A Statistical Theory of Overfitting for Imbalanced Classification

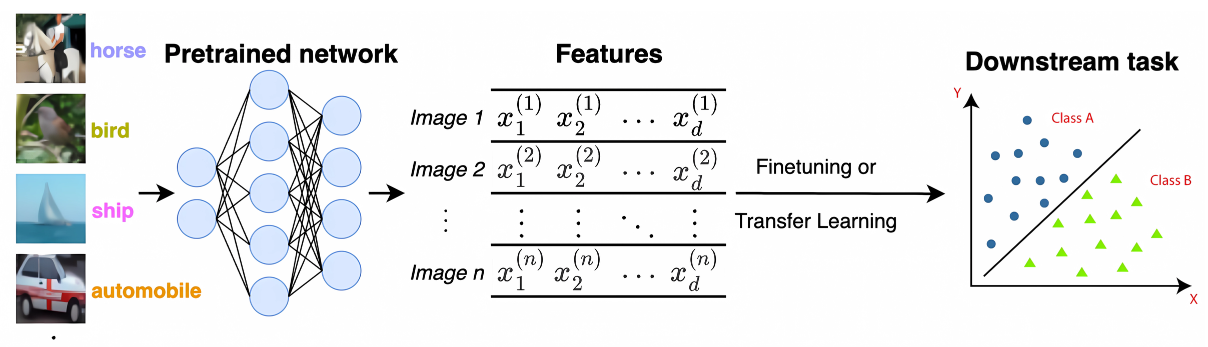
Jingyang Lyu*, Kangjie Zhou†, Yiqiao Zhong*

*Department of Statistics, University of Wisconsin–Madison, †Department of Statistics, Columbia University

arXiv GitHub

Challenge in Imbalanced Classification

Training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{\boldsymbol{x},y}$. Features $\boldsymbol{x}_i \in \mathbb{R}^d$. For binary labels $y_i \in \{\pm 1\}$.



Class imbalance. $\mathbb{P}(y_i = +1) < \mathbb{P}(y_i = -1)$. (WLOG, assume "+1" is minority)

Challenges of high dimensions: a brief summary of high dimensional statistical theory:

	Low dimensions	High dimensions
Parameter estimation	$\left\langle \frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|}, \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} \right\rangle \approx 1$	$\left\langle \frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|}, \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} \right\rangle < 1$
Generalization	Training error ≈ Test error	Training error < Test error
Distribution of logits	1D projection of $P_{\boldsymbol{x}}$	Skewed/distorted 1D projection of $P_{\boldsymbol{x}}$

Challenges of data imbalance: minority classes have poor training/test errors, classical theory and finite-sample correction fail in high dimensions, the practice is heuristic-driven and ad hoc...

Key Questions

[Q1]. Mathematically **characterize overfitting** in high-dim imbalanced classification ?
[Q2]. What are the adverse effects of overfitting, particularly on the **minority class** ?
[Q3]. What are the consequences for **uncertainty quantification**, such as calibration ?

Setup of Theory

Theoretical tools: consider a **two-component Gaussian mixture model (2-GMM)**

① Minority: $\mathbb{P}(y_i = +1) = \pi$, Majority: $\mathbb{P}(y_i = -1) = 1 - \pi$, ② $\boldsymbol{x}_i \,|\, y_i \sim \mathsf{N}(y_i \boldsymbol{\mu}, \mathbf{I}_d)$. (1)

Focus on linear classifier $\widehat{y}(\boldsymbol{x}) = 2\mathbb{1}\{\widehat{f}(\boldsymbol{x}) > 0\} - 1$ with $\widehat{f}(\boldsymbol{x}) = \langle \boldsymbol{x}, \widehat{\boldsymbol{\beta}} \rangle + \widehat{\beta}_0$, where $\widehat{\boldsymbol{\beta}}$, $\widehat{\beta}_0$ are estimated by two standard approaches: (generalized) logistic regression and support vector machines (SVMs). ($\ell(x)$ is strictly convex decreasing, including $\log(1 + e^{-x})$.)

logistic regression:
$$\underset{\boldsymbol{\beta} \in \mathbb{R}^d, \beta_0 \in \mathbb{R}}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^n \ell\big(y_i(\langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle + \beta_0)\big), \tag{2a}$$

SVM: (max-margin classifier)
$$\underset{\boldsymbol{\beta} \in \mathbb{R}^d, \beta_0, \kappa \in \mathbb{R}}{\text{maximize}} \quad \kappa, \quad \text{subject to} \quad y_i(\langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle + \beta_0) \ge \kappa, \quad \forall i \in [n], \quad \|\boldsymbol{\beta}\|_2 \le 1. \tag{2b}$$

These two classifiers are closely related by **inductive bias** on separable data. Our theory can also be extended to multiple classes and non-isotropic covariance.

Characterizing Overfitting via Empirical Logit Distribution

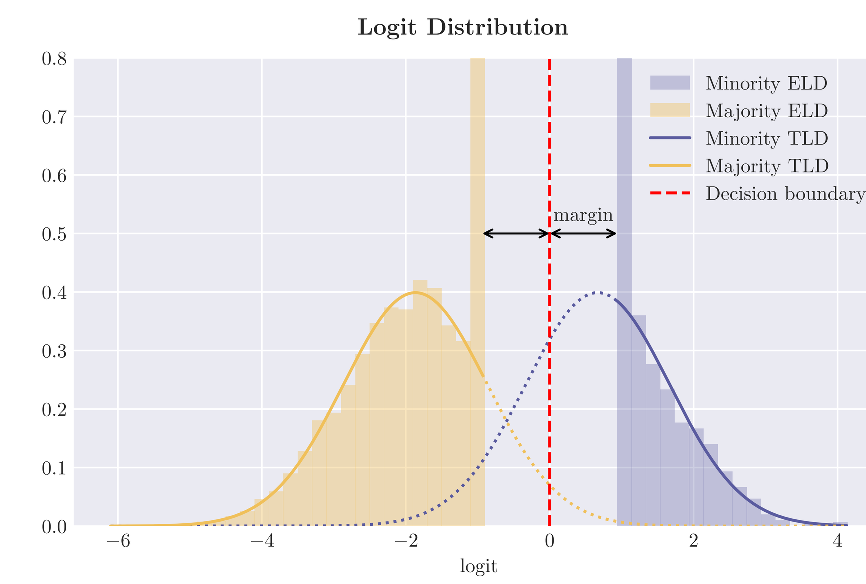
Empirical logit distribution (ELD), or *training logit distribution* is defined as the empirical distribution of label-logit pairs in the training set. ***Testing logit distribution (TLD)*** is defined as the distribution of the label-logit pair for a test data point.

$$\text{ELD:} \quad \widehat{\nu}_n = \frac{1}{n} \sum_{i=1}^n \delta_{(y_i, \widehat{f}(\boldsymbol{x}_i))}, \qquad \text{TLD:} \quad \widehat{\nu}_n^{\text{test}} = \mathtt{Law}\big(y_{\text{test}}, \widehat{f}(\boldsymbol{x}_{\text{test}})\big), \tag{3}$$

Let $\delta_{\boldsymbol{a}}$ be delta measure supported at $\boldsymbol{a}$, and $\mathtt{Law}$ be the distribution of random variables/vectors.

The **discrepancy** between train/test accuracies is known as **overfitting**, which can be analyzed via ELD and TLD. For separable data, see simulation in Figure 2 (2-GMM) and real-data examples in Figure 3 (pretrained neural network).

Takeaway: overfitting = "truncation". (for non-separable data: "shrinking/skewing" effect)



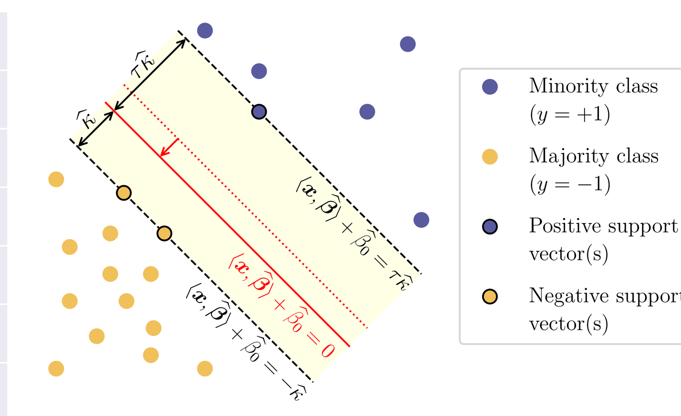


Figure 1. Margin-rebalanced SVM.

Figure 2. **Empirical logit distribution (ELD) and testing logit distribution (TLD).** We train a max-margin classifier (SVM) $\widehat{f}$ on synthetic data from a 2-component Gaussian mixture model. Colors indicate labels y_i and x-axis indicates logits $\widehat{f}(\boldsymbol{x}_i)$. **ELD for both classes:** the *rectified Gaussian* distribution (histogram). **TLD for both classes:** Gaussian distribution (curve). **Overfitting effect:** The density areas below the dotted curves are overlapping in TLD ⇒ test errors > 0; but they are "pushed" to respective margin boundaries in ELD ⇒ separability and training errors = 0.

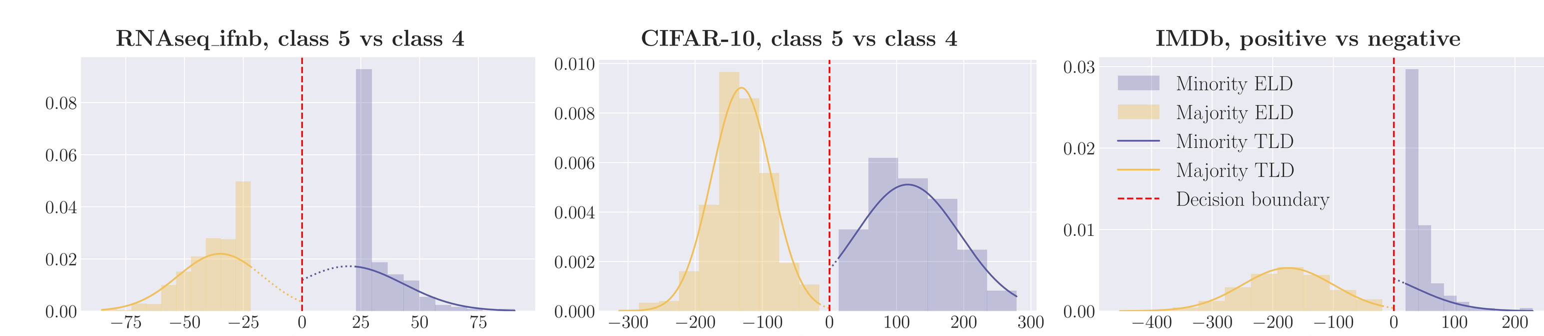


Figure 3. **ELD & TLD for real data**. **Left: IFNB single-cell** RNA-seq dataset (tabular data). **Middle: CIFAR-10** preprocessed by the pretrained **ResNet-18** for feature extraction (image data). **Right: IMDb** movie review preprocessed by **BERT** base model (110M) for feature extraction (text data).

Theoretical results: variational characterization of ELD vs. TLD

Let $(\widehat{\boldsymbol{\beta}}, \widehat{\beta}_0, \widehat{\kappa})$ be trained from (2b), where $\widehat{\kappa}$ is the **margin**. Denote $\widehat{\rho} := \langle \frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|}, \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|} \rangle$. On a test point $(\boldsymbol{x}_{\text{test}}, y_{\text{test}}) \sim P_{\boldsymbol{x},y}$, we consider the minority and majority errors

$$\mathrm{Err}_+ := \mathbb{P}\left(\widehat{f}(\boldsymbol{x}_{\text{test}}) \le 0 \,\middle|\, y_{\text{test}} = +1\right), \qquad \mathrm{Err}_- := \mathbb{P}\left(\widehat{f}(\boldsymbol{x}_{\text{test}}) > 0 \,\middle|\, y_{\text{test}} = -1\right). \tag{4}$$

Theorem (Separable data). Consider 2-GMM with $n/d \to \delta \in (0, \infty)$ as $n, d \to \infty$. There is a critical threshold $\delta_c = \delta_c(\pi, \|\boldsymbol{\mu}\|)$, such that if $\delta < \delta_c$, as $n, d \to \infty$:

a. Phase transition. $\mathbb{P}\{\text{training set is linearly separable}\} \to 1$.

b. Parameter convergence. $(\widehat{\rho}, \widehat{\beta}_0, \widehat{\kappa}) \xrightarrow{\text{p}} (\rho^*, \beta_0^*, \kappa^*)$, where $(\rho^*, \beta_0^*, \kappa^*)$ is unique optimal solution to the following **variational problem**: (we have $\rho^* > 0$, $\beta_0^* < 0$)
$$\underset{\rho \in [-1,1], \beta_0 \in \mathbb{R}, \kappa > 0, \xi \in \mathcal{L}^2}{\text{maximize}} \quad \kappa, \quad \text{s.t.} \quad \rho\|\boldsymbol{\mu}\| + G + Y\beta_0 + \sqrt{1 - \rho^2}\,\xi \ge \kappa, \quad \mathbb{E}[\xi^2] \le 1/\delta. \tag{5}$$
(Let $\mathcal{L}^2$ denote all square integrable random variables in $(\Omega, \mathcal{F}, \mathbb{P})$, and $(Y, G) \sim P_y \times \mathsf{N}(0,1)$ where $P_y = \mathtt{Law}(y_i)$. Note that ξ is an unknown random variable (function) to be optimized.)

c. Asymptotic errors. $\mathrm{Err}_\pm \to \Phi\left(-\rho^*\|\boldsymbol{\mu}\| \mp \beta_0^*\right)$. $(\Phi(t) = \mathbb{P}(\mathsf{N}(0,1) \le t))$

d. ELD convergence. The empirical (training) logit distribution $\widehat{\nu}_n$ has limit ν_*:
$$W_2(\widehat{\nu}_n, \nu_*) \xrightarrow{\text{p}} 0, \qquad \text{where } \nu_* := \mathtt{Law}\left(Y, Y\max\{\kappa^*, \rho^*\|\boldsymbol{\mu}\| + G + Y\beta_0^*\}\right).$$
TLD convergence. The testing logit distribution $\widehat{\nu}_n^{\text{test}}$ has limit ν_*^{test}:
$$\widehat{\nu}_n^{\text{test}} \xrightarrow{w} \nu_*^{\text{test}}, \qquad \text{where } \nu_*^{\text{test}} := \mathtt{Law}\left(Y, Y(\rho^*\|\boldsymbol{\mu}\| + G + Y\beta_0^*)\right).$$

Rebalancing margin is crucial

Mainstay: take a hyperparameter $\tau > 0$ and consider the **margin-rebalanced SVM**
$$\underset{\boldsymbol{\beta} \in \mathbb{R}^d, \beta_0, \kappa \in \mathbb{R}}{\text{maximize}} \quad \kappa, \quad \text{subject to} \quad \widetilde{y}_i(\langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle + \beta_0) \ge \kappa, \quad \forall i \in [n], \quad \|\boldsymbol{\beta}\| \le 1, \tag{6}$$
where $\widetilde{y}_i = \tau^{-1}$ if $y_i = +1$, otherwise $\widetilde{y}_i = -1$. This **shifts the decision boundary** as shown in Figure 1. For imbalanced classification, it is common to consider the ***balanced error*** $\mathrm{Err_b} := (\mathrm{Err}_+ + \mathrm{Err}_-)/2$. We conduct analysis under two regimes:

(i) proportional regime, where $n/d \to \delta \in (0, \infty)$ as $n, d \to \infty$. Denote Err_+^*, Err_-^*, $\mathrm{Err}_\mathrm{b}^*$ as the limits of Err_+, Err_-, $\mathrm{Err_b}$ as $n \to \infty$, respectively.

Proposition (Optimal τ in proportional regime). Define the optimal margin ratio which minimizes the asymptotic balanced error as $\tau^{\text{opt}} := \arg\min_\tau \mathrm{Err}_\mathrm{b}^*$. When $\tau = \tau^{\text{opt}} > 0$, we have $\beta_0^* = 0$, $\mathrm{Err}_+^* = \mathrm{Err}_-^* = \mathrm{Err}_\mathrm{b}^*$. (roughly speaking, $\tau^{\text{opt}} \asymp \sqrt{1/\pi}$)

A critical observation is that, changing τ only has an effect on $\widehat{\beta}_0$ but not $\widehat{\boldsymbol{\beta}}$. When $\tau = \tau^{\text{opt}}$, we have **monotone trends** of the errors, see summary in Table 1.

(ii) high imbalance regime, in the sense $\pi \asymp d^{-a}$, $\|\boldsymbol{\mu}\|^2 \asymp d^b$, $n \asymp d^{c+1}$ as $d \to \infty$.

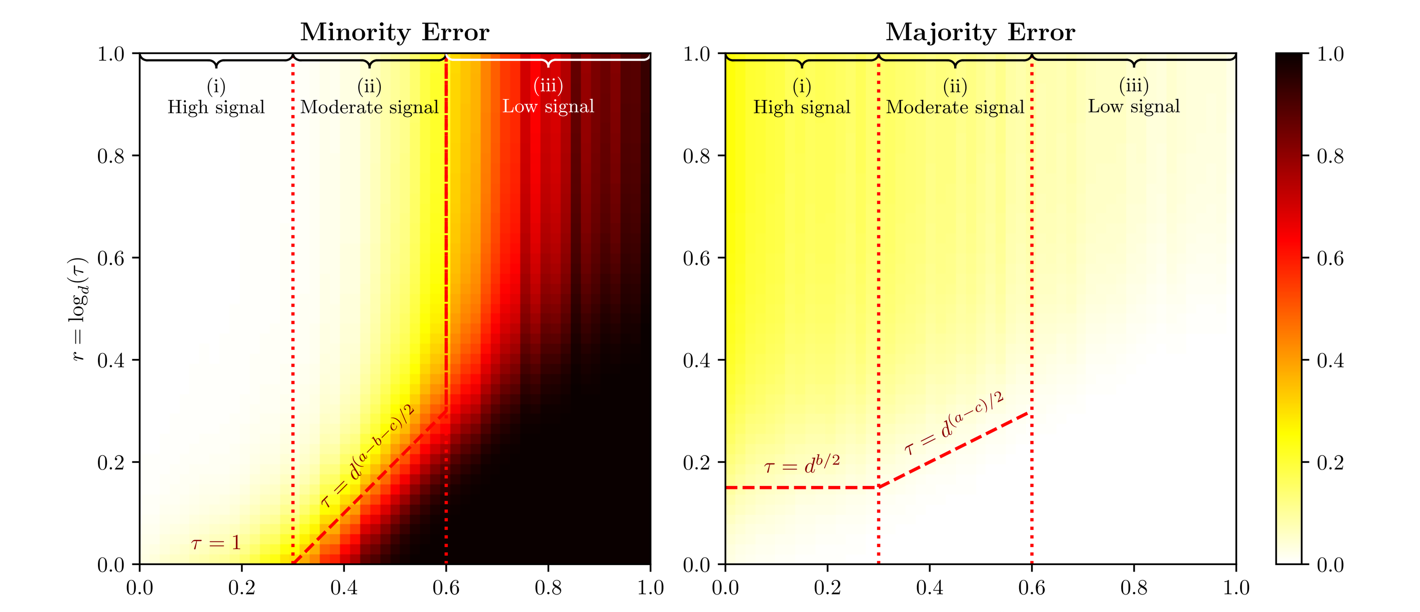


Figure 4. **Phase transition in high imbalance regime**. 2-GMM simulation under different settings of parameters (a, c) and $\tau = d^r$ $(b = 0.3)$. **Left:** minority accuracy is (i) high for any τ under high signal, (ii) high for $\tau \gg d^{(a-b-c)/2}$ under moderate signal, but (iii) low for any τ under low signal. **Right:** majority accuracy is close to 1 under high/moderate signal as long as τ is not too large.

Theorem (High imbalance). Consider 2-GMM as $d \to \infty$. Suppose $a - c < 1$.

(i) High signal: $a - c < b$. If take $1 \le \tau_d \ll d^{b/2}$, then $\mathrm{Err}_+ = o(1)$ and $\mathrm{Err}_- = o(1)$.

(ii) Moderate signal: $b < a - c < 2b$. If $d^{a-b-c} \ll \tau_d \ll d^{(a-c)/2}$, then $\mathrm{Err}_+ = o(1)$ and $\mathrm{Err}_- = o(1)$. If naively take $\tau_d \asymp 1$, then $\mathrm{Err}_+ = 1 - o(1)$ and $\mathrm{Err}_- = o(1)$.

(iii) Low signal: $a - c > 2b$. For any τ_d, we have $\mathrm{Err_b} \ge \frac{1}{2} - o(1)$.

Consequences for confidence estimation and calibration

Confidence: prediction probability, i.e., $\widehat{p}(\boldsymbol{x}) := \sigma(\widehat{f}(\boldsymbol{x}))$ where $\sigma(t) = (1 + e^{-t})^{-1}$.
Calibration: quantity uncertainty, measure faithfulness of prediction probabilities.

$\widehat{p}_0(\boldsymbol{x}) := \mathbb{P}(y = 1 \,|\, \widehat{p}(\boldsymbol{x}))$, $p^*(\boldsymbol{x}) := \mathbb{P}(y = 1 \,|\, \boldsymbol{x})$. Some popular miscalibration metrics: **calibration error** $\mathrm{CalErr}(\widehat{p}) := \mathbb{E}[(\widehat{p}(\boldsymbol{x}) - \widehat{p}_0(\boldsymbol{x}))^2]$, **mean squared error** $\mathrm{MSE}(\widehat{p}) := \mathbb{E}[(\mathbb{1}_{y=1} - \widehat{p}(\boldsymbol{x}))^2]$, and **confidence estimation error** $\mathrm{ConfErr}(\widehat{p}) := \mathbb{E}[(\widehat{p}(\boldsymbol{x}) - p^*(\boldsymbol{x}))^2]$.

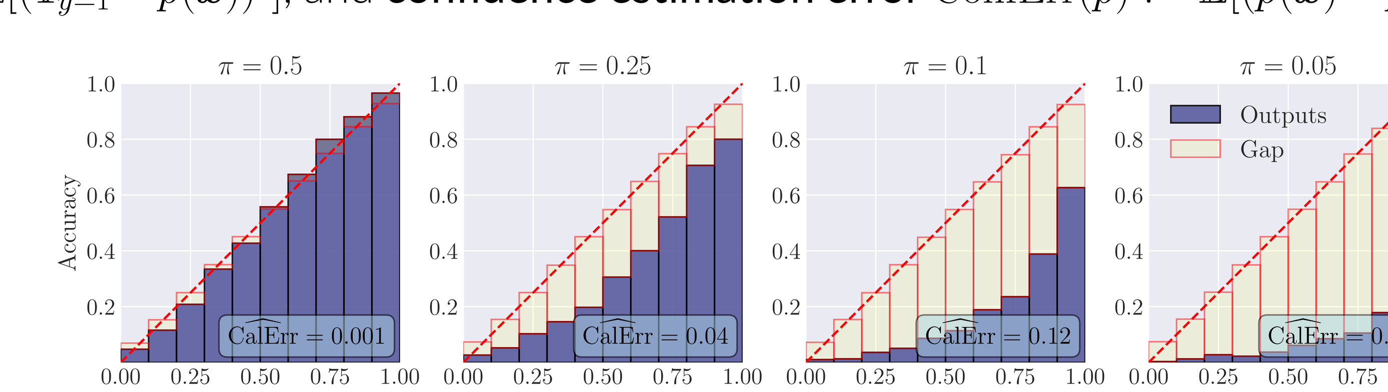


Figure 5. **Reliability diagrams: imbalance worsens calibration**. In 2-GMM simulations, we train SVMs and obtain confidence $\widehat{p}(\boldsymbol{x})$. For each p (x-axis), we calculate $\mathbb{P}(y = 1 \,|\, \widehat{p}(\boldsymbol{x}) = p)$ (y-axis) based on a test set. As imbalance increases (smaller π), the classifier becomes more miscalibrated.

Theoretical results:	$\mathrm{Err}_+^*, \mathrm{Err}_-^*, \mathrm{Err}_\mathrm{b}^*$	CalErr^*	MSE^*	$\mathrm{ConfErr}^*$
imbalance ratio π ↑	↓		↓	↓
signal strength $\|\boldsymbol{\mu}\|_2$ ↑	↓	↓	↓	
aspect ratio $n/d \to \delta$ ↑	↓	↓	↓	↓

Table 1. Monotonicity of test errors and miscalibration metrics on model parameters.