Label-Imbalanced and Group-Sensitive Classification under Overparameterization

Ganesh Ramachandra Kini[1], Orestis Paraskevas[1], Samet Oymak[2], Christos Thrampoulidis[3,1]

[1]Department of Electrical and Computer Engineering, UC Santa Barbara
[2]Department of Electrical and Computer Engineering, UC Riverside
[3]Department of Electrical and Computer Engineering, University of British Columbia

March 2, 2021

Abstract

Label-imbalanced and group-sensitive classification seeks to appropriately modify standard training algorithms to optimize relevant metrics such as *balanced error* and/or *equal opportunity*. For label imbalances, recent works have proposed a *logit-adjusted loss* modification to standard empirical risk minimization. We show that this might be ineffective in general and, in particular so, in the overparameterized regime where training continues in the *zero training-error regime*. Specifically for binary linear classification of a separable dataset, we show that the modified loss converges to the max-margin SVM classifier despite the logit adjustment. Instead, we propose a more general *vector-scaling loss* that directly relates to the *cost-sensitive SVM* (CS-SVM), thus favoring larger margin to the minority class. Through an insightful sharp asymptotic analysis for a Gaussian-mixtures data model, we demonstrate the efficacy of CS-SVM in balancing the errors of the minority/majority classes. Our analysis also leads to a simple strategy for optimally tuning the involved margin-ratio parameter. Then, we show how our results extend naturally to binary classification with sensitive groups, thus treating the two common types of imbalances (label/group) in a unifying way. We corroborate our theoretical findings with numerical experiments on both synthetic and real-world datasets.

1 Introduction

1.1 Motivation

Equitable learning in the presence of imbalances in the data is a rather classical problem in the ML community [HM19]. However, it has seen a surge of interest over the past few years as we aspire to use ML algorithms to create automated decision rules in increasingly more applications that directly involve people [BS16]. Two common types of imbalances that have attracted particular attention are those present in *label-imbalanced* and *group-sensitive* classification. In the first type, examples from a target class are heavily outnumbered by examples from the rest of the classes. The standard metric of average classification error is insensitive to such *long-tail label distributions* and better classical alternatives exist, e.g. see [JK19]. In our paper, we focus on the notion of *balanced error*. In the second type, the broad goal is to ensure fairness with respect to a protected underrepresented group (e.g. gender or race). There are several intuitive notions of fairness and [KMR16, FSV16] showed that there is no universal fairness metric. Thus, here too, several fairness metrics have been proposed, e.g. [CKP09, HPS16, ZVGRG17, WM19]. In our paper, we focus on *Equal Opportunity* which favors same true positive rates across groups [HPS16].

Methods to cope with class/group imbalances are broadly categorized into data-level and algorithm-level ones. Of interest to us are *cost-sensitive methods* within the latter category and specifically approaches that modify the loss function during training to account for varying class/group penalties, e.g. [DOBD+18, CWG+19] and references therein. For label-imbalanced classification, [MJR+20]

considered the following *logit-adjusted loss* as a modification to cross-entropy:

$$\ell(y, f(\mathbf{x})) = \omega_y \log(1 + \sum_{y' \neq y} e^{\iota_{yy'}} \cdot e^{(f_{y'}(\mathbf{x}) - f_y(\mathbf{x}))}), \tag{1}$$

In this paper, we ask: *How does the classifier trained by optimizing the loss in (1) behave in the overparameterized regime where training continues in the zero training-error regime? How effective is the classifier in terms of balanced error in label-imbalanced settings?* Finally,

Can we design a provably better loss for this regime?

We also study related questions for group-sensitive classification using the same set of tools. Specifically, the loss in (1) admits a natural modification that makes it also relevant to this setting. We ask: *What are potential benefits of a more principled loss function in overparameterized group-sensitive classification in terms of Equal Opportunity?*

1.2 Contributions

This work focuses on binary label-imbalanced and group sensitive classification in the overparameterized regime and makes multiple key contributions as summarized below.

• **A new *vector-scaling* loss:** We propose the following *vector-scaling* loss (*VS-loss*, in short):

$$\begin{aligned} \ell(y, f(\mathbf{x})) &= \omega_y \log(1 + \sum_{y' \neq y} e^{\iota_{yy'}} e^{\Delta_y (f_{y'}(\mathbf{x}) - f_y(\mathbf{x}))}) \\ &= -\omega_y \log \frac{e^{\Delta_y f_y(\mathbf{x}) + \iota_y}}{\sum_{y' \in [k]} e^{\Delta_y f_{y'}(\mathbf{x}) + \iota_{y'}}}, \end{aligned} \tag{2}$$

as a modification of standard cross-entropy loss appropriate for imbalanced k-class datasets. In addition to the weights ω_y and to the label-dependent offset parameters $\iota_{yy'} = \iota_{y'} - \iota_y$ in (1), the VS-loss in (2) introduces *multiplicative* scaling factors $\Delta_y > 0$ to the logits (in red for emphasis). Both theoretically and empirically we demonstrate the beneficial role of these new parameters when training continues in the zero training-error regime.

• **Connection to cost-sensitive SVM:** We focus on binary classification. For linear predictors and separable datasets, we argue that optimizing the loss in (1) with gradient descent leads to a classifier whose direction converges to the max-margin SVM solution *irrespective* of the choice of the parameters ω_y and ι_y. Instead, the VS loss leads to the solution of another old friend: the *cost-sensitive SVM (CS-SVM)* with margin-ratio parameter $\delta = \Delta_{-1}/\Delta_{+1}$; see (7).

• **CS-SVM through a modern lens:** For an insightful Gaussian mixtures model (GMM), we present formulae that sharply predict the classification and balanced errors of CS-SVM (and thus, of the VS-loss as well) in a *high-dimensional* separable regime. Our formulae are explicit in terms of data geometry, class priors, parameterization ratio and tuning parameter δ. Additionally, we identify a key structural property of CS-SVM that together with our asymptotic theory lead to an explicit formula for the optimal margin ratio $\delta_\star$ that minimizes the balanced error. For example, we show that $\delta_\star$ not only depends on the class priors, but is also sensitive to the parameterization ratio.

• **Group-sensitive SVM:** We propose natural modifications to (2) and to CS-SVM for classification in the presence of sensitive groups. We then extend the GMM to the capture the presence of imbalanced groups, and for binary labels we develop a sharp analysis of our algorithms under sufficient overparameterization.

• **Numerical experiments:** We present numerical experiments that corroborate our findings above. Also, using our sharp analysis we study key tradeoffs between balanced error / equal opportunity and misclassification error.

1.3 Connections to related literature

Logit-adjusted loss: The idea of the *weights* ω_y is rather old [XM89], but becomes ineffective under overparameterization [BL19]. This deficiency together with the trend for overparameterized models has led to the idea of the pairwise label-based offset parameters $\iota_{yy'}$ as seen in (1). These offset

parameters enter (1) in an *additive* way with respect to the logits $f_y(\mathbf{x})$. For example, the following choices for $\iota_{yy'}$ have been proposed: $1/\mathbb{P}(y)^{1/4}$ [CWG+19], $\log \mathbb{P}(y')$ [TWL+20] and $\log \frac{\mathbb{P}(y')}{\mathbb{P}(y)}$ [MJR+20]. The latter is shown by [MJR+20] to lead to a classifier that is Fisher consistent. However, Fisher consistency is only relevant in the large sample size limit. Perhaps surprisingly, we argue that *the additive offsets $\iota_{yy'}$ might be ineffective when data are separable and propose the more general loss (2).*

Relation to vector-scaling calibration: The multiplicative weighting in our loss is reminiscent of the vector scaling (VS) [GPSW17], which inspired our naming. VS is a *post-hoc procedure* that modifying the logits $\mathbf{v}$ of a neural net *after training* via $\mathbf{v} \to \boldsymbol{\Delta} \odot \mathbf{v} + \boldsymbol{\iota}$ where $\odot$ is the Hadamard product. Related to us, [ZCO20] shows that VS can improve calibration for imbalanced classes. As important distinctions of our VS-loss compared to VS calibration, note that the multiplicative scalings in (2): (i) are part of the loss, thus they directly affect training; (ii) are applied to the margins and *not* individually to the logits. Finally, (inspired by the more general matrix-scaled calibration [GPSW17]), we can define a matrix-scaled loss modifying the margins as $\Delta_{yy'}(f_{y'}(\mathbf{x}) - f_y(\mathbf{x})) + i_{yy'}$. Exploring potential benefits of this is left for future work.

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

Cost-sensitive SVM: [MSV10] arrived to CS-SVM by properly extending the SVM hinge-loss to guarantee Fisher consistency. In Section 5.1 we give a different interpretation to CS-SVM by connecting to our VS-loss. We also interpret CS-SVM as "post-hoc weight normalization" to SVM.

2 Problem setup

2.1 Data models

Let $\{(\mathbf{x}_i, g_i, y_i)\}_{i=1}^n$ be a sequence of n i.i.d. training 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, $\mathcal{Y} \in \{\pm 1\}$ the set of binary labels and $\mathcal{G} = [K]$ refers to the group membership among $K \geq 1$ groups.

Gaussian mixtures model (GMM). For the purpose of our sharp asymptotic analysis and some of our synthetic experiments, we further specify the following Gaussian-mixture type generative model for the data distribution $\mathcal{D}$. For the label $y \in \{\pm 1\}$, we assume

$$\mathbb{P}\{y = +1\} = 1 - \mathbb{P}\{y = -1\} = \pi \in (0, 1).$$

The group membership is decided conditionally on the label such that for all $j \in [K]$,

$$\mathbb{P}\{g = j | y = +1\} = p_{+,j} \quad \text{and} \quad \mathbb{P}\{g = j | y = -1\} = p_{-,j},$$

with $\sum_{j\in[K]} p_{+,j} = \sum_{j\in[K]} p_{-,j} = 1$. Finally, within each class, each group is associated with a mean vector $\boldsymbol{\mu}_{\pm 1,j} \in \mathbb{R}^d$ and the conditional of the feature $\mathbf{x}$ given its label y and group g is a multivariate Gaussian of mean $\boldsymbol{\mu}_{y,g}$ and covariance $\boldsymbol{\Sigma}$, i.e.,

$$\mathbf{x} | (y, g) \sim \mathcal{N}(\boldsymbol{\mu}_{y,g}, \boldsymbol{\Sigma}). \tag{3}$$

For *label-imbalances*, we assume all examples belong to the same group ($K = 1$), and an imbalanced setting in which $\pi \ll 1 - \pi$. For *imbalances with respect to group-membership*, we focus on two groups ($K = 2$) with equal priors for the positive and negative class labels, i.e. $\pi = \frac{1}{2}, p_{+,j} = p_{-,j} = p_j$, $j = 1, 2$ and imbalance $p := p_1 \ll p_2 = 1 - p$. See Figure 1 for a graphical illustration in $\mathbb{R}^2$.

2.2 Balanced error and Equal Opportunity measures

We consider linear classifiers parameterized by a decision hyperplane $\mathbf{w} \in \mathbb{R}^d$ and an intercept/offset $b \in \mathbb{R}$. Given a new sample $\mathbf{x}$ we decide class membership $\hat{y} \in \{\pm 1\}$ as $\hat{y} = \text{sign}(\mathbf{w}^T \mathbf{x} + b)$. The (standard)

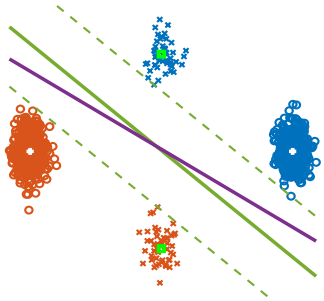


Figure 1: Visualizing the Gaussian mixture model of Section 2.1 with $K = 2$ imbalanced groups in the two-dimensional space ($d = 2$). Different colors (resp., markers) correspond to different class (resp., group) membership. Examples in the minority group correspond to cross markers (×). The means of the majority / minority groups are depicted in white / green markers. The purple line illustrates our group-sensitive SVM classifier (8) that forces larger margin to the minority group examples in relation to standard SVM in green.

risk or **misclassification error** of such a classifier is

$$\mathcal{R} := \mathcal{R}\big((\mathbf{w}, b)\big) = \mathbb{P}\left\{\hat{y} \neq y\right\}.$$

This can be written as $\mathcal{R} = \pi\mathcal{R}_+ + (1-\pi)\mathcal{R}_-$ in terms of the **class-conditional risks**

$$\mathcal{R}_\pm = \mathbb{P}\left\{\hat{y} \neq y \,|y = \pm 1\right\},$$

and as $\mathcal{R} = \pi \sum_{j\in[K]} p_{+,j}\mathcal{R}_{+,j} + (1-\pi)\sum_{j\in[K]} p_{-,j}\mathcal{R}_{-,j}$ in terms of the **group-conditional risks**

$$\mathcal{R}_{\pm,j} = \mathbb{P}\left\{\hat{y} \neq y \,|y = \pm 1, g = j\right\}, \quad j \in [K].$$

The misclassification error is a rather poor measure of performance in class/group imbalances. Here, we focus on the following two popular alternatives for label- and group- imbalances, respectively. The **balanced error** simply averages the conditional risks of the two classes:

$$\mathcal{R}_{\text{bal}} := (\mathcal{R}_+ + \mathcal{R}_-)/2.$$

Assuming $K = 2$, the constraint of Equal Opportunity (EO) is satisfied [HPS16] if $\mathcal{R}_{+,1} = \mathcal{R}_{+,2}$. More generally, we consider the **difference of equal opportunity (DEO)**

$$\mathcal{R}_{\text{deo}} := \mathcal{R}_{+,1} - \mathcal{R}_{+,2}.$$

2.3 Overparameterization

We assume that the learning model $(\mathbf{w}, b)$ is overparameterized enough to perfectly fit the data such that the training error $\mathcal{R}_{\text{train}} = \frac{1}{n}\sum_{i\in[n]} \mathbb{1}[\text{sign}(\mathbf{w}^{\text{T}}\mathbf{x}_{\text{i}} + \text{b}) \neq \text{y}_{\text{i}}]$ is zero. Equivalently, the training data are *linearly separable*, i.e.

$$\exists (\mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R} \;\text{ s.t. }\; y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \;\; \forall i \in [n]. \tag{4}$$

Notation. We use $\xrightarrow{P}$ to denote convergence in probability and let $Q(\cdot)$ be the tail distribution function of the standard normal distribution. We define $(x)_- := \min\{x, 0\}$. We denote $\mathbb{1}[\mathcal{E}]$ the indicator function of an event $\mathcal{E}$. We let $\mathcal{S}_2^r$ denote the unit sphere in $\mathbb{R}^r$. Finally, let $[K] := \{1, \ldots, K\}$.

3 Algorithms

Below, we first present a more general version of the VS-loss that also applies to group-sensitive classification and then we specialize it to binary problems, which is the focus of this paper. Also, we recall the CS-SVM and introduce a natural extension to account for imbalanced groups.

VS-loss. For binary classification (2) simplifies to

$$\ell(y, f(\mathbf{x})) = \omega_y \cdot \log\left(1 + e^{\iota_y} \cdot e^{-\Delta_y y f(\mathbf{x})}\right), \tag{5}$$

for hyperparameters $\iota_\pm \in \mathbb{R}, \omega_\pm > 0, \Delta_\pm > 0$. For linear classification we set $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ or $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}+b$. The logit-adjusted loss is a special case of (5) with $\Delta_\pm = 1$. When the goal is to (additionally) ensure fairness with respect to a *sensitive group* within the classes, we can naturally extend the VS-loss by introducing in (2)/(5) parameters $(\Delta_{y,g}, \iota_{y,g}, w_{y,g})$ (rather than (Δ_y, ι_y, w_y)) that depend *both* on the class and group membership (specified by y and g, respectively). For binary datasets, we simply have

$$\ell(y, f(\mathbf{x}), g) = \omega_{y,g} \cdot \log\left(1 + e^{\iota_{y,g}} \cdot e^{-\Delta_{y,g} y f(\mathbf{x})}\right). \tag{6}$$

CS-SVM. The cost-sensitive (hard-margin) SVM (CS-SVM for short) [MSV10] produces such a classifier $(\hat{\mathbf{w}}_\delta, \hat{b}_\delta)$ as follows:

$$\min_{\mathbf{w},b} \|\mathbf{w}\|_2 \text{ sub. to } \begin{cases} \mathbf{w}^T\mathbf{x}_i + b \geq \delta & , y_i = +1 \\ \mathbf{w}^T\mathbf{x}_i + b \leq -1 & , y_i = -1 \end{cases}, i \in [n], \tag{7}$$

for some $\delta \in \mathbb{R}_+$ that denotes the ratio of margins. Note that the "standard form" of hard-margin SVM corresponds to (7) with $\delta = 1$. Onwards, we refer to (7) with $\delta = 1$ simply as SVM, while the acronym CS-SVM is reserved for the general values of δ. The CS-SVM naturally allows tuning $\delta > 1$ (resp. $\delta < 1$) to favor a larger margin $\delta/\|\hat{\mathbf{w}}_\delta\|_2$ for the minority class vs $1/\|\hat{\mathbf{w}}_\delta\|_2$ for the majority class, which is an intuitive means to balance the error between the two. Thus, $\delta \to +\infty$ (resp. $\delta \to 0$) corresponds to the extreme scenario where the decision boundary starts right at the boundary of class $y = -1$ (resp. $y = +1$).

Group-sensitive SVM. For simplicity, we focus on two protected groups $K = 2$ and leave (natural) extensions to future studies. We propose the following group-sensitive version of CS-SVM, which we refer to as (hard-margin) GS-SVM for short:

$$\min_{\mathbf{w},b} \|\mathbf{w}\|_2 \text{ s.t. } \begin{cases} y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq \delta, & g_i = 1 \\ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, & g_i = 2 \end{cases}, \ i \in [n]. \tag{8}$$

By tuning $\delta > 1$, GS-SVM favors larger margin for the sensitive group $g = 1$. Once again, refined versions of (8) are also possible for instances where the classes are also imbalanced themselves by modifying the constraints to $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq \delta_{y_i,1}$ for $g_i = 1$, and, $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq \delta_{y_i,2}$ for $g_i = 2$ for positive $\delta_{1,1}, \delta_{1,2}, \delta_{-1,1}, \delta_{-1,2}$. Both the hard-margin CS-SVM and GS-SVM are feasible if and only if the data are linearly separable. However, we caution that the role of the hyper-parameters in GS-SVM is in general harder to interpret as "margin-ratios" since its constraints may or may not be active depending on the data geometry.

4 Vector-scaling loss: Motivational examples

4.1 Imbalanced classification of GMM

We start with an experiment on synthetic data in Figure 2. We generated a binary Gaussian-mixture dataset of $n = 100$ examples in $\mathbb{R}^{d=300}$ with data means sampled independently from the Gaussian distribution and normalized such that $\|\boldsymbol{\mu}_{+1}\|_2 = 4, \|\boldsymbol{\mu}_{-1}\|_2 = 2$ and 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 $\mathbf{w} \in \mathbb{R}^p$ using only the first p features, i.e. $f(\mathbf{x}) = \mathbf{w}^T\tilde{\mathbf{x}}$ with $\tilde{\mathbf{x}} = \mathbf{x}(1:p) \in \mathbb{R}^p$. This allows us to investigate performance with a varying parameterization [1] ratio $\gamma = p/n$. We train the model $\mathbf{w}$ using loss $\mathcal{L}_n(\mathbf{w}) = \sum_{i=1}^n \ell(y_i, \mathbf{w}^T\tilde{\mathbf{x}}_i)$ with ℓ chosen to be either our proposed VS-loss or the logit-adgusted (LA) loss. Specifically, we train the VS-loss in (5) for $f(\mathbf{x}) = \mathbf{w}^T\tilde{\mathbf{x}}$ and the following choice of parameters:

$$\omega_\pm = 1, \ \iota_\pm = 0 \quad \text{and} \quad \Delta_y = \delta_\star^{-1}\mathbb{1}[y = +1] + \mathbb{1}[y = -1].$$

[1] Simple models like this have been recently used in [HMRT19, BHX19, DKT19, KT20] for analytic studies of double descent [BMM18, BRT19, MM19, NKB+19] in the misclassification error. Although it is not our focus here, Figure 2 reveals that the balanced error of the VS-loss undergoes a similar double descent.

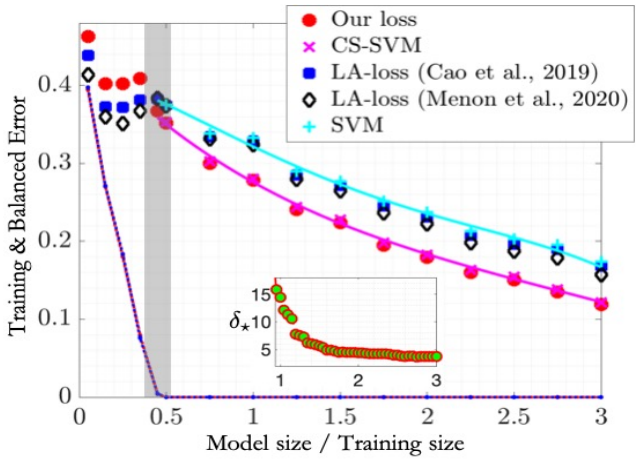


Figure 2: This figure highlights the benefits of our theory-inspired VS-loss (red markers) over the logit-adjusted loss of [CWG+19](blue markers) and [MJR+20](black-markers). We trained a mismatched linear model with varying model size p on a binary Gaussian-mixture dataset of $n = 100$ examples in $\mathbb{R}^{d=300}$. x-axis is the parameterization ratio p/n. The prior of the minority class is set to $\pi = 0.1$. The other details are provided in Sec. 4.1. The shaded region highlights the transition to zero training error: on the right side of it data become linearly separable. In this separable regime, we train additionally using SVM (cyan plus marker) and cost-sensitive SVM (magenta cross). The inset displays the margin-ratio parameter δ that we used to tune the VS-loss and CS-SVM. The solid lines depict theoretical predictions.

Here, $\delta_\star > 0$ was set to the value shown in the inset plot; see explanation later. As discussed, the LA-loss is a special case of the VS-loss with $\Delta_y = 1$. Thus, for training with the LA-loss we used (5) with

$$\Delta_\pm = 1,\ \omega_\pm = 1 \quad \text{and} \quad \iota_y = \pi^{-1/4}\mathbb{1}[y = +1] + (1-\pi)^{-1/4}\mathbb{1}[y = -1],$$

as suggested by [CWG+19] (see blue markers in the figure) and

$$\Delta_\pm = 1,\ \omega_\pm = 1 \quad \text{and} \quad \iota_y = \log\big(\frac{1-\pi}{\pi}\big)\mathbb{1}[y = +1] + \log\big(\frac{\pi}{1-\pi}\big)\mathbb{1}[y = -1],$$

as suggested by [MJR+20] (see black markers in the figure). In all cases, we ran (normalized) gradient descent with a varying learning rate normalized by the gradient of the loss at each iteration (see Appendix A.6 for details). For each value of p, we ran 25 independent experiments and reported averages of the *balanced test error* on a test set of size 10^4 generated from the same distribution as the training set. The reported values are shown in red/blue/black markers. We also plot the (average) training error for each one of the loss. Observe that for all losses the training error is zero for parameterization ratio $\gtrsim 0.45$. The shaded region highlights the transition to the overparameterized regime where the data model size p is large enough to drive the training error to zero (eqv., to make the training data separable).

VS-loss vs LA-loss. The experiment above reveals the following clear message:

Our VS-loss has better balanced-error performance compared to the logit-adjusted loss when both trained to zero training error.

In addition to proposing the VS-loss, we will also give an analytical explanation for this behavior. Our analysis will reveal the crucial role of the multiplicative scaling factors Δ_y in (5). It will also demonstrate that the particular choice of $\iota_\pm$ above is irrelevant in the overparameterized regime: any choice is as good as $\iota_\pm = \pm 1$ (i.e. standard logistic loss) when data are separable. We emphasize that our conclusions hold for the overparameterized regime (i.e. $\mathcal{R}_{\text{train}} = 0$). We see in Figure 2 that when data are not separable, the LA-loss can outperform the VS-loss for the specific choice of parameters. Thus, all three sets of hyperparameters are useful in the formulated VS-loss in (5): $\omega_\pm$ and $\iota_\pm$ help balancing the minority/majority errors in the underparameterized regime, while $\Delta_\pm$ is responsible for good performance in the overparameterized regime.

An intuitive explanation via connection to max-margin classifiers. We see in Figure 2 that for $p \geq 50$ the model is large enough to separate the data. In this regime, we train additionally the

SVM and the cost-sensitive (CS) SVM classifiers. Specifically, we solve the optimization in (7) for $b = 0$ (no offset), for feature vectors $\tilde{\mathbf{x}}_i$ of "reduced" size p, and, $\delta = 1$ for SVM (cyan) and $\delta = \delta_\star$ for CS-SVM (magenta). We observe the following:

When data are separable, the performance of the VS-loss matches the performance of CS-SVM.

Moreover, the performance of the logit-adjusted loss is the same as the performance of SVM. We will formally explain this behavior in Section 5.1.

Tuning $\delta_\star$. The inset plot shows the chosen values for the hyperparameter δ. Observe that the value depends on the parameterization ratio p/n. These values of $\delta_\star$ were chosen based on our sharp asymptotic theory in Section 5.3. Our theory in the same section sharply predicts the generalization behavior of CS-SVM for any $\delta > 0$. The predictions for $\delta = 1$ (SVM) and $\delta = \delta_\star$ (GS-SVM) are shown in cyan and magenta solid lines respectively. Notice the perfect match with the numerical values of the Monte Carlo simulations.

4.2 Group-sensitive classification

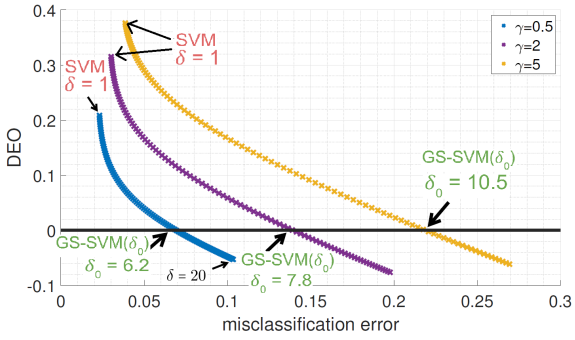


Figure 3: This figure highlights the benefit of our group-sensitive SVM over regular SVM in terms of Equal Opportunity. For a Gaussian mixture model with a sensitive group of prior $p = 0.05$, it depicts the trade-off between the misclassification error and DEO of a linear GS-SVM classifier with different parameter values $\delta \geq 1$ and three distinct parameterization levels $\gamma = d/n$. In this setting, with appropriate tuning of δ, GS-SVM can achieve *zero* DEO.

Our second example concerns a group-sensitive binary classification setting. Specifically, we consider the GMM of Section 2.1 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 $\pi = 1/2$. Our goal is to investigate the effect of the parameter δ of the group-sensitive SVM (8) in terms of the DEO (see Section 2.2). *How much better is GS-SVM compared to standard SVM?*

We answer this question for the GMM model in Figure 3 by deriving the complete tradeoff between DEO and misclassification error of GS-SVM as the parameter δ increases starting from 1 (for which value it coincides with the SVM). To obtain this tradeoff, we establish in Section 6.1 sharp predictions for the generalization performance of GS-SVM both in terms of DEO and misclassification error. We observe that the largest DEO and the smallest misclassification error are achieved by the SVM ($\delta = 1$). Initially, the true-positive error of the minority group is much higher than that of the majority group. But, with increasing δ, misclassification error is traded-off for reduction in absolute value of DEO until a specific $\delta_0 = \delta_0(\gamma)$ giving $\mathcal{R}_{\text{deo}} = 0$, before starting to increase again as the error in the majority group now begins to grow larger. We also observe that the value of δ_0 increases as a function of the parameterization level $\gamma = d/n$ in this example.

5 Label-imbalanced classification

5.1 Connection between VS-loss and CS-SVM

Let $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i \in [n]}$ be a collection of feature/vector pairs in a binary classification setting with imbalance $\pi = \mathbb{P}(y_i = +1) \ll 1 - \pi$. (Here, the $\mathbf{x}_i$'s can be either raw features or outputs of some feature

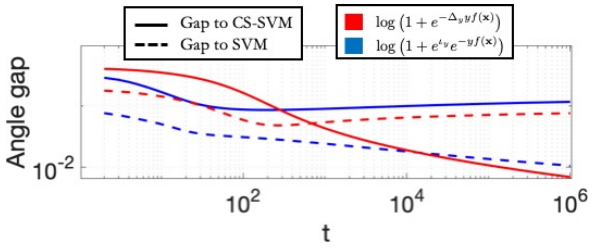


Figure 4: Convergence of gradient-descent iterates $\mathbf{w}_t, t \geq 1$ on the loss in (5) with $f(x) = \mathbf{w}^T\mathbf{x}$ (i.e. no intercept $b = 0$) for two set of parameter choices: (1) In red, $\omega_y = 1, \iota_y = 0, \Delta_y = \frac{1}{\delta}\mathbb{1}[y = 1] + \mathbb{1}[y = -1]$; (2) In blue, $\omega_y = 1, \iota_y = \delta\mathbb{1}[y = 1] + \mathbb{1}[y = -1], \Delta_y = 1$. In both cases we chose $\delta = 4$. We plotted the angle gap $1 - \frac{\hat{\mathbf{w}}^T\mathbf{w}_t}{\|\mathbf{w}_t\|_2\|\hat{\mathbf{w}}\|_2}$ of $\mathbf{w}_t$ to $\hat{\mathbf{w}}$, for two values of $\hat{\mathbf{w}}$: (1) In solid lines, $\hat{\mathbf{w}}$ is the CS-SVM solution in (7) with parameter δ and $b = 0$; (2) In dashed lines, $\hat{\mathbf{w}}$ is the standard SVM solution with $b = 0$. 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 $\pi = 0.2$. The learning rate of GD was set constant to 0.1 for all iterations t.

map $\mathbf{x}_i = \phi(\mathbf{z}_i)$ applied to raw features $\mathbf{z}_i$.) Consider training a binary linear classifier with the VS-loss in (5) with $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ (aka no intercept $b = 0$). Suppose that the set $\mathcal{T}$ is linearly separable and denote $\mathcal{S}_{\text{sep}} = \left\{\mathbf{w} \,:\, \|\mathbf{w}\|_2 = 1 \text{ and } y_i(\mathbf{w}^T\mathbf{x}_i) \geq 1, \forall i \in [n]\right\}$ the set of unit-norm linear separators. It is not hard to see that for any $\mathbf{w}_{\text{sep}} \in \mathcal{S}_{\text{sep}}$ the classifier $\mathbf{w} = t\,\mathbf{w}_{\text{sep}}$ makes the loss in (5) approach zero as $t \to \infty$. Which of these candidate solutions is preferred? To answer this, we consider solving a sequence of norm-constrained loss-minimization problems —each one having a unique solution $\mathbf{w}_R$— that approach the original unconstrained optimization by increasing the constraint threshold $R > 0$. As R becomes large, we prove that the direction of $\mathbf{w}_R$ converges to the direction of the CS-SVM solution $\hat{\mathbf{w}}_\delta$ for $\delta = \Delta_-/\Delta_+$.

Proposition 1 (Implicit bias of VS-loss) *Consider the VS-loss $\mathcal{L}_n(\mathbf{w}) := \sum_{i\in[n]} \ell(y_i, \mathbf{w}^T\mathbf{x}_i)$ with ℓ defined in (5) for positive (but otherwise arbitrary) parameters $\Delta_\pm, \omega_\pm \geq 0$ and arbitrary $\iota_\pm$. Assume there is at least one example from each of the two classes. Define the norm-constrained optimal classifier $\mathbf{w}_R = \arg\min_{\|\mathbf{w}\|_2 \leq R} \mathcal{L}_n(\mathbf{w})$. Assume that the training dataset is linearly separable. Let $\hat{\mathbf{w}}_\delta$ be the solution of (7) without intercept (i.e. $b = 0$) and $\delta = \Delta_-/\Delta_+$. 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, Proposition 1 makes clear that the values of the weights ω_y and ι_y for $y \in \{\pm 1\}$ become irrelevant/ineffective in the separable regime as they all result in the same SVM solutions. On the other hand, the incorporation of the additional parameters $\Delta_\pm$ leads to the same classifier as that of CS-SVM, thus favoring solutions that move the classifier towards the majority class. In our comparison, we removed the intercept terms. Note that intercept can be accounted for by enlarging the features via $\mathbf{x} \to [\mathbf{x}\ 1]$ and adjusting the objective in (7) to $\|\mathbf{w}\|_2^2 + b^2$. The proof of the proposition is given in Appendix B.

For a numerical illustration of Proposition 1 refer to Figure 4 where we plot the angle gap $1 - \frac{\hat{\mathbf{w}}^T\mathbf{w}_t}{\|\mathbf{w}_t\|_2\|\hat{\mathbf{w}}\|_2}$ between gradient-descent (GD) outputs $\mathbf{w}_t, t = 1, 2, \ldots$ for either the VS-loss (in red) or the LA-loss (in blue) and the solution $\hat{\mathbf{w}}$ to either the CS-SVM (in solid lines) or to the SVM (in dashed lines). We observe that the GD outcomes for the VS-loss (resp., LA-loss) converge in direction to the CS-SVM (resp., SVM) solution. From [SHN+18, JT18], where the authors studied the implicit bias of GD for standard logistic regression (i.e,. $\iota_y = 1$ in the LA-loss), we know that it converges to the SVM solution. Figure 4 reveals that the LA-loss with $\iota_{+1} = 4, \iota_{-1} = 1$ (the exact value chosen arbitrarily here) converges to the same solution. We note that our Proposition 1 is suggestive of this behavior of GD —and the result is similar in nature to those of [SHN+18, JT18, GLSS18, CB20]— but it does not directly address the GD iterations, which is left for future work. Perhaps an even more exciting future direction is investigating the algorithmic behavior / implicit bias of the multiclass VS-loss. As a final remark, for the experiments in Figure 4 we kept a constant learning rate for all iterations. Significantly faster convergence is observed when implementing a normalized GD scheme at which the iterates are normalized with the (vanishing) norm of the gradient of the loss at previous iterations; see [NLG+19] and the numerical study in Appendix A.6.

5.2 Connection to post-hoc weight normalization.

The next lemma allows us to also view CS-SVM as an appropriate "post-hoc weight normalization"-approach.

Lemma 1 *Let* $(\hat{\mathbf{w}}_1, \hat{b}_1)$ *be the hard-margin SVM solution. Fix any* $\delta > 0$ *in* (7) *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* (7).

Thus, classification using (7) is equivalent to the following. First learn $(\hat{\mathbf{w}}_1, \hat{b}_1)$ via standard hard-margin SVM, and then simply predict: $\hat{y} = \text{sign}\big((\hat{\mathbf{w}}_1^T\mathbf{x} + \hat{b}_1) + \frac{\delta-1}{\delta+1}\big)$. 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) [STK99, WC03]. Here, we have shown the equivalence of CS-SVM to BM-SVM for a specific choice of the boundary shift. The proof of Lemma 1 presented in Appendix C shows the desired using the KKT conditions of (7).

5.3 Sharp asymptotics for CS-SVM

The structural results in Proposition 1 and Lemma 1 regarding CS-SVM hold for arbitrary binary linearly-separable training datasets $\{(\mathbf{x}_i, y_i)\}_{i\in[n]}$. In this section, under additional distributional assumptions, we establish a sharp theory for CS-SVM in the large-system limit.

Data model: We consider the GMM in Section 2.1 with $K = 1$ and priors $(\pi, 1-\pi)$ for the two classes. We let $\pi < 1-\pi$, so that class +1 is the minority class. We consider the case $\mathbf{\Sigma} = \mathbf{I}_d$ in (3) however supplementary explains further extensions. Let $\mathbf{M} = \begin{bmatrix}\boldsymbol{\mu}_+ & \boldsymbol{\mu}_-\end{bmatrix}$ be the matrix of means and consider the eigen-decomposition of its Grammian:

$$\mathbf{M}^T\mathbf{M} = \mathbf{V}\mathbf{S}^2\mathbf{V}^T, \quad \mathbf{S} \succ \mathbf{0}_{r\times r}, \mathbf{V} \in \mathbb{R}^{2\times r}, r \in \{1, 2\}, \tag{9}$$

with $\mathbf{S}$ an $r \times r$ diagonal positive-definite matrix and $\mathbf{V}$ an orthonormal matrix obeying $\mathbf{V}^T\mathbf{V} = \mathbf{I}_r$. Finally, let $\mathbf{e}_1 = [1\,,\,0]^T$ and $\mathbf{e}_2 = [0\,,\,1]^T$ be the standard basis vectors in $\mathbb{R}^2$.

Learning regime: Our results hold in the large-system limit. Specifically, we study the proportional high-dimensional regime in which both n and d grow to infinity at a fixed rate $\gamma = d/n$. As mentioned earlier, we focus on a regime in which the system is sufficiently overparameterized such that the training error $\mathcal{R}_{\text{train}}$ is zero (equivalently data are linearly separable). For the data model above, it turns out that linear separability undergoes sharp phase-transition. Specifically, there exists $\gamma_\star := \gamma_\star(\mathbf{V}, \mathbf{S}, \pi) \geq 1/2$ such that the training data are linear separable as per (4) with probability approaching one provided that $\gamma > \gamma_\star$. We defer the formal statement of this result and the explicit definition of $\gamma_\star$ in terms of the problem parameters $\mathbf{V}, \mathbf{S}$ and π to Appendix E.4. The result is a small extension to a setting with possibly different class means $\boldsymbol{\mu}_+ \neq \boldsymbol{\mu}_-$ of similar phase transitions established recently in [DKT19, KA20] (also [CS+20, MRSY19] for related results for discriminative models). Onwards, we assume that $\gamma > \gamma_\star$; thus, CS-SVM is feasible with probability approaching 1 for all $\delta > 0$.

Now that we have described the data model and the asymptotic regime, we are ready to present our precise characterization of the balanced error of CS-SVM. For this, we will also need a few definitions as follows. For an arbitrary $\delta > 0$, define random variables as follows.

$$\begin{cases} G \sim \mathcal{N}(0,1) \\ Y \text{ symmetric Bernoulli with } \mathbb{P}\{Y = +1\} = \pi \\ E_Y = \mathbf{e}_1\mathbb{1}[Y = 1] - \mathbf{e}_2\mathbb{1}[Y = -1] \\ \Delta_Y = \delta\cdot\mathbb{1}[Y = +1] + \mathbb{1}[Y = -1] \end{cases}$$

With these further define function $\eta_\delta : \mathbb{R}_{\geq 0} \times \mathcal{S}_2^+ \times \mathbb{R} \to \mathbb{R}$:

$$\eta_\delta(\tilde{q}, \tilde{\boldsymbol{\rho}}, \tilde{b}) := \mathbb{E}\Big[\big(G + E_Y^T\mathbf{V}\mathbf{S}\tilde{\boldsymbol{\rho}} + \frac{\tilde{b}Y - \Delta_Y}{\tilde{q}}\big)_-^2\Big] - (1 - \|\tilde{\boldsymbol{\rho}}\|_2^2)\gamma.$$

Theorem 1 (Balanced error of CS-SVM) *Consider the Gaussian mixture data model with* $K = 1$, *priors* $(\pi, 1-\pi)$ *for the two classes and eigen-decomposition of Grammian matrix as in* (9). *Let*

$\mathcal{R}_{bal} = \frac{\mathcal{R}_+ + \mathcal{R}_-}{2}$ *be the balanced error of the CS-SVM classifier in* (7) *with a fixed margin-ratio* $\delta > 0$. *Let* $(q_\delta, \boldsymbol{\rho}_\delta, b_\delta)$ *be the unique triplet satisfying*

$$\eta_\delta(q_\delta, \boldsymbol{\rho}_\delta, b_\delta) = 0 \quad \textit{and} \quad (\boldsymbol{\rho}_\delta, b_\delta) := \arg \min_{\|\tilde{\boldsymbol{\rho}}\|_2 \le 1, \tilde{b} \in \mathbb{R}} \eta_\delta(q_\delta, \tilde{\boldsymbol{\rho}}, \tilde{b}). \tag{10}$$

With these define

$$\overline{\mathcal{R}}_+ := Q\left(\mathbf{e}_1^T \mathbf{V}\mathbf{S}\boldsymbol{\rho}_\delta + b_\delta/q_\delta\right) \qquad \textit{and} \qquad \overline{\mathcal{R}}_- := Q\left(-\mathbf{e}_2^T \mathbf{V}\mathbf{S}\boldsymbol{\rho}_\delta - b_\delta/q_\delta\right).$$

Then, in the limit of $n, d \to \infty$ *with* $d/n = \gamma > \gamma_\star$, *it holds that* $\mathcal{R}_+ \xrightarrow{P} \overline{\mathcal{R}}_+$ *and* $\mathcal{R}_i \xrightarrow{P} \overline{\mathcal{R}}_-$. *In particular, the balanced error* $\mathcal{R}_{bal}$ *converges in probability as follows:* $\mathcal{R}_{bal} \xrightarrow{P} \overline{\mathcal{R}}_{bal} := \left(\overline{\mathcal{R}}_+ + \overline{\mathcal{R}}_-\right)/2$.

Theorem 1 characterizes the asymptotic classification performance of CS-SVM in terms of three key parameters $(q_\delta, \boldsymbol{\rho}_\delta, b_\delta)$ which can be found by numerically solving (10). Note that the function η_δ is parameterized in terms of the margin ratio δ, the prior probability π, the parameterization ratio γ and the eigenstructure of the Gram matrix of the means in (9). The theorem's proof reveals the following specific role of the three key parameters $(q_\delta, \boldsymbol{\rho}_\delta, b_\delta)$:

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

Thus, b_δ is the asymptotic value of the intercept, q_δ^{-1} is the asymptotic value of the classifier's margin $\frac{1}{\|\mathbf{w}_\delta\|_2}$ to the majority class, and $\boldsymbol{\rho}_\delta$ characterizes the asymptotic alignment of the classifier's hyperplane with the class means. The proof of the theorem uses the convex Gaussian min-max theorem (CGMT) framework [Sto13a, TOH15]; see Appendix E for background, related works and the proof. We remark that, as discussed in supplementary (a) our results lead to simpler expressions when the means are antipodal ($\pm\boldsymbol{\mu}$) and (b) our theory allows for extending results to general covariance model ($\boldsymbol{\Sigma} \neq \mathbf{I}$).

5.3.1 Optimal δ-tuning

The parameter δ in (7) 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 $\mathcal{R}_\pm(\delta) := \mathcal{R}_\pm\big((\hat{\mathbf{w}}_\delta, \hat{b}_\delta)\big)$? Thanks to Theorem 1, 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

$$\begin{aligned} \arg\min_{\delta>0} \;& Q(\mathbf{e}_1^T \mathbf{V}\mathbf{S}\boldsymbol{\rho}_\delta + b_\delta/q_\delta) + Q(-\mathbf{e}_2^T \mathbf{V}\mathbf{S}\boldsymbol{\rho}_\delta - b_\delta/q_\delta) \\ \text{sub. to} \quad & (q_\delta, \boldsymbol{\rho}_\delta, b_\delta) \text{ defined as (10).} \end{aligned} \tag{11}$$

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 (10) for each candidate value of δ. Instead, we can exploit the structural property of CS-SVM given in Lemma (1) to rewrite (11) in a more convenient form. Specifically, we can show (see Appendix D for details) that (11) is equivalent to the following *explicit minimization*:

$$\arg\min_{\delta>0} \; Q\Big(\ell_+ + \big(\frac{\delta-1}{\delta+1}\big)q_1^{-1}\Big) + Q\Big(\ell_- - \big(\frac{\delta-1}{\delta+1}\big)q_1^{-1}\Big), \tag{12}$$

where we defined $\ell_+ := \mathbf{e}_1^T \mathbf{V}\mathbf{S}\boldsymbol{\rho}_1 + b_1/q_1$, $\ell_- := -\mathbf{e}_2^T \mathbf{V}\mathbf{S}\boldsymbol{\rho}_1 - b_1/q_1$, and, $(q_1, \boldsymbol{\rho}_1, b_1)$ are as defined in Theorem 1 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 $\left(Q(\ell_+) + Q(\ell_-)\right)/2$. To summarize, we have shown that one can optimally tune δ to minimize the *asymptotic* balanced error by minimizing the objective in (12) that only depends on the parameters $(q_1, \boldsymbol{\rho}_1, b_1)$ characterizing the asymptotic performance of SVM. In fact, in Appendix D we obtain explicit formulas for the optimal value $\delta_\star$ in (12) as follows

$$\delta_\star := \left(\ell_- - \ell_+ + 2q_1^{-1}\right)/\left(\ell_+ - \ell_- + 2q_1^{-1}\right)_+, \tag{13}$$

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 (7) 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. $\mathcal{R}_+ = \mathcal{R}_- = Q(\frac{\ell_- + \ell_+}{2})$.

6 Group-sensitive classification

Recall from Section 2.1 that we focus on $K = 2$ with equal prior probabilities for the positive and negative labels $p_{+,1} = p_{-,1} = p, p_{+,2} = p_{-,2} = 1-p$. and an imbalance $p \ll 1-p$. Our results here naturally extend to $K > 2$. We consider a setting with only group imbalances. Hence, we may use the VS-loss in (6) with $\Delta_{y,g} = \Delta_g$ for $g = 1, 2$. In particular, since we are interested on separable data, we will focus on the GS-SVM in (8) (where similar to Proposition 1 for label imbalances, we map $\delta = \Delta_2/\Delta_1$).

6.1 Sharp asymptotics for the GS-SVM

In this section, we characterize the DEO of GS-SVM for data generated from the GMM. Thanks to our unifying approach, the analysis is at a high level similar to that of Section 5.3, but there are differences to account for (both in the phase-transition threshold and the generalization formulas) since now each class itself is a mixture of Gaussians.

Data model: For the feature distribution, we let

$$\mathbf{x} \,|\, (y, g) \sim \mathcal{N}(y\boldsymbol{\mu}_g, \mathbf{I}_d)$$

where (for simplicity) $\pm\boldsymbol{\mu}_g$ are the means of groups $g = 1, 2$ with positive/negative labels. As in Section 5.3, let $\mathbf{M} = \begin{bmatrix}\boldsymbol{\mu}_1 & \boldsymbol{\mu}_2\end{bmatrix}$ be the matrix of means of the two groups and assume the eigenvalue decomposition $\mathbf{M}^T\mathbf{M} = \mathbf{V}\mathbf{S}^2\mathbf{V}^T, \mathbf{S} \succ \mathbf{0}_{r\times r}, \mathbf{V} \in \mathbb{R}^{2\times r}, r \in \{1, 2\}$.

Learning regime: The learning regime is similar to the setting of Sec. 5.3. Again, there exists a phase transition threshold $\widetilde{\gamma}_\star := \gamma_\star(\mathbf{V}, \mathbf{S}, \pi, p) \ge 1/2$ such that the training data are separable if and only if $\gamma > \widetilde{\gamma}_\star$ (see Appendix F.2 for exact statements). We assume this for feasibility of (8).

Before stating the main result of this section, we need the following definitions. Fix $\delta > 0$. Define $G, Y, S, \Delta_S \in \mathbb{R}$, and $E_S \in \mathbb{R}^{2\times 1}$ as follows:

$$\begin{cases} G \sim \mathcal{N}(0,1) \\ Y \text{ symmetric Bernoulli with } \mathbb{P}\{Y = +1\} = \pi \\ S \text{ takes values 1 or 2 with probabilities } p \text{ and } 1-p, \text{ respectively} \\ E_S = \mathbf{e}_1 \mathbb{1}[S = 1] + \mathbf{e}_2 \mathbb{1}[S = 2] \\ \Delta_S = \delta \cdot \mathbb{1}[S = 1] + 1 \cdot \mathbb{1}[S = 2]. \end{cases}$$

With these define function $\widetilde{\eta}_\delta : \mathbb{R}_{\ge 0} \times \mathcal{S}^r \times \mathbb{R} \to \mathbb{R}$ as $\eta_\delta(q_\delta, \boldsymbol{\rho}_\delta, b_\delta) := \mathbb{E}\big(G + E_S^T\mathbf{V}\mathbf{S}\tilde{\boldsymbol{\rho}} + \frac{\tilde{b}Y - \Delta_S}{\tilde{q}}\big)_-^2 - (1 - \|\tilde{\boldsymbol{\rho}}\|_2^2)\gamma$.

Theorem 2 (Equal Opportunity of GS-SVM) *Consider the GMM with feature distribution and priors as specified in the 'Data model' above. Let $(\widetilde{q}_\delta, \widetilde{\boldsymbol{\rho}}_\delta, \widetilde{b}_\delta)$ be the unique triplet satisfying (10) 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, the difference of equal opportunity (DEO) satisfies in the same limiting regime: $\mathcal{R}_{\text{deo}} \xrightarrow{P} Q\big(\mathbf{e}_1^T\mathbf{V}\mathbf{S}\widetilde{\boldsymbol{\rho}}_\delta + \widetilde{b}_\delta/\widetilde{q}_\delta\big) - Q\big(\mathbf{e}_2^T\mathbf{V}\mathbf{S}\widetilde{\boldsymbol{\rho}}_\delta + \widetilde{b}_\delta/\widetilde{q}_\delta\big)$.*

The theorem directly implies sharp formulas for both the DEO and the misclassification error by expressing them in terms of the conditional errors as in Sec. 2.2. Using these sharp characterizations allowed us to study the tradeoff between EO and accuracy in Figure 3. In view of the formulas, the requirement for EO translates to finding a parameter δ_0 such that $Q\big(\mathbf{e}_1^T\mathbf{V}\mathbf{S}\widetilde{\boldsymbol{\rho}}_{\delta_0} + \widetilde{b}_{\delta_0}/\widetilde{q}_{\delta_0}\big) = Q\big(\mathbf{e}_2^T\mathbf{V}\mathbf{S}\widetilde{\boldsymbol{\rho}}_{\delta_0} + \widetilde{b}_{\delta_0}/\widetilde{q}_{\delta_0}\big)$. While, we cannot find an explicit formula (as we did for $\delta_\star$ in (13)), we can search for δ_0 numerically. Interestingly, Figure 3 shows that such values exist in our setting even though the GS-SVM does not directly impose EO constraints as several related works, e.g. [OA18, DOBD+18]. The proof of Theorem 2, which is similar to that of Theorem 1, is presented in Appendix F.

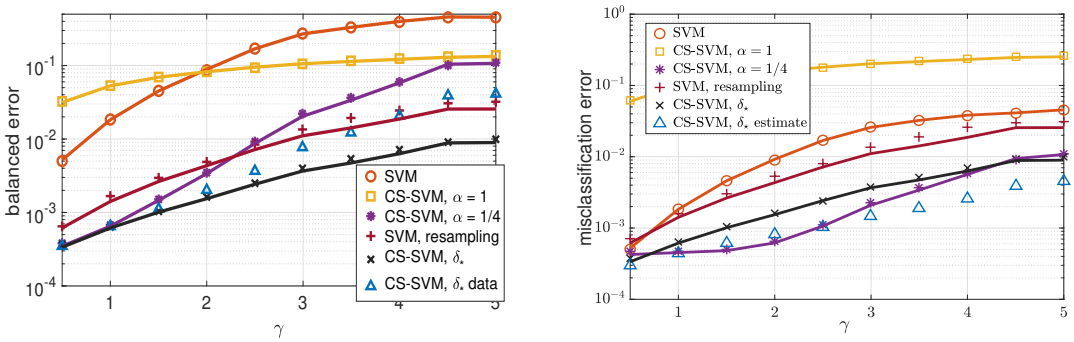


Figure 5: Balanced (Left) and misclassification (Right) errors as a function of the parameterization ratio $\gamma = d/n$ for the following algorithms: SVM with and without majority class resampling, CS-SVM with different choices of $\delta = \left(\frac{1-\pi}{\pi}\right)^{\alpha}, \pi = 0.05$ and $\delta = \delta_\star$ (cf. Eqn. (13)) plotted for different values of $\gamma = d/n$. Solid lines show the theoretical values thanks to Theorem 1 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.

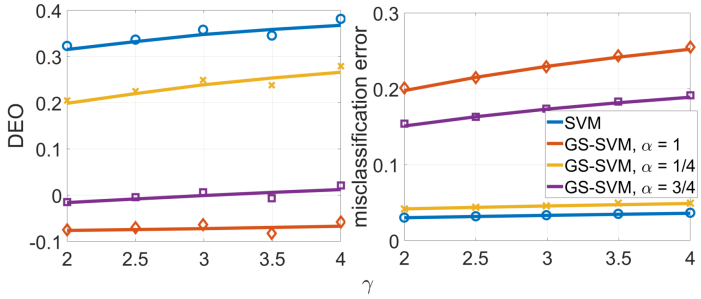


Figure 6: DEO and misclassification error of SVM and GS-SVM with different choices of $\delta = \left(\frac{1-p}{p}\right)^{\alpha}$ for minority group prior $p = 0.05$ plotted against $\gamma = d/n$. Solid lines show the theoretical values and the discrete markers represent empirical errors over 100 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.

7 Numerical experiments

Our numerical results presented in this section corroborate and further justify our theoretical results presented in the previous sections.

7.1 Validity of theoretical performance analysis

In Figures 5 and 6, we demonstrate how our Theorems 1 and 2 provide remarkably precise prediction of the GMM performance even when dimensions are in the order of hundreds. Moreover, both figures demonstrate the clear advantage of CS/GS-SVM over regular SVM and naive resampling strategies in terms of balanced error and equal opportunity, respectively.

In both figures, solid lines show theoretical values and the discrete markers represent simulated error probabilities. For the simulations we fixed $d = 500$. 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 by taking simple average of the class-conditional test errors. The empirical probabilities were computed by averaging over 100 independent realizations of the training and test datasets.

Additionally, Figure 5 validates the explicit formula that we derived in (13) for $\delta_\star$ minimizing

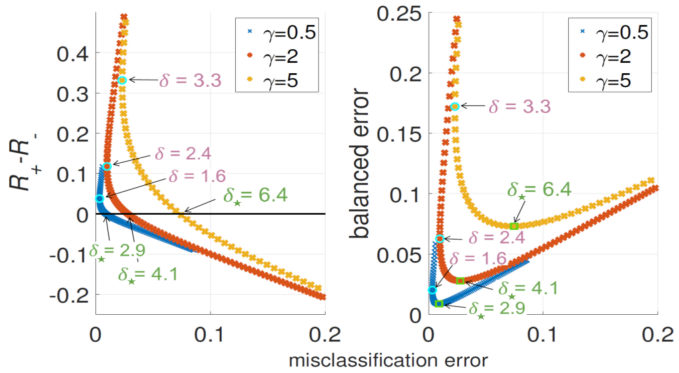


Figure 7: Scatter plots of the trade-off between the misclassification error and error imbalance ($\mathcal{R}_+ - \mathcal{R}_-$) and the trade-off between the misclassification error and balanced error achieved by CS-SVM as a function of δ for different values of γ, for GMM data with $\|\boldsymbol{\mu}_+\| = 3, \boldsymbol{\mu}_- = -\boldsymbol{\mu}_+$ and $\pi = 0.05$. δ is varied in the region $\delta \geq 1$. See text for details.

the balanced error. In Figure 5, observe that the CS-SVM with $\delta = \delta_\star$ ('$\times$' markers) not only does it minimize the balanced error (as predicted in Section 5.3.1), but it also leads to better misclassification error compared to SVM ('$\circ$' markers) for all depicted values of γ. The figure also shows the performance of a classifier that uses a data-dependent heuristic of computing $\delta_\star$ ('$\triangle$' markers); see Appendix A.3.1 for details. The heuristic appears to be accurate for small values of γ and is still better in terms of balanced error compared to the other two heuristic choices of $\delta = (\frac{1-\pi}{\pi})^\alpha, \alpha = 1/4, 1$. Finally, in addition to the δ-tuned CS-SVM and SVM, we compare with a naive but popular scheme of training a standard max-margin SVM after randomly subsampling the majority class examples to retain equal number of examples for both the classes. The error performance analysis of such a scheme is a straightforward extension of our analysis and is discussed in Appendix A.4. Observe that SVM with resampling outperforms SVM without resampling in terms of balanced error, but the optimally tuned CS-SVM is superior to both.

7.2 Tradeoffs

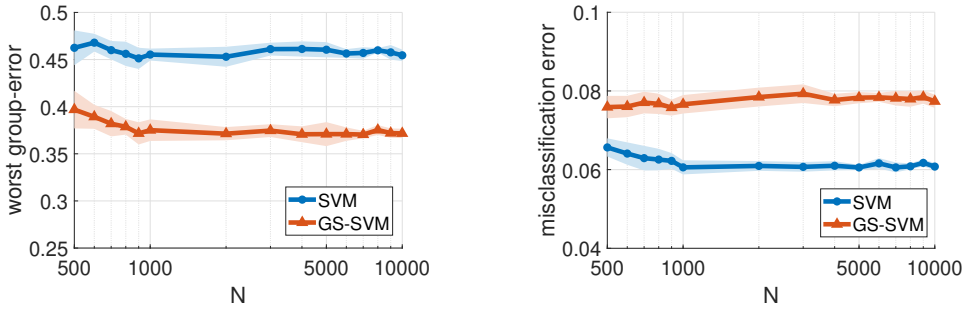
In Figure 3 we discussed tradeoffs of GS-SVM via Theorem 2. Here, in Figure 7 we use our Theorem 1 to present a detailed study of tradeoffs between $\mathcal{R}_{\text{bal}}$ and $\mathcal{R}_+ - \mathcal{R}_-$ vs misclassification error $\mathcal{R}$ for GMM binary data as we increase the margin-ratio parameter δ starting from 1 for three values of $\gamma = d/n$. Focusing on the right subfigure note that both $\mathcal{R}_{\text{bal}}$ and $\mathcal{R}$ vary in a way that there are unique δs minimizing each (shown in green and magenta, respectively). Interestingly, $\mathcal{R}$ is minimized at some non-trivial $\delta \neq 1$. The values minimizing $\mathcal{R}_{\text{bal}}$ coincide with our formula in (13). We also confirm in the left subplot that at $\delta_\star$ it holds that $\mathcal{R}_+ = \mathcal{R}_-$ exactly as predicted in Section 5.3.1.

7.3 Tuning of δ

The theoretical formula (13) for $\delta_\star$ was derived for a GMM and also evaluating it requires knowledge of the true means. In Appendix A.3.1, we describe a data-dependent heuristic inspired by the analytic formula (13) and we examine the heuristic's performance on a synthetic example (see triangle markers in Fig. 5) and on an instance of the MNIST dataset (see Appendix A.3). 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). However, to keep exposition coherent, in our experiments we employed our theoretically-inspired tuning strategy and leave further investigations to future.

7.4 Implications for datasets with spurious correlations

Our results on group-sensitive classification are also relevant in settings when there are *spurious correlations* in the data, such as strong associations between label and background in image classification, e.g. [SKHL19, SRKL20, XEIM20]. Such a setting is easy to understand in the Waterbirds dataset



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

Figure 8: Figures showing the benefit of GS-SVM compared to SVM in achieving smaller *worst-group error* without significant loss on the misclassification error, in the Waterbirds dataset [SRKL20] where there are *spurious correlations* in the data. See text for details.

by [SKHL19, SRKL20], where the goal is to classify images portraying either 'waterbirds' or 'land-birds', while their background —either 'water background' or 'land background'— can be spuriously correlated with the type of birds. In other words, there can be images in the training/test sets depicting 'waterbirds' in 'land background' and 'land-birds' in 'water background'. Formally, the label y of an example in the Waterbirds dataset belongs to $\mathcal{Y} = \{+1, -1\} \equiv \{\text{waterbird}, \text{landbird}\}$. Also, each example belongs to a group $g = (y, a)$ where a is an attribute taking values in $\mathcal{A} = \{+1, -1\} \equiv \{\text{water background}, \text{land background}\}$. Thus, we have a total of *four* groups. Out of these, the groups $(+1, -1)$, $(-1, +1)$ correspond to the minority groups. Specifically, letting $\hat{p}_{(y,a)}$ denote the empirical probability of each group (y, a) calculated over the training data set (i.e. $p_{(y,a)} = \frac{1}{n} \sum_{i \in [n]} \mathbb{1}[(y_i, a_i) = (y, a)]$), we computed:

$$p_{+1,+1} = 0.22 \quad p_{+1,-1} = 0.012 \quad p_{-1,+1} = 0.038 \quad \text{and} \quad p_{-1,-1} = 0.73.$$

In their study, [SRKL20] demonstrated that overparameterization exacerbates such spurious correlations. Specifically, they showed for weighted logistic regression, that while it reduces the misclassification error, it results in a large *worst-group error* in the overparameterized regime (e.g., see Fig. 3 in [SRKL20]). In their analysis, they observed that this is because weighted logistic loss in the separable regime behaves like SVM, which is insensitive to groups.

Motivated by our results, we repeat the experiment of [SRKL20] and compare the *worst-group error* performance (i.e., $\mathcal{R}_{\text{worst}} := \max_{y \in \{\pm 1\}, a \in \{\pm 1\}} \{\mathcal{R}_{(y,a)}\}$ where $\mathcal{R}_{(y,a)}$ is the conditional risk of group $g = (y, a)$) of standard SVM to that of our GS-SVM. Specifically, we trained the following instance of GS-SVM:

$$\min_{\mathbf{w}} \quad \|\mathbf{w}\|_2 \qquad \text{sub. to} \quad y_i(\mathbf{x}_i^T \mathbf{w} + b) \geq \delta_{g_i}, \ i \in [n]. \tag{14}$$

where $\delta_{g_i} = \delta_{(y_i, a_i)} = (\frac{1}{\hat{p}_{(y,a)}})^4$ and $\mathbf{x}_i, i \in [n]$ are N-dimensional random projections of the ResNet18 features used in [SRKL20]. Here, $n = 4795$, N took a range of values from 500 to 10000 and the raw feature dimension was $d = 512$. The curves show the average value of errors over 10 realizations of the random projection matrix along with standard deviations depicted using shaded error-bars. Figure 8 confirms that GS-SVM consistently outperforms standard SVM in the overparameterized regime in terms of a fairness metric such as the worst-group error. In fact, we observe that this gain comes without significant losses on the misclassification error.

7.5 Additional results

Additional numerical experiments are presented in Appendix A:

(i) further details on Figure 2

(ii) numerical comparisons of the VS-loss vs the LA-loss on a group-sensitive GMM

(iii) experiments showing that VS-loss/CS-SVM outperform the LA-loss/SVM on a binary classification instance of the MNIST dataset

(iv) conditional group errors of SVM, GS-SVM (14), and, SVM with subsampling for the Waterbirds dataset

(v) further numerical illustrations supporting Proposition 1 and discussing faster convergence of *normalized GD* [NLG+19] compared to GD with fixed step-size.

8 Future Work

We presented a unified study of learning from imbalanced data where imbalances can be across different groups or classes. To optimize key metrics of interest, we proposed new loss functions and provided insightful theoretical analysis to shed light on the interplay of problem variables.

This work opens up a wealth of exciting future research opportunities. For instance, we suspect our proposed vector-scaling loss can benefit a diverse range of practical applications in NLP and computer vision. All these applications also motivate a theoretical understanding for multiclass problems. What can be said about algorithmic convergence and performance of the multiclass VS-loss? When it comes to group-sensitive learning, it is of broad interest to extend our theory to other fairness metrics of interest beyond equal opportunity. Ideally, our precise asymptotic theory could help contrast different definitions and assess their pros/cons.

Acknowledgments

G. Kini, O. Paraskevas and C. Thrampoulidis are partially supported by the NSF under Grant Numbers CCF-2009030 and HDR-1934641. S. Oymak is partially supported by the NSF award CNS-1932254 and by the NSF CAREER award CCF-2046816.

References

[AG82] 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.

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

[AKLZ20] 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.

[ALMT14] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

[ASH19] Ehsan Abbasi, Fariborz Salehi, and Babak Hassibi. Universality in learning from linear measurements. In *Advances in Neural Information Processing Systems*, pages 12372–12382, 2019.

[BBEKY13] 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.

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

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

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

[BM11] 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.

[BM12] Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *Information Theory, IEEE Transactions on*, 58(4):1997–2017, 2012.

[BMM18] 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.

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

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

[CB20] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

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

[CM19] Michael Celentano and Andrea Montanari. Fundamental barriers to high-dimensional regression with convex penalties. *arXiv preprint arXiv:1903.10603*, 2019.

[CRPW12] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

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

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

[DJM11] David Donoho, Iain Johnstone, and Andrea Montanari. Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *arXiv preprint arXiv:1111.1041*, 2011.

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

[DM15] David L Donoho and Andrea Montanari. Variance breakdown of huber (m)-estimators: n/p \rightarrow m\in (1,\infty). *arXiv preprint arXiv:1503.02106*, 2015.

[DM16] 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.

[DMM09] 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.

[DMM11] David L Donoho, Arian Maleki, and Andrea Montanari. The noise-sensitivity phase transition in compressed sensing. *Information Theory, IEEE Transactions on*, 57(10):6920–6941, 2011.

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

[Don06] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[DT05] David L Donoho and Jared Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9452–9457, 2005.

[EK18] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1-2):95–175, 2018.

[ESAP+20] Melikasadat Emami, Mojtaba Sahraee-Ardakan, Parthe Pandit, Sundeep Rangan, and Alyson Fletcher. Generalization error of generalized linear models in high dimensions. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2892–2901. PMLR, 13–18 Jul 2020.

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

[GLSS18] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information Processing Systems*, 31:9461–9471, 2018.

[Gor85] Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

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

[HM19] Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58, 2019.

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

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

[JK19] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

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

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

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

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

[KT21] Ganesh 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.

[LS20] 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.

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

[MKL+20] 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.

[MLC19] Xiaoyi Mai, Zhenyu Liao, and Romain Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3361. IEEE, 2019.

[MM18] Léo Miolane and Andrea Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv preprint arXiv:1811.01212*, 2018.

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

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

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

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

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

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

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

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

[OT17] Samet Oymak and Joel A Tropp. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 7(3):337–446, 2017.

[OTH13] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized lasso: A precise analysis. *arXiv preprint arXiv:1311.0830*, 2013.

[RV06] M. Rudelson and R. Vershynin. Sparse reconstruction by convex relaxation: Fourier and gaussian measurements. In *40th Annual Conference on Information Sciences and Systems*, pages 207–212, 2006.

[RV18] Cynthia Rush and Ramji Venkataramanan. Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 64(11):7264–7286, 2018.

[SAH18] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. A precise analysis of phasemax in phase retrieval. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 976–980. IEEE, 2018.

[SAH19] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. *arXiv preprint arXiv:1906.03761*, 2019.

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

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

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

[STK99] Grigoris Karakoulas John Shawe-Taylor and Grigoris Karakoulas. Optimizing classifiers for imbalanced training sets. *Advances in neural information processing systems*, 11(11):253, 1999.

[Sto09a] Mihailo Stojnic. Block-length dependent thresholds in block-sparse compressed sensing. *arXiv preprint arXiv:0907.3679*, 2009.

[Sto09b] Mihailo Stojnic. Various thresholds for ℓ_1-optimization in compressed sensing. *arXiv preprint arXiv:0907.3666*, 2009.

[Sto13a] Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

[Sto13b] Mihailo Stojnic. A performance analysis framework for socp algorithms in noisy compressed sensing. *arXiv preprint arXiv:1304.0002*, 2013.

[Sto13c] Mihailo Stojnic. Regularly random duality. *arXiv preprint arXiv:1303.7295*, 2013.

[Sto13d] Mihailo Stojnic. Upper-bounding ℓ_1-optimization weak thresholds. *arXiv preprint arXiv:1303.7289*, 2013.

[TAH15] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems*, pages 3420–3428, 2015.

[TAH18] 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.

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

[TPT20a] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Fundamental limits of ridge-regularized empirical risk minimization in high dimensions. *arXiv preprint arXiv:2006.08917*, 2020.

[TPT20b] 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.

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

[TXH18] 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.

[WC03] 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.

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

[WWM19] Shuaiwen Wang, Haolei Weng, and Arian Maleki. Does slope outperform bridge regression? *arXiv preprint arXiv:1909.09345*, 2019.

[XEIM20] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

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

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

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

A Additional numerical results

A.1 VS-loss vs LA-loss for a label-imbalanced GMM (Additional results on the experiment of Figure 2)

In Figure 9 we have implemented the same experiment as in Figure 2 detailed in Section 4.1. Additionally to the balanced error, we present results for the misclassification error and the two conditional errors. Moreover, we ran simulations for one more set of parameters for the LA-loss. Specifically, we use the suggestion of [MJR+20] and set $\iota_y = \log \frac{1-\pi_y}{\pi_y}$. As promised by Proposition 1 the specific choice of the additive offset parameters ι_y are irrelevant in the separable regime: for all choices the performance is eventually the same as that of SVM. While the performance of the VS-loss is clearly better in terms of balanced error compared to the LA-loss in the separable regime, the offsets ι_y improve the performance in the non-separable regime. Specifically, the figure confirms experimentally the superiority of the tuning of the LA-loss in [MJR+20] compared to that in [CWG+19] (but only in the underparameterized regime).

In all cases, we report both the results of Monte Carlo simulations, as well as, the theoretical formulas predicted by Theorem 1. As promised, the theorem sharply predicts the conditional error probabilities of both the minority and the majority class. Note the almost perfect match with the numerical averages despite the relatively small problem dimension ($d = 300$).

As noted in Section 4.1, we observe in the 'Top Left' of the figure that the VS-loss results in a better balanced error in the separable regime (where $\mathcal{R}_{\text{train}} = 0$) compared to the 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 explain these observations showing that VS-loss 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 model; see also [HMRT19, DKT19].

For the numerical experiments in Figure 9 we minimized the VS-loss and the LA-loss in the separable regime using normalized gradient descent. Specifically, we use an in increasing learning rate that is appropriately normalized by the norm of the loss gradient for faster convergence. Please refer to Figure 14 and Section A.6 for a discussion on the advantage of this over a constant learning rate. In the experiments of Figure 9 we ran normalized GD 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 VS-loss.

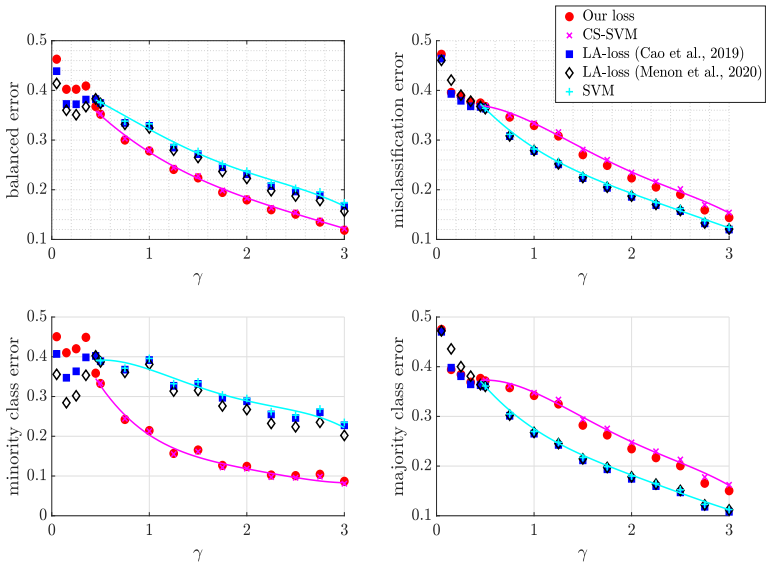


Figure 9: Performance of the VS-loss vs the Logit-adjusted loss for a label-imbalanced GMM with missing features. The experiment confirms that the specific choice of the offset parameters ι_y in the LA-loss are irrelevant in the separable regime: for all choices the performance is eventually the same as that of SVM. Instead, our VS-loss has the same improved balanced-error performance as the CS-SVM. The experimental setting is identical to that of Figure 2. *Top Left:* balanced error $\mathcal{R}_{\text{bal}}$. *Top Right:* misclassification error $\mathcal{R}$. *Bottom Left:* majority class error $\mathcal{R}_-$. *Bottom Right:* minority class error $\mathcal{R}_+$. Solid lines correspond to theoretical formulas obtained thanks to Theorem 1.

A.2 VS-loss vs LA-loss for a group-sensitive GMM

In Figure 10 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 of Section 6 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. A.6) on

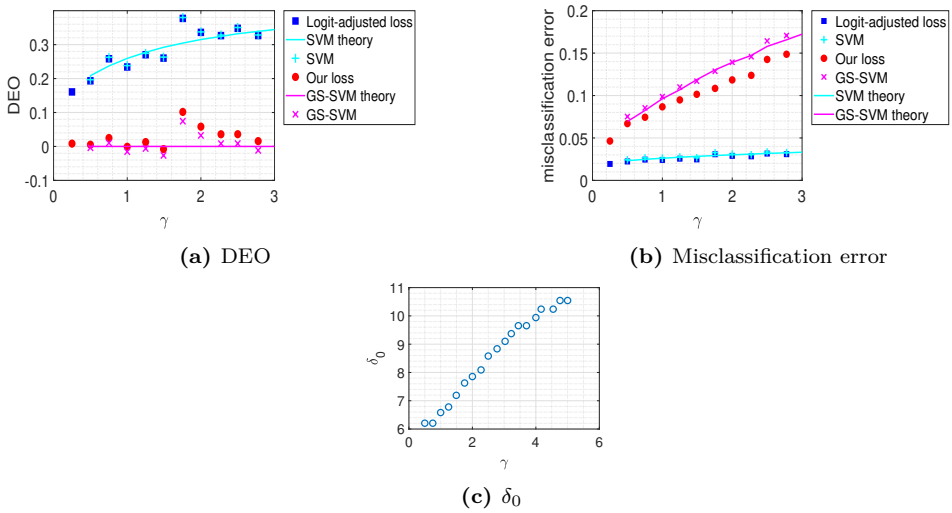
- our VS-loss $\ell(y, \mathbf{w}^T\mathbf{x}, g) := \log(1 + e^{-\Delta_g y(\mathbf{w}^T\mathbf{x})})$ with $\Delta_g = \delta_0 \mathbb{1}[g = 1] + \mathbb{1}[g = 2]$.
- the LA-loss modified for group-sensitive classification $\ell(y, \mathbf{w}^T\mathbf{x}, g) := \log(1 + e^{\iota_g} e^{y(\mathbf{w}^T\mathbf{x})})$ 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 [CWG+19], 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 in (8) 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 2 for the DEO of GS-SVM for any value of δ and performed a grid-search giving us the desired δ_0; see Figure 10 for the values of δ_0 for different values of γ.

Figure 10(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.



(a) DEO **(b)** Misclassification error **(c)** δ_0

Figure 10: 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 2) produced by the corresponding GS-SVM is 0. The solid lines depict theoretical predictions obtained by Theorem 2. The empirical probabilities were computed by averaging over 25 independent realizations of the training and test data.

A.3 Experiments on the MNIST dataset

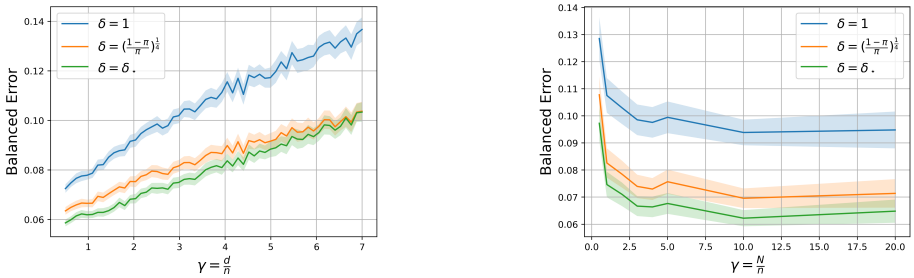
We complement our experiments on synthetic GMM data with additional results on the MNIST dataset.

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 11 as shaded regions.

We ran two experiments. In the first one depicted in Figure 11(a), we trained linear classifiers using the standard SVM (blue), the CS-SVM with a heuristic value $\delta = (\frac{1-\pi}{\pi})^{\frac{1}{4}}$ (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 (13) for the GMM; see Section A.3.1 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 $\delta = (\frac{1-\pi}{\pi})^{\frac{1}{4}}$.

Next, in Figure 11(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. [HMRT19, BLLT20, MM19] and that again the estimated optimal $\delta_\star$ results in tuning of CS-SVM that outperforms the other depicted choices.

In Figure 12 we repeat the experiment of Figure 11(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 (5) 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



(a) Linear classifier **(b)** Random-features classifier

Figure 11: A comparison of CS-SVM balanced error against the overparameterization ratio γ, for the standard hard margin SVM ($\delta = 1$), for a heuristic $\delta = (\frac{1-\pi}{\pi})^{\frac{1}{4}}$ and for our *approximation* of the optimal δ ($\delta = \tilde{\delta}_\star$) obtained by the data-dependent heuristic in Section A.3.1. 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.

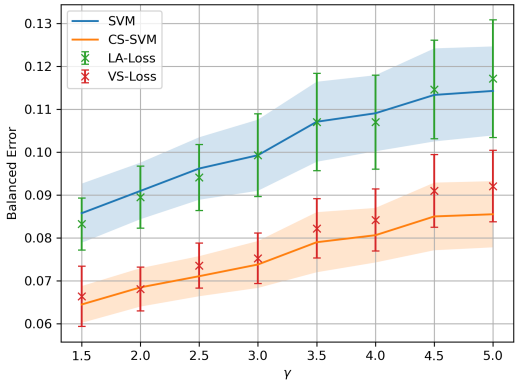


Figure 12: In the overparameterized regime, our VS loss converges to the CS-SVM classifier, while the LA-loss converges to the inferior —in terms of balanced-error performance— SVM. The experiment was performed on the MNIST dataset in a one-vs-rest classification task where the goal is to separate the minority class containing images of the digit 7 from the majority class containing images of all other digits. See text for details.

Section A.3.1 for $\tilde{\delta}_\star$). In a similar manner, LA-loss is defined using the same formula (5), but with parameters $\Delta_\pm = 1$, $\omega_\pm = 1$ and $\iota_+ = \pi^{-1/4}, \iota_- = (1-\pi)^{-1/4}$ (as suggested in [CWG+19]).

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.

A.3.1 Data-dependent heuristic to estimate $\delta_\star$

In Section 5.3 we derived an explicit expression for the optimal $\delta = \delta_\star$ that minimizes the error of CS-SVM for a GMM. The optimal value of $\delta_\star$ in Eqn. (13) not only relies on a GMM but also its computation requires knowledge of the correlation of the SVM classifier with the true means of the classes. In this section, we propose a *data-dependent heuristic* to estimate $\delta_\star$.

Recall from (13) the formula $\delta_\star := (\ell_- - \ell_+ + 2q_1^{-1})/(\ell_+ - \ell_- + 2q_1^{-1})_+$, where $\ell_+ := \mathbf{e}_1^T \mathbf{V}\mathbf{S}\boldsymbol{\rho}_1 + b_1/q_1$

and $\ell_- := -\mathbf{e}_2^T \mathbf{V}\mathbf{S}\boldsymbol{\rho}_1 - b_1/q_1$. Also, according to Theorem 1 and for $\delta = 1$ it holds that

$$(\|\hat{\mathbf{w}}_1\|_2, \hat{\mathbf{w}}_1^T \boldsymbol{\mu}_+/\|\hat{\mathbf{w}}_1\|_2, \hat{\mathbf{w}}_1^T \boldsymbol{\mu}_-/\|\hat{\mathbf{w}}_1\|_2, \hat{b}_1) \xrightarrow{P} (q_1, \mathbf{e}_1^T \mathbf{V}\mathbf{S}\boldsymbol{\rho}_1, \mathbf{e}_2^T \mathbf{V}\mathbf{S}\boldsymbol{\rho}_1, b_1). \tag{15}$$

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. Hence, we can run SVM on the data and estimate q_1 and b_1 using (15). Unfortunately, to estimate $\boldsymbol{\rho}_1$ it appears from (15) that we also 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 we want to do that on a balanced validation set that we create by splitting the training data. After that, we train an instance of the CS-SVM model with $\delta = 1$ (aka 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:

$$\tilde{\delta}_\star := \big(\tilde{\ell_-} - \tilde{\ell}_+ + 2\|\hat{\mathbf{w}}_1\|_2^{-1}\big)/\big(\tilde{\ell_+} - \tilde{\ell_-} + 2\|\hat{\mathbf{w}}_1\|_2^{-1}\big)_+, \quad \text{with} \quad \tilde{\ell_+} := \frac{\hat{\mathbf{w}}_1^T \tilde{\boldsymbol{\mu}}_+ + \hat{b}_1}{\|\hat{\mathbf{w}}_1\|_2}, \quad \tilde{\ell_-} := -\frac{\hat{\mathbf{w}}_1^T \tilde{\boldsymbol{\mu}}_- + \hat{b}_1}{\|\hat{\mathbf{w}}_1\|_2}. \tag{16}$$

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 11 and 5. 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). However, to keep exposition coherent, in all our experiments we employed our theoretically-inspired tuning strategies and leave further investigations to future.

A.4 Max-margin SVM with random majority class undersampling

A popular technique that learns a linear classifier aiming at good balanced error when the training data in label-imbalanced datasets is to randomly undersample the examples from the majority class, followed by max-margin SVM. The asymptotic performance of this scheme under a GMM can be analyzed using Theorem 1 as we explain 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:

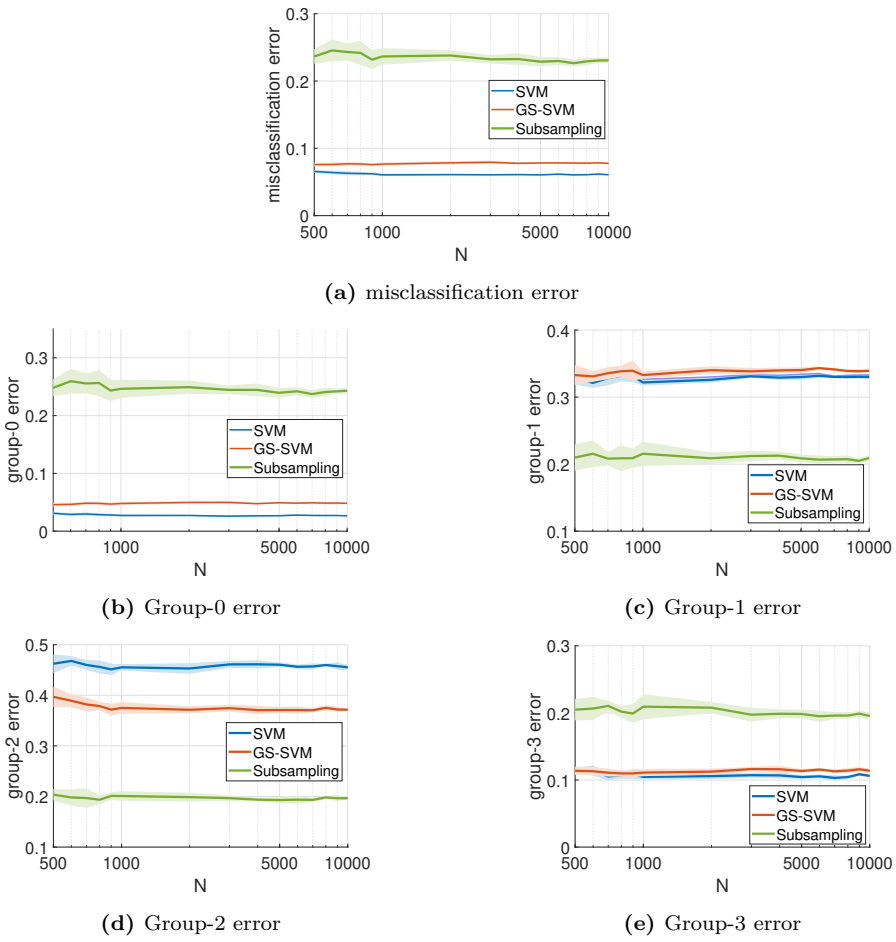
$$\begin{aligned}
&\mathcal{R}_{+,\text{undersampling}}(\gamma, \pi) \xrightarrow{P} \overline{\mathcal{R}}_{+,\text{undersampling}}(\gamma, \pi) = \overline{\mathcal{R}}_+\big(\frac{\gamma}{2\pi}, 0.5\big) \\
&\mathcal{R}_{-,\text{undersampling}}(\gamma, \pi) \xrightarrow{P} \overline{\mathcal{R}}_{-,\text{undersampling}}(\gamma, \pi) = \overline{\mathcal{R}}_{+,\text{undersampling}}(\gamma, \pi).
\end{aligned} \tag{17}$$

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 $\overline{\mathcal{R}}_+\big(\frac{\gamma}{2\pi}, 0.5\big)$ is the asymptotic conditional risk of a *balanced* dataset with overparameterization ratio $\frac{\gamma}{2\pi}$. This is computed as instructed in Theorem 1 for the assignments $\gamma \leftarrow \frac{\gamma}{2\pi}$ and $\pi \leftarrow 1/2$ in the formulas therein.

Our numerical simulations in Figure 5 verify the above formulas.

A.5 Further results on the Waterbirds dataset

In Section 7.4 we presented numerical results showing that GS-SVM (14) with $\delta_{(y,a)} = \big(\frac{1}{p_{(y,a)}}\big)^4$ consistently outperforms standard SVM in terms of *worst-group error* without significant losses in *misclassification error*. In [SRKL20] the authors further compared the performance of weighted empirical risk minimization (ERM) (aka SVM in the separable regime) against ERM with subsampling. Specifically, they demonstrated that the latter achieves a low *worst-group error*. In this section, we show numerical results for *conditional group errors* of SVM, GS-SVM, as well as, SVM with subsampling (corresponding to ERM with subsampling). For the latter, we subsampled the training data such that the resulting set has equal number of examples from every group. In particular, we chose 56 examples from every group, as this is the size of the smallest group *Group-2*. For the purpose of demonstration, we arbitrarily subsampled the training set once. Now, we run standard SVM on the



(a) misclassification error

(b) Group-0 error

(c) Group-1 error

(d) Group-2 error

(e) Group-3 error

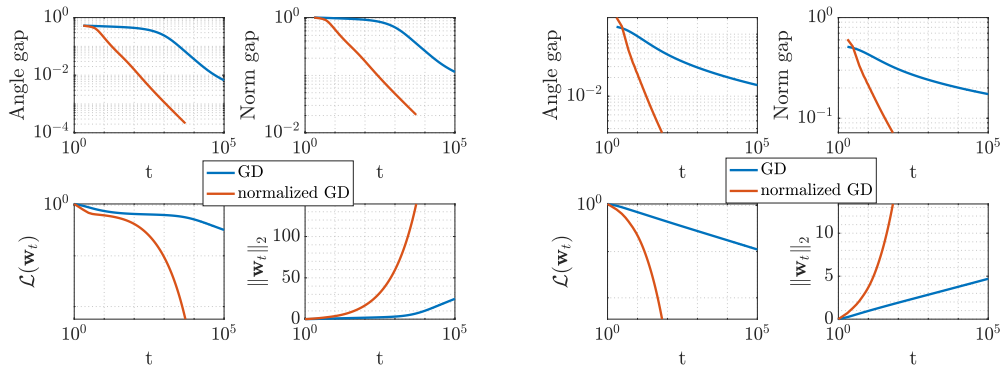
Figure 13: Figures showing misclassification errors and conditional group errors achieved by SVM (blue), GS-SVM with heuristic tuning $\delta_{(y,a)} = \left(\frac{1}{p_{(y,a)}}\right)^4$ (red), and, SVM with subsampling (green) for the Waterbirds dataset of [SRKL20] with *spurious correlations*. The GS-SVM has lower worst-case error (see Group-2, subfigure (d)) compared to the SVM without significant increase on the misclassification error (see subfigure (a)). The SVM with subsampling has the best worst-group error performance, but the price paid is a an up to 5-fold increase in the errors of the majority Group-0 (see subfigure (b)). This also leads to a significant increase of the misclassification error in subfigure (a). See text for more details.

resulting (smaller) dataset, which does not possess group-imbalance. Recall, in the original dataset, Group-0 and Group-3 were the *majority groups* with 3498 and 1057 examples respectively, while Group-1 and Group-2 were the *minority groups* with 184 and 56 examples, respectively. The *worst group errors* shown in Figure 8 corresponded to Group-2.

Our results are shown in Figures 13. Similar to the observation in [SRKL20], SVM with subsampling achieves low *worst group error*, lower than both SVM and GS-SVM with $\delta_{(y,a)} = \left(\frac{1}{p_{(y,a)}}\right)^4$. Specifically, note the low errors for Groups 2 (Figures 13(b) and 13(c) in) and 3 (minority groups) with SVM with subsampling. However, this happens at a significant cost paid by the majority Groups- 1 and 3 (Figures 13(a) and 13(d) in). This results in the misclassification error increasing by a factor close to 5 in comparison to the standard SVM without subsampling (Figure 13(a)). We expect that, with better tuning of the parameters $\delta_{(y,a)}$, the GS-SVM on the full dataset can help achieve even lower *group errors* for the minority groups without hurting the *majority group errors* significantly. We leave this such investigations to future work.

A.6 Implicit bias of the VS-loss (Numerical illustration of Proposition 1)

Figure 14 complements 4 in demonstrating numerically the validity of Proposition 1. The proposition establishes a connection between the norm-constrained VS-loss and the CS-SVM in the limit of



(a) VS-loss vs CS-SVM **(b)** LA-loss vs SVM

Figure 14: Convergence properties of gradient-descent (blue) and normalized gradient-descent (red) iterates $\mathbf{w}_t, t \geq 1$ on: on the loss in (5) with $f(x) = \mathbf{w}^T\mathbf{x}$ (i.e. no intercept $b = 0$) 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 VS-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 (7) with parameter δ; (b) $\hat{\mathbf{w}}$ is the standard SVM solution with $b = 0$. 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 [NLG+19].

increasing model weights. In order to empirically demonstrate this, we solve the unconstrained VS-loss in (5) using gradient descent (GD). Specifically, we generate data from a GMM with class imbalance $\pi = 0.1$ and we run two experiments for two choices of parameters in (5) corresponding to our VS-loss (with non-trivial multiplicative weights) and the LA-loss (cf. (1)); see the figure's caption for details. For each iterate outcome $\mathbf{w}_t$ of GS, 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.

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 both cases, the loss $\mathcal{L}(\mathbf{w}_t)$ is driven to zero and the norm of the weights $\|\mathbf{w}_t\|_2$ to infinity with increasing t. Several recent works [SHN+18, JT18, GLSS18, CB20] have studied in detail the convergence properties of GD on *standard* logistic loss. We suspect that a formal analysis is also possible for the VS-loss using similar tools, but we leave such investigations to future work.

In Figure 14, we also study (curves in red) the convergence properties of *normalized GD*. Following [NLG+19], we implement 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: $\eta_t = \frac{1}{\|\nabla\mathcal{L}(\widehat{\mathbf{w}})\|_2\sqrt{t+1}}$. [NLG+19] 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.

B Connection of VS-loss to CS-SVM

B.1 A more general version of Proposition 1

We will state and prove a more general theorem to which Proposition 1 is a corollary. The new theorem also shows that the group-sensitive adjusted VS-loss in (6) converges to the GS-SVM, which we analyzed in Section 6.

As in Section 2.1, let $\{(\mathbf{x}_i, g_i, y_i)\}_{i=1}^n$ be a sequence of n i.i.d. training samples from a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{G} \times \mathcal{Y}$.

Consider the VS-loss empirical risk minimization

$$\mathcal{L}(\mathbf{w}) := \sum_{i\in[n]} \ell(y_i, \mathbf{w}^T\mathbf{x}_i, g_i) := \omega_i \log\left(1 + e^{\iota_i} \cdot e^{-\Delta_i y_i(\mathbf{w}^T\mathbf{x}_i)}\right). \tag{18}$$

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 (6).

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

$$\widehat{\mathbf{w}} := \arg\min_{\mathbf{w}} \|\mathbf{w}\|_2 \quad \text{subject to} \quad y_i(\mathbf{w}^T\mathbf{x}_i) \geq 1/\Delta_i. \tag{19}$$

The following theorem connects (18) to (19).

Theorem 3 (On implicit bias of the VS-loss: General result) *Define the norm-constrained optimal classifier*

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

with the loss $\mathcal{L}_n$ is defined in (18). 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, (7) is feasible. Moreover, letting $\widehat{\mathbf{w}}$ be the solution of (7), it holds that

$$\lim_{R\to\infty} \mathbf{w}_R/\|\mathbf{w}_R\|_2 = \widehat{\mathbf{w}}/\|\widehat{\mathbf{w}}\|_2. \tag{20}$$

Proof of Proposition 1. Before proving Theorem 3, note that Proposition 1 follows as a corollary by choosing $\omega_i = \omega_{y_i}$, $\iota_i = \iota_{y_i}$ and $\Delta_i = \Delta_{y_i}$. Indeed for this choice the loss in (18) reduces to (5). Also, (19) reduces to (7). The latter follows from the equivalence of the following two optimization problems:

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

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 an immediate corollary of Theorem 3 we get an analogue of Proposition 1 for the group-imbalance data setting of Section 6. Specifically, under the setting of Section 6, with only group imbalances and $K = 2$, we may use the VS-loss in (18) with margin parameters $\Delta_i = \Delta_g, g = 1, 2$. Then, 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 in (8) with $\delta = \Delta_2/\Delta_1$.

B.1.1 Proof of Theorem 3

There are two statements to prove and we show them in the order in which they appear in the theorem's statement.

Linear separability $\implies$ feasibility of (19). 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 (19). To check this, note that

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

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

Proof of (20). First, we will argue that for any $R > 0$ the solution to the constrained VS-loss minimization is on the boundary, i.e.

$$\|\mathbf{w}_R\|_2 = R. \tag{21}$$

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

(19) (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 $\widetilde{\mathbf{w}}^T \nabla \mathcal{L}(\hat{\mathbf{w}}) = 0$. On the other hand, by positivity of $\omega_i, \Delta_i, \forall i \in [n]$:

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

which leads to a contradiction.

Now, suppose that (20) is not true. This means that there is some $\epsilon > 0$ such that there is always an arbitrarily large $R > 0$ such that $\frac{\mathbf{w}_R^T \hat{\mathbf{w}}}{\|\mathbf{w}_R\|_2 \|\hat{\mathbf{w}}\|_2} \leq 1 - \epsilon$. Equivalently, (in view of (21)):

$$\frac{\mathbf{w}_R^T \hat{\mathbf{w}}}{R \|\hat{\mathbf{w}}\|_2} \leq 1 - \epsilon. \tag{22}$$

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.

$$\mathcal{L}(\hat{\mathbf{w}}_R) < \mathcal{L}(\mathbf{w}_R), \qquad \text{for sufficiently large } R. \tag{23}$$

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:

$$y_i \hat{\mathbf{w}}_R^T \mathbf{x}_i = \frac{R}{\|\hat{\mathbf{w}}\|_2} y_i \hat{\mathbf{w}}^T \mathbf{x}_i \geq \frac{R}{\|\hat{\mathbf{w}}\|_2} (1/\Delta_i) =: \frac{\bar{R}}{\Delta_i}, \tag{24}$$

where the inequality follows from feasibility of $\hat{\mathbf{w}}$ in (19) and we set $\bar{R} := R/\|\hat{\mathbf{w}}\|_2$. Then, using (24) it follows immediately that

$$\begin{aligned}
\mathcal{L}(\hat{\mathbf{w}}_R) &= \sum_{i=1}^n \omega_i \log\left(1 + e^{\iota_i} e^{-\Delta_i y_i \hat{\mathbf{w}}_R^T \mathbf{x}_i}\right) \\
&\leq \sum_{i=1}^n \omega_i \log\left(1 + e^{\iota_i} e^{-\frac{\bar{R}}{\Delta_i} \Delta_i}\right) \\
&= \sum_{i=1}^n \omega_i \log\left(1 + e^{\iota_i} e^{-\bar{R}}\right) \\
&\leq \omega_{\max} n e^{\iota_{\max} - \bar{R}}.
\end{aligned} \tag{25}$$

In the first inequality above we used (24) 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 (22), 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 (22). Thus, it must be that $\bar{\mathbf{w}} \neq \hat{\mathbf{w}}$. From these and strong convexity of the objective function in (19), it follows that $\bar{\mathbf{w}}$ must be *infeasible* for (7). Thus, there exists at least one example $\mathbf{x}_j$, $j \in [n]$ and $\epsilon > 0$ such that

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

But then

$$y_j \mathbf{w}_R^T \mathbf{x}_j \leq \bar{R}(1 - \epsilon)(1/\Delta_i), \tag{26}$$

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

$$\begin{aligned}\mathcal{L}(\mathbf{w}_R) &\geq \omega_j \log\left(1 + e^{\iota_j - \Delta_j y_j \mathbf{w}_R^T \mathbf{x}_j}\right) \\ &\geq \omega_j \log\left(1 + e^{\iota_{y_j} - \bar{R}\Delta_j \frac{(1-\epsilon)}{\Delta_j}}\right) \\ &\geq \omega_{\min} \log\left(1 + e^{\iota_{\min} - \bar{R}(1-\epsilon)}\right). \end{aligned} \tag{27}$$

The second inequality follows fron (26) and non-negativity of $\Delta_\pm, \omega_\pm$.

To finish the proof we compare (27) against (25). 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.

C Structural property of CS-SVM: Proof of Lemma 1

From optimality of $(\hat{\mathbf{w}}_1, \hat{b}_1)$, convexity of (7) and the KKT-conditions, there exist dual variables $\beta_i, i \in [n]$ such that:

$$\begin{aligned}&\hat{\mathbf{w}}_1 = \sum_{i\in[n]} y_i \beta_i \mathbf{x}_i, \quad \sum_{i\in[n]} y_i \beta_i = 0, \\ &\forall i \in [n] \ : \ \beta_i\left(\mathbf{x}_i^T \hat{\mathbf{w}}_1 + \hat{b}_1\right) = \beta_i y_i, \quad \beta_i \geq 0.\end{aligned} \tag{28}$$

Let $(\hat{\mathbf{w}}_\delta, \hat{b}_\delta)$ defined as in the statement of the lemma and further define $\epsilon_i := \left(\frac{\delta+1}{2\|\hat{\mathbf{w}}_1\|_2}\right)$, $i \in [n]$. Then, it can be checked using (28) that the following conditions hold

$$\begin{aligned}&\hat{\mathbf{w}}_2 = \sum_{i\in[n]} y_i \epsilon_i \mathbf{x}_i, \quad \sum_{i\in[n]} y_i \epsilon_i = 0, \\ &\forall i \in [n] \ : \ \epsilon_i\left(\mathbf{x}_i^T \hat{\mathbf{w}}_2 + \hat{b}_2\right) = \epsilon_i \cdot \begin{cases} \delta & \text{, if } y_i = +1 \\ -1 & \text{, if } y_i = -1 \end{cases}, \quad \epsilon_i \geq 0.\end{aligned} \tag{29}$$

It can also be verified that (29) are the KKT conditions of the CS-SVM with parameter δ. This proves that $(\hat{\mathbf{w}}_2, \hat{b}_2)$ is optimal in (7) as desired.

D On optimal tuning of CS-SVM: Appendix for Section 5.3.1

We state, prove and further discuss the following result on optimality of $\delta_\star$ originally presented in Section 5.3.1.

Proposition 2 (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 1. Further let $(q_1, \boldsymbol{\rho}_1, b_1)$ the solution to (10) for $\delta = 1$. Finally, define*

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

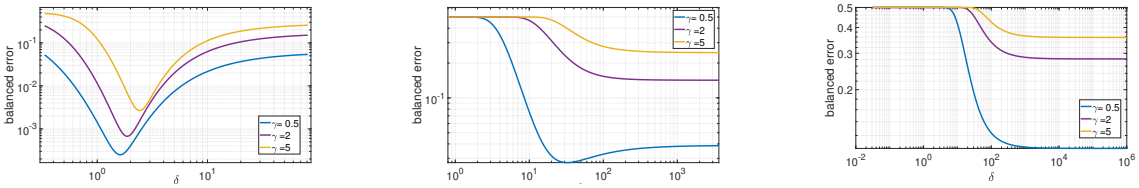
Then, for all $\delta > 0$ it holds that

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

where $\delta_\star$ is defined as

$$\delta_\star = \begin{cases} \frac{\ell_- - \ell_+ + 2q_1^{-1}}{\ell_+ - \ell_- + 2q_1^{-1}} & \text{if } \ell_+ + \ell_- \geq 0 \text{ and } \ell_+ - \ell_- + 2q_1^{-1} > 0, \\ \to \infty & \text{if } \ell_+ + \ell_- \geq 0 \text{ and } \ell_+ - \ell_- + 2q_1^{-1} \leq 0, \\ \to 0 & \text{if } \ell_+ + \ell_- < 0. \end{cases} \tag{30}$$

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



(a) $\pi = 0.1$ **(b)** $\pi = 0.0025$ **(c)** $\pi = 0.0001$

Figure 15: Graphical illustration of the result of Proposition 2: 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 1. 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 Proposition 2. Specifically, note the following for the three different priors. (a) For all values of γ, the minimum is attained (cf. first branch of (30)). (b) For $\gamma = 2, 5$ the minimum is approached in the limit $\delta \to \infty$ (cf. second branch of (30)), but it is attained for $\gamma = 0.5$ (c) The minimum is always approached as $\delta_\star \to \infty$.

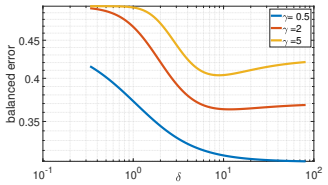


Figure 16: An example showing the dependence of $\delta_\star$ on the data geometry. The above figure is similar to Fig 15 but with a smaller $\|\mu_+\| = \|\mu_-\| = 1$, and for $\pi = 0.1$. While in Fig 15, the value of $\delta_\star$, whenever finite, can be seen to increase with increase in γ, for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

D.1 Proof of Proposition 2

As discussed in Section 5.3.1 the proof proceeds in two steps:
(i) First, starting from (11), we prove (12).
(ii) Second, we analytically solve (12) to derive the explicit expression for $\delta_\star$ in (30).

D.1.1 Proof of (12)

Fix any $\delta > 0$. From Lemma 1,

$$\hat{\mathbf{w}}_\delta = \Big(\frac{\delta+1}{2}\Big)\hat{\mathbf{w}}_1 \quad \text{and} \quad \hat{b}_\delta = \Big(\frac{\delta+1}{2}\Big)\hat{b}_1 + \Big(\frac{\delta-1}{2}\Big). \tag{31}$$

Recall from Theorem 1 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, $q_\delta, \rho_\delta, b_\delta$ and q_1, ρ_1, b_1 are as defined in Theorem 1. Thus, from (31) we find that

$$\boldsymbol{\rho}_\delta = \boldsymbol{\rho}_1, \qquad q_\delta = \Big(\frac{\delta+1}{2}\Big)q_1 \qquad \text{and} \qquad b_\delta = \Big(\frac{\delta+1}{2}\Big)b_1 + \Big(\frac{\delta-1}{2}\Big). \tag{32}$$

Hence, it holds:

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

A similar expression can be written for the conditional error of class -1. Putting these together shows (12), as desired.

D.1.2 Proof of (30)

Recall from (12) that we now need to solve the following constrained minimization where for convenience we call $a = \ell_+$, $b = \ell_-$ and $c = q_1^{-1}$:

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

We define a new variable $x = \frac{\delta-1}{\delta+1}c$. The constraint $\delta > 0$ then writes $x \le c$. 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 $a+b \ge 0$, then $1 > f(x) \ge f(x_\star)$ and $x_\star$ is the unique minimum.
- If $a+b < 0$, then $1 < f(x) \le f(x_\star)$ and $x_\star$ is the unique maximum.

Thus, we conclude with the following:

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

Equivalently,

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

This shows (30). The remaining statement of the proposition is easy to prove requiring simple algebra manipulations.

E Asymptotic analysis of CS-SVM

E.1 Proof of Theorem 1: Preliminaries

The main goal of this section is proving Theorem 1. Recall the data model of Section 5.3. Also, for fixed $\delta > 0$, let $(\hat{\mathbf{w}}, \hat{b})$ be the solution to the CS-SVM in (7). In the following sections, we will prove the following convergence properties for the solution of the CS-SVM:

$$(\|\hat{\mathbf{w}}\|_2, \frac{\hat{\mathbf{w}}^T\boldsymbol{\mu}_+}{\|\hat{\mathbf{w}}\|_2}, \frac{\hat{\mathbf{w}}^T\boldsymbol{\mu}_-}{\|\hat{\mathbf{w}}\|_2}, \hat{b}) \xrightarrow{P} (q_\delta, \mathbf{e}_1^T\mathbf{V}\mathbf{S}\boldsymbol{\rho}_\delta, \mathbf{e}_2^T\mathbf{V}\mathbf{S}\boldsymbol{\rho}_\delta, b_\delta). \tag{33}$$

where $q_\delta, \boldsymbol{\rho}_\delta$ and b_δ are as defined in the theorem's statement. In this section, we show how to use (33) 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 $z \sim \mathcal{N}(0,1)$. Thus, the class conditional can be expressed explicitly in terms of the three summary quantities on the left hand side of (33) 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 (33) in the expression above.

In order to prove the key convergence result in (33) we rely on the convex Gaussian min-max theorem (CGMT) framework. We give some necessary background before we proceed with the proof.

E.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. [DMM11, BM12, Sto13a, Sto13b, Sto13c, OTH13, EK18, BBEKY13, DM16, DM15, TAH15, TAH18, TXH18, SAH18, MM18, SAH19, MLC19, WWM19, CM19, ASH19, WWM19]. The origins of these works trace back to the study of sharp phase transitions in compressed sensing [Don06, RV06, Sto09b, Sto09a, DJM11, DMM11, DT05, ALMT14, OT17, CRPW12, Sto13d] 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 [BM11, DMM09, RV18, ESAP+20], and, (b) the convex Gaussian min-max theorem (CGMT) framework [Sto13a, Sto13c, TOH15]. More recently, these powerful tools have proved very useful for the analysis of linear classifiers [SAH19, MRSY19, DKT19, KT20, MKL+20, LS20, TPT20b, CS+20, AKLZ20, TPT20a]. Theorems 1 and 2 rely on the CGMT and contribute to this line of work. Specifically, our results are most closely related to [DKT19, MRSY19] who first studied max-margin type classifiers.

CGMT framework: Specifically, we rely on the CGMT framework. Here, we only summarize the framework's essential ideas and refer the reader to [TOH15, TAH18] for more details and precise statements. Consider the following two Gaussian processes:

$$X_{\mathbf{w},\mathbf{u}} := \mathbf{u}^T\mathbf{A}\mathbf{w} + \psi(\mathbf{w},\mathbf{u}), \tag{34a}$$

$$Y_{\mathbf{w},\mathbf{u}} := \|\mathbf{w}\|_2\mathbf{h}_n^T\mathbf{u} + \|\mathbf{u}\|_2\mathbf{h}_d^T\mathbf{w} + \psi(\mathbf{w},\mathbf{u}), \tag{34b}$$

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:

$$\Phi(\mathbf{A}) = \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} X_{\mathbf{w},\mathbf{u}}, \tag{35a}$$

$$\phi(\mathbf{h}_n,\mathbf{h}_d) = \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} Y_{\mathbf{w},\mathbf{u}}. \tag{35b}$$

According to the first first statement of the CGMT Theorem 3 in [TOH15] (this is only a slight reformulation of Gordon's original comparison inequality [Gor85]), for any $c \in \mathbb{R}$, it holds:

$$\mathbb{P}\left\{\Phi(\mathbf{A}) < c\right\} \le 2\,\mathbb{P}\left\{\phi(\mathbf{h}_n,\mathbf{h}_d) < c\right\}. \tag{36}$$

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 [TOH15]:

$$\mathbb{P}\left\{|\Phi(\mathbf{A}) - \nu| > t\right\} \le 2\,\mathbb{P}\left\{|\phi(\mathbf{h}_n,\mathbf{h}_d) - \nu| > t\right\}. \tag{37}$$

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 (7) as a PO in the form of (35a). 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 (33). For the other parts, we will employ the following "deviation argument" of the CGMT framework [TOH15]. For arbitrary $\epsilon > 0$, consider the desired set

$$\mathcal{S} := \left\{(\mathbf{v},c) \;\middle|\; \max\left\{\left|\|\mathbf{v}\|_2 - q_\delta\right|, \left|\frac{\mathbf{v}^T\boldsymbol{\mu}_+}{\|\mathbf{v}\|_2} - \mathbf{e}_1^T\mathbf{V}\mathbf{S}\boldsymbol{\rho}_\delta\right|, , \left|\frac{\mathbf{v}^T\boldsymbol{\mu}_-}{\|\mathbf{v}\|_2} - \mathbf{e}_2^T\mathbf{V}\mathbf{S}\boldsymbol{\rho}_\delta\right|, \left|c - b_\delta\right|\right\} \le \epsilon\right\}. \tag{38}$$

Our goal towards (33) 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 $\mathcal{S}$:

$$\Phi_{\mathcal{S}^c}(\mathbf{A}) := \min_{(\mathbf{w},b)\in\mathcal{S}^c} \|\mathbf{w}\|_2 \text{ sub. to} \begin{cases} \mathbf{w}^T\mathbf{x}_i + b \ge \delta & , y_i = +1 \\ \mathbf{w}^T\mathbf{x}_i + b \le -1 & , y_i = -1 \end{cases}, i \in [n], \tag{39}$$

As per Theorem 6.1(iii) in [TAH18] it will suffice to find costants $\bar{\phi}, \bar{\phi}_S$ and $\eta > 0$ such that the following three conditions hold:

$$\begin{cases} \text{(i)} & \bar{\phi}_S \geq \bar{\phi} + 3\eta \\ \text{(ii)} & \phi(\mathbf{h}_n, \mathbf{h}_d) \leq \bar{\phi} + \eta \quad \text{with overwhelming probability} \\ \text{(iii)} & \phi_{\mathcal{S}^c}(\mathbf{h}_n, \mathbf{h}_d) \geq \bar{\phi}_S - \eta \quad \text{with overwhelming probability,} \end{cases} \tag{40}$$

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 (39).

To prove these conditions for the AO of the CS-SVM, in the next section we follow the principled machinery of [TAH18] 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".

E.3 Proof of Theorem 1: Proof of (33)

Let $(\hat{\mathbf{w}}, \hat{b})$ be solution pair to the CS-SVM in (7) for some fixed margin-ratio parameter $\delta > 0$, which we rewrite here expressing the constraints in matrix form:

$$\min_{\mathbf{w},b} \|\mathbf{w}\|_2 \;\; \text{sub. to} \begin{cases} \mathbf{w}^T\mathbf{x}_i + b \geq \delta, & y_i = +1 \\ -(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, & y_i = -1 \end{cases}, \; i \in [n] \;\; = \;\; \min_{\mathbf{w},b} \|\mathbf{w}\|_2 \;\; \text{sub. to } \mathbf{D}_{\mathbf{y}}(\mathbf{X}\mathbf{w} + b\mathbf{1}_n) \geq \boldsymbol{\delta}_{\mathbf{y}}, \tag{41}$$

where we have used the notation

$$\mathbf{X}^T = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix}, \; \mathbf{y} = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^T,$$
$$\mathbf{D}_{\mathbf{y}} = \operatorname{diag}(\mathbf{y}) \text{ and } \boldsymbol{\delta}_{\mathbf{y}} = \begin{bmatrix} \delta\mathbb{1}[y_1 = +1] + \mathbb{1}[y_1 = -1] & \cdots & \delta\mathbb{1}[y_n = +1] + \mathbb{1}[y_n = -1] \end{bmatrix}^{\mathrm{T}}.$$

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 $\mathcal{N}(0,1)$:

$$\mathbf{X} = \mathbf{Y}\mathbf{M}^T + \mathbf{Z}. \tag{42}$$

Following the CGMT strategy [TOH15], we express (41) in a min-max form to bring it in the form of the PO as follows:

$$\begin{aligned} &\min_{\mathbf{w},b} \max_{\mathbf{u}\leq 0} \; \frac{1}{2}\|\mathbf{w}\|_2^2 + \mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{X}\mathbf{w} + b(\mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{1}_n) - \mathbf{u}^T\boldsymbol{\delta}_{\mathbf{y}} \\ =&\min_{\mathbf{w},b} \max_{\mathbf{u}\leq 0} \; \frac{1}{2}\|\mathbf{w}\|_2^2 + \mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{Z}\mathbf{w} + \mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{Y}\mathbf{M}^T\mathbf{w} + b(\mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{1}_n) - \mathbf{u}^T\boldsymbol{\delta}_{\mathbf{y}}. \end{aligned} \tag{43}$$

where in the last line we used (42) and $\mathbf{D}_{\mathbf{y}}\mathbf{D}_{\mathbf{y}} = \mathbf{I}_n$. We immediately recognize that the last optimization is in the form of a PO (cf. (35a)) and the corresponding AO (cf. (35b)) is as follows:

$$\min_{\mathbf{w},b} \max_{\mathbf{u}\leq 0} \; \frac{1}{2}\|\mathbf{w}\|_2^2 + \|\mathbf{w}\|_2\mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{h}_n + \|\mathbf{D}_{\mathbf{y}}\mathbf{u}\|_2\mathbf{h}_d^T\mathbf{w} + \mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{Y}\mathbf{M}^T\mathbf{w} + b(\mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{1}_n) - \mathbf{u}^T\boldsymbol{\delta}_{\mathbf{y}}. \tag{44}$$

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 [TOH15], we need boundedness of the constraint sets. Thus, we restrict the minimization in (44) and (43) to a bounded set $\|\mathbf{w}\|_2^2 + b^2 \leq R$ for (say) $R := 2\left(q_\delta^2 + b_\delta^2\right)$. 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 [DKT19].

For the maximization, we follow the recipe in App. A of [DKT19] who analyzed the standard SVM. Specifically, combining Remark 3 of [DKT19] 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 (44). Towards simplifying this auxiliary optimization, note that $\mathbf{D}_{\mathbf{y}}\mathbf{h}_n \sim \mathbf{h}_n$ by rotational invariance of the Gaussian measure. Also, $\|\mathbf{D}_{\mathbf{y}}\mathbf{u}\|_2 = \|\mathbf{u}\|_2$. Thus, we can express the AO in the following more convenient form:

$$\min_{\|\mathbf{w}\|_2^2+b^2\le R} \max_{\mathbf{u}\le 0} \frac{1}{2}\|\mathbf{w}\|_2^2 + \|\mathbf{w}\|_2\mathbf{u}^T\mathbf{h}_n + \|\mathbf{u}\|_2\mathbf{h}_d^T\mathbf{w} + \mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{Y}\mathbf{M}^T\mathbf{w} + b(\mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{1}_n) - \mathbf{u}^T\boldsymbol{\delta}_{\mathbf{y}}. \tag{45}$$

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

$$\begin{aligned}
&\min_{\|\mathbf{w}\|_2^2+b^2\le R} \max_{\beta\ge 0} \frac{1}{2}\|\mathbf{w}\|_2^2 + \beta\left(\left\|\left(\|\mathbf{w}\|_2\mathbf{h}_n + \mathbf{D}_{\mathbf{y}}\mathbf{Y}\mathbf{M}^T\mathbf{w} + b\,\mathbf{D}_{\mathbf{y}}\mathbf{1}_n - \boldsymbol{\delta}_{\mathbf{y}}\right)_-\right\|_2 - \mathbf{h}_d^T\mathbf{w}\right)\\
=&\min_{\|\mathbf{w}\|_2^2+b^2\le R} \frac{1}{2}\|\mathbf{w}\|_2^2 \quad \text{sub. to } \left\|\left(\|\mathbf{w}\|_2\mathbf{h}_n + \mathbf{D}_{\mathbf{y}}\mathbf{Y}\mathbf{M}^T\mathbf{w} + b\,\mathbf{D}_{\mathbf{y}}\mathbf{1}_n - \boldsymbol{\delta}_{\mathbf{y}}\right)_-\right\|_2 \le \mathbf{h}_d^T\mathbf{w}.
\end{aligned}$$

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 $\mathbf{M}^T = \mathbf{V}\mathbf{S}\mathbf{U}^T$ and project $\mathbf{w}$ on the columns of $\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2\end{bmatrix} \in \mathbb{R}^{d\times 2}$ 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. $\mathbf{S} \succ \mathbf{0}_{2\times 2}$. 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

$$\begin{aligned}
&\min_{\mu_1^2+\mu_2^2+\|\mathbf{w}^\perp\|_2^2+b^2\le R} \frac{1}{2}(\mu_1^2+\mu_2^2+\alpha^2)\\
&\text{sub. to} \quad \left\|\left(\sqrt{\mu_1^2+\mu_2^2+\alpha^2}\mathbf{h}_n + \mathbf{D}_{\mathbf{y}}\mathbf{Y}\mathbf{V}\mathbf{S}\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} + b\,\mathbf{D}_{\mathbf{y}}\mathbf{1}_n - \boldsymbol{\delta}_{\mathbf{y}}\right)_-\right\|_2 \le \mu_1(\mathbf{h}_d^T\mathbf{u}_1) + \mu_2(\mathbf{h}_d^T\mathbf{u}_2) + \mathbf{h}_d^T\mathbf{U}^\perp\mathbf{w}^\perp.
\end{aligned}$$

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

$$\begin{aligned}
&\min_{\mu_1^2+\mu_2^2+\alpha^2+b^2\le R} \frac{1}{2}(\mu_1^2+\mu_2^2+\alpha^2)\\
&\text{sub. to} \quad \left\|\left(\sqrt{\mu_1^2+\mu_2^2+\alpha^2}\mathbf{h}_n + \mathbf{D}_{\mathbf{y}}\mathbf{Y}\mathbf{V}\mathbf{S}\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} + b\,\mathbf{D}_{\mathbf{y}}\mathbf{1}_n - \boldsymbol{\delta}_{\mathbf{y}}\right)_-\right\|_2 \le \mu_1(\mathbf{h}_d^T\mathbf{u}_1) + \mu_2(\mathbf{h}_d^T\mathbf{u}_2) + \alpha\|\mathbf{h}_d^T\mathbf{U}^\perp\|_2.
\end{aligned}$$

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

$$\begin{aligned}
&\min_{\substack{q\ge\sqrt{\mu_1^2+\mu_2^2+\alpha^2}\\ q^2+b^2\le R}} \frac{1}{2}q^2\\
&\text{sub. to} \quad \left\|\left(q\mathbf{h}_n + \mathbf{D}_{\mathbf{y}}\mathbf{Y}\mathbf{V}\mathbf{S}\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} + b\,\mathbf{D}_{\mathbf{y}}\mathbf{1}_n - \boldsymbol{\delta}_{\mathbf{y}}\right)_-\right\|_2 \le \mu_1(\mathbf{h}_d^T\mathbf{u}_1) + \mu_2(\mathbf{h}_d^T\mathbf{u}_2) + \alpha\|\mathbf{h}_d^T\mathbf{U}^\perp\|_2.
\end{aligned} \tag{46}$$

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 $q \ge \sqrt{\mu_1^2+\mu_2^2+\alpha^2} \iff q \ge \|\begin{bmatrix}\mu_1 & \mu_2 & \alpha\end{bmatrix}\|_2$ is a conic second-order cone constraint, and, (ii) the function

$$\mathcal{L}_n(q,\mu_1,\mu_2,\alpha,b) := \frac{1}{\sqrt{n}}\left\|\left(q\mathbf{h}_n + \mathbf{D}_{\mathbf{y}}\mathbf{Y}\mathbf{V}\mathbf{S}\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} + b\,\mathbf{D}_{\mathbf{y}}\mathbf{1}_n - \boldsymbol{\delta}_{\mathbf{y}}\right)_-\right\|_2 - \mu_1\frac{\mathbf{h}_d^T\mathbf{u}_1}{\sqrt{n}} - \mu_2\frac{\mathbf{h}_d^T\mathbf{u}_2}{\sqrt{n}} - \alpha\frac{\|\mathbf{h}_d^T\mathbf{U}^\perp\|_2}{\sqrt{n}} \tag{47}$$

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

$$\mathcal{L}_n(q,\mu_1,\mu_2,\alpha,b) \xrightarrow{P} L(q,\mu_1,\mu_2,\alpha,b) := \sqrt{\mathbb{E}\big(qG + E_Y^T \mathbf{V}\mathbf{S}\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} + bY - \Delta_Y\big)_-^2} - \alpha\sqrt{\gamma}, \tag{48}$$

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 [AG82]. Therefore, the convergence in (48) is in fact uniform in the compact feasible set of (46).

Consider then the deterministic high-probability equivalent of (46) 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:

$$\begin{aligned} &\min_{q^2+b^2\leq R, q>0, \|\boldsymbol{\rho}\|_2\leq 1} \ \frac{1}{2}q^2 \\ &\text{sub. to} \quad \mathbb{E}\Big[\Big(G + E_Y^T\mathbf{V}\mathbf{S}\boldsymbol{\rho} + \frac{bY-\Delta_Y}{q}\Big)_-^2\Big] \leq \big(1-\|\boldsymbol{\rho}\|_2^2\big)\gamma. \end{aligned} \tag{49}$$

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, (49) becomes

$$\min\left\{q^2 \ \Big|\ 0\leq q\leq\sqrt{R} \quad \text{and} \quad \min_{b^2\leq R-q^2, \|\boldsymbol{\rho}\|_2\leq 1} \eta_\delta(q,\boldsymbol{\rho},b) \leq 0\right\}. \tag{50}$$

We will prove that

$$\text{the function } f(q) := \min_{b, \|\boldsymbol{\rho}\|_2\leq 1} \eta_\delta(q,\boldsymbol{\rho},b) \text{ is strictly decreasing.} \tag{51}$$

Let q_δ be as in the statement of the theorem such that $f(q_\delta) = 0$. Then, we have the following relations

$$f(q) \leq 0 \ \Rightarrow \ f(q) \leq f(q_\delta) \ \Rightarrow \ q \geq q_\delta.$$

Thus, the minimizers in (50) 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 (40)). 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 (33). 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 (38) has a cost that is strictly larger than q_δ^2 (i.e. the cost of the unconstrained PO). As per the GCMT this can be done again by showing that the statement is true for the correspondingly constrained AO (i.e. show condition (iii) in (40)). With the exact same simplifications as above, the latter program simplifies to (46) with the additional constraints:

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

Also, using the uniform convergence in (48), it suffices to study the deterministic equivalent (50) with the additional constraints above. Now, we can show the desired (cf. condition (i) in (40)) again by exploiting (51). This part of the argument is identical to Section C.3.5 in [DKT19] and we omit the details.

To complete the proof, it remains to show (51). We do so by combining the following three observations to show that $\frac{\mathrm{d}f}{\mathrm{d}q} < 0$.

First,

$$\frac{\partial \eta_\delta}{\partial q} = \frac{2}{q^2}\mathbb{E}\Big[(G + E_Y^T \mathbf{V}\mathbf{S}\boldsymbol{\rho} + \frac{bY - \Delta_Y}{q})_- \cdot \Delta_Y\Big] - \frac{2b}{q^2}\mathbb{E}\Big[(G + E_Y^T \mathbf{V}\mathbf{S}\boldsymbol{\rho} + \frac{bY - \Delta_Y}{q})_- \cdot Y\Big]$$
$$< -\frac{2b}{q^2}\mathbb{E}\Big[(G + E_Y^T \mathbf{V}\mathbf{S}\boldsymbol{\rho} + \frac{bY - \Delta_Y}{q})_- \cdot Y\Big] \tag{52}$$

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

$$\frac{\partial \eta_\delta}{\partial b} = 0 \iff \mathbb{E}\Big[(G + E_Y^T \mathbf{V}\mathbf{S}\boldsymbol{\rho} + \frac{bY - \Delta_Y}{q})_- \cdot Y\Big] = 0. \tag{53}$$

Third, by the envelope theorem

$$\frac{\mathrm{d}f}{\mathrm{d}q} = \frac{\partial \eta_\delta}{\partial q}\Big|_{\boldsymbol{\rho}^\star, b^\star}. \tag{54}$$

The desired inequality $\frac{\mathrm{d}f}{\mathrm{d}q} < 0$ follows directly by successively applying (54), (52) and (53).

We show how Theorem 1 for the isotropic case can still be applied in the general case. Assume $\boldsymbol{\Sigma} \succ 0$. 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 $\mathbf{z}_i := \boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i = y_i\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_{y_i} + \mathbf{h}_i$. Let

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

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

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

Similar derivation holds for $\mathcal{R}_-$. This completes the proof of the claim.

E.4 Phase transition of CS-SVM

Here, we present a formula for the threshold $\gamma_\star$ such that the CS-SVM of (7) is feasible (resp., infeasible) with overwhelming probability provided that $\gamma > \gamma_\star$ (resp., $\gamma < \gamma_\star$). The first, simple but key, observation is that the phase-transition threshold $\gamma_\star$ of feasibility of the GS-SVM is the same as the threshold of feasibility of the standard SVM for the same model; see Section B.1.1. The feasibility threshold of the standard SVM under the data model of Section 5.3 can be derived immediately from [KT21] who studied separability phase-transitions for the more general multiclass Gaussian mixture model. Specifically, the result below is a small extension to a setting with possibly different class means $\boldsymbol{\mu}_+ \neq \boldsymbol{\mu}_-$ of similar phase transitions established recently in [DKT19, MKL+20] (also [CS+20, MRSY19] for related results for discriminative models).

Proposition 3 ([KT21]) *Consider the same data model and notation as in Theorem 1 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:

$$\gamma_\star := \min_{\mathbf{t}\in\mathbb{R}^r, b\in\mathbb{R}} \mathbb{E}\left[\left(\sqrt{1 + \|\mathbf{t}\|_2^2}\,G + E_Y^T\mathbf{V}\mathbf{S}\mathbf{t} - bY\right)_-^2\right]. \tag{55}$$

Then, the following hold:

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

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 $\delta > 0$.

F Asymptotic analysis of GS-SVM

The theorem below is a slightly more general version of Theorem 2 allowing for different noise variances for the minority/majority groups. Specifically, assume that for $y \in \{\pm 1\}$ and $g \in \{1, 2\}$,

$$\mathbf{x} \,|\, (y, g) \sim \mathcal{N}(y\boldsymbol{\mu}_g, \sigma_g \mathbf{I}_d), \tag{56}$$

for σ_1^2, σ_2^2 the noise variances of the minority and the majority groups, respectively.

Theorem 4 (Sharp asymptotics of GS-SVM: different noise levels) *Consider the GMM with feature distribution and priors as specified in the 'Data model' above. Fix $\delta > 0$. Define $G, Y, S, \Delta_S, \Sigma_S \in \mathbb{R}$, and $E_S \in \mathbb{R}^{2\times 1}$ 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; $E_S = \mathbf{e}_1 \mathbb{1}[S = 1] + \mathbf{e}_2 \mathbb{1}[S = 2]$; $\Delta_S = \delta \cdot \mathbb{1}[S = 1] + 1 \cdot \mathbb{1}[S = 2]$ 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*

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

Let $(\widetilde{q}_\delta, \widetilde{\boldsymbol{\rho}}_\delta, \widetilde{b}_\delta)$ be the unique triplet satisfying (10) 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 $\mathcal{R}_{\pm,i} \xrightarrow{P} Q\big(\mathbf{e}_i^T \mathbf{V}\mathbf{S}\widetilde{\boldsymbol{\rho}}_\delta \pm \widetilde{b}_\delta / \widetilde{q}_\delta\big)$,

F.1 Proof of Theorem 4

The proof of Theorem 4 also relies on the CGMT framework and is very similar to the proof of Theorem 1. 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 E.3 can be followed mutatis mutandis.

Let $(\hat{\mathbf{w}}, \hat{b})$ be solution pair to the GS-SVM in (8) for some fixed parameter $\delta > 0$, which we rewrite here expressing the constraints in matrix form:

$$\min_{\mathbf{w},b} \|\mathbf{w}\|_2 \;\text{ sub. to }\; \begin{cases} y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq \delta, & g_i = 1 \\ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, & g_i = 2 \end{cases}, \; i \in [n] \;=\; \min_{\mathbf{w},b} \|\mathbf{w}\|_2 \;\text{ sub. to }\; \mathbf{D}_\mathbf{y}(\mathbf{X}\mathbf{w} + b\mathbf{1}_n) \geq \boldsymbol{\delta}_\mathbf{g}, \tag{57}$$

where we have used the notation

$$\mathbf{X}^T = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix}, \; \mathbf{y} = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^T,$$
$$\mathbf{D}_\mathbf{y} = \text{diag}(\mathbf{y}) \text{ and } \boldsymbol{\delta}_\mathbf{g} = \begin{bmatrix} \delta\mathbb{1}[g_1 = 1] + \mathbb{1}[g_1 = 2] & \cdots & \delta\mathbb{1}[g_n = 1] + \mathbb{1}[g_n = 2] \end{bmatrix}^\mathrm{T}.$$

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

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

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 $\mathcal{N}(0,1)$:

$$\mathbf{X} = \mathbf{D}_\mathbf{y}\mathbf{G}\mathbf{M}^T + \mathbf{D}_\sigma\mathbf{Z}. \tag{58}$$

As usual, we express (8) in a min-max form to bring it in the form of the PO as follows:

$$\begin{aligned}&\min_{\mathbf{w},b}\max_{\mathbf{u}\le 0}\ \frac{1}{2}\|\mathbf{w}\|_2^2+\mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{X}\mathbf{w}+b(\mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{1}_n)-\mathbf{u}^T\boldsymbol{\delta}_{\mathbf{g}}\\ =&\min_{\mathbf{w},b}\max_{\mathbf{u}\le 0}\ \frac{1}{2}\|\mathbf{w}\|_2^2+\mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{D}_\sigma\mathbf{Z}\mathbf{w}+\mathbf{u}^T\mathbf{G}\mathbf{M}^T\mathbf{w}+b(\mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{1}_n)-\mathbf{u}^T\boldsymbol{\delta}_{\mathbf{g}}.\end{aligned}\tag{59}$$

where in the last line we used (58) and $\mathbf{D}_{\mathbf{y}}\mathbf{D}_{\mathbf{y}}=\mathbf{I}_n$. We immediately recognize that the last optimization is in the form of a PO and the corresponding AO is as follows:

$$\min_{\mathbf{w},b}\max_{\mathbf{u}\le 0}\ \frac{1}{2}\|\mathbf{w}\|_2^2+\|\mathbf{w}\|_2\mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{D}_\sigma\mathbf{h}_n+\|\mathbf{D}_{\mathbf{y}}\mathbf{D}_\sigma\mathbf{u}\|_2\mathbf{h}_d^T\mathbf{w}+\mathbf{u}^T\mathbf{G}\mathbf{M}^T\mathbf{w}+b(\mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{1}_n)-\mathbf{u}^T\boldsymbol{\delta}_{\mathbf{g}}.\tag{60}$$

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 E.3 we consider the one-sided constrained AO in (60). 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, $\|\mathbf{D}_{\mathbf{y}}\mathbf{D}_\sigma\mathbf{u}\|_2=\|\mathbf{D}_\sigma\mathbf{u}\|_2$. Thus, we can express the AO in the following more convenient form:

$$\begin{aligned}&\min_{\|\mathbf{w}\|_2^2+b^2\le R}\max_{\mathbf{u}\le 0}\ \frac{1}{2}\|\mathbf{w}\|_2^2+\|\mathbf{w}\|_2\mathbf{u}^T\mathbf{D}_\sigma\mathbf{h}_n+\|\mathbf{D}_\sigma\mathbf{u}\|_2\mathbf{h}_d^T\mathbf{w}+\mathbf{u}^T\mathbf{G}\mathbf{M}^T\mathbf{w}+b(\mathbf{u}^T\mathbf{D}_{\mathbf{y}}\mathbf{1}_n)-\mathbf{u}^T\boldsymbol{\delta}_{\mathbf{g}}\\ =&\min_{\|\mathbf{w}\|_2^2+b^2\le R}\max_{\mathbf{v}\le 0}\ \frac{1}{2}\|\mathbf{w}\|_2^2+\|\mathbf{w}\|_2\mathbf{v}^T\mathbf{h}_n+\|\mathbf{v}\|_2\mathbf{h}_d^T\mathbf{w}+\mathbf{v}^T\mathbf{D}_\sigma^{-1}\mathbf{G}\mathbf{M}^T\mathbf{w}+b(\mathbf{v}^T\mathbf{D}_\sigma^{-1}\mathbf{D}_{\mathbf{y}}\mathbf{1}_n)-\mathbf{v}^T\mathbf{D}_\sigma^{-1}\boldsymbol{\delta}_{\mathbf{g}},\end{aligned}$$

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 (45) in Section E.3. Following analogous steps, omitted here for brevity, we obtain the following scalarized AO:

$$\min_{\substack{q\ge\sqrt{\mu_1^2+\mu_2^2+\alpha^2}\\ q^2+b^2\le R}}\ \frac{1}{2}q^2\tag{61}$$

$$\text{sub. to}\quad\frac{1}{\sqrt{n}}\Big\|\big(q\mathbf{h}_n+\mathbf{D}_\sigma^{-1}\mathbf{G}\mathbf{V}\mathbf{S}\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}+b\,\mathbf{D}_\sigma^{-1}\mathbf{D}_y\mathbf{1}_n-\mathbf{D}_\sigma^{-1}\boldsymbol{\delta}_{\mathbf{g}}\big)_-\Big\|_2-\mu_1\frac{\mathbf{h}_d^T\mathbf{u}_1}{\sqrt{n}}-\mu_2\frac{\mathbf{h}_d^T\mathbf{u}_2}{\sqrt{n}}-\alpha\frac{\|\mathbf{h}_d^T\mathbf{U}^\perp\|_2}{\sqrt{n}}\le 0.$$

where as in Section E.3 we have decomposed the matrix of means $\mathbf{M}=\mathbf{U}\mathbf{S}\mathbf{V}^T$ 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

$$\bar{L}(q,\mu_1,\mu_2,\alpha,b):=\sqrt{\mathbb{E}\big(qG+\Sigma_S^{-1}E_S^T\mathbf{V}\mathbf{S}\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}+b\,\Sigma_S^{-1}Y-\Sigma_S^{-1}\Delta_S\big)_-^2}-\alpha\sqrt{\gamma},\tag{62}$$

where the random variables G,E_S,Y,Δ_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. (49)):

$$\min_{q^2+b^2\le R,q>0,\|\boldsymbol{\rho}\|_2\le 1}\ \frac{1}{2}q^2\tag{63}$$

$$\text{sub. to}\quad\mathbb{E}\big(G+\Sigma_S^{-1}E_S^T\mathbf{V}\mathbf{S}\boldsymbol{\rho}+\frac{b\,\Sigma_S^{-1}Y-\Sigma_S^{-1}\Delta_S}{q}\big)_-^2\le\big(1-\|\boldsymbol{\rho}\|_2^2\big)\gamma.$$

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, (63) becomes

$$\min\left\{q^2\ \middle|\ 0\le q\le\sqrt{R}\quad\text{and}\quad\min_{b^2\le R-q^2,\|\boldsymbol{\rho}\|_2\le 1}\widetilde{\eta}_\delta(q,\boldsymbol{\rho},b)\le 0\right\}.\tag{64}$$

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

F.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 B.1.1. The feasibility threshold of the standard SVM under the data model of Section 6 can be derived from [KT21], who study the separability question for the more general case of a multiclass mixture of mixtures Gaussian model. We give their result characterizing $\widetilde{\gamma}_\star$ below.

Proposition 4 *Consider the same data model and notation as in Theorem 2 and consider the event*

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

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

$$\widetilde{\gamma}_\star := \min_{\mathbf{t} \in \mathbb{R}^r, b \in \mathbb{R}} \mathbb{E}\left[\left(\sqrt{1 + \|\mathbf{t}\|_2^2}\, G + E_S^T \mathbf{V}\mathbf{S}\mathbf{t} - bY \right)_-^2 \right]. \tag{65}$$

Then, the following hold:

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

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 $\delta > 0$.