsampling (green) for the Waterbirds dataset. GS-SVM has lower worst-case error (Sub-group-2) compared to the SVM without significant increase on the misclassification error. SVM with subsampling has the best worst-group error performance, but also worst misclassification error in subfigure (a).

56 examples from every sub-group (this is the size of the smallest sub-group) and ran SVM on the resulting (smaller), now balanced, dataset. Figure 8 reports missclassification error, as well as, *conditional sub-group errors.* Recall that in the original dataset, sub-groups- 0 and 3 were the *majority* with 3498 and 1057 examples respectively, while sub-groups 1 and 2 were the *minorities* with 184 and 56 examples, respectively. We find the following:

- Consistent with [63] SVM with subsampling achieves low *worst case sub-group error*, lower than both SVM and GS-SVM (at least, when tuned with $\delta_{(y,a)} = \left(\frac{1}{p_{(y,a)}}\right)^4$).

- Specifically, note the very low errors achieved by SVM with subsampling for minority sub-groups 2 and 3.

- However, the gain comes at a significant cost paid for the majority sub-groups- 1 and 3 resulting in an increase of the misclassification error by more than 3− times compared to standard SVM and GS-SVM.

We expect that, with more careful tuning of the hyper-parameters $\delta_{(y,g)}$, GS-SVM can eventually achieve even lower *sub-group errors* for the minority sub-groups without hurting the *majority sub-group errors* significantly. We leave this to future work.

C   Additional numerical results

C.1   Multiplicative vs Additive adjustments for label-imbalanced GMM

In Figure 9 we show a more complete version of Figure 2(a), where we additionally report standard and per-class accuracies. We minimized the CDT/LA losses in the separable regime with normalized gradient descent (GD), which uses increasing learning rate appropriately normalized by the loss-gradient norm for faster convergence; refer to Figure 14 and Section D.4 for the advantages over constant learning rate. Here, normalized GD was ran until the norm of the gradient of the loss becomes less than 10^{-8}. We observed empirically that the GD on the LA-loss reaches the stopping criteria faster compared to the CDT-loss. This is in full agreement with the CIFAR-10 experiments in Section 3.1 and theoretical findings in Section 3.1.

In all cases, we reported both the results of Monte Carlo simulations, as well as, the theoretical formulas predicted by Theorem 2. As promised, the theorem sharply predicts the conditional error probabilities of minority/majority class despite the moderate dimension of $d = 300$.
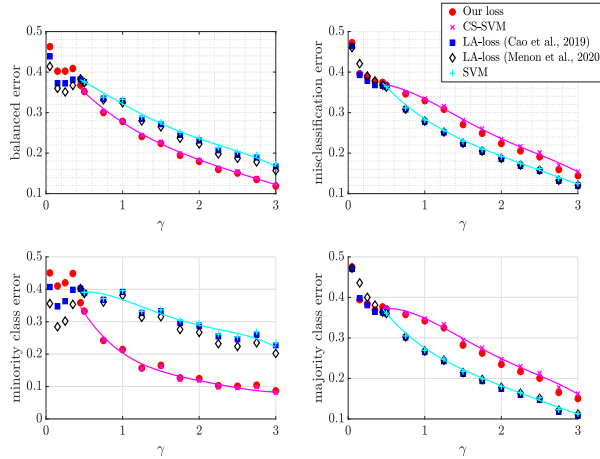
Figure 9: Performance of CDT vs LA loss for label-imbalanced GMM with missing features. (Top Left) is same as Figure 2(a). The other three plots show: (Top Right) misclassification error $\mathcal{R}$; (Bottom Left) majority class error $\mathcal{R}_-$; (Bottom Right) minority class error $\mathcal{R}_+$. Throughout, solid lines correspond to theoretical formulas obtained thanks to Theorem 2.

As noted in Section 3.1, CDT-loss results in better balanced error (see 'Top Left') in the separable regime (where $\mathcal{R}_{\text{train}} = 0$) compared to LA-loss. This naturally comes at a cost, as the role of the two losses is reversed in terms of the misclassification error (see 'Top Right'). The two bottom figures better explain these, by showing that VS sacrifices the error of majority class for a significant drop in the error of the minority class. All types of errors decrease with increasing overparameterization ratio γ due to the mismatch feature model.

Finally, while balanced-error performance of CDT-loss is clearly better compared to the LA-loss in the separable regime, the additive offsets ι_y's improve performance in the non-separable regime. Specifically, the figure confirms experimentally the superiority of the tuning of the LA-loss in [46] compared to that in [13] (but only in the underparameterized regime). Also, it confirms our message: VS-loss that combines the best of two worlds by using both additive and multiplicative adjustments.

C.2 Multiplicative vs Additive adjustments with ℓ_2-regularized GD

In this section we shed more light on the experiments presented in Figure 2(b,c), by studying the effect of ℓ_2-regularization. Specifically, we repeat here the experiment of Fig. 2(b) with $p = d = 50, n = 30$. We train with CE, CDT, and LA-losses in TPT with a weight-decay implementation of ℓ_2-regularization, that is GD with update step: $\mathbf{w}_{t+1} = (1 - \beta)\mathbf{w}_t - \eta\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}_t)$, where β is the weight-decay factor and we used $\beta \in \{0, 10^{-3}, 10^{-2}\}$.

For our discussion, recall our findings in Section 3.1: (i) CDT-loss trained without regularization in TPT converges to CS-SVM, thus achieving better balanced error than LA-loss converging to SVM; (ii) however, at the beginning of training, multiplicative adjustment of CDT-loss can hurt the balanced error; (iii) Additive adjustments on the other hand helped in the beginning of GD iterations but were not useful deep in TPT.

We now turn our focus to the behavior of training in presence of ℓ_2-regularization. The weight-decay factor was kept small enough to still achieve zero training error. A few interesting observations are summarized below:

- The classifier norm plateaus when trained with regularization (while it increases logarithmically without regularization; see Theorem 5). The larger the weight decay factor, the earlier the norm saturates; see Fig. 10(b) and (d).

- Suppose a classifier is trained with a small, but non-zero, weight decay factor in TPT, and the resulting classifier has a norm saturating at some value $\zeta > 0$. The final balanced error performance of such a classifier closely matches the balanced error
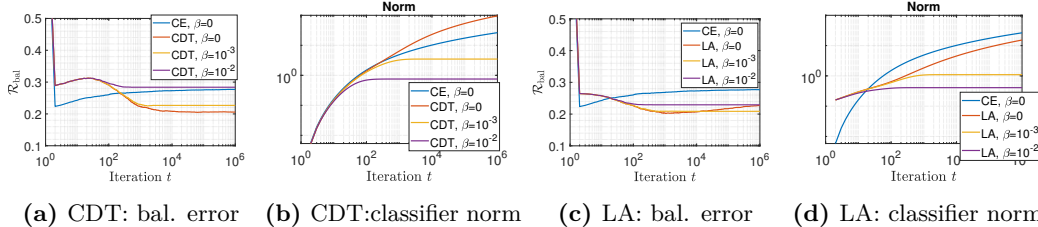
(a) CDT: bal. error **(b)** CDT:classifier norm **(c)** LA: bal. error **(d)** LA: classifier norm

Figure 10: Training dynamics of a linear classifier trained with gradient descent on LA and CDT losses, with and without weight decay (parameter β).

produced by a classifier trained without regularization but with training stopped early at that iteration for which the classifier-norm is equal to ζ; compare for example, the value of yellow curve (CDT, $\beta = 10^{-3}$) at $t = 10^6$ with the value of the red curve (CDT, $\beta = 0$) at around $t = 300$ in Fig. 10(c) and (d). [3]

- If early-stopped (appropriately) before entering TPT, LA-loss can give better balanced performance than CDT-loss. In view of the above mentioned mapping between weight-decay and training epoch, the use of weight decay results in same behavior. Overall, this supports that VS-loss, combining both additive and multiplicative adjustments is a better choice for a wide range of ℓ_2- regularization parameters.

C.3 Additional information on Figures 2(b),(c) and 3(a),(b)

Figures 2(b,c). We generate data from a binary GMM with $d = 50, n = 30$ and $\pi = 0.1$. We generate mean vectors as random iid Gaussian vector and scale their norms to 5 and 1, respectively. For training, we use gradient descent with constant learning rate 0.1 and fixed number of 10^6 iterations. The balanced test error in Figure 2(b) is computed by Monte Carlo on a balanced test set of 10^5 samples. Figure 2(c) measures the angle gap of GD outputs $\mathbf{w}^t$ to the solution $\hat{\mathbf{w}}_\delta$ of CS-SVM in (4) with $\delta = \delta_\star$ and $\mathbf{h}(\mathbf{x}_i) = \mathbf{x}_i$.

Figures 3(a,b). In (a), we generated GMM data with $\|\boldsymbol{\mu}_+\| = 3, \boldsymbol{\mu}_- = -\boldsymbol{\mu}_+$ and $\pi = 0.05$. In (b), we considered the GMM of Section 4 with $\|\boldsymbol{\mu}_{y,g}\| = 3, y \in \{\pm 1\}, g \in \{1, 2\}$ and $\boldsymbol{\mu}_{+,1} \perp \boldsymbol{\mu}_{+,2} \in \mathbb{R}^d$, sensitive group prior $p = 0.05$ and equal class priors $\pi = 1/2$.

C.4 VS-loss vs LA-loss for a group-sensitive GMM

In Figure 11 we test the performance of our theory-inspired VS-loss against the logit-adjusted (LA)-loss in a group-sensitive classification setting with data from a Gaussian mixture model with a minority and and a majority group. Specifically, we generated synthetic data from the model with class prior $\pi = 1 - \pi = 1/2$, minority group membership prior $p = 0.05$ (for group $g = 1$) and $\boldsymbol{\mu}_1 = 3\mathbf{e}_1, \boldsymbol{\mu}_2 = 3\mathbf{e}_2 \in \mathbb{R}^{500}$. We trained homogeneous linear classifiers based on a varying number of training sample $n = d/\gamma$. For each value of n (eqv. γ) we ran normalized gradient descent (see Sec. D.4) on

- CDT-loss $\ell(y, \mathbf{w}^T\mathbf{x}, g) := \log(1 + e^{-\Delta_g y(\mathbf{w}^T\mathbf{x})})$ with $\Delta_g = \delta_0 \mathbb{1}[g = 1] + \mathbb{1}[g = 2]$.
- the LA-loss modified for group-sensitive classification $\ell(y, \mathbf{w}^T\mathbf{x}, g) := \log(1 + e^{\iota_g} e^{y(\mathbf{w}^T\mathbf{x})})$ with $\iota_g = p^{-1/4}\mathbb{1}[g = 1] + (1 - p)^{-1/4}\mathbb{1}[g = 2]$. This value for ι is inspired by [13], but that paper only considered applying the LA-loss in label-imbalanced settings.

For $\gamma > 0.5$ where data are necessarily separable, we also ran the standard SVM and the GS-SVM with $\delta = \delta_0$.

Here, we chose the parameter δ_0 such that the GS-SVM achieves zero DEO. To do this, we used the theoretical predictions of Theorem 7 for the DEO of GS-SVM for any value of δ

---

[3]See also [60, 2] for the connection between gradient-descent and regularization solution paths.

(a) DEO

(b) Misclassification error
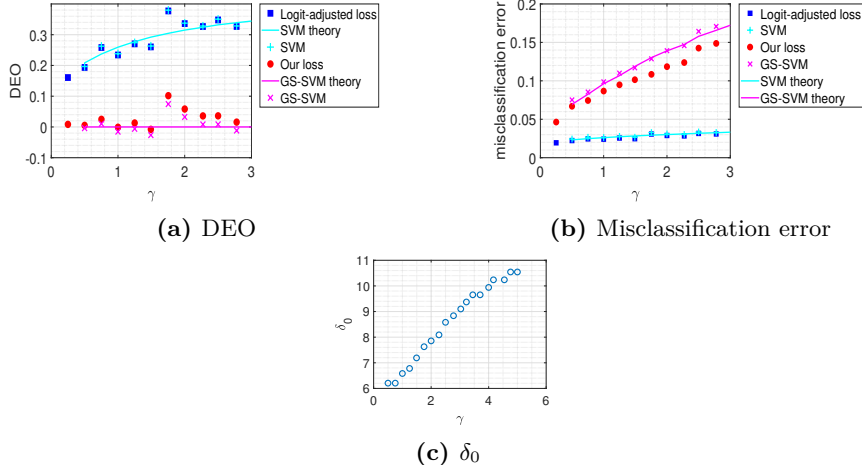

(c) δ_0

Figure 11: This figure highlights the benefits of our theory-inspired VS-loss and GS-SVM over regular SVM and logit-adjusted loss in a group-sensitive classification setting. We trained a linear model with varying number n of examples in $\mathbb{R}^{d=100}$, of a binary Gaussian-mixture dataset with two groups. x-axis is the parameterization ratio d/n. Data were generated from a GMM with prior $p = 0.05$ for the minority group. For $\gamma > 0.5$, we train additionally using SVM (cyan plus marker) and group-sensitive SVM (magenta cross). The plot (c) displays the parameter $\delta = \delta_0$ that we used to tune the VS-loss and GS-SVM. These values were obtained through a grid search from the theoretical prediction such that the theoretical $\mathcal{R}_{\text{deo}}$ (cf. Theorem 7) produced by the corresponding GS-SVM is 0. The solid lines depict theoretical predictions obtained by Theorem 7. The empirical probabilities were computed by averaging over 25 independent realizations of the training and test data.

and performed a grid-search giving us the desired δ_0; see Figure 11 for the values of δ_0 for different values of γ.

Figure 11(a) verifies that the GS-SVM achieves DEO (very close to) zero on the generated data despite the finite dimensions in the simulations. On the other hand, SVM has worse DEO performance. In fact, the DEO of SVM increases with γ, while that of GS-SVM stays zero by appropriately tuning δ_0.

The figure further confirms the message of Theorem 3: In the separable regime, GD on logit-adjusted loss converges to the standard SVM performance, whereas GD on our VS-loss converges to the corresponding GS-SVM solution, thus allowing to tune a suitable δ that can trade-off misclassification error to smaller DEO magnitudes. The stopping criterion of GD was a tolerance value on the norm of the gradient. The match between empirical values and the theoretical predictions improves with increase in the dimension, more Monte-Carlo averaging and a stricter stopping criterion for GD.

C.5 Validity of theoretical performance analysis

Figures 12 and 13 demonstrate that our Theorems 2 and 7 provide remarkably precise prediction of the GMM performance even when dimensions are in the order of hundreds. Moreover, both figures show the clear advantage of CS/GS-SVM over regular SVM and naive resampling strategies in terms of balanced error and equal opportunity, respectively.

The reported values for the misclassification error and the balanced error / DEO were computed over 10^5 test samples drawn from the same distribution as the training examples.

Additionally, Figure 12 validates the explicit formula that we derive in Equation (33) for $\delta_\star$ minimizing the balanced error. Specifically, observe that CS-SVM with $\delta = \delta_\star$ ('×' markers) not only minimizes balanced error (as predicted in Section E.3), but also leads to better misclassification error compared to SVM for all depicted values of γ. The figure also shows the performance of our data-dependent heuristic of computing $\delta_\star$ introduced in Section E.1.1. The heuristic appears to be accurate for small values of γ and is still better in terms of balanced
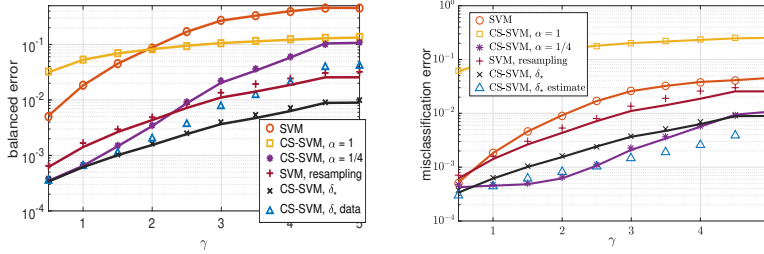
24

Figure 12: Balanced (Left) and misclassification (Right) errors as a function of the parameterization ratio $\gamma = d/n$ for the following algorithms: SVM with and without majority class resampling, CS-SVM with different choices of $\delta = \left(\frac{1-\pi}{\pi}\right)^{\alpha}$, $\pi = 0.05$ and $\delta = \delta_\star$ (cf. Eqn. (33)) plotted for different values of $\gamma = d/n$. Solid lines show the theoretical values thanks to Theorem 2 and the discrete markers represent empirical errors over 100 realizations of the dataset. Data were generated from a GMM with $\boldsymbol{\mu}_+ = 4\mathbf{e}_1, \boldsymbol{\mu}_- = -\boldsymbol{\mu}_+ \in \mathbb{R}^{500}$, and $\pi = 0.05$. SVM with resampling outperforms SVM without resampling in terms of balanced error, but the optimally tuned CS-SVM is superior to both in terms of both balanced and misclassification errors for all values of γ.
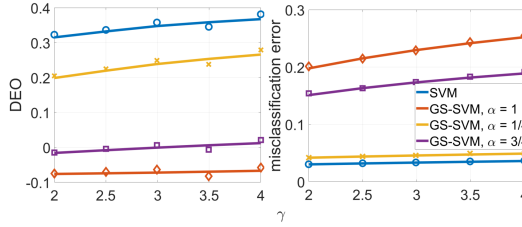

Figure 13: DEO and misclassification error of SVM and GS-SVM with different choices of $\delta = \left(\frac{1-p}{p}\right)^{\alpha}$ for minority group prior $p = 0.05$ plotted against $\gamma = d/n$. Solid lines show the theoretical values and the discrete markers represent empirical errors over 100 realizations of the dataset. Data generated from a GMM with $\boldsymbol{\mu}_{+,1} = 3\mathbf{e}_1, \boldsymbol{\mu}_{+,2} = 3\mathbf{e}_2 \in \mathbb{R}^{500}$. While SVM has the least misclassification error, it suffers from a high DEO. By trading off misclassification error, it is possible to tune GS-SVM (specifically, $\alpha = 0.75$) so that it achieves DEO close to 0 for all the values of γ considered here.

error compared to the other two heuristic choices of $\delta = \left(\frac{1-\pi}{\pi}\right)^{\alpha}, \alpha = 1/4, 1$. Finally, we also evaluated the SVM+subsampling algorithm; see Section C.5.1 below for the algorithm's description and performance analysis. Observe that SVM+resampling outperforms SVM without resampling in terms of balanced error, but the optimally tuned CS-SVM is superior to both.

C.5.1 Max-margin SVM with random majority class undersampling

For completeness, we briefly discuss here SVM combined with undersampling, a popular technique that first randomly undersamples majority examples and only then trains max-margin SVM. The asymptotic performance of this scheme under GMM can be analyzed using Theorem 2 as explained below.

Suppose the majority class is randomly undersampled to ensure equal size of the two classes. This increases the effective overparameterization ratio by a factor of $\frac{1}{2\pi}$ (in the asymptotic limits). In particular, the conditional risks converge as follows:

$$\mathcal{R}_{+,\text{undersampling}}(\gamma, \pi) \xrightarrow{P} \overline{\mathcal{R}}_{+,\text{undersampling}}(\gamma, \pi) = \overline{\mathcal{R}}_+\left(\frac{\gamma}{2\pi}, 0.5\right)$$

$$\mathcal{R}_{-,\text{undersampling}}(\gamma, \pi) \xrightarrow{P} \overline{\mathcal{R}}_{-,\text{undersampling}}(\gamma, \pi) = \overline{\mathcal{R}}_{+,\text{undersampling}}(\gamma, \pi). \tag{7}$$

Above, $\mathcal{R}_{+,\text{undersampling}}$ and $\mathcal{R}_{-,\text{undersampling}}$ are the class-conditional risks of max-margin SVM after random undersampling of the majority class to ensure equal number of training examples from the two classes. The risk $\overline{\mathcal{R}}_+\left(\frac{\gamma}{2\pi}, 0.5\right)$ is the asymptotic conditional risk of

a *balanced* dataset with overparameterization ratio $\frac{\gamma}{2\pi}$. This is computed as instructed in Theorem 2 for the assignments $\gamma \leftarrow \frac{\gamma}{2\pi}$ and $\pi \leftarrow 1/2$ in the formulas therein.

Our numerical simulations in Figure 12 verify the above formulas.

D Margin properties and implicit bias of VS-loss

D.1 A more general version and proof of Theorem 1

We will state and prove a more general theorem to which Theorem 1 is a corollary. The new theorem also shows that the group-sensitive adjusted VS-loss in (3) converges to the GS-SVM.

Remark 3. *Theorem 1 and the content of this section are true for* arbitrary *linear models* $\mathbf{f_w}(\mathbf{x}) = \langle h(\mathbf{x}), \mathbf{w}\rangle$ *and feature maps* $h : \mathcal{X} \to \mathbb{R}^p$. *To lighten notation in the proofs, we assume for simplicity that h is the identity map, that is* $\mathbf{h}(\mathbf{x}) = \mathbf{x}$. *For the general case, just substitute the raw features* $\mathbf{x}_i \in \mathcal{X}$ *below with their feature representation* $h(\mathbf{x}_i) \in \mathbb{R}^p$.

Consider the VS-loss empirical risk minimization (cf. (1) with $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$):

$$\mathcal{L}(\mathbf{w}) \coloneqq \sum_{i\in[n]} \ell(y_i, \mathbf{w}^T\mathbf{x}_i, g_i) \coloneqq \omega_i \log\left(1 + e^{\iota_i} \cdot e^{-\Delta_i y_i(\mathbf{w}^T\mathbf{x}_i)}\right). \tag{8}$$

for strictly positive (but otherwise arbitrary) parameters $\Delta_i, \omega_i > 0$ and arbitrary ι_i. For example, setting $\omega_i = \omega_{y_i,g_i}, \Delta_i = \Delta_{y_i,g_i}$ and $\iota_i = \iota_{y_i,g_i}$ recovers the general form of our binary VS-loss in (3).

Also, consider the following general cost-sensitive SVM (to which both the CS-SVM and the GS-SVM are special instances)

$$\hat{\mathbf{w}} \coloneqq \arg\min_{\mathbf{w}} \|\mathbf{w}\|_2 \quad \text{subject to} \ \ y_i(\mathbf{w}^T\mathbf{x}_i) \geq 1/\Delta_i, \forall i \in [n]. \tag{9}$$

First, we state the following simple facts about the cost-sensitive max-margin classifier in (9). The proof of this claim is rather standard and is included in Section D.1.3 for completeness.

Lemma 1. *Assume that the training dataset is linearly separable, i.e.* $\exists \mathbf{w}$ *such that* $y_i(\mathbf{w}^T\mathbf{x}_i) \geq 1$ *for all* $i \in [n]$. *Then,* (9) *is feasible. Moreover, letting* $\hat{\mathbf{w}}$ *be the solution of* (9), *it holds that*

$$\frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|_2} = \arg\max_{\|\mathbf{w}\|_2=1} \min_{i\in[n]} \Delta_i y_i \mathbf{x}_i^T \mathbf{w}. \tag{10}$$

Next, we state the main result of this section connecting the VS-loss in (8) to the max-margin classifier in (9). After its statement, we show how it leads to Theorem 1; its proof is given later in Section D.1.2.

Theorem 3 (Margin properties of VS-loss: General result). *Define the norm-constrained optimal classifier*

$$\mathbf{w}_R \coloneqq \arg\min_{\|\mathbf{w}\|_2 \leq R} \mathcal{L}(\mathbf{w}), \tag{11}$$

with the loss $\mathcal{L}$ *as defined in* (8) *for positive (but otherwise arbitrary) parameters* $\Delta_i, \omega_i > 0$ *and arbitrary* ι_i. *Assume that the training dataset is linearly separable and let* $\hat{\mathbf{w}}$ *be the solution of* (9). *Then, it holds that*

$$\lim_{R\to\infty} \frac{\mathbf{w}_R}{\|\mathbf{w}_R\|_2} = \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|_2}. \tag{12}$$

D.1.1 Proof of Theorem 1

Theorem 1 is a corollary of Theorem 3 by setting $\omega_i = \omega_{y_i}$, $\iota_i = \iota_{y_i}$ and $\Delta_i = \Delta_{y_i}$. Indeed for this choice the loss in Equation (8) reduces to that in Equation (1). Also, (9) reduces to (4).

The latter follows from the equivalence of the following two optimization problems:

$$\left\{ \arg\min_{\mathbf{w}} \|\mathbf{w}\|_2 \quad \text{subject to} \quad \mathbf{w}^T\mathbf{x}_i \begin{cases} \geq 1/\Delta_+ & y_i = +1 \\ \leq -1/\Delta_- & y_i = -1 \end{cases} \right\}$$

$$= \left\{ \arg\min_{\mathbf{v}} \|\mathbf{v}\|_2 \quad \text{subject to} \quad \mathbf{v}^T\mathbf{x}_i \begin{cases} \geq \Delta_-/\Delta_+ & y_i = +1 \\ \leq -1 & y_i = -1 \end{cases} \right\},$$

which can be verified simply by a change of variables $\mathbf{v}/\Delta_- \leftrightarrow \mathbf{w}$ and $\Delta_- > 0$.

The case of group-sensitive VS-loss. As another immediate corollary of Theorem 3 we get an analogue of Theorem 1 for a group-imbalance data setting with $K = 2$ and balanced classes. Then, we may use the VS-loss in (8) with margin parameters $\Delta_i = \Delta_g, g = 1, 2$. From Theorem 3, we know that in the separable regime and in the limit of increasing weights, the classifier $\mathbf{w}_R$ (normalized) will converge to the solution of the GS-SVM with $\delta = \Delta_2/\Delta_1$.

D.1.2   Proof of Theorem 3

First, we will argue that for any $R > 0$ the solution to the constrained VS-loss minimization is on the boundary, i.e.

$$\|\mathbf{w}_R\|_2 = R. \tag{13}$$

We will prove this by contradiction. Assume to the contrary that $\mathbf{w}_R$ is a point in the strict interior of the feasible set. It must then be by convexity that $\nabla\mathcal{L}(\mathbf{w}_R) = 0$. Let $\widetilde{\mathbf{w}}$ be any solution feasible in (9) (which exists as shown above) such that $y_i(\mathbf{x}_i^T\widetilde{\mathbf{w}}) \geq 1/\Delta_i$. On one hand, we have $\widetilde{\mathbf{w}}^T\nabla\mathcal{L}(\mathbf{w}_R) = 0$. On the other hand, by positivity of $\omega_i, \Delta_i, \forall i \in [n]$:

$$\widetilde{\mathbf{w}}^T\nabla\mathcal{L}(\mathbf{w}_R) = \sum_{i\in[n]} \underbrace{\frac{-\omega_i\Delta_i e^{-\Delta_i y_i \mathbf{x}_i^T\mathbf{w}_R + \iota_i}}{1 + e^{\iota_i}e^{-\Delta_i y_i \mathbf{x}_i^T\mathbf{w}_R}}}_{<0} \underbrace{y_i\widetilde{\mathbf{w}}^T\mathbf{x}_i}_{>0} < 0, \tag{14}$$

which leads to a contradiction.

Now, suppose that (12) is not true. This means that there is some $\epsilon_0 > 0$ such that there is always an arbitrarily large $R > 0$ such that $\frac{\mathbf{w}_R^T\hat{\mathbf{w}}}{\|\mathbf{w}_R\|_2\|\hat{\mathbf{w}}\|_2} \leq 1 - \epsilon_0$. Equivalently, (in view of (13)):

$$\frac{\mathbf{w}_R^T\hat{\mathbf{w}}}{R\|\hat{\mathbf{w}}\|_2} \leq 1 - \epsilon_0. \tag{15}$$

Towards proving a contradiction, we will show that, in this scenario using $\hat{\mathbf{w}}_R = R\frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|_{\ell_2}}$ yields a strictly smaller VS-loss (for sufficiently large $R > 0$), i.e.

$$\mathcal{L}(\hat{\mathbf{w}}_R) < \mathcal{L}(\mathbf{w}_R), \qquad \text{for sufficiently large } R. \tag{16}$$

We start by upper bounding $\mathcal{L}(\hat{\mathbf{w}}_R)$. To do this, we first note from definition of $\hat{\mathbf{w}}_R$ the following margin property:

$$y_i\hat{\mathbf{w}}_R^T\mathbf{x}_i = \frac{R}{\|\hat{\mathbf{w}}\|_2}y_i\hat{\mathbf{w}}^T\mathbf{x}_i \geq \frac{R}{\|\hat{\mathbf{w}}\|_2}(1/\Delta_i) =: \frac{\bar{R}}{\Delta_i}, \tag{17}$$

where the inequality follows from feasibility of $\hat{\mathbf{w}}$ in (9) and we set $\bar{R} := R/\|\hat{\mathbf{w}}\|_2$. Then, using (17) it follows immediately that

$$\begin{aligned}
\mathcal{L}(\hat{\mathbf{w}}_R) &= \sum_{i=1}^{n} \omega_i \log\left(1 + e^{\iota_i}e^{-\Delta_i y_i\hat{\mathbf{w}}_R^T\mathbf{x}_i}\right) \\
&\leq \sum_{i=1}^{n} \omega_i \log\left(1 + e^{\iota_i}e^{-\frac{\bar{R}}{\Delta_i}\Delta_i}\right) \\
&= \sum_{i=1}^{n} \omega_i \log\left(1 + e^{\iota_i}e^{-\bar{R}}\right) \\
&\leq \omega_{\max} n e^{\iota_{\max} - \bar{R}}.
\end{aligned} \tag{18}$$

In the first inequality above we used (17) and non-negativity of $\omega_i, \Delta_i \geq 0$. In the last line, we have called $\omega_{\max} := \max_{i \in [n]} \omega_i > 0$ and $\iota_{\max} := \max_{i \in [n]} \iota_i > 0$.

Next, we lower bound $\mathcal{L}(\mathbf{w}_R)$. To do this, consider the vector

$$\bar{\mathbf{w}} = \frac{\|\hat{\mathbf{w}}\|_{\ell_2}}{R} \mathbf{w}_R = \mathbf{w}_R / \bar{R}.$$

By feasibility of $\mathbf{w}_R$ (i.e. $\|\mathbf{w}_R\|_2 \leq R$), note that $\|\bar{\mathbf{w}}\|_2 \leq \|\hat{\mathbf{w}}\|_2$. Also, from (15), we know that $\bar{\mathbf{w}} \neq \hat{\mathbf{w}}$. Indeed, if it were $\bar{\mathbf{w}} = \hat{\mathbf{w}} \iff \hat{\mathbf{w}}/\|\hat{\mathbf{w}}\|_2 = \mathbf{w}_R/R$, then

$$\frac{\hat{\mathbf{w}}^T \mathbf{w}_R}{R\|\hat{\mathbf{w}}\|_2} = 1,$$

which would contradict (15). Thus, it must be that $\bar{\mathbf{w}} \neq \hat{\mathbf{w}}$. From these and strong convexity of the objective function in (9), it follows that $\bar{\mathbf{w}}$ must be *infeasible* for (4). Thus, there exists at least one example $\mathbf{x}_j$, $j \in [n]$ and $\epsilon > 0$ such that

$$y_j \bar{\mathbf{w}}^T \mathbf{x}_j \leq (1 - \epsilon)(1/\Delta_j).$$

But then

$$y_j \mathbf{w}_R^T \mathbf{x}_j \leq \bar{R}(1 - \epsilon)(1/\Delta_j), \tag{19}$$

which we can use to lower bound $\mathcal{L}(\mathbf{w}_R)$ as follows:

$$\begin{aligned} \mathcal{L}(\mathbf{w}_R) &\geq \omega_j \log\left(1 + e^{\iota_j - \Delta_j y_j \mathbf{w}_R^T \mathbf{x}_j}\right) \\ &\geq \omega_j \log\left(1 + e^{\iota_{y_j} - \bar{R}\Delta_j \frac{(1-\epsilon)}{\Delta_j}}\right) \\ &\geq \omega_{\min} \log\left(1 + e^{\iota_{\min} - \bar{R}(1-\epsilon)}\right). \end{aligned} \tag{20}$$

The second inequality follows fron (19) and non-negativity of $\Delta_{\pm}, \omega_{\pm}$.

To finish the proof we compare (20) against (18). If $\epsilon \geq 1$, clearly $\mathcal{L}(\hat{\mathbf{w}}_R) < \mathcal{L}(\mathbf{w}_R)$ for sufficiently large R. Otherwise $e^{-\bar{R}(1-\epsilon)} \to 0$ with $R \to \infty$. Hence,

$$\mathcal{L}(\mathbf{w}_R) \geq \omega_{\min} \log\left(1 + e^{\iota_{\min} - \bar{R}(1-\epsilon)}\right) \geq 0.5 \omega_{\min} e^{\iota_{\min} - \bar{R}(1-\epsilon)}.$$

Thus, again

$$\mathcal{L}(\hat{\mathbf{w}}_R) < \mathcal{L}(\mathbf{w}_R) \impliedby \omega_{\max} n e^{\iota_{\max} - \bar{R}} < 0.5 \omega_{\min} e^{\iota_{\min} - \bar{R}(1-\epsilon)} \iff e^{\bar{R}\epsilon} > \frac{2n\omega_{\max}}{\omega_{\min}} e^{\iota_{\max} - \iota_{\min}},$$

because the right side is true by picking R arbitrarily large.

D.1.3 Proof of Lemma 1

The proof of Lemma 1 is standard, but included here for completeness. The lemma has two statements and we prove them in the order in which they appear.

Linear separability $\implies$ feasibility of (9). Assume $\mathbf{w}$ such that $y_i(\mathbf{w}^T \mathbf{x}_i) \geq 1$ for all $i \in [n]$, which exists by assumption. Define $M := \max_{i \in [n]} \frac{1}{\Delta_i} > 0$ and consider $\widetilde{\mathbf{w}} = M\mathbf{w}$. Then, we claim that $\widetilde{\mathbf{w}}$ is feasible for (9). To check this, note that

$$\begin{aligned} y_i = +1 &\implies \mathbf{x}_i^T \widetilde{\mathbf{w}} = M(\mathbf{x}_i^T \mathbf{w}) \geq M \geq 1/\Delta_i \quad \text{since } \mathbf{x}_i^T \mathbf{w} \geq 1, \\ y_i = -1 &\implies \mathbf{x}_i^T \widetilde{\mathbf{w}} = M(\mathbf{x}_i^T \mathbf{w}) \leq -M \leq -1/\Delta_i \quad \text{since } \mathbf{x}_i^T \mathbf{w} \leq -1. \end{aligned}$$

Thus, $y_i(\mathbf{x}_i^T \widetilde{\mathbf{w}}) \geq 1/\Delta_i$ for all $i \in [n]$, as desired.

Proof of (10). For the sake of contradiction let $\widetilde{\mathbf{w}} \neq \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|_2}$ be the solution to the max-min optimization in the RHS of (10). Specifically, this means that $\|\widetilde{\mathbf{w}}\|_2 = 1$ and

$$\tilde{m} := \min_{i \in [n]} \Delta_i y_i \mathbf{x}_i^T \widetilde{\mathbf{w}} > \min_{i \in [n]} \Delta_i y_i \mathbf{x}_i^T \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|_2} =: m.$$

We will prove that the vector $\mathbf{w}' := \widetilde{\mathbf{w}}/\widetilde{m}$ is feasible in (9) and has smaller ℓ_2-norm than $\hat{\mathbf{w}}$ contradicting the optimality of the latter. First, we check feasibility. Note that, by definition of $\widetilde{m}$, for any $i \in [n]$:

$$\Delta_i y_i \mathbf{x}_i^T \mathbf{w}' = \frac{\Delta_i y_i \mathbf{x}_i^T \widetilde{\mathbf{w}}}{\widetilde{m}} \geq 1,$$

Second, we show that $\|\mathbf{w}'\|_2 < \|\hat{\mathbf{w}}\|_2$:

$$\|\mathbf{w}'\|_2 = \frac{\|\widetilde{\mathbf{w}}\|_2}{\widetilde{m}} = \frac{1}{\widetilde{m}} < \frac{1}{m} = \frac{\|\hat{\mathbf{w}}\|_2}{\min_{i \in [n]} \Delta_i y_i \mathbf{x}_i^T \hat{\mathbf{w}}} \leq \|\hat{\mathbf{w}}\|_2,$$

where the last inequality follows by feasibility of $\hat{\mathbf{w}}$ in (9). This completes the proof of the lemma.

D.2 Multiclass extension

In this section, we present a natural extension of Theorem 3 to the multiclass VS-loss in (2). Here, let we let the label set $\mathcal{Y} = \{1, 2, \ldots, C\}$ for a C-class classification setting and consider the cross-entropy VS-loss:

$$\mathcal{L}(\mathbf{W}) := \sum_{i \in [n]} \ell(y_i, \mathbf{w}_1^T \mathbf{x}_i, \ldots, \mathbf{w}_K^T \mathbf{x}_i) = \sum_{i \in [n]} \omega_{y_i} \log \left(1 + \sum_{\substack{y' \in [C] \\ y' \neq y_i}} e^{\iota_{y'} - \iota_{y_i}} e^{-(\Delta_{y_i} \mathbf{w}_{y_i}^T \mathbf{x}_i - \Delta_{y'} \mathbf{w}_{y'}^T \mathbf{x}_i)} \right),$$

$$(21)$$

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_C] \in \mathbb{R}^{C \times d}$ and $\mathbf{w}_y$ is the classifier corresponding to class $y \in [C]$. We will also consider the following multiclass version of the CS-SVM in (9):

$$\hat{\mathbf{W}} = \arg \min_{\mathbf{W}} \|\mathbf{W}\|_F \quad \text{subject to} \quad \mathbf{x}_i^T (\Delta_{y_i} \mathbf{w}_{y_i} - \Delta_{y'} \mathbf{w}_{y'}) \geq 1, \ \forall y' \neq y_i \in [C] \ \text{ and } \ \forall i \in [n].$$

$$(22)$$

Similar to Lemma 1, it can be easily checked that (22) is feasible provided that the training data are separable, in the sense that

$$\exists \mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K] \text{ suc that } \mathbf{x}_i^T (\mathbf{w}_{y_i} - \mathbf{w}_{y'}) \geq 1, \forall y' \in [C], y' \neq y_i \ \text{ and } \ \forall i \in [n]. \quad (23)$$

Moreover, it holds that

$$\hat{\mathbf{W}}/\|\hat{\mathbf{W}}\|_F = \arg \max_{\|\mathbf{W}\|_F = 1} \min_{i \in [n]} \min_{y' \neq y_i} \mathbf{x}_i^T (\Delta_{y_i} \mathbf{w}_{y_i} - \Delta_{y'} \mathbf{w}_{y'}).$$

The theorem below is an extension of Theorem 3 to multiclass classification.

Theorem 4 (Margin properties of VS-loss: Multiclass). *Consider a C-class classification problem and define the norm-constrained optimal classifier*

$$\mathbf{W}_R = \arg \min_{\|\mathbf{W}\|_F \leq R} \mathcal{L}(\mathbf{W}), \quad (24)$$

with the loss $\mathcal{L}$ as defined in (21) for positive (but otherwise arbitrary) parameters $\Delta_y, \omega_y > 0, y \in [C]$ and arbitrary $\iota_y, y \in [C]$. Assume that the training dataset is linearly separable as in (23) and let $\hat{\mathbf{W}}$ be the solution of (22). Then, it holds that

$$\lim_{R \to \infty} \frac{\mathbf{W}_R}{\|\mathbf{W}_R\|_F} = \frac{\hat{\mathbf{W}}}{\|\hat{\mathbf{W}}\|_2}. \quad (25)$$

Proof. The proof follows the same steps as in the proof of Theorem 3. Thus, we skip some details and outline only the basic calculations needed.

It is convenient to introduce the following notation, for $\ell \in [C]$:

$$p(\ell | \mathbf{x}, y, \mathbf{W}) := \frac{e^{\iota_y} e^{\Delta_y \mathbf{x}^T \mathbf{w}_\ell}}{\sum_{y' \in [C]} e^{\iota_{y'}} e^{\Delta_{y'} \mathbf{x}^T \mathbf{w}_{y'}}}.$$

29

In this notation, $\mathcal{L}(\mathbf{W}) = -\sum_{i\in[n]}\log\big(p(y_i|\mathbf{x}_i,y_i,\mathbf{W})\big)$ and for all $\ell\in[C]$ it holds that
$$\nabla_{\mathbf{w}_\ell}\mathcal{L}(\mathbf{W}) = \sum_{i\in[n]} \omega_{y_i}\Delta_{y_i}\left(p(\ell|\mathbf{x}_i,y_i,\mathbf{W}) - \mathbb{1}[y_i=\ell]\right)\mathbf{x}_i.$$

Thus, for any $\widetilde{\mathbf{W}}$ that is feasible in (22)
$$\begin{aligned}
\sum_{\ell\in[C]}\widetilde{\mathbf{w}}_\ell^T\nabla_{\mathbf{w}_\ell}\mathcal{L}(\mathbf{W}) &= \sum_{i\in[n]}\sum_{\ell\in[C]}\omega_{y_i}\Delta_{y_i}\left(p(\ell|\mathbf{x}_i,y_i,\mathbf{W}) - \mathbb{1}[y_i=\ell]\right)\mathbf{x}_i^T\widetilde{\mathbf{w}}_\ell\\
&= \sum_{i\in[n]}\sum_{\ell\neq y_i}\omega_{y_i}\Delta_{y_i}p(\ell|\mathbf{x}_i,y_i,\mathbf{W})\mathbf{x}_i^T\widetilde{\mathbf{w}}_\ell - \omega_{y_i}\Delta_{y_i}\left(1-p(y_i|\mathbf{x}_i,y_i,\mathbf{W})\right)\mathbf{x}_i^T\widetilde{\mathbf{w}}_{y_i}\\
&= \sum_{i\in[n]}\underbrace{-\omega_{y_i}\Big(\sum_{\ell\neq y_i}p(\ell|\mathbf{x}_i,y_i,\mathbf{W})\Big)}_{<0}\underbrace{\Delta_{y_i}\mathbf{x}_i^T\left(\widetilde{\mathbf{w}}_\ell-\widetilde{\mathbf{w}}_{y_i}\right)}_{>0} < 0,
\end{aligned}$$

where in the third line we used that $\sum_{\ell\in[C]}p(\ell|\mathbf{x},y,\mathbf{W}) = 1$. With the above it can be shown following the exact same argument as in the proof of (13) for the binary case that $\|\mathbf{W}_R\|_F = R$, the minimizer of (24) satisfies the constraint with equality.

The proof continues with a contradiction argument similar to the binary case. Assume the desired (25) does not hold. We will then show that for $\hat{\mathbf{W}}_R = \frac{R}{\|\hat{\mathbf{W}}\|_F}\hat{\mathbf{W}}$ and sufficiently large $R > 0$: $\mathcal{L}(\hat{\mathbf{W}}_R) < \mathcal{L}(\mathbf{W}_R)$.

Using feasibility of $\hat{\mathbf{W}}$ in (22) and defining $\omega_{\max} := \max_{y\in[C]}\omega_y$ and $\iota_{\max} = \max_{y\neq y'\in[C]}\iota_{y'}-\iota_y$, it can be shown similar to (18) that
$$\begin{aligned}
\mathcal{L}(\hat{\mathbf{W}}_R) &= \sum_{i\in[n]}\omega_{y_i}\log\Big(1+\sum_{\substack{y'\in[C]\\y'\neq y_i}}e^{\iota_{y'}-\iota_{y_i}}e^{-(R/\|\mathbf{W}\|_F)(\Delta_{y_i}\hat{\mathbf{w}}_{y_i}^T\mathbf{x}_i-\Delta_{y'}\hat{\mathbf{w}}_{y'}^T\mathbf{x}_i)}\Big),\\
&\leq n\omega_{\max}\log\Big(1+(K-1)e^{\iota_{\max}}e^{-R/\|\hat{\mathbf{W}}\|_F}\Big) \leq n(K-1)e^{\iota_{\max}}e^{-R/\|\hat{\mathbf{W}}\|_F}. \quad (26)
\end{aligned}$$

Next, by contradiction assumption and strong convexity of (22), for $\bar{\mathbf{W}} = \frac{\|\hat{\mathbf{W}}\|_2}{R}\mathbf{W}_R$, there exist $\epsilon > 0$ and at least one $j\in[n]$ and $y'\neq y_j$ such that $\mathbf{x}_j^T(\Delta_{y_j}\bar{\mathbf{w}}_j-\Delta_{y'}\bar{\mathbf{w}}_{y'}) \leq (1-\epsilon)$. With this, we can show similar to (20) that
$$\mathcal{L}(\mathbf{W}_R) \geq \log\Big(1+e^{\iota_{y'}-\iota_{y_j}}e^{R/\|\hat{\mathbf{W}}\|_F(1-\epsilon)}\Big). \quad (27)$$

The proof is complete by showing that for sufficiently large R the RHS of (27) is larger than the LHS of (26) leading to a contradiction. We omit the details for brevity. $\qquad\square$

D.3 Implicit bias of Gradient flow with respect to VS-loss

Theorem 3 does not consider the effect of the optimization algorithm. Instead here, we study gradient flow (the limit of gradient descent for infinitesimal step-size) and characterize its implicit bias when applied to the VS-loss. Similar, to Theorem 3, we find that the iterations of gradient flow converge to the solution of a corresponding CS-SVM. For simplicity, we consider a VS-type adjusted exponential loss $\ell(t) = e^{-t}$, rather than logistic loss $\ell(t) = \log(1+e^{-t})$. Recent work makes it clear that both loss functions have similar implicit biases and similar lines of arguments are used to analyze the convergence properties [29, 28]. Thus, one would expect that insights also apply to logistic loss.

Theorem 5 (Implicit bias of the gradient flow). *Consider the gradient flow iteration $\dot{\mathbf{w}}_t = -\nabla\mathcal{L}(\mathbf{w}_t)$, on the exponential VS-loss $\mathcal{L}(\mathbf{w}) = \sum_{i\in[n]}\omega_i\exp(-\Delta_i y_i\mathbf{x}_i^T\mathbf{w}+\iota_i)$. Recall that $\hat{\mathbf{w}}$ is the solution to the CS-SVM in (9). For almost every dataset which is linearly separable and any starting point $\mathbf{w}_0$ the gradient flow iterates will behave as $\mathbf{w}(t) = \hat{\mathbf{w}}\log(t)+\boldsymbol{\rho}_t$ with a bounded residual $\boldsymbol{\rho}_t$ so that $\lim_{t\to\infty}\frac{\mathbf{w}_t}{\|\mathbf{w}_t\|_2} = \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|_2}$.*

Note that [66] previously studied the implicit bias of the gradient flow on standard CE or exponential loss. The theorem above studies the gradient flow applied to the VS-loss and its proof is similar to [66].

Proof. Let $\mathcal{S} \subset [n]$ be the set of indices such that $\forall i \in \mathcal{S} : \Delta_i y_i \mathbf{x}_i^T \hat{\mathbf{w}} = 1$, i.e. the set of support vectors of the CS-SVM. By KKT conditions (eg. see Equation (38)), there exist $\epsilon_i > 0$ such that $\hat{\mathbf{w}} = \sum_{i \in \mathcal{S}} \epsilon_i y_i \mathbf{x}_i$. Moreover, by [66, Lemma 12], for almost all datasets it is true that $|\mathcal{S}| \leq d$ and $i \in \mathcal{S} \implies \epsilon_i > 0$. Thus, for almost all datasets we can define vector $\widetilde{\mathbf{w}}$ satisfying the following equation $\omega_i \Delta_i \exp(-\Delta_i y_i \mathbf{x}_i^T \widetilde{\mathbf{w}} + \iota_i) = \epsilon_i, \forall i \in [S]$. Note then that

$$\hat{\mathbf{w}} = \sum_{i \in [S]} \omega_i \Delta_i e^{-\Delta_i y_i \mathbf{x}_i^T \widetilde{\mathbf{w}} + \iota_i} y_i \mathbf{x}_i \tag{28}$$

Let us define $\mathbf{r}_t = \boldsymbol{\rho}_t - \widetilde{\mathbf{w}} = \mathbf{w}_t - \log(t)\hat{\mathbf{w}} - \widetilde{\mathbf{w}}$. It suffices to show that $\|\mathbf{r}(t)\|_2$ is bounded, since that would automatically give $\boldsymbol{\rho}_t$ is bounded. By the gradient flow equation, we have that

$$\dot{\mathbf{r}}_t = -\nabla \mathcal{L}(\mathbf{w}_t) - \frac{\hat{\mathbf{w}}}{t} = \sum_{i \in [n]} \omega_i \Delta_i y_i e^{-\Delta_i y_i \mathbf{x}_i^T \mathbf{w}_t + \iota_i} \mathbf{x}_i - \frac{\hat{\mathbf{w}}}{t}.$$

Therefore,

$$\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|\mathbf{r}_t\|_2^2 = \dot{\mathbf{r}}_t^T \mathbf{r}_t = \sum_{i \in [n]} \omega_i \Delta_i y_i e^{-\Delta_i y_i \mathbf{x}_i^T \mathbf{w}_t + \iota_i} \mathbf{x}_i^T \mathbf{r}_t - \frac{1}{t} \hat{\mathbf{w}}^T \mathbf{r}_t$$

$$= \underbrace{\sum_{i \in \mathcal{S}} \omega_i \Delta_i y_i e^{-\Delta_i y_i \mathbf{x}_i^T \mathbf{w}_t + \iota_i} \mathbf{x}_i^T \mathbf{r}_t - \frac{1}{t} \hat{\mathbf{w}}^T \mathbf{r}_t}_{:=A} + \underbrace{\sum_{i \notin \mathcal{S}} \omega_i \Delta_i y_i e^{-\Delta_i y_i \mathbf{x}_i^T \mathbf{w}_t + \iota_i} \mathbf{x}_i^T \mathbf{r}_t}_{:=B} \tag{29}$$

We now study the two terms A and B separately. In doing so, recall that $\mathbf{w}_t = \mathbf{r}_t + \log(t)\hat{\mathbf{w}} + \widetilde{\mathbf{w}}$. Hence, using the fact that $\Delta_i y_i \mathbf{x}_i^T \hat{\mathbf{w}} = \begin{cases} = 1 & i \in \mathcal{S} \\ \geq m > 1 & i \notin \mathcal{S} \end{cases}$, it holds that

$$\exp\left(-\Delta_i y_i \mathbf{x}_i^T \mathbf{w}_t + \iota_i\right) \begin{cases} = \frac{1}{t} \cdot \exp\left(-\Delta_i y_i \mathbf{x}_i^T \mathbf{r}_t\right) \cdot \exp\left(-\Delta_i y_i \mathbf{x}_i^T \widetilde{\mathbf{w}} + \iota_i\right) & i \in \mathcal{S} \\ \leq \frac{1}{t^m} \cdot \exp\left(-\Delta_i y_i \mathbf{x}_i^T \mathbf{r}_t\right) \cdot \exp\left(-\Delta_i y_i \mathbf{x}_i^T \widetilde{\mathbf{w}} + \iota_i\right) & i \notin \mathcal{S} \end{cases}$$

Using this and (28), the term A becomes

$$A = \frac{1}{t} \sum_{i \in [S]} \omega_i e^{-\Delta_i y_i \mathbf{x}_i^T \widetilde{\mathbf{w}} + \iota_i} \cdot e^{-\Delta_i y_i \mathbf{x}_i^T \mathbf{r}_t} \Delta_i y_i \mathbf{x}_i^T \mathbf{r}_t - \frac{1}{t} \sum_{i \in [S]} \omega_i \Delta_i e^{-\Delta_i y_i \mathbf{x}_i^T \widetilde{\mathbf{w}} + \iota_i} y_i \mathbf{x}_i^T \mathbf{r}_t$$

$$= \frac{1}{t} \sum_{i \in [S]} \omega_i e^{-\Delta_i y_i \mathbf{x}_i^T \widetilde{\mathbf{w}} + \iota_i} \cdot \left(e^{-\Delta_i y_i \mathbf{x}_i^T \mathbf{r}_t} \Delta_i y_i \mathbf{x}_i^T \mathbf{r}_t - \Delta_i y_i \mathbf{x}_i^T \mathbf{r}_t\right) \leq 0,$$

since $\forall x, x \geq xe^{-x}$.

Similarly, for term B:

$$B \leq \frac{1}{t^m} \sum_{i \notin \mathcal{S}} \omega_i e^{-\Delta_i y_i \mathbf{x}_i^T \widetilde{\mathbf{w}} + \iota_i} \cdot e^{-\Delta_i y_i \mathbf{x}_i^T \mathbf{r}_t} \cdot \Delta_i y_i \mathbf{x}_i^T \mathbf{r}_t \leq \frac{1}{t^m} \sum_{i \notin \mathcal{S}} \omega_i e^{-\Delta_i y_i \mathbf{x}_i^T \widetilde{\mathbf{w}} + \iota_i}, \tag{30}$$

since $\forall x, xe^{-x} \leq 1$.

To finish the proof it only takes now using the above bounds on A, B and integrating both sides of Equation (29). This gives that for all $t_0, t > t_0$, there exists finite constant C such that $\|\mathbf{r}_t\|^2 \leq \|\mathbf{r}_{t_0}\|^2 + C$ where it was critical that $m > 1$ in (30) for the corresponding integral to be finite. This proves that $\|\mathbf{r}_t\|_2$ is bounded as desired. □

We note that the above proof is a straightforward extension of [66] for analysis of CDT, with a simple rescaling of the features of the training set according to the labels, however the analysis for VS-loss with additive logit-adjustments (although similar) cannot be obtained as a special case of [66].

D.4 Numerical illustrations of Theorems 1 and 5

Figure 14 numerically demonstrate the validity of Theorems 1 and 5. Here, we solved the VS-loss in Equation (1) using gradient descent (GD) for GMM data with class imbalance
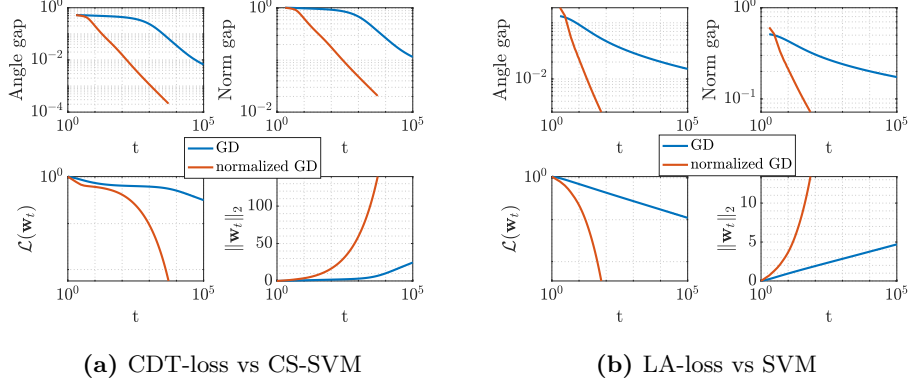
(a) CDT-loss vs CS-SVM   **(b)** LA-loss vs SVM

Figure 14: Convergence properties of GD (blue) and normalized GD (red) iterates $\mathbf{w}_t, t \geq 1$ on VS-loss with $f_{\mathbf{w}}(x) = \mathbf{w}^T \mathbf{x}$ for two set of parameter choices: (a) $\omega_y = 1, \iota_y = 0, \Delta_y = \delta \mathbb{1}[y = 1] + \mathbb{1}[y = -1]$ (aka CDT-loss) with $\delta = 20$; (b) $\omega_y = 1, \iota_y = \pi^{-1/4} \mathbb{1}[y = 1] + (1 - \pi)^{-1/4} \mathbb{1}[y = -1], \Delta_y = 1$ (aka LA-loss). We plotted the angle gap $1 - \frac{\hat{\mathbf{w}}^T \mathbf{w}_t}{\|\mathbf{w}_t\|_2 \|\hat{\mathbf{w}}\|_2}$ and norm gap $\|\frac{\mathbf{w}_t}{\|\mathbf{w}_t\|_2} - \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|_2}\|_2$ of $\mathbf{w}_t$ to $\hat{\mathbf{w}}$, for two values of $\hat{\mathbf{w}}$ for the two subfigures as follows: (a) $\hat{\mathbf{w}}$ is the CS-SVM solution in (4) with parameter δ; (b) $\hat{\mathbf{w}}$ is the standard SVM solution. Data were generated from a Gaussian mixture model with $\boldsymbol{\mu}_1 = 2\mathbf{e}_1, \boldsymbol{\mu}_2 = -3\mathbf{e}_1 \in \mathbb{R}^{220}$, $n = 100$ and $\pi = 0.1$. For (standard) GD we used a constant rate $\eta_t = 0.1$. For normalized GD, we used $\eta_t = \frac{1}{\sqrt{t}\|\nabla \mathcal{L}(\mathbf{w}_t)\|_2}$ as suggested in [51].

$\pi = 0.1$. We ran two experiments for two choices of parameters in (1) corresponding to CDT-loss (with non-trivial multiplicative weights) and the LA-loss (with non-trivial additive weights); see the figure's caption for details. For each iterate outcome $\mathbf{w}_t$ of GD, we report the (i) angle and (ii) vector-norm gap to CS-SVM and SVM for the VS-loss and LA-loss, respectively, as well as, the (iii) value of the loss $\mathcal{L}(\mathbf{w}_t)$ and the (iv) norm of the weights $\|\mathbf{w}_t\|_2$ at current iteration. Observe that the loss $\mathcal{L}(\mathbf{w}_t)$ is driven to zero and the norm of the weights $\|\mathbf{w}_t\|_2$ increases to infinity with increasing t.

The experiment confirms that the VS-loss converges (aka angle/norm gap vanishes) to the CS-SVM solution, while the LA-loss converges to the SVM.

In Figure 14, we also study (curves in red) the convergence properties of *normalized GD*. Following [51], we implemented a version of normalized GD that uses a variable learning rate η_t at iteration t normalized by the gradient of the loss as follows: $\eta_t = \frac{1}{\|\nabla \mathcal{L}(\widetilde{\mathbf{w}})\|_2 \sqrt{t+1}}$. [51] (see also [29]) demonstrated that this normalization speeds up the convergence of standard logistic loss to SVM. Figure 14 suggests that the same is true for convergence of the VS-loss to the CS-SVM.

E   Optimal tuning of CS-SVM

E.1   An explicit formula for optimal tuning

The parameter δ in the CS-SVM constraints in (4) aims to shift the decision space towards the majority class so that it better balances the conditional errors of the two classes. But, how to best choose δ to achieve that? That is, how to find $\arg\min_\delta \mathcal{R}_+(\delta) + \mathcal{R}_-(\delta)$ where $\mathcal{R}_{\pm}(\delta) := \mathcal{R}_{\pm}((\hat{\mathbf{w}}_\delta, \hat{b}_\delta))$? Thanks to Theorem 2, we can substitute this hard, data-dependent parameter optimization problem with an analytic form that only depends on the problem parameters π, γ and $\mathbf{M}$. Specifically, we seek to solve the following optimization problem

$$\arg\min_{\delta > 0} \;\; Q(\mathbf{e}_1^T \mathbf{V S} \boldsymbol{\rho}_\delta + b_\delta/q_\delta) + Q(-\mathbf{e}_2^T \mathbf{V S} \boldsymbol{\rho}_\delta - b_\delta/q_\delta)$$

$$\text{sub. to} \quad (q_\delta, \boldsymbol{\rho}_\delta, b_\delta) \text{ defined as } (42). \tag{31}$$

Compared to the original data-dependent problem, the optimization above has the advantage that it is explicit in terms of the problem parameters. However, as written, the optimization

is still cumbersome as even a grid search over possible values of δ requires solving the non-linear equation (42) for each candidate value of δ. Instead, we can exploit a structural property of CS-SVM (see Lemma 2 in Section E.2) to rewrite (31) in a more convenient form. Specifically, we will show in Section E.3 that (31) is equivalent to the following *explicit minimization*:

$$\arg\min_{\delta>0} \; Q\Big(\ell_+ + \big(\frac{\delta-1}{\delta+1}\big)q_1^{-1}\Big) + Q\Big(\ell_- - \big(\frac{\delta-1}{\delta+1}\big)q_1^{-1}\Big), \tag{32}$$

where we defined $\ell_+ := \mathbf{e}_1^T \mathbf{VS}\boldsymbol{\rho}_1 + b_1/q_1$, $\ell_- := -\mathbf{e}_2^T \mathbf{VS}\boldsymbol{\rho}_1 - b_1/q_1$, and, $(q_1, \boldsymbol{\rho}_1, b_1)$ are as defined in Theorem 2 for $\delta = 1$. In other words, $(q_1, \boldsymbol{\rho}_1, b_1)$ are the parameters related to the standard hard-margin SVM, for which the balanced error is then given by $\big(Q(\ell_+) + Q(\ell_-)\big)/2$. To summarize, we have shown that one can optimally tune δ to minimize the *asymptotic* balanced error by minimizing the objective in (32) that only depends on the parameters $(q_1, \boldsymbol{\rho}_1, b_1)$ characterizing the asymptotic performance of SVM. In fact, we obtain explicit formulas for the optimal value $\delta_\star$ in (32) as follows

$$\delta_\star := \big(\ell_- - \ell_+ + 2q_1^{-1}\big)\big/\big(\ell_+ - \ell_- + 2q_1^{-1}\big)_+, \tag{33}$$

where it is understood that when the denominator is zero (i.e. $\ell_+ - \ell_- + 2q_1^{-1} \le 0$) then $\delta_\star \to \infty$. When $\ell_+ - \ell_- + 2q_1^{-1} > 0$, setting $\delta = \delta_\star$ in (4) not only achieves minimum balanced error among all other choices of δ, but also it achieves perfect balancing between the conditional errors of the two classes, i.e. $\mathcal{R}_+ = \mathcal{R}_- = Q(\frac{\ell_- + \ell_+}{2})$.

Formally, we have the following result.

Theorem 6 (Optimal tuning of CS-SVM)**.** *Fix $\gamma > \gamma_\star$. Let $\overline{\mathcal{R}}_{bal}(\delta)$ denote the asymptotic balanced error of the CS-SVM with margin-ratio parameter $\delta > 0$ as specified in Theorem 2. Further let $(q_1, \boldsymbol{\rho}_1, b_1)$ the solution to (42) for $\delta = 1$. Finally, define*

$$\ell_+ := \mathbf{e}_1^T \mathbf{VS}\boldsymbol{\rho}_1 + b_1/q_1, \quad \ell_- := -\mathbf{e}_2^T \mathbf{VS}\boldsymbol{\rho}_1 - b_1/q_1,$$

Then, for all $\delta > 0$ it holds that

$$\overline{\mathcal{R}}_{bal}(\delta) \ge \overline{\mathcal{R}}_{bal}(\delta_\star)$$

where $\delta_\star$ is defined as

$$\delta_\star = \begin{cases} \frac{\ell_- - \ell_+ + 2q_1^{-1}}{\ell_+ - \ell_- + 2q_1^{-1}} & \text{if } \ell_+ + \ell_- \ge 0 \text{ and } \ell_+ - \ell_- + 2q_1^{-1} > 0, \\ \to \infty & \text{if } \ell_+ + \ell_- \ge 0 \text{ and } \ell_+ - \ell_- + 2q_1^{-1} \le 0, \\ \to 0 & \text{if } \ell_+ + \ell_- < 0. \end{cases} \tag{34}$$

Specifically, if $\ell_+ + \ell_- \ge 0$ and $\ell_+ - \ell_- + 2q_1^{-1} > 0$ hold, then the following two hold: (i) $\overline{\mathcal{R}}_{bal}(\delta_\star) = Q\big((\ell_- + \ell_+)/2\big)$, and, (ii) the asymptotic conditional errors are equal, i.e. $\mathcal{R}_+(\delta_\star) = \mathcal{R}_-(\delta_\star)$.

See Figures 15c and 16 for numerical illustrations of the formula in Theorem 6, specifically how $\delta_\star$ depends on π and γ.

E.1.1 Data-dependent heuristic to estimate $\delta_\star$

It is natural to ask if formula (34) can be used for tuning in practice. To answer this, observe that evaluating the formula requires knowledge of the true means, which are typically unknown. In this section, we propose a *data-dependent heuristic* to estimate $\delta_\star$. More generally, tuning δ (or Δ_y in VS-loss) requires a train-validation split by creating a balanced validation set from the original training data which would help assess balanced risk. Since there is only a single hyperparameter we expect this approach to work well with fairly small validation data (without hurting the minority class sample size).

Recall from Equation (33) that $\delta_\star := \big(\ell_- - \ell_+ + 2q_1^{-1}\big)\big/\big(\ell_+ - \ell_- + 2q_1^{-1}\big)_+$, where $\ell_+ := \mathbf{e}_1^T \mathbf{VS}\boldsymbol{\rho}_1 + b_1/q_1$ and $\ell_- := -\mathbf{e}_2^T \mathbf{VS}\boldsymbol{\rho}_1 - b_1/q_1$. Also, according to Theorem 2 and for $\delta = 1$ it holds that

$$\big(\|\hat{\mathbf{w}}_1\|_2, \hat{\mathbf{w}}_1^T \boldsymbol{\mu}_+/\|\hat{\mathbf{w}}_1\|_2, \hat{\mathbf{w}}_1^T \boldsymbol{\mu}_-/\|\hat{\mathbf{w}}_1\|_2, \hat{b}_1\big) \xrightarrow{P} (q_1, \mathbf{e}_1^T \mathbf{VS}\boldsymbol{\rho}_1, \mathbf{e}_2^T \mathbf{VS}\boldsymbol{\rho}_1, b_1). \tag{35}$$

The first key observation here is that $\hat{\mathbf{w}}_1, \hat{b}_1$ are the solutions to SVM, thus they are data-dependent quantities to which we have access to. Hence, we can simply run SVM and
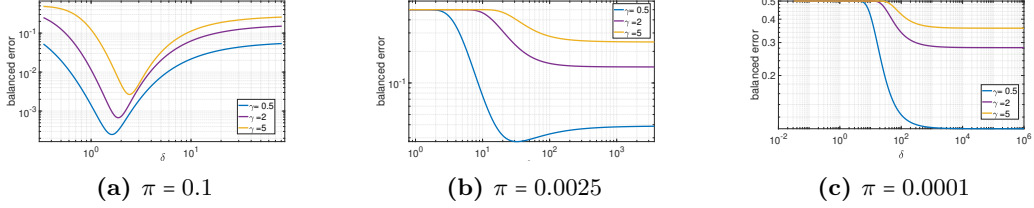
(a) $\pi = 0.1$ **(b)** $\pi = 0.0025$ **(c)** $\pi = 0.0001$

Figure 15: Graphical illustration of the result of Theorem 6: Balanced errors of CS-SVM against the margin-ratio parameter δ for a GMM of antipodal means with $\|\mu_+\| = \|\mu_-\| = 4$ and different minority class probabilities π. The balanced error is computed using the formulae of Theorem 2. For each case, we studied three different values of γ. The value $\delta_\star$ at which the curves attain (or approach) their minimum are predicted by Theorem 6. Specifically, note the following for the three different priors. (a) For all values of γ, the minimum is attained (cf. first branch of (34)). (b) For $\gamma = 2, 5$ the minimum is approached in the limit $\delta \to \infty$ (cf. second branch of (34)), but it is attained for $\gamma = 0.5$ (c) The minimum is always approached as $\delta_\star \to \infty$.
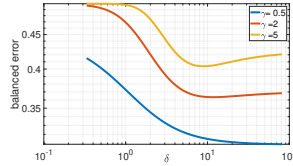

Figure 16: An example showing the dependence of $\delta_\star$ on the data geometry. The above figure is similar to Fig 15 but with a smaller $\|\mu_+\| = \|\mu_-\| = 1$, and for $\pi = 0.1$. While in Fig 15, the value of $\delta_\star$, whenever finite, can be seen to increase with increase in γ, for the current setting, it is observed to decrease. Note also that $\delta_\star \to \infty$ for $\gamma = 0.5$, but finite for $\gamma = 2, 5$.

estimate q_1 and b_1 using Equation (35). Unfortunately, to further estimate $\boldsymbol{\rho}_1$ we need knowledge of the data means. When this is not available, we propose approximating the data means by a simple average of the features, essentially pretending that the data follow a GMM.

Concretely, our recipe for approximating the optimal δ is as follows. First, using the training set we calculate the empirical means for the two classes, $\tilde{\boldsymbol{\mu}}_+$ and $\tilde{\boldsymbol{\mu}}_-$. (Ideally, this can be done on a balanced validation set.) Then, we train standard SVM on the same set of data and keep track of the coefficients $\hat{\mathbf{w}}_1$ and the intercept $\hat{b}_1$. Then, we can reasonably approximate the optimal δ as:

$$\tilde{\delta}_\star := \frac{\tilde{\ell}_- - \tilde{\ell}_+ + 2\|\hat{\mathbf{w}}_1\|_2^{-1}}{\left(\tilde{\ell}_+ - \tilde{\ell}_- + 2\|\hat{\mathbf{w}}_1\|_2^{-1}\right)_+}, \quad \text{with } \tilde{\ell}_+ := \frac{\hat{\mathbf{w}}_1^T \tilde{\boldsymbol{\mu}}_+ + \hat{b}_1}{\|\hat{\mathbf{w}}_1\|_2}, \quad \tilde{\ell}_- := -\frac{\hat{\mathbf{w}}_1^T \tilde{\boldsymbol{\mu}}_- + \hat{b}_1}{\|\hat{\mathbf{w}}_1\|_2}. \tag{36}$$

We expect this data-dependent theory-driven heuristic to perform reasonably well on data that resemble the GMM. For example, this is confirmed by our experiments in Figures 5 and 12. More generally, we propose tuning δ with a train-validation split by creating a balanced validation set from the original training data which would help assess balanced risk. Since there is only a single hyperparameter we expect this approach to work well with a fairly small validation data (without hurting the minority class sample size).

E.2 CS-SVM as post-hoc weight normalization

We need the lemma below to prove Theorem 6. But the results is interesting on its own right as it allows us to view CS-SVM as an appropriate "post-hoc weight normalization"-approach.

Lemma 2. *Let $(\hat{\mathbf{w}}_1, \hat{b}_1)$ be the hard-margin SVM solution. Fix any $\delta > 0$ in (4) and define: $\hat{\mathbf{w}}_\delta := \left(\frac{\delta+1}{2}\right)\hat{\mathbf{w}}_1$ and $\hat{b}_\delta := \left(\frac{\delta+1}{2}\right)\hat{b}_1 + \left(\frac{\delta-1}{2}\right)$. Then, $(\hat{\mathbf{w}}_\delta, \hat{b}_\delta)$ is optimal in (4).*

Thus, classification using (4) is equivalent to the following. First learn $(\hat{\mathbf{w}}_1, \hat{b}_1)$ via standard hard-margin SVM, and then simply predict: $\hat{y} = \text{sign}\left((\hat{\mathbf{w}}_1^T \mathbf{x} + \hat{b}_1) + \frac{\delta-1}{\delta+1}\right)$. The term $\frac{\delta-1}{\delta+1}$ can be

seen as an additive form of post-hoc weight normalization to account for class imbalances. In the literature this post-hoc adjustment of the threshold b of standard SVM is often referred to as boundary-movement SVM (BM-SVM) [65, 76]. Here, we have shown the equivalence of CS-SVM to BM-SVM for a specific choice of the boundary shift. The proof of Lemma 2 presented in Appendix E.2 shows the desired using the KKT conditions of (4).

Proof. From optimality of $(\hat{\mathbf{w}}_1, \hat{b}_1)$, convexity of (4) and the KKT-conditions, there exist dual variables $\beta_i, i \in [n]$ such that:

$$\hat{\mathbf{w}}_1 = \sum_{i \in [n]} y_i \beta_i \mathbf{x}_i, \quad \sum_{i \in [n]} y_i \beta_i = 0, \tag{37}$$

$$\forall i \in [n] \ : \ \beta_i \big(\mathbf{x}_i^T \hat{\mathbf{w}}_1 + \hat{b}_1 \big) = \beta_i y_i, \quad \beta_i \geq 0.$$

Let $(\hat{\mathbf{w}}_\delta, \hat{b}_\delta)$ defined as in the statement of the lemma and further define $\epsilon_i := \big(\frac{\delta+1}{2} \big) \beta_i, \ i \in [n]$. Then, it only takes a few algebra steps using (37) to check that the following conditions hold:

$$\hat{\mathbf{w}}_\delta = \sum_{i \in [n]} y_i \epsilon_i \mathbf{x}_i, \quad \sum_{i \in [n]} y_i \epsilon_i = 0, \tag{38}$$

$$\forall i \in [n] \ : \ \epsilon_i \big(\mathbf{x}_i^T \hat{\mathbf{w}}_\delta + \hat{b}_\delta \big) = \epsilon_i \cdot \begin{cases} \delta & \text{, if } y_i = +1 \\ -1 & \text{, if } y_i = -1 \end{cases}, \quad \epsilon_i \geq 0.$$

It can also be verified that (38) are the KKT conditions of the CS-SVM with parameter δ. This proves that $(\hat{\mathbf{w}}_\delta, \hat{b}_\delta)$ is optimal in (4) as desired. $\qquad \square$

E.3 Proof of Theorem 6

As discussed in the section above the proof proceeds in two steps:

(i) First, starting from (31), we prove (32).

(ii) Second, we analytically solve (32) to derive the explicit expression for $\delta_\star$ in (34).

Proof of (32). Fix any $\delta > 0$. From Lemma 2,

$$\hat{\mathbf{w}}_\delta = \big(\frac{\delta+1}{2} \big) \hat{\mathbf{w}}_1 \quad \text{and} \quad \hat{b}_\delta = \big(\frac{\delta+1}{2} \big) \hat{b}_1 + \big(\frac{\delta-1}{2} \big). \tag{39}$$

Recall from Theorem 2 that $\|\hat{\mathbf{w}}_\delta\|_2 \xrightarrow{P} q_\delta$, $\|\hat{\mathbf{w}}_1\|_2 \xrightarrow{P} q_1$, $\hat{b}_\delta \xrightarrow{P} b_\delta$, $\hat{b}_1 \xrightarrow{P} b_1$, and, for $i = 1, 2$: $\frac{\hat{\mathbf{w}}_\delta^T \boldsymbol{\mu}_i}{\|\hat{\mathbf{w}}_\delta\|_2} \xrightarrow{P} \mathbf{e}_i^T \mathbf{V} \mathbf{S} \boldsymbol{\rho}_\delta$ and $\frac{\hat{\mathbf{w}}_1^T \boldsymbol{\mu}_i}{\|\hat{\mathbf{w}}_1\|_2} \xrightarrow{P} \mathbf{e}_i^T \mathbf{V} \mathbf{S} \boldsymbol{\rho}_1$. Here, $q_\delta, \boldsymbol{\rho}_\delta, b_\delta$ and $q_1, \boldsymbol{\rho}_1, b_1$ are as defined in Theorem 2. Thus, from (39) we find that

$$\boldsymbol{\rho}_\delta = \boldsymbol{\rho}_1, \qquad q_\delta = \big(\frac{\delta+1}{2} \big) q_1 \qquad \text{and} \qquad b_\delta = \big(\frac{\delta+1}{2} \big) b_1 + \big(\frac{\delta-1}{2} \big). \tag{40}$$

Hence, it holds:

$$Q \big(\mathbf{e}_1^T \mathbf{V} \mathbf{S} \boldsymbol{\rho}_\delta + b_\delta / q_\delta \big) = Q \Big(\underbrace{\mathbf{e}_1^T \mathbf{V} \mathbf{S} \boldsymbol{\rho}_\delta + b_1 / q_1}_{= \ell_+} + \frac{\delta-1}{\delta+1} q_1^{-1} \Big).$$

A similar expression can be written for the conditional error of class -1. Putting these together shows (32), as desired.

Proof of (34). Recall from (32) that we now need to solve the following constrained minimization where for convenience we call $a = \ell_+$, $b = \ell_-$ and $c = q_1^{-1}$:

$$\min_{\delta > 0} \ Q \Big(a + \frac{\delta-1}{\delta+1} c \Big) + Q \Big(b - \frac{\delta-1}{\delta+1} c \Big).$$

We define a new variable $x = \frac{\delta-1}{\delta+1} c$. The constraint $\delta > 0$ then writes $x \leq c$. This is because the function $\delta \in (0, \infty) \mapsto \frac{\delta-1}{\delta+1}$ is onto the interval $(-1, 1)$.

Thus, we equivalently need to solve

$$\min_{-c < x < c} f(x) := Q(a + x) + Q(b - x).$$

Define function $f(x) = Q(a + x) + Q(b - x)$ for some $a, b \in \mathbb{R}$. Direct differentiation gives $\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{1}{\sqrt{2\pi}} \left(e^{-(b-x)^2/2} - e^{-(a+x)^2/2} \right)$. Furthermore, note that $\lim_{x \to \pm\infty} f(x) = 1$. With thes and some algebra it can be checked that $f(\cdot)$ behaves as follows depending on the sign of $a + b$. Denote $x_\star = (b - a)/2$.

- If $a + b \geq 0$, then $1 > f(x) \geq f(x_\star)$ and $x_\star$ is the unique minimum.
- If $a + b < 0$, then $1 < f(x) \leq f(x_\star)$ and $x_\star$ is the unique maximum.

Thus, we conclude with the following:

$$\arg \inf_{-c < x < c} f(x) = \begin{cases} x_\star & \text{if } a + b \geq 0 \text{ and } b - a < 2c, \\ c & \text{if } a + b \geq 0 \text{ and } b - a \geq 2c, \\ -c & \text{if } a + b < 0. \end{cases}$$

Equivalently,

$$\arg \inf_{\delta > -1} \ Q\left(\ell_+ + \frac{\delta - 1}{\delta + 1} q_1^{-1} \right) + Q\left(\ell_- - \frac{\delta - 1}{\delta + 1} q_1^{-1} \right) = \begin{cases} \frac{\ell_- - \ell_+ + 2q_1^{-1}}{\ell_+ - \ell_- + 2q_1^{-1}} & \text{if } \ell_+ + \ell_- \geq 0 \text{ and } \ell_+ - \ell_- + 2q_1^{-1} > 0, \\ \infty & \text{if } \ell_+ + \ell_- \geq 0 \text{ and } \ell_+ - \ell_- + 2q_1^{-1} \leq 0, \\ 0 & \text{if } \ell_+ + \ell_- < 0. \end{cases}$$

This shows (34). The remaining statement of the theorem is easy to prove requiring simple algebra manipulations.

F Asymptotic analysis of CS-SVM

F.1 Preliminaries

The main goal of this appendix is proving Theorem 2. For fixed $\delta > 0$, let $(\hat{\mathbf{w}}, \hat{b})$ be the solution to the CS-SVM in (4). (See also (48) below.) In the following sections, we will prove the following convergence properties for the solution of the CS-SVM:

$$\left(\|\hat{\mathbf{w}}\|_2, \frac{\hat{\mathbf{w}}^T \boldsymbol{\mu}_+}{\|\hat{\mathbf{w}}\|_2}, \frac{\hat{\mathbf{w}}^T \boldsymbol{\mu}_-}{\|\hat{\mathbf{w}}\|_2}, \hat{b} \right) \xrightarrow{P} (q_\delta, \mathbf{e}_1^T \mathbf{V} \mathbf{S} \boldsymbol{\rho}_\delta, \mathbf{e}_2^T \mathbf{V} \mathbf{S} \boldsymbol{\rho}_\delta, b_\delta). \tag{41}$$

where the triplet $(q_\delta, \boldsymbol{\rho}_\delta, b_\delta)$ is as defined in the theorem's statement, that is, the unique triplet satisfying

$$\eta_\delta(q_\delta, \boldsymbol{\rho}_\delta, b_\delta) = 0 \quad \text{and} \quad (\boldsymbol{\rho}_\delta, b_\delta) := \arg \min_{\|\boldsymbol{\rho}\|_2 \leq 1, b \in \mathbb{R}} \eta_\delta(q_\delta, \boldsymbol{\rho}, b). \tag{42}$$

In this section, we show how to use (41) to derive the asymptotic limit of the conditional class probabilities.

Consider the class conditional $\mathcal{R}_+ = \mathbb{P}\left\{ (\mathbf{x}^T \hat{\mathbf{w}} + b) < 0 \,|\, y = +1 \right\}$. Recall that conditioned on $y = +1$, we have $\mathbf{x} = \boldsymbol{\mu}_+ + \mathbf{z}$ for $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Thus, the class conditional can be expressed explicitly in terms of the three summary quantities on the left hand side of (41) as follows:

$$\begin{aligned} \mathcal{R}_+ &= \mathbb{P}\left\{ (\mathbf{x}^T \hat{\mathbf{w}} + \hat{b}) < 0 \,|\, y = +1 \right\} = \mathbb{P}\left\{ \mathbf{z}^T \hat{\mathbf{w}} + \boldsymbol{\mu}_+^T \hat{\mathbf{w}} + \hat{b} < 0 \,|\, y = +1 \right\} \\ &= \mathbb{P}\left\{ \mathbf{z}^T \hat{\mathbf{w}} > \boldsymbol{\mu}_+^T \hat{\mathbf{w}} + \hat{b} \right\} \\ &= \mathbb{P}_{G \sim \mathcal{N}(0,1)}\left\{ G\|\hat{\mathbf{w}}\|_2 > \boldsymbol{\mu}_+^T \hat{\mathbf{w}} + \hat{b} \right\} = \mathbb{P}_{G \sim \mathcal{N}(0,1)}\left\{ G > \frac{\boldsymbol{\mu}_+^T \hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|_2} + \frac{\hat{b}}{\|\hat{\mathbf{w}}\|_2} \right\} \\ &= Q\left(\frac{\boldsymbol{\mu}_+^T \hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|_2} + \frac{\hat{b}}{\|\hat{\mathbf{w}}\|_2} \right). \end{aligned}$$

Then, the theorem's statement follows directly by applying (41) in the expression above.

In order to prove the key convergence result in (41) we rely on the convex Gaussian min-max theorem (CGMT) framework. We give some necessary background before we proceed with the proof.

F.2 Background and related literature

Related works: Our asymptotic analysis of the CS-SVM fits in the growing recent literature on sharp statistical performance asymptotics of convex-based estimators, e.g. [9, 18, 72, 74] and references therin. The origins of these works trace back to the study of sharp phase transitions in compressed sensing, e.g. see [72] for historical remarks and performance analysis of the LASSO estimator for sparse signal recovery. That line of work led to the development of two analysis frameworks: (a) the approximate message-passing (AMP) framework [8, 19], and, (b) the convex Gaussian min-max theorem (CGMT) framework [67, 73]. More recently, these powerful tools have proved very useful for the analysis of linear classifiers [64, 49, 15, 34, 47, 39, 70, 12, 4, 69]. Theorems 2 and 7 rely on the CGMT and contribute to this line of work. Specifically, our results are most closely related to [15] who first studied max-margin type classifiers together with [49].

CGMT framework: Specifically, we rely on the CGMT framework. Here, we only summarize the framework's essential ideas and refer the reader to [73, 72] for more details and precise statements. Consider the following two Gaussian processes:

$$X_{\mathbf{w},\mathbf{u}} := \mathbf{u}^T \mathbf{A} \mathbf{w} + \psi(\mathbf{w}, \mathbf{u}), \tag{43a}$$

$$Y_{\mathbf{w},\mathbf{u}} := \|\mathbf{w}\|_2 \mathbf{h}_n^T \mathbf{u} + \|\mathbf{u}\|_2 \mathbf{h}_d^T \mathbf{w} + \psi(\mathbf{w}, \mathbf{u}), \tag{43b}$$

where: $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{h}_n \in \mathbb{R}^n$, $\mathbf{h}_d \in \mathbb{R}^d$, they all have entries iid Gaussian; the sets $\mathcal{S}_{\mathbf{w}} \subset \mathbb{R}^d$ and $\mathcal{S}_{\mathbf{u}} \subset \mathbb{R}^n$ are compact; and, $\psi : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$. For these two processes, define the following (random) min-max optimization programs, which are refered to as the *primary optimization* (PO) and the *auxiliary optimization* (AO) problems:

$$\Phi(\mathbf{A}) = \min_{\mathbf{w} \in \mathcal{S}_{\mathbf{w}}} \max_{\mathbf{u} \in \mathcal{S}_{\mathbf{u}}} X_{\mathbf{w},\mathbf{u}}, \tag{44a}$$

$$\phi(\mathbf{h}_n, \mathbf{h}_d) = \min_{\mathbf{w} \in \mathcal{S}_{\mathbf{w}}} \max_{\mathbf{u} \in \mathcal{S}_{\mathbf{u}}} Y_{\mathbf{w},\mathbf{u}}. \tag{44b}$$

According to the first statement of the CGMT Theorem 3 in [73] (this is only a slight reformulation of Gordon's original comparison inequality [21]), for any $c \in \mathbb{R}$, it holds:

$$\mathbb{P}\{\Phi(\mathbf{A}) < c\} \leq 2\,\mathbb{P}\{\phi(\mathbf{h}_n, \mathbf{h}_d) < c\}. \tag{45}$$

In other words, a high-probability lower bound on the AO is a high-probability lower bound on the PO. The premise is that it is often much simpler to lower bound the AO rather than the PO. However, the real power of the CGMT comes in its second statement, which asserts that if the PO is *convex* then the AO in can be used to tightly infer properties of the original PO, including the optimal cost and the optimal solution. More precisely, if the sets $\mathcal{S}_{\mathbf{w}}$ and $\mathcal{S}_{\mathbf{u}}$ are convex and *bounded*, and ψ is continuous *convex-concave* on $\mathcal{S}_{\mathbf{w}} \times \mathcal{S}_{\mathbf{u}}$, then, for any $\nu \in \mathbb{R}$ and $t > 0$, it holds [73]:

$$\mathbb{P}\{|\Phi(\mathbf{A}) - \nu| > t\} \leq 2\,\mathbb{P}\{|\phi(\mathbf{h}_n, \mathbf{h}_d) - \nu| > t\}. \tag{46}$$

In words, concentration of the optimal cost of the AO problem around q^* implies concentration of the optimal cost of the corresponding PO problem around the same value q^*. Asymptotically, if we can show that $\phi(\mathbf{h}_n, \mathbf{h}_d) \xrightarrow{P} q^*$, then we can conclude that $\Phi(\mathbf{A}) \xrightarrow{P} q^*$.

In the next section, we will show that we can indeed express the CS-SVM in (4) as a PO in the form of (44a). Thus, the argument above will directly allow us to determine the asymptotic limit of the optimal cost of the CS-SVM. In our case, the optimal cost equals $\|\hat{\mathbf{w}}\|_2$; thus, this shows the first part of (41). For the other parts, we will employ the following

"deviation argument" of the CGMT framework [73]. For arbitrary $\epsilon > 0$, consider the desired set

$$\mathcal{S} := \left\{ (\mathbf{v}, c) \ \Big| \ \max\left\{ \|\mathbf{v}\|_2 - q_\delta|, \left| \frac{\mathbf{v}^T \boldsymbol{\mu}_+}{\|\mathbf{v}\|_2} - \mathbf{e}_1^T \mathbf{V} \mathbf{S} \boldsymbol{\rho}_\delta \right|, , \left| \frac{\mathbf{v}^T \boldsymbol{\mu}_-}{\|\mathbf{v}\|_2} - \mathbf{e}_2^T \mathbf{V} \mathbf{S} \boldsymbol{\rho}_\delta \right|, |c - b_\delta| \right\} \le \epsilon \right\}. \quad (47)$$

Our goal towards (41) is to show that with overwhelming probability $(\mathbf{w}, b) \in \mathcal{S}$. For this, consider the following constrained CS-SVM that further constraints the feasible set to the complement $\mathcal{S}^c$ of $\mathcal{S}$:

$$\Phi_{\mathcal{S}^c}(\mathbf{A}) := \min_{(\mathbf{w},b) \in \mathcal{S}^c} \|\mathbf{w}\|_2 \text{ sub. to } \begin{cases} \mathbf{w}^T \mathbf{x}_i + b \ge \delta & , y_i = +1 \\ \mathbf{w}^T \mathbf{x}_i + b \le -1 & , y_i = -1 \end{cases}, i \in [n], \quad (48)$$

As per Theorem 6.1(iii) in [72] it will suffice to find costants $\bar{\phi}, \bar{\phi}_S$ and $\eta > 0$ such that the following three conditions hold:

$$\begin{cases} \text{(i)} & \bar{\phi}_S \ge \bar{\phi} + 3\eta \\ \text{(ii)} & \phi(\mathbf{h}_n, \mathbf{h}_d) \le \bar{\phi} + \eta \quad \text{with overwhelming probability} \\ \text{(iii)} & \phi_{\mathcal{S}^c}(\mathbf{h}_n, \mathbf{h}_d) \ge \bar{\phi}_S - \eta \quad \text{with overwhelming probability,} \end{cases} \quad (49)$$

where $\phi_{\mathcal{S}^c}(\mathbf{h}_n, \mathbf{h}_d)$ is the optimal cost of the constrained AO corresponding to the constrained PO in (48).

To prove these conditions for the AO of the CS-SVM, in the next section we follow the principled machinery of [72] that allows simplifying the AO from a (random) optimization over vector variables to an easier optimization over only few scalar variables, termed the "scalarized AO".

F.3 Proof of Theorem 2

Let $(\hat{\mathbf{w}}, \hat{b})$ be solution pair to the CS-SVM in (4) for some fixed margin-ratio parameter $\delta > 0$, which we rewrite here expressing the constraints in matrix form:

$$\min_{\mathbf{w},b} \|\mathbf{w}\|_2 \text{ sub. to } \begin{cases} \mathbf{w}^T \mathbf{x}_i + b \ge \delta, \ y_i = +1 \\ -(\mathbf{w}^T \mathbf{x}_i + b) \ge 1, \ y_i = -1 \end{cases}, i \in [n] = \min_{\mathbf{w},b} \|\mathbf{w}\|_2 \text{ sub. to } \mathbf{D}_\mathbf{y}(\mathbf{X}\mathbf{w} + b\mathbf{1}_n) \ge \boldsymbol{\delta}_\mathbf{y}, \quad (50)$$

where we have used the notation

$$\mathbf{X}^T = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix}, \ \mathbf{y} = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^T,$$

$$\mathbf{D}_\mathbf{y} = \text{diag}(\mathbf{y}) \text{ and } \boldsymbol{\delta}_\mathbf{y} = \begin{bmatrix} \delta \mathbb{1}[y_1 = +1] + \mathbb{1}[y_1 = -1] & \cdots & \delta \mathbb{1}[y_n = +1] + \mathbb{1}[y_n = -1] \end{bmatrix}^T.$$

We further need to define the following one-hot-encoding of the labels:

$$\mathbf{y}_i = \mathbf{e}_1 \mathbb{1}[y_i = 1] + \mathbf{e}_2 \mathbb{1}[y_i = -1], \quad \text{and} \quad \mathbf{Y}_{n \times 2}^T = \begin{bmatrix} \mathbf{y}_1 & \cdots & \mathbf{y}_n \end{bmatrix}.$$

where recall that $\mathbf{e}_1, \mathbf{e}_2$ are standard basis vectors in $\mathbb{R}^2$.

With these, notice for later use that under our model, $\mathbf{x}_i = \boldsymbol{\mu}_{y_i} + \mathbf{z}_i = \mathbf{M}\mathbf{y}_i + \mathbf{z}_i, \ \mathbf{z}_i \sim \mathcal{N}(0, 1)$. Thus, in matrix form with $\mathbf{Z}$ having entries $\mathcal{N}(0, 1)$:

$$\mathbf{X} = \mathbf{Y}\mathbf{M}^T + \mathbf{Z}. \quad (51)$$

Following the CGMT strategy [73], we express (50) in a min-max form to bring it in the form of the PO as follows:

$$\min_{\mathbf{w},b} \max_{\mathbf{u} \le 0} \frac{1}{2} \|\mathbf{w}\|_2^2 + \mathbf{u}^T \mathbf{D}_\mathbf{y} \mathbf{X} \mathbf{w} + b(\mathbf{u}^T \mathbf{D}_\mathbf{y} \mathbf{1}_n) - \mathbf{u}^T \boldsymbol{\delta}_\mathbf{y}$$

$$= \min_{\mathbf{w},b} \max_{\mathbf{u} \le 0} \frac{1}{2} \|\mathbf{w}\|_2^2 + \mathbf{u}^T \mathbf{D}_\mathbf{y} \mathbf{Z} \mathbf{w} + \mathbf{u}^T \mathbf{D}_\mathbf{y} \mathbf{Y} \mathbf{M}^T \mathbf{w} + b(\mathbf{u}^T \mathbf{D}_\mathbf{y} \mathbf{1}_n) - \mathbf{u}^T \boldsymbol{\delta}_\mathbf{y}. \quad (52)$$

where in the last line we used (51) and $\mathbf{D_y D_y} = \mathbf{I}_n$. We immediately recognize that the last optimization is in the form of a PO (cf. (44a)) and the corresponding AO (cf. (44b)) is as follows:

$$\min_{\mathbf{w},b} \max_{\mathbf{u} \leq 0} \frac{1}{2}\|\mathbf{w}\|_2^2 + \|\mathbf{w}\|_2 \mathbf{u}^T \mathbf{D_y h}_n + \|\mathbf{D_y u}\|_2 \mathbf{h}_d^T \mathbf{w} + \mathbf{u}^T \mathbf{D_y Y M}^T \mathbf{w} + b(\mathbf{u}^T \mathbf{D_y 1}_n) - \mathbf{u}^T \boldsymbol{\delta_y}.$$
(53)

where $\mathbf{h}_n \sim \mathcal{N}(0, \mathbf{I}_n)$ and $\mathbf{h}_d \sim \mathcal{N}(0, \mathbf{I}_d)$.

In order to apply the CGMT in [73], we need boundedness of the constraint sets. Thus, we restrict the minimization in (53) and (52) to a bounded set $\|\mathbf{w}\|_2^2 + b^2 \leq R$ for (say) $R := 2(q_\delta^2 + b_\delta^2)$. This will allow us to show that the solutions $\hat{\mathbf{w}}_R, \hat{b}_R$ of this constrained PO satisfy $\hat{\mathbf{w}}_R \xrightarrow{P} q_\delta$ and $\hat{b}_R \xrightarrow{P} b_\delta$. Thus, with overwhelming probability, $\|\hat{\mathbf{w}}_R\|_2^2 + \hat{b}_R^2 < R$. From this and convexity of the PO, we can argue that the minimizers $\hat{\mathbf{w}}, \hat{b}$ of the original unconstrained problem satisfy the same convergence properties. Please see also Remark 4 in App. A of [15].

For the maximization, we follow the recipe in App. A of [15] who analyzed the standard SVM. Specifically, combining Remark 3 of [15] together with (we show this next) the property that the AO is reduced to a convex program, it suffices to consider the unconstrained maximization.

Thus, in what follows we consider the one-sided constrained AO in (53). Towards simplifying this auxiliary optimization, note that $\mathbf{D_y h}_n \sim \mathbf{h}_n$ by rotational invariance of the Gaussian measure. Also, $\|\mathbf{D_y u}\|_2 = \|\mathbf{u}\|_2$. Thus, we can express the AO in the following more convenient form:

$$\min_{\|\mathbf{w}\|_2^2 + b^2 \leq R} \max_{\mathbf{u} \leq 0} \frac{1}{2}\|\mathbf{w}\|_2^2 + \|\mathbf{w}\|_2 \mathbf{u}^T \mathbf{h}_n + \|\mathbf{u}\|_2 \mathbf{h}_d^T \mathbf{w} + \mathbf{u}^T \mathbf{D_y Y M}^T \mathbf{w} + b(\mathbf{u}^T \mathbf{D_y 1}_n) - \mathbf{u}^T \boldsymbol{\delta_y}. \quad (54)$$

We are now ready to proceed with simplification of the AO. First we optimize over the direction of $\mathbf{u}$ and rewrite the AO as

$$\min_{\|\mathbf{w}\|_2^2 + b^2 \leq R} \max_{\beta \geq 0} \frac{1}{2}\|\mathbf{w}\|_2^2 + \beta \left(\left\| \left(\|\mathbf{w}\|_2 \mathbf{h}_n + \mathbf{D_y Y M}^T \mathbf{w} + b \mathbf{D_y 1}_n - \boldsymbol{\delta_y} \right)_- \right\|_2 - \mathbf{h}_d^T \mathbf{w} \right)$$

$$= \min_{\|\mathbf{w}\|_2^2 + b^2 \leq R} \frac{1}{2}\|\mathbf{w}\|_2^2 \quad \text{sub. to } \left\| \left(\|\mathbf{w}\|_2 \mathbf{h}_n + \mathbf{D_y Y M}^T \mathbf{w} + b \mathbf{D_y 1}_n - \boldsymbol{\delta_y} \right)_- \right\|_2 \leq \mathbf{h}_d^T \mathbf{w}.$$

Above, $(\cdot)_-$ acts elementwise to the entries of its argument.

Now, we wish to further simplify the above by minimizing over the direction of $\mathbf{w}$ in the space orthogonal to $\mathbf{M}$. To see how this is possible consider the SVD $\mathbf{M}^T = \mathbf{VSU}^T$ and project $\mathbf{w}$ on the columns of $\mathbf{U} = [\mathbf{u}_1 \quad \mathbf{u}_2] \in \mathbb{R}^{d \times 2}$ as follows:

$$\mathbf{w} = \mathbf{u}_1(\mathbf{u}_1^T \mathbf{w}) + \mathbf{u}_2(\mathbf{u}_2^T \mathbf{w}) + \mathbf{w}^\perp,$$

where $\mathbf{w}^\perp = \mathbf{U}^\perp \mathbf{w}$, $\mathbf{U}^\perp$ is the orthogonal complement of $\mathbf{U}$. For simplicity we will assume here that $\mathbf{M}$ is full column rank, i.e. $\mathbf{S} > \mathbf{0}_{2 \times 2}$. The argument for the case where $\mathbf{M}$ is rank 1 is very similar.

Let us denote $\mathbf{u}_i^T \mathbf{w} := \mu_i, i = 1, 2$ and $\|\mathbf{w}^\perp\|_2 := \alpha$. In this notation, the AO becomes

$$\min_{\mu_1^2 + \mu_2^2 + \|\mathbf{w}^\perp\|_2^2 + b^2 \leq R} \frac{1}{2}(\mu_1^2 + \mu_2^2 + \alpha^2)$$

$$\text{sub. to } \left\| \left(\sqrt{\mu_1^2 + \mu_2^2 + \alpha^2}\, \mathbf{h}_n + \mathbf{D_y Y V S}\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix} + b\, \mathbf{D_y 1}_n - \boldsymbol{\delta_y} \right)_- \right\|_2$$

$$\leq \mu_1(\mathbf{h}_d^T \mathbf{u}_1) + \mu_2(\mathbf{h}_d^T \mathbf{u}_2) + \mathbf{h}_d^T \mathbf{U}^\perp \mathbf{w}^\perp.$$

At this point, we can optimize over the direction of $\mathbf{w}^\perp$ which leads to

$$\min_{\mu_1^2 + \mu_2^2 + \alpha^2 + b^2 \leq R} \frac{1}{2}(\mu_1^2 + \mu_2^2 + \alpha^2)$$

$$\text{sub. to } \left\| \left(\sqrt{\mu_1^2 + \mu_2^2 + \alpha^2}\, \mathbf{h}_n + \mathbf{D_y Y V S}\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix} + b\, \mathbf{D_y 1}_n - \boldsymbol{\delta_y} \right)_- \right\|_2$$

$$\leq \mu_1(\mathbf{h}_d^T \mathbf{u}_1) + \mu_2(\mathbf{h}_d^T \mathbf{u}_2) + \alpha \|\mathbf{h}_d^T \mathbf{U}^\perp\|_2.$$

As a last step in the simplification of the AO, it is convenient to introduce an additional variable $q = \sqrt{\mu_1^2 + \mu_2^2 + \alpha^2}$. It then follows that the minimization above is equivalent to the following

$$\min_{\substack{q \geq \sqrt{\mu_1^2 + \mu_2^2 + \alpha^2} \\ q^2 + b^2 \leq R}} \frac{1}{2} q^2 \tag{55}$$

$$\text{sub. to} \quad \left\| \left(q\mathbf{h}_n + \mathbf{D_y YVS} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + b\,\mathbf{D_y 1}_n - \boldsymbol{\delta_y} \right)_- \right\|_2 \leq \mu_1(\mathbf{h}_d^T \mathbf{u}_1) + \mu_2(\mathbf{h}_d^T \mathbf{u}_2) + \alpha \|\mathbf{h}_d^T \mathbf{U}^\perp\|_2.$$

In this formulation it is not hard to check that the optimization is jointly convex in its variables $(\mu_1, \mu_2, \alpha, b, q)$. To see this note that: (i) the constraint $q \geq \sqrt{\mu_1^2 + \mu_2^2 + \alpha^2} \iff q \geq \| \begin{bmatrix} \mu_1 & \mu_2 & \alpha \end{bmatrix} \|_2$ is a second-order cone constraint, and, (ii) the function

$$\mathcal{L}_n(q, \mu_1, \mu_2, \alpha, b) := \frac{1}{\sqrt{n}} \left\| \left(q\mathbf{h}_n + \mathbf{D_y YVS} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + b\,\mathbf{D_y 1}_n - \boldsymbol{\delta_y} \right)_- \right\|_2$$

$$- \mu_1 \frac{\mathbf{h}_d^T \mathbf{u}_1}{\sqrt{n}} - \mu_2 \frac{\mathbf{h}_d^T \mathbf{u}_2}{\sqrt{n}} - \alpha \frac{\|\mathbf{h}_d^T \mathbf{U}^\perp\|_2}{\sqrt{n}} \tag{56}$$

is also convex since $\|(\cdot)_-\|_2 : \mathbb{R}^n \to \mathbb{R}$ is itslef convex and is composed here with an affine function.

Now, by law of large numbers, notice that for fixed $(q, \mu_1, \mu_2, \alpha, b)$, $\mathcal{L}_n$ converges in probability to

$$\mathcal{L}_n(q, \mu_1, \mu_2, \alpha, b) \xrightarrow{P} L(q, \mu_1, \mu_2, \alpha, b) := \sqrt{\mathbb{E}\left(qG + E_Y^T \mathbf{VS} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + bY - \Delta_Y \right)_-^2} - \alpha\sqrt{\gamma}, \tag{57}$$

where the random variables G, E_Y, Y, Δ_Y are as in the statement of the theorem. But convergence of convex functions is uniform over compact sets as per Cor. II.I in [3]. Therefore, the convergence in (57) is in fact uniform in the compact feasible set of (55).

Consider then the deterministic high-probability equivalent of (55) which is the following convex program:

$$\min_{\substack{q \geq \sqrt{\mu_1^2 + \mu_2^2 + \alpha^2} \\ q^2 + b^2 \leq R \\ L(q, \mu_1, \mu_2, \alpha, b) \leq 0}} \frac{1}{2} q^2.$$

Since q is positive and the constraint $q \geq \sqrt{\mu_1^2 + \mu_2^2 + \alpha^2}$ must be active at the optimum, it is convenient to rewrite this in terms of new variables $\boldsymbol{\rho} = \begin{bmatrix} \boldsymbol{\rho}_1 \\ \boldsymbol{\rho}_2 \end{bmatrix} := \begin{bmatrix} \mu_1/q \\ \mu_2/q \end{bmatrix}$ as follows:

$$\min_{q^2 + b^2 \leq R, q > 0, \|\boldsymbol{\rho}\|_2 \leq 1} \frac{1}{2} q^2 \tag{58}$$

$$\text{sub. to} \quad \mathbb{E}\left[\left(G + E_Y^T \mathbf{VS} \boldsymbol{\rho} + \frac{bY - \Delta_Y}{q} \right)_-^2 \right] \leq \left(1 - \|\boldsymbol{\rho}\|_2^2 \right) \gamma.$$

Now, recall the definition of the function η_δ in the statement of the theorem and observe that the constraint above is nothing but

$$\eta_\delta(q, \boldsymbol{\rho}, b) \leq 0.$$

Thus, (58) becomes

$$\min \left\{ q^2 \; \middle| \; 0 \leq q \leq \sqrt{R} \quad \text{and} \quad \min_{b^2 \leq R - q^2, \|\boldsymbol{\rho}\|_2 \leq 1} \eta_\delta(q, \boldsymbol{\rho}, b) \leq 0 \right\}. \tag{59}$$

We will prove that

$$\text{the function } f(q) := \min_{b, \|\boldsymbol{\rho}\|_2 \leq 1} \eta_\delta(q, \boldsymbol{\rho}, b) \text{ is strictly decreasing.} \tag{60}$$

40

Before that, let us see how this completes the proof of the theorem. Let q_δ be as in the statement of the theorem, that is such that $f(q_\delta) = 0$. Then, we have the following relations

$$f(q) \le 0 \ \Rightarrow \ f(q) \le f(q_\delta) \ \Rightarrow \ q \ge q_\delta.$$

Thus, the minimizers in (59) are $(q_\delta, \boldsymbol{\rho}_\delta, b_\delta)$, where we also recall that we have set $R > q_\delta^2 + b_\delta^2$.

With all these, we have shown that the AO converges in probability to q_δ^2 (cf. condition (ii) in (49)). From the CGMT, the same is true for the PO. Now, we want to use the same machinery to prove that the minimizers $(\hat{\mathbf{w}}, \hat{b})$ of the PO satisfy (41). To do this, as explained in the previous section, we use the standard strategy of the CGMT framework , i.e., to show that the PO with the additional constraint $(\mathbf{w}, b) \in \mathcal{S}^c$ for the set $\mathcal{S}$ in (47) has a cost that is strictly larger than q_δ^2 (i.e. the cost of the unconstrained PO). As per the CGMT this can be done again by showing that the statement is true for the correspondingly constrained AO (i.e. show condition (iii) in (49)). With the exact same simplifications as above, the latter program simplifies to (55) with the additional constraints:

$$|q - q_\delta| > \epsilon \,, \ \left|\mu_i/q - \boldsymbol{\rho}_{\delta,i}\right| > \epsilon, \ i = 1, 2 \,, \ |b - b_\delta| > \epsilon.$$

Also, using the uniform convergence in (57), it suffices to study the deterministic equivalent (59) with the additional constraints above. Now, we can show the desired (cf. condition (i) in (49)) again by exploiting (60). This part of the argument is similar to Section C.3.5 in [15] and we omit the details.

Proof of (60): To complete the proof, it remains to show (60). Specifically, we show that $\frac{\mathrm{d}f}{\mathrm{d}q} < 0$ by combining the following three observations.

First,

$$\frac{\partial \eta_\delta}{\partial q} = \frac{2}{q^2} \mathbb{E}\Big[\big(G + E_Y^T \mathbf{V}\mathbf{S}\boldsymbol{\rho} + \frac{bY - \Delta_Y}{q}\big)_- \cdot \Delta_Y\Big] - \frac{2b}{q^2} \mathbb{E}\Big[\big(G + E_Y^T \mathbf{V}\mathbf{S}\boldsymbol{\rho} + \frac{bY - \Delta_Y}{q}\big)_- \cdot Y\Big]$$

$$< -\frac{2b}{q^2} \mathbb{E}\Big[\big(G + E_Y^T \mathbf{V}\mathbf{S}\boldsymbol{\rho} + \frac{bY - \Delta_Y}{q}\big)_- \cdot Y\Big] \tag{61}$$

where for the inequality we observed that $(\cdot)_-$ is always non-positive, its argument has non-zero probability measure on the negative real axis, and Δ_Y are positive random variables.

Second, letting $\boldsymbol{\rho}^\star := \boldsymbol{\rho}^\star(q)$ and $b^\star := b^\star(q)$ the minimizers of $\eta_\delta(q, \boldsymbol{\rho}, b)$, it follows from first-order optimality conditions that

$$\frac{\partial \eta_\delta}{\partial b} = 0 \iff \mathbb{E}\Big[\big(G + E_Y^T \mathbf{V}\mathbf{S}\boldsymbol{\rho}^* + \frac{b^* Y - \Delta_Y}{q}\big)_- \cdot Y\Big] = 0. \tag{62}$$

Third, by the envelope theorem

$$\frac{\mathrm{d}f}{\mathrm{d}q} = \frac{\partial \eta_\delta}{\partial q}\Big|_{\boldsymbol{\rho}^\star, b^\star}. \tag{63}$$

The desired inequality $\frac{\mathrm{d}f}{\mathrm{d}q} < 0$ follows directly by successively applying (63), (61) and (62).

Uniqueness of triplet $(q_\delta, \boldsymbol{\rho}_\delta, b_\delta)$**.** First, we prove that the minimizers $\boldsymbol{\rho}_\delta, b_\delta$ are unique. This follows because $\eta_\delta(q, \boldsymbol{\rho}, b)$ is jointly strictly convex in $(\boldsymbol{\rho}, b)$ for fixed q. To see this note that the function $x \mapsto (x)_-^2$ is strictly convex for $x < 0$ and that the random variable $G + E_Y^T \mathbf{V}\mathbf{S}\boldsymbol{\rho} + (bY - \Delta_Y)/q$ has strictly positive measure on the real line (thus, also in the negative axis). Next, consider q_δ, which was defined such that $f(q_\delta) = 0$ for the function $f(\cdot)$ in (60). From (60) we know that $f(\cdot)$ is strictly decreasing. Thus, it suffices to prove that the function has a zero crossing in $(0, \infty)$, which we do by proving $\lim_{q \to 0} f(q) = \infty$ and $\lim_{q \to \infty} f(q) < 0$. Specifically, we have

$$\lim_{q \to 0} f(q) \ge \lim_{q \to 0} \min_{b \in \mathbb{R}, \|\boldsymbol{\rho}\|_2 \le 1} \mathbb{E}\Big[\big(G + E_Y^T \mathbf{V}\mathbf{S}\boldsymbol{\rho} + \frac{bY - \Delta_Y}{q}\big)_-^2\Big] - \gamma$$

$$\ge \lim_{q \to 0} \min_{b \in \mathbb{R}, \|\boldsymbol{\rho}\|_2 \le 1} \mathbb{E}\Big[\big(G + E_Y^T \mathbf{V}\mathbf{S}\boldsymbol{\rho} + \frac{bY - \Delta_Y}{q}\big)_-^2 \mathbb{1}\big[G + E_Y^T \mathbf{V}\mathbf{S}\boldsymbol{\rho} + (b/q)Y \le 0\big]\Big] - \gamma$$

$$\ge \lim_{q \to 0} \min_{b \in \mathbb{R}, \|\boldsymbol{\rho}\|_2 \le 1} 1/q^2 - \gamma = \infty,$$

41

where in the last inequality we used the facts that $x \mapsto (x)_-^2$ is decreasing and the event $\{G + E_Y^T \mathbf{V} \mathbf{S} \boldsymbol{\rho} + (b/q)Y \le 0 \le 0\}$ has non-zero measure for all $\|\boldsymbol{\rho}\|_2 \le 1, b \in \mathbb{R}$, as well as, $\Delta_Y \ge 1$ (because $\delta > 1$). Moreover,

$$
\begin{aligned}
\lim_{q \to \infty} f(q) = \lim_{1/q \to 0^+} f(q) &= \lim_{1/q \to 0^+} \min_{b \in \mathbb{R}, \|\boldsymbol{\rho}\|_2 \le 1} \mathbb{E}\Big[\big(G + E_Y^T \mathbf{V} \mathbf{S} \boldsymbol{\rho} + \frac{bY - \Delta_Y}{q}\big)_-^2\Big] - (1 - \|\boldsymbol{\rho}\|_2^2)\gamma \\
&= \lim_{1/q \to 0^+} \min_{\tilde{b} \in \mathbb{R}, \|\boldsymbol{\rho}\|_2 \le 1} \mathbb{E}\Big[\big(G + E_Y^T \mathbf{V} \mathbf{S} \boldsymbol{\rho} + \tilde{b}Y - \frac{\Delta_Y}{q}\big)_-^2\Big] - (1 - \|\boldsymbol{\rho}\|_2^2)\gamma \\
&\le \min_{\tilde{b} \in \mathbb{R}, \|\boldsymbol{\rho}\|_2 \le 1} \mathbb{E}\Big[\big(G + E_Y^T \mathbf{V} \mathbf{S} \boldsymbol{\rho} + \tilde{b}Y\big)_-^2\Big] - (1 - \|\boldsymbol{\rho}\|_2^2)\gamma \\
&\le \min_{\tilde{b} \in \mathbb{R}, \|\boldsymbol{\rho}\|_2 \le 1} (1 - \|\boldsymbol{\rho}\|_2^2) \cdot \Big(\mathbb{E}\Big[\big((G + E_Y^T \mathbf{V} \mathbf{S} \boldsymbol{\rho} + \tilde{b}Y)/\sqrt{1 - \|\boldsymbol{\rho}\|_2^2}\big)_-^2\Big] - \gamma\Big) \\
&\le \min_{\tilde{b} \in \mathbb{R}, \|\boldsymbol{\rho}\|_2 \le 1} \mathbb{E}\Big[\big((G + E_Y^T \mathbf{V} \mathbf{S} \boldsymbol{\rho} + \tilde{b}Y)/\sqrt{1 - \|\boldsymbol{\rho}\|_2^2}\big)_-^2\Big] - \gamma \\
&\le \min_{\breve{b} \in \mathbb{R}, \mathbf{t} \in \mathbb{R}^r} \mathbb{E}\Big[\big(\sqrt{1 + \|\mathbf{t}\|_2^2}\,G + E_Y^T \mathbf{V} \mathbf{S} \mathbf{t} + \breve{b}Y\big)_-^2\Big] - \gamma = \gamma_\star - \gamma < 0,
\end{aligned}
$$

where to get the penultimate inequality we used the change of variables $\mathbf{t} = \boldsymbol{\rho}/\sqrt{1 - \|\boldsymbol{\rho}\|_2^2}$ and $\breve{b} = \tilde{b}/\sqrt{1 - \|\boldsymbol{\rho}\|_2^2}$. Also, in the last line above, we used the definition of the phase-transition threshold $\gamma_\star$ in Equation (64) and the theorem's assumption that $\gamma > \gamma_\star$ (aka separable regime).

We note that similar uniqueness argument was presented in [15] for the special case of antipodal means, *no* intercept and $\delta = 1$.

F.4 Antipodal means and non-isotropic data

Antipodal means. In the special case of antipodal means of equal energy $\boldsymbol{\mu}_+ = -\boldsymbol{\mu}_- = \boldsymbol{\mu}$ with $s := \|\boldsymbol{\mu}\|_2$, the formulas of Theorem 2 simplify as we have $r = 1$ with $\mathbf{S} = s\sqrt{2}$ and $\mathbf{V} = [1/\sqrt{2}, -1\sqrt{2}]^T$. Now, the function η_δ can be written as $\mathbb{E}\big[(G + \tilde{\rho}s + \frac{\tilde{b}}{q}Y - \frac{1}{q}\Delta_Y)_-^2\big] - \big(1 - \tilde{\rho}^2\big)\gamma$. The asymptotic performance of SVM for this special geometry of the means has been recently studied in [15, 47]. We extend this to the CS-SVM classifier, to general means for the two classes and to $\boldsymbol{\Sigma} \neq \mathbf{I}$.

Non-isotropic data. We show how Theorem 2 for the isotropic case can still be applied in the general case $\boldsymbol{\Sigma} \neq \mathbf{I}$. Assume $\boldsymbol{\Sigma} > 0$. Write $\mathbf{x}_i = y_i \boldsymbol{\mu}_{y_i} + \boldsymbol{\Sigma}^{1/2} \mathbf{h}_i$ for $\mathbf{h}_i \sim \mathcal{N}(0, \mathbf{I}_d)$. Consider whitened features $\mathbf{z}_i := \boldsymbol{\Sigma}^{-1/2} \mathbf{x}_i = y_i \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\mu}_{y_i} + \mathbf{h}_i$ and let

$$
(\hat{\mathbf{w}}, \hat{b}) = \arg\min_{\mathbf{w}, b} \frac{1}{n} \sum_{i \in [n]} \ell(y_i(\mathbf{x}_i^T \mathbf{w} + b)),
$$

$$
(\hat{\mathbf{v}}, \hat{c}) = \arg\min_{\mathbf{v}, c} \frac{1}{n} \sum_{i \in [n]} \ell(y_i(\mathbf{z}_i^T \mathbf{v} + c)).
$$

Clearly, $\hat{\mathbf{w}} = \boldsymbol{\Sigma}^{-1/2} \hat{\mathbf{v}}$ and $\hat{b} = \hat{c}$. Thus,

$$
\begin{aligned}
\mathcal{R}_+\big((\hat{\mathbf{w}}, \hat{b})\big) = \mathbb{P}\{(\mathbf{x}^T \hat{\mathbf{w}} + \hat{b}) < 0 \,|\, y = +1\} &= \mathbb{P}\{\boldsymbol{\mu}_+^T \hat{\mathbf{w}} + \hat{\mathbf{w}}^T \boldsymbol{\Sigma}^{1/2} \mathbf{h} + \hat{b} < 0\} = Q\left(\frac{\boldsymbol{\mu}_+^T \hat{\mathbf{w}} + \hat{b}}{\|\boldsymbol{\Sigma}^{1/2} \hat{\mathbf{w}}\|_2}\right) \\
&= Q\Big(\frac{\boldsymbol{\mu}_+^T \boldsymbol{\Sigma}^{-1/2} \hat{\mathbf{v}} + \hat{c}}{\|\hat{\mathbf{v}}\|_2}\Big) = \mathbb{P}\{(\mathbf{z}^T \hat{\mathbf{v}} + \hat{c}) < 0 \,|\, y = +1\} \\
&= \mathcal{R}_+\big((\hat{\mathbf{v}}, \hat{c})\big)
\end{aligned}
$$

Similar derivation holds for $\mathcal{R}_-$. Thus, we can just apply Theorem 2 for $\mathbf{S}, \mathbf{V}$ given by the eigendecomposition of the new Grammian $\mathbf{M}^T \boldsymbol{\Sigma}^{-1} \mathbf{M}$.

F.5 Phase transition of CS-SVM

Here, we present a formula for the threshold $\gamma_\star$ such that the CS-SVM of (4) is feasible (resp., infeasible) with overwhelming probability provided that $\gamma > \gamma_\star$ (resp., $\gamma < \gamma_\star$). The
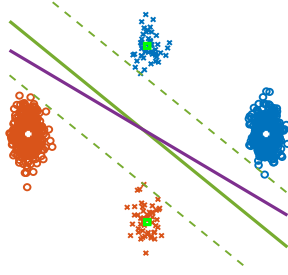
Figure 17: Visualizing the Gaussian mixture model of Section 4 with $K = 2$ imbalanced groups in the two-dimensional space ($d = 2$). Different colors (resp., markers) correspond to different class (resp., group) membership. Examples in the minority group correspond to cross markers ($\times$). The means of the majority / minority groups are depicted in white / green markers. The purple line illustrates the group-sensitive SVM (GS-SVM) classifier that forces larger margin to the minority group examples in relation to standard SVM in green.

first observation is that the phase-transition threshold $\gamma_\star$ of feasibility of the CS-SVM is the same as the threshold of feasibility of the standard SVM for the same model; see Section D.1.3. Then, the desired result follows [35] who very recently established separability phase-transitions for the more general multiclass Gaussian mixture model

Proposition 1 ([35]). *Consider the same data model and notation as in Theorem 2 and define the event*

$$\mathcal{E}_{\mathrm{sep},n} := \left\{ \exists (\mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R} \quad s.t. \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad \forall i \in [n] \right\}.$$

Define threshold $\gamma_\star := \gamma_\star(\mathbf{V}, \mathbf{S}, \pi)$ as follows:

$$\gamma_\star := \min_{\mathbf{t} \in \mathbb{R}^r, b \in \mathbb{R}} \mathbb{E}\left[\left(\sqrt{1 + \|\mathbf{t}\|_2^2}\, G + E_Y^T \mathbf{V} \mathbf{S} \mathbf{t} - bY \right)_-^2 \right]. \tag{64}$$

Then, the following hold:

$$\gamma > \gamma_\star \Rightarrow \lim_{n \to \infty} \mathbb{P}(\mathcal{E}_{\mathrm{sep},n}) = 1 \quad and \quad \gamma < \gamma_\star \Rightarrow \lim_{n \to \infty} \mathbb{P}(\mathcal{E}_{\mathrm{sep},n}) = 0.$$

In words, the data are linearly separable (with overwhelming probability) if and only if $\gamma > \gamma_\star$. Furthermore, if this condition holds, then CS-SVM is feasible (with overwhelming probability) for any value of $\delta > 0$.

G   Asymptotic analysis of GS-SVM

In Theorem 2 we derived the asymptotic generalization performance of CS-SVM under the Gaussian mixture data model. Here, we state the counterpart result for GS-SVM with an appropriate Gaussian mixture data model with group imbalances, which we repeat here for convenience.

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 $\boldsymbol{\Sigma}_{y,g}$, that is, $\mathbf{x} | (y, g) \sim \mathcal{N}(\boldsymbol{\mu}_{y,g}, \boldsymbol{\Sigma}_{y,g})$. We focus on two groups $K = 2$ with $p_{+,1} = p_{-,1} = p < 1 - p = p_{+,2} = p_{-,2}, j = 1, 2$ and $\mathbf{x} | (y, g) \sim \mathcal{N}(y \boldsymbol{\mu}_g, \sigma_g \mathbf{I}_d)$, for σ_1^2, σ_2^2 the noise variances of the minority and the majority groups, respectively. As before, 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 $h(\mathbf{x}) = \mathbf{x}$.

43

Learning regime. Again, as in Theorem 2, we focus on a regime where training data are linearly separable. Specifically, there exists threshold $\widetilde{\gamma}_\star := \widetilde{\gamma}_\star(\mathbf{V}, \mathbf{S}, \pi, p) \leq 1/2$, such that GMM data with groups are linearly separable with probability approaching one provided that $\gamma > \widetilde{\gamma}_\star$ (see Section G.2). We assume $\gamma > \widetilde{\gamma}_\star$, so that GS-SVM is feasible with probability approaching 1.

Although similar in nature, the result below differs to Theorem 2 since now each class itself is a Gaussian mixture.

Theorem 7 (Sharp asymptotics of GS-SVM)**.** *Consider the GMM with feature distribution and priors as specified in the 'Data model' above. Fix $\delta > 0$ (corresponding to group VS-loss with $\Delta_{y,g} = \Delta_g, g = 1, 2$ such that $\delta = \Delta_2/\Delta_1$). Define $G, Y, S, \widetilde{\Delta}_S, \Sigma_S \in \mathbb{R}$, and $\widetilde{E}_S \in \mathbb{R}^{2 \times 1}$ as follows: $G \sim \mathcal{N}(0, 1)$; Y is a symmetric Bernoulli with $\mathbb{P}\{Y = +1\} = \pi$; S takes values 1 or 2 with probabilities p and $1 - p$, respectively; $\widetilde{E}_S = \mathbf{e}_1 \mathbb{1}[S = 1] + \mathbf{e}_2 \mathbb{1}[S = 2]$; $\widetilde{\Delta}_S = \delta \cdot \mathbb{1}[S = 1] + 1 \cdot \mathbb{1}[S = 2]$ and $\Sigma_S = \sigma_1 \mathbb{1}[S = 1] + \sigma_2 \mathbb{1}[S = 2]$. With these define function $\widetilde{\eta}_\delta : \mathbb{R}_{\geq 0} \times \mathcal{S}^r \times \mathbb{R} \to \mathbb{R}$ as*

$$\widetilde{\eta}_\delta(q, \boldsymbol{\rho}, b) := \mathbb{E}\big(G + \Sigma_S^{-1} \widetilde{E}_S^T \mathbf{V} \mathbf{S} \boldsymbol{\rho} + \frac{b \Sigma_S^{-1} Y - \Sigma_S^{-1} \widetilde{\Delta}_S}{q}\big)_-^2 - (1 - \|\boldsymbol{\rho}\|_2^2)\gamma.$$

Let $(\widetilde{q}_\delta, \widetilde{\boldsymbol{\rho}}_\delta, \widetilde{b}_\delta)$ be the unique triplet satisfying (42) but with η_δ replaced with the function $\widetilde{\eta}_\delta$ above. Then, in the limit of $n, d \to \infty$ with $d/n = \gamma > \widetilde{\gamma}_\star$ it holds for $i = 1, 2$ that $\mathcal{R}_{\pm,i} \xrightarrow{P} Q\big(\mathbf{e}_i^T \mathbf{V} \mathbf{S} \widetilde{\boldsymbol{\rho}}_\delta \pm \widetilde{b}_\delta/\widetilde{q}_\delta\big)$. In particular, $\mathcal{R}_{\mathrm{deo}} \xrightarrow{P} Q\big(\mathbf{e}_1^T \mathbf{V} \mathbf{S} \widetilde{\boldsymbol{\rho}}_\delta + \widetilde{b}_\delta/\widetilde{q}_\delta\big) - Q\big(\mathbf{e}_2^T \mathbf{V} \mathbf{S} \widetilde{\boldsymbol{\rho}}_\delta + \widetilde{b}_\delta/\widetilde{q}_\delta\big)$.

G.1 Proof of Theorem 7

The proof of Theorem 7 also relies on the CGMT framework and is very similar to the proof of Theorem 2. To avoid repetitions, we only present the part that is different. As we will show the PO is slightly different as now we are dealing with a classification between mixtures of mixtures of Gaussians. We will derive the new AO and will simplify it to a point from where the same steps as in Section F.3 can be followed mutatis mutandis.

Let $(\hat{\mathbf{w}}, \hat{b})$ be solution pair to the GS-SVM for some fixed parameter $\delta > 0$, which we rewrite here expressing the constraints in matrix form:

$$\min_{\mathbf{w}, b} \|\mathbf{w}\|_2 \quad \text{sub. to} \quad \begin{cases} y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq \delta, \ g_i = 1 \\ y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \ g_i = 2 \end{cases}, \ i \in [n] \quad = \quad \min_{\mathbf{w}, b} \|\mathbf{w}\|_2 \quad \text{sub. to} \quad \mathbf{D_y}\big(\mathbf{X}\mathbf{w} + b\mathbf{1}_n\big) \geq \boldsymbol{\delta}_{\mathbf{g}},$$

(65)

where we have used the notation

$$\mathbf{X}^T = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix}, \ \mathbf{y} = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^T,$$

$$\mathbf{D_y} = \mathrm{diag}(\mathbf{y}) \ \text{ and } \ \boldsymbol{\delta}_{\mathbf{g}} = \begin{bmatrix} \delta\mathbb{1}[g_1 = 1] + \mathbb{1}[g_1 = 2] & \cdots & \delta\mathbb{1}[g_n = 1] + \mathbb{1}[g_n = 2] \end{bmatrix}^{\mathrm{T}}.$$

We further need to define the following one-hot-encoding for group membership:

$$\mathbf{g}_i = \mathbf{e}_1 \mathbb{1}[g_i = 1] + \mathbf{e}_2 \mathbb{1}[g_i = 2], \quad \text{and} \quad \mathbf{G}_{n \times 2}^T = \begin{bmatrix} \mathbf{g}_1 & \cdots & \mathbf{g}_n \end{bmatrix}.$$

where recall that $\mathbf{e}_1, \mathbf{e}_2$ are standard basis vectors in $\mathbb{R}^2$. Finally, let

$$\mathbf{D}_\sigma = \mathrm{diag}\big(\begin{bmatrix} \sigma_{g_1} & \cdots & \sigma_{g_n} \end{bmatrix}\big).$$

With these, notice for later use that under our model, $\mathbf{x}_i = y_i \boldsymbol{\mu}_{g_i} + \sigma_{g_i} \mathbf{z}_i = y_i \mathbf{M} \mathbf{g}_i + \sigma_{g_i} \mathbf{z}_i$, $\mathbf{z}_i \sim \mathcal{N}(0, 1)$. Thus, in matrix form with $\mathbf{Z}$ having entries $\mathcal{N}(0, 1)$:

$$\mathbf{X} = \mathbf{D_y} \mathbf{G} \mathbf{M}^T + \mathbf{D}_\sigma \mathbf{Z}.$$

(66)

As usual, we express the GS-SVM program in a min-max form to bring it in the form of the PO as follows:

$$\min_{\mathbf{w}, b} \max_{\mathbf{u} \leq 0} \frac{1}{2} \|\mathbf{w}\|_2^2 + \mathbf{u}^T \mathbf{D_y} \mathbf{X} \mathbf{w} + b(\mathbf{u}^T \mathbf{D_y} \mathbf{1}_n) - \mathbf{u}^T \boldsymbol{\delta}_{\mathbf{g}}$$

$$= \min_{\mathbf{w}, b} \max_{\mathbf{u} \leq 0} \frac{1}{2} \|\mathbf{w}\|_2^2 + \mathbf{u}^T \mathbf{D_y} \mathbf{D}_\sigma \mathbf{Z} \mathbf{w} + \mathbf{u}^T \mathbf{G} \mathbf{M}^T \mathbf{w} + b(\mathbf{u}^T \mathbf{D_y} \mathbf{1}_n) - \mathbf{u}^T \boldsymbol{\delta}_{\mathbf{g}}.$$

(67)

where in the last line we used (66) and $\mathbf{D_y D_y = I_n}$. We immediately recognize that the last optimization is in the form of a PO and the corresponding AO is as follows:

$$\min_{\mathbf{w},b} \max_{\mathbf{u}\le 0} \frac{1}{2}\|\mathbf{w}\|_2^2 + \|\mathbf{w}\|_2 \mathbf{u}^T \mathbf{D_y D_\sigma h}_n + \|\mathbf{D_y D_\sigma u}\|_2 \mathbf{h}_d^T \mathbf{w} + \mathbf{u}^T \mathbf{GM}^T \mathbf{w} + b(\mathbf{u}^T \mathbf{D_y 1}_n) - \mathbf{u}^T \boldsymbol{\delta_g}. \tag{68}$$

where $\mathbf{h}_n \sim \mathcal{N}(0,\mathbf{I}_n)$ and $\mathbf{h}_d \sim \mathcal{N}(0,\mathbf{I}_d)$.

As in Section F.3 we consider the one-sided constrained AO in (68). Towards simplifying this auxiliary optimization, note that $\mathbf{D}_y \mathbf{h}_n \sim \mathbf{h}_n$ by rotational invariance of the Gaussian measure. Also, $\|\mathbf{D_y D_\sigma u}\|_2 = \|\mathbf{D_\sigma u}\|_2$. Thus, we can express the AO in the following more convenient form:

$$\min_{\|\mathbf{w}\|_2^2+b^2\le R} \max_{\mathbf{u}\le 0} \frac{1}{2}\|\mathbf{w}\|_2^2 + \|\mathbf{w}\|_2 \mathbf{u}^T \mathbf{D_\sigma h}_n + \|\mathbf{D_\sigma u}\|_2 \mathbf{h}_d^T \mathbf{w} + \mathbf{u}^T \mathbf{GM}^T \mathbf{w} + b(\mathbf{u}^T \mathbf{D_y 1}_n) - \mathbf{u}^T \boldsymbol{\delta_g}$$

$$= \min_{\|\mathbf{w}\|_2^2+b^2\le R} \max_{\mathbf{v}\le 0} \frac{1}{2}\|\mathbf{w}\|_2^2 + \|\mathbf{w}\|_2 \mathbf{v}^T \mathbf{h}_n + \|\mathbf{v}\|_2 \mathbf{h}_d^T \mathbf{w} + \mathbf{v}^T \mathbf{D}_\sigma^{-1} \mathbf{GM}^T \mathbf{w} + b(\mathbf{v}^T \mathbf{D}_\sigma^{-1} \mathbf{D_y 1}_n) - \mathbf{v}^T \mathbf{D}_\sigma^{-1} \boldsymbol{\delta_g},$$

where in the second line we performed the change of variables $\mathbf{v} \leftrightarrow \mathbf{D}_\sigma \mathbf{u}$ and used positivity of the diagonal entries of $\mathbf{D}_\sigma$ to find that $\mathbf{u} \le 0 \iff \mathbf{v} \le 0$.

Notice that the optimization in the last line above is very similar to the AO (54) in Section F.3. Following analogous steps, omitted here for brevity, we obtain the following scalarized AO:

$$\min_{\substack{q\ge\sqrt{\mu_1^2+\mu_2^2+\alpha^2}\\q^2+b^2\le R}} \frac{1}{2}q^2 \tag{69}$$

$$\text{sub. to} \quad \frac{1}{\sqrt{n}}\left\|\left(q\mathbf{h}_n + \mathbf{D}_\sigma^{-1}\mathbf{GVS}\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix} + b\,\mathbf{D}_\sigma^{-1}\mathbf{D}_y\mathbf{1}_n - \mathbf{D}_\sigma^{-1}\boldsymbol{\delta_g}\,\right)_-\right\|_2$$

$$- \mu_1 \frac{\mathbf{h}_d^T \mathbf{u}_1}{\sqrt{n}} - \mu_2 \frac{\mathbf{h}_d^T \mathbf{u}_2}{\sqrt{n}} - \alpha\frac{\|\mathbf{h}_d^T \mathbf{U}^\perp\|_2}{\sqrt{n}} \le 0.$$

where as in Section F.3 we have decomposed the matrix of means $\mathbf{M} = \mathbf{USV}^T$ and μ_1,μ_2,α above represent $\mathbf{u}_1^T\mathbf{w}$, $\mathbf{u}_1^T\mathbf{w}$ and $\|\mathbf{w}^\perp\|_2$. Now, by law of large numbers, notice that for fixed (q,μ_1,μ_2,α,b), the functional in the constraint above converges in probability to

$$\bar{L}(q,\mu_1,\mu_2,\alpha,b) := \sqrt{\mathbb{E}\Big(qG + \Sigma_S^{-1}\widetilde{E}_S^T\mathbf{VS}\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix} + b\Sigma_S^{-1}Y - \Sigma_S^{-1}\widetilde{\Delta}_S\Big)_-^2} - \alpha\sqrt{\gamma}, \tag{70}$$

where the random variables $G,\widetilde{E}_S,Y,\widetilde{\Delta}_S$ and Σ_S are as in the statement of the theorem. Thus, the deterministic equivalent (high-dimensional limit) of the AO expressed in variables $\boldsymbol{\rho} = \begin{bmatrix}\boldsymbol{\rho}_1\\\boldsymbol{\rho}_2\end{bmatrix} := \begin{bmatrix}\mu_1/q\\\mu_2/q\end{bmatrix}$ becomes (cf. Eqn. (58)):

$$\min_{q^2+b^2\le R,q>0,\|\boldsymbol{\rho}\|_2\le 1} \frac{1}{2}q^2 \tag{71}$$

$$\text{sub. to} \quad \mathbb{E}\Big(G + \Sigma_S^{-1}\widetilde{E}_S^T\mathbf{VS}\boldsymbol{\rho} + \frac{b\Sigma_S^{-1}Y - \Sigma_S^{-1}\widetilde{\Delta}_S}{q}\Big)_-^2 \le \big(1 - \|\boldsymbol{\rho}\|_2^2\big)\gamma.$$

Now, recall the definition of the function $\widetilde{\eta}_\delta$ in the statement of the theorem and observe that the constraint above is nothing but

$$\widetilde{\eta}_\delta(q,\boldsymbol{\rho},b) \le 0.$$

Thus, (71) becomes

$$\min\left\{q^2 \;\Big|\; 0\le q\le\sqrt{R} \quad\text{and}\quad \min_{b^2\le R-q^2,\|\boldsymbol{\rho}\|_2\le 1}\widetilde{\eta}_\delta(q,\boldsymbol{\rho},b)\le 0\right\}. \tag{72}$$

The remaining steps of the proof are very similar to those in Section F.3 and are omitted.

G.2 Phase transition of GS-SVM

The phase-transition threshold $\widetilde{\gamma}_\star$ of feasibility of the GS-SVM is the same as the threshold of feasibility of the standard SVM for the same model (see Section D.1.2). But, the feasibility threshold of SVM under the group GMM with $K = 2$ groups is different from that of Section F.5 for $K = 1$, since now each class is itself a mixture of Gaussians. We derive the desired result from [35], who recently studied the separability question for the more general case of a multiclass mixture of mixtures of Gaussians.

Proposition 2. *Consider the same data model and notation as in Theorem 7 and consider the event*

$$\mathcal{E}_{\mathrm{sep},n} := \left\{ \exists (\mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R} \quad s.t. \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad \forall i \in [n] \right\}.$$

Define threshold $\gamma_\star := \gamma_\star(\mathbf{V}, \mathbf{S}, \pi)$ as follows:

$$\widetilde{\gamma}_\star := \min_{\mathbf{t} \in \mathbb{R}^r, b \in \mathbb{R}} \mathbb{E}\left[\left(\sqrt{1 + \|\mathbf{t}\|_2^2}\, G + \widetilde{E}_S^T \mathbf{V} \mathbf{S} \mathbf{t} - bY \right)_-^2 \right]. \tag{73}$$

Then, the following hold:

$$\gamma > \widetilde{\gamma}_\star \Rightarrow \lim_{n \to \infty} \mathbb{P}(\mathcal{E}_{\mathrm{sep},n}) = 1 \qquad and \qquad \gamma < \widetilde{\gamma}_\star \Rightarrow \lim_{n \to \infty} \mathbb{P}(\mathcal{E}_{\mathrm{sep},n}) = 0.$$

In words, the data are linearly separable with overwhelming probability if and only if $\gamma > \widetilde{\gamma}_\star$. Furthermore, if this condition holds, then GS-SVM is feasible with overwhelming probability for any value of $\delta > 0$.

Checklist

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]