WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

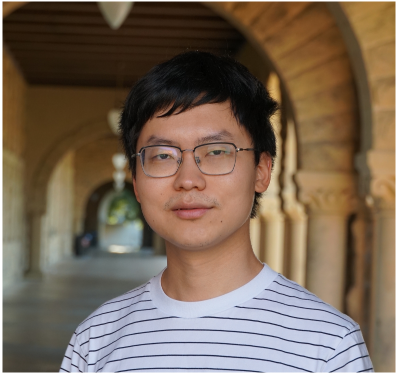
A statistical theory of overfitting for imbalanced classification

Jingyang Lyu[*] Kangjie Zhou[†] Yiqiao Zhong[*]

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

Department of Statistics
SCHOOL OF COMPUTER, DATA & INFORMATION SCIENCES
UNIVERSITY OF WISCONSIN–MADISON

Collaborators



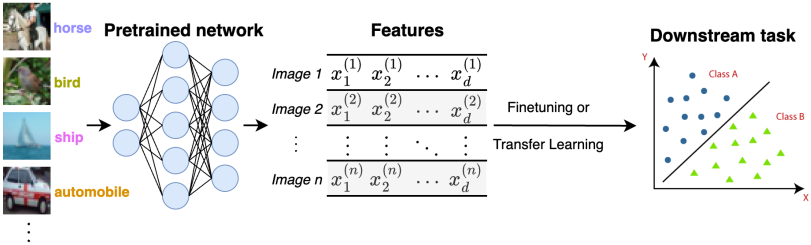
Kangjie Zhou, postdoc at Columbia U

Yiqiao Zhong, UW–Madison

Paper: https://arxiv.org/abs/2502.11323

Contents

Challenge 1: High dimensionality



High dimensional features are everywhere:

- Finetuning a classification layer in deep learning
- Linear probing, interpretability of LLMs
- Single-cell omics

Challenge 1: High dimensionality

	Low dimensions	**High dimensions**
Parameter estimation	$\left\langle \frac{\widehat{\boldsymbol{\beta}}}{\lVert\widehat{\boldsymbol{\beta}}\rVert}, \frac{\boldsymbol{\beta}}{\lVert\boldsymbol{\beta}\rVert} \right\rangle \approx 1$	$\left\langle \frac{\widehat{\boldsymbol{\beta}}}{\lVert\widehat{\boldsymbol{\beta}}\rVert}, \frac{\boldsymbol{\beta}}{\lVert\boldsymbol{\beta}\rVert} \right\rangle < 1$
Generalization	Train error $\approx$ Test error	Train error $<$ Test error

Table: Qualitative comparison for linear classification, $\boldsymbol{\beta}$ is the slope parameter vector.

The advances of high-dimensional statistics in the past 15 years.

- El Karoui el al. (2013), Donoho and Montanari (2016), Sur and Candés (2019)
- Double descent and benign overfitting: Belkin et al. (2019), Bartlett el al. (2020)
- Many more . . .

Q: New angles for the (overfitting) effects of dimensionality?

Challenge 2: Data imbalance

Real-world datasets are generally **imbalanced**.

- **Sentiment analysis.**

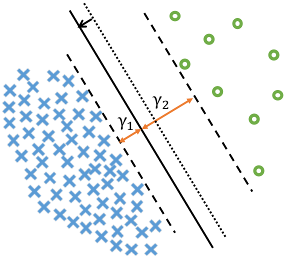
Dataset	Tweets	#Negative	#Positive
Stanford Twitter Test Set (STS-Test) [9]	359	177	182
Sanders Dataset (Sanders) [17]	1224	654	570
Obama McCain Debate (OMD) [7]	1906	1196	710
Health Care Reform (HCR) [22]	1922	1381	541
Stanford Gold Standard (STS-Gold) [17]	2034	632	1402
Sentiment Strength Twitter Dataset (SSTD) [23]	2289	1037	1252
The Dialogue Earth Weather Dataset (WAB) [3]	5495	2580	2915
The Dialogue Earth Gas Prices Dataset (GASP) [3]	6285	5235	1050
Semeval Dataset (Semeval) [14]	7535	2186	5349

Figure: Twitter datasets used for sentiment analysis [Saif et al. 2015]

- **Industrial fault detection** (failures $\ll$ normal operations)
- **Healthcare and medical diagnosis** (rare disease/genetic markers, privacy issue)

Challenge 2: Data imbalance

- **Minority classes** have worse training and testing errors.
- The classical asymptotic theory or finite-sample analysis is inaccurate in high dimensions.
- The practice is heuristic-driven and ad hoc.
  - Re-sampling: oversampling the minority or under-sampling the majority
  - Re-weighting: assigning higher weights for minority classes
  - Synthetic data: SMOTE (2002), Mixup (2018)
  - **Margin adjustment**: popular in deep learning.



Cao et al. 2019

Q: How to quantify the impact of factors (imbalance ratio, SNR, dimension) on accuracy?

Goals of this talk

Goal 1. Provide a new angle of **characterizing overfitting** for imbalanced classification.

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

Goal 2. Quantify the **adverse effects** of overfitting, esp. for the minority class.

$$\begin{array}{r} \boxed{\text{dimensions v.s. sample size}} \\ \boxed{\text{degree of imbalance}} \\ \boxed{\text{signal strength}} \end{array} \quad \overset{?}{\Longrightarrow} \quad \begin{array}{l} \boxed{\text{test accuracy}} \\ \boxed{\text{uncertainty quantification}} \end{array}$$

Contents

Binary classification

- Training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \overset{\text{i.i.d.}}{\sim} P_{\boldsymbol{x},y}$.
  - $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \{+1, -1\}$
- **Imbalance ratio:** denote $\pi = \mathbb{P}(y_i = +1)$.
  - The classification is imbalanced if $\pi < 1/2$.

$$y_i = \begin{cases} +1 & \text{with prob } \quad \pi \quad \text{(minority)} \\ -1 & \text{with prob } 1 - \pi \quad \text{(majority)} \end{cases}$$

- Build a classifier based on $f : \mathbb{R}^d \to \mathbb{R}$.

  For a point $\boldsymbol{x}$, the predicted label is

$$\widehat{y}(\boldsymbol{x}) = \begin{cases} +1 & \text{if } f(\boldsymbol{x}) > 0 \\ -1 & \text{if } f(\boldsymbol{x}) \le 0 \end{cases}$$

Two linear classifiers

- We focus on two linear classifiers.

$$
\begin{aligned}
\textbf{(logistic regression)} \quad & \underset{\boldsymbol{\beta}\in\mathbb{R}^d, \beta_0\in\mathbb{R}}{\text{minimize}} \quad && \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i(\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle + \beta_0)\big), \\
\textbf{(SVM)} \quad & \underset{\boldsymbol{\beta}\in\mathbb{R}^d,\, \beta_0, \kappa\in\mathbb{R}}{\text{maximize}} \quad && \kappa, \\
& \text{subject to} \quad && y_i(\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle + \beta_0) \ge \kappa, \quad \forall\, 1 \le i \le n, \\
& && \|\boldsymbol{\beta}\|_2 \le 1.
\end{aligned}
$$

- **Connection** (inductive bias): when training data is linear separable,

$$
\text{SVM} \;=\; \text{Max-margin classifier} \;=\; \text{Ridgeless logistic regression}
$$

Contents

Logits and empirical logit distribution

For any classifier $\widehat{y}(\boldsymbol{x}) = 2\mathbb{1}\{\widehat{f}(\boldsymbol{x}) > 0\} - 1$ (e.g., SVM, neural network, language model)

- **Logit** for point $\boldsymbol{x}$: $\widehat{f}(\boldsymbol{x})$
- **Margin**: $\widehat{\kappa}_n = \min_{1 \le i \le n} y_i \widehat{f}(\boldsymbol{x}_i)$
  - When $\widehat{\kappa}_n > 0$, the training set is **linearly separable**.

Definition (Empirical logit distribution, or ELD)

For any binary classifier $\widehat{y}(\boldsymbol{x})$ built on $\widehat{f}(\boldsymbol{x})$, the *empirical logit distribution* is defined as

$$\widehat{\nu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(y_i, \widehat{f}(\boldsymbol{x}_i))} \tag{2}$$

where $\delta_{\boldsymbol{a}}$ denotes the delta measure supported at point $\boldsymbol{a}$.

Empirical logit distribution v.s. testing logit distribution

Let $(\boldsymbol{x}_{\text{test}}, y_{\text{test}}) \sim P_{\boldsymbol{x},y}$ be a new data point.

- **Overfitting** can be characterized by discrepancy between

$$\underbrace{\widehat{\nu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(y_i, \widehat{f}(\boldsymbol{x}_i))}}_{\substack{\text{empirical logit distribution} \\ \text{(“training” logit distribution)}}} \qquad \text{and} \qquad \underbrace{\widehat{\nu}_n^{\text{test}} = \mathtt{Law}\left(y_{\text{test}}, \widehat{f}(\boldsymbol{x}_{\text{test}})\right)}_{\text{testing logit distribution}}$$

- Note: both $\widehat{\nu}_n$, $\widehat{\nu}_n^{\text{test}}$ are random measures.
  - Since $\widehat{f}$ depends on training set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$.

Empirical phenomenon: Simulation

Settings:

1. Generate a **(linearly) separable** training set from a Gaussian mixture model (GMM):

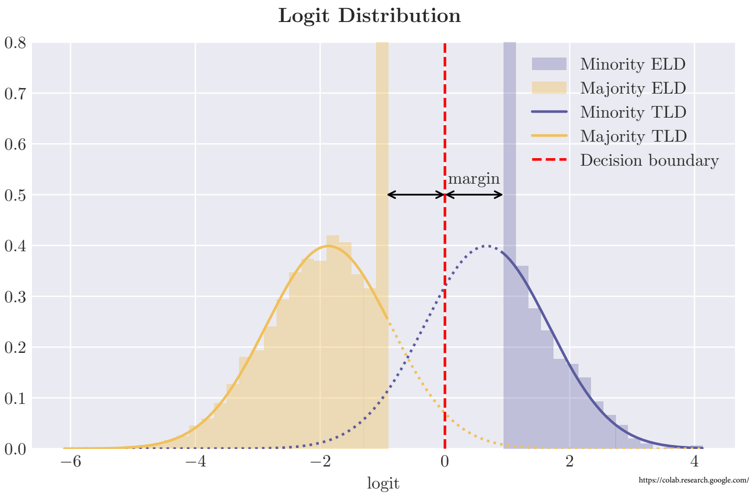
$$y_i = \begin{cases} +1, & \text{w.p. } \quad \pi \quad \text{(minority)} \\ -1, & \text{w.p. } 1-\pi \quad \text{(majority)} \end{cases}, \qquad \boldsymbol{x}_i \mid y_i \sim \mathcal{N}(y_i \boldsymbol{\mu}, \mathbf{I}_d), \qquad i = 1, 2, \dots, n.$$

2. Train a **max-margin classifier (SVM)**: $\Longrightarrow$ $\widehat{\boldsymbol{\beta}}, \widehat{\beta}_0, \widehat{\kappa}$

$$\begin{aligned} \underset{\boldsymbol{\beta} \in \mathbb{R}^d, \beta_0 \in \mathbb{R}, \kappa \in \mathbb{R}}{\text{maximize}} \quad & \kappa, \\ \text{subject to} \quad & y_i(\langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle + \beta_0) \ge \kappa, \quad \forall\, 1 \le i \le n, \\ & \|\boldsymbol{\beta}\|_2 \le 1. \end{aligned}$$

3. Compare empirical / testing logit distribution for $\widehat{f}(\boldsymbol{x}) = \langle \boldsymbol{x}, \widehat{\boldsymbol{\beta}} \rangle + \widehat{\beta}_0$.

Empirical phenomenon: Simulation



Code

https://colab.research.google.com/drive/1DQ1HvOtRCb8fXw6TBF3-Pc9Ozdv5kEM5?usp=sharing

Figure: Empirical (training) and testing logit distribution for binary Gaussian mixture model

Empirical phenomenon: tubular data

RNA-seq `ifnb` dataset with logistic regression ($\pi = 0.2$)

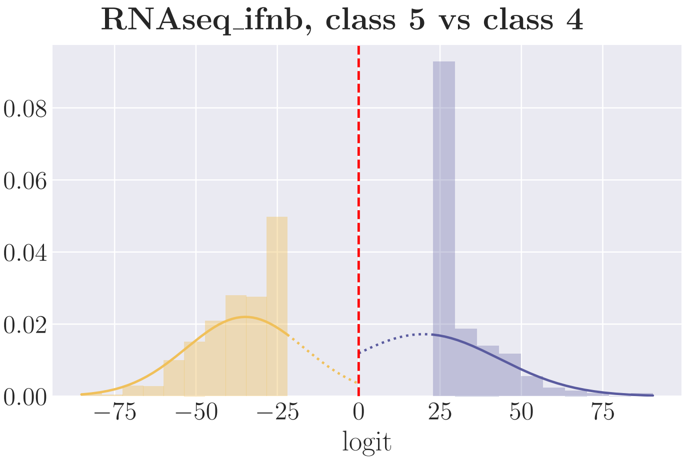


Figure: Empirical (training) and testing logit distribution for single-cell dataset

Empirical phenomenon: image data

ResNet-18 trained on CIFAR-10 ($\pi = 0.1$)

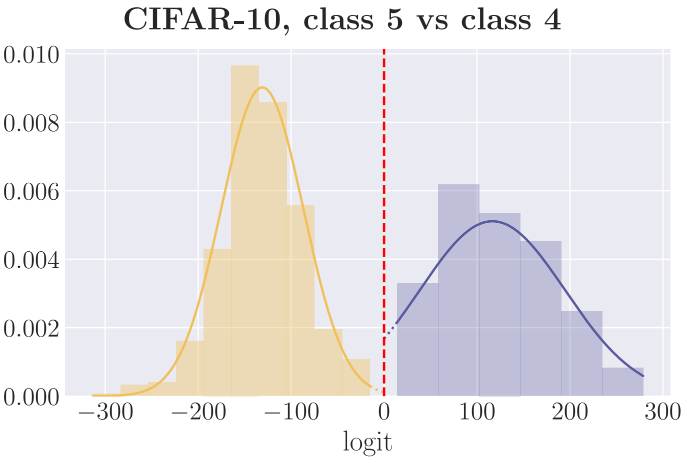


Figure: Empirical (training) and testing logit distribution for CIFAR-10 dataset

Empirical phenomenon: text data

BERT(110M) trained on IMDb movie reviews ($\pi = 0.02$)

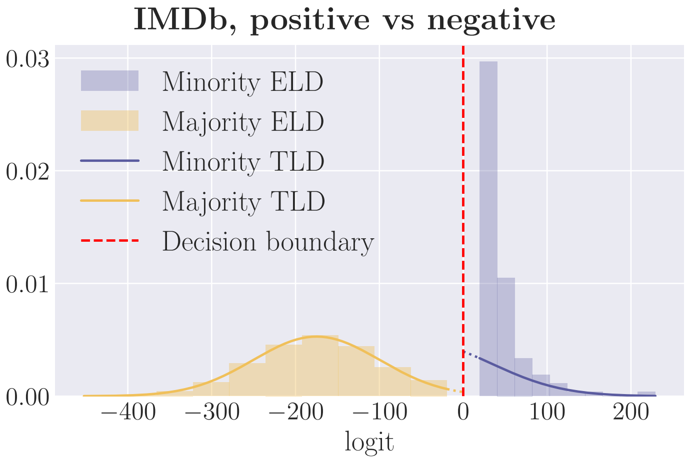


Figure: Empirical (training) and testing logit distribution for IMDb dataset

Theoretical foundation

Consider GMM with asymptotic regime $n/d \to \delta \in (0, \infty)$.

- Recall $\widehat{\kappa} = \min_{1 \le i \le n} y_i(\langle \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}} \rangle + \widehat{\beta}_0)$. Denote $\widehat{\rho} = \left\langle \frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|}, \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|} \right\rangle$. ($\|\widehat{\boldsymbol{\beta}}\| = 1$ when separable)
- We may expect $(\widehat{\rho}, \widehat{\beta}_0, \widehat{\kappa})$ converge to some limit $(\rho^*, \beta_0^*, \kappa^*)$ as $n, d \to \infty$.

Let $(\boldsymbol{x}_{\mathrm{test}}, y_{\mathrm{test}})$ be a new testing point, then

$$
\begin{aligned}
y_{\mathrm{test}} \left(\langle \boldsymbol{x}_{\mathrm{test}}, \widehat{\boldsymbol{\beta}} \rangle + \widehat{\beta}_0 \right) &= y_{\mathrm{test}} \left\langle y_{\mathrm{test}} \boldsymbol{\mu} + \mathcal{N}(\mathbf{0}, \mathbf{I}_d),\ \widehat{\boldsymbol{\beta}} \right\rangle + y_{\mathrm{test}} \widehat{\beta}_0 \\
&= \widehat{\rho}\, \|\boldsymbol{\mu}\| + \left\langle \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \widehat{\boldsymbol{\beta}} \right\rangle + y_{\mathrm{test}} \widehat{\beta}_0 \\
&\approx \rho^* \|\boldsymbol{\mu}\| + G + Y \beta_0^*, \qquad \text{where } (Y, G) \sim P_y \times \mathcal{N}(0, 1).
\end{aligned}
$$

Theoretical foundation

For a testing point $(\boldsymbol{x}_{\text{test}}, y_{\text{test}})$,

$$y_{\text{test}} \left(\langle \boldsymbol{x}_{\text{test}}, \widehat{\boldsymbol{\beta}} \rangle + \widehat{\beta}_0 \right) \approx \rho^* \|\boldsymbol{\mu}\| + G + Y \beta_0^* . \tag{$\widehat{\nu}_n^{\text{test}}$}$$

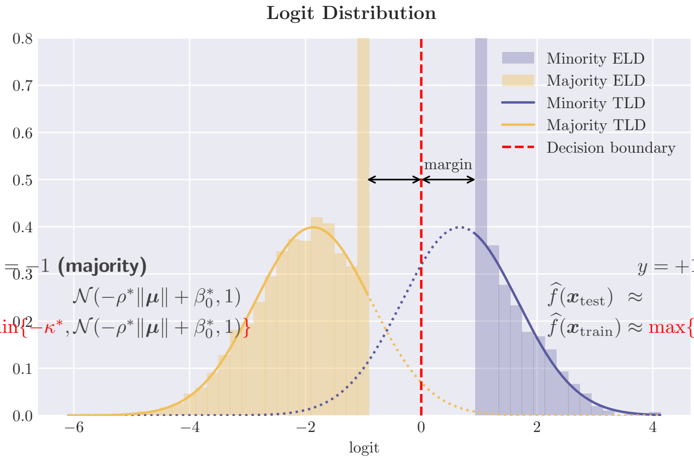
However, for a training point $(\boldsymbol{x}_i, y_i)$,

- There is a **distortion effect** on the distribution due to dependence between $(\boldsymbol{x}_i, y_i)$ and $\widehat{f}$.

$$y_i \left(\langle \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}} \rangle + \widehat{\beta}_0 \right) \approx \max\left\{ \kappa^*, \rho^* \|\boldsymbol{\mu}\| + G + Y \beta_0^* \right\} . \tag{$\widehat{\nu}_n$}$$

Key takeaway

Overfitting = "Truncation"



$y = -1$ **(majority)**

$$\widehat{f}(\boldsymbol{x}_{\text{test}}) \approx \mathcal{N}(-\rho^*\|\boldsymbol{\mu}\| + \beta_0^*, 1)$$

$$\widehat{f}(\boldsymbol{x}_{\text{train}}) \approx \min\{-\kappa^*, \mathcal{N}(-\rho^*\|\boldsymbol{\mu}\| + \beta_0^*, 1)\}$$

$y = +1$ **(minority)**

$$\widehat{f}(\boldsymbol{x}_{\text{test}}) \approx \mathcal{N}(\rho^*\|\boldsymbol{\mu}\| + \beta_0^*, 1)$$

$$\widehat{f}(\boldsymbol{x}_{\text{train}}) \approx \max\{\kappa^*, \mathcal{N}(\rho^*\|\boldsymbol{\mu}\| + \beta_0^*, 1)\}$$

Theoretical foundation

Theorem (Separable regime, simplified ver.)

Consider GMM with asymptotic regime $n/d \to \delta \in (0, \infty)$.

(a) **(Phase transition)** *There is a critical threshold* $\delta_c = \delta_c(\|\boldsymbol{\mu}\|, \pi)$, *such that*

$$\mathbb{P}\left\{\textit{training set is linearly separable}\right\} \to 1, \qquad \textit{if} \;\; \delta < \delta_c.$$

(b) **(Parameter convergence)** *If* $\delta < \delta_c$, *then* $(\widehat{\rho}, \widehat{\beta}_0, \widehat{\kappa}) \xrightarrow{\mathrm{p}} (\rho^*, \beta_0^*, \kappa^*)$, *where* $(\rho^*, \beta_0^*, \kappa^*)$ *is the unique solution of the following variational optimization problem:*

$$\begin{array}{ll} \underset{\rho \in [-1,1], \beta_0 \in \mathbb{R}, \kappa > 0, \xi \in \mathcal{L}^2}{\text{maximize}} & \kappa, \\ \textit{subject to} & \rho\|\boldsymbol{\mu}\| + G + Y\beta_0 + \sqrt{1-\rho^2}\,\xi \ge \kappa, \qquad \mathbb{E}[\xi^2] \le 1/\delta. \end{array}$$

(c) **(ELD convergence)** *If* $\delta < \delta_c$, *denote* $\nu^* = \max\{\kappa^*, \rho^*\|\boldsymbol{\mu}\| + G + Y\beta_0^*\}$. *Then* $W_2(\widehat{\nu}_n, \nu^*) \xrightarrow{\mathrm{p}} 0$.

Theoretical foundation: remarks

$$\begin{array}{lll} \underset{\boldsymbol{\beta}\in\mathbb{R}^d,\beta_0\in\mathbb{R},\kappa\in\mathbb{R}}{\text{maximize}} & \kappa, & \\ \text{subject to} & \forall\, 1 \le i \le n \quad y_i(\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle + \beta_0) \ge \kappa, \qquad \|\boldsymbol{\beta}\|_2 \le 1. & \text{(A)} \end{array}$$

$$\begin{array}{lll} \underset{\rho\in[-1,1],\beta_0\in\mathbb{R},\kappa>0,\xi\in\mathcal{L}^2}{\text{maximize}} & \kappa, & \\ \text{subject to} & \rho\|\boldsymbol{\mu}\| + G + Y\beta_0 + \sqrt{1-\rho^2}\,\xi \ge \kappa, \qquad \mathbb{E}[\xi^2] \le 1/\delta. & \text{(B)} \end{array}$$

- In (B), it can be shown that $\sqrt{1-\rho^2}\,\xi = (\kappa - \rho\,\|\boldsymbol{\mu}\| - G - Y\beta_0)_+$ $\qquad (t)_+ = \max\{0, t\}$.

  $\Rightarrow$ The random variable ξ represents the **overfitting effect** in high dimensions.
- In (B), $\beta_0^* < 0$. The mean of minority testing logits is *closer to margin* than majority.

  $\Rightarrow$ **Overfitting hurts minority** class more than majority.

Theoretical foundations: non-separable regime

Logistic regression: we obtained similar variational formulation in the limit.

$$\underset{\rho\in[-1,1],R\geq 0,\beta_0\in\mathbb{R},\xi\in\mathcal{L}^2}{\text{minimize}} \quad \mathbb{E}\left[\ell\left(\rho\,\|\boldsymbol{\mu}\|_2\,R + RG + Y\beta_0 + R\sqrt{1-\rho^2}\xi\right)\right],$$

$$\text{subject to} \quad \mathbb{E}\left[\xi^2\right] \leq 1/\delta.$$

$$\mathrm{prox}_{\lambda\ell}(x) := \underset{t\in\mathbb{R}}{\arg\min}\left\{\ell(t) + \tfrac{1}{2\lambda}(t-x)^2\right\}$$

Proximal operator instead of truncation characterizes overfitting effects.

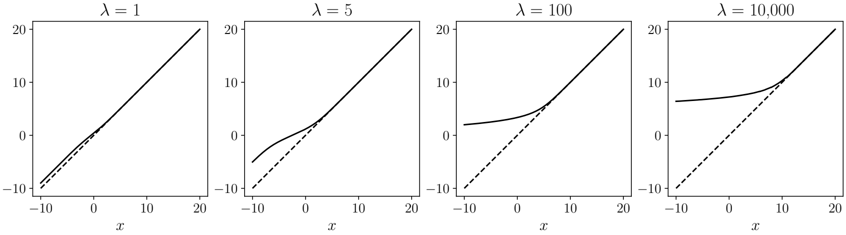


Figure: Plots of proximal operator $x \mapsto \mathrm{prox}_{\lambda\ell}(x)$ where λ represents the strength of overfitting.

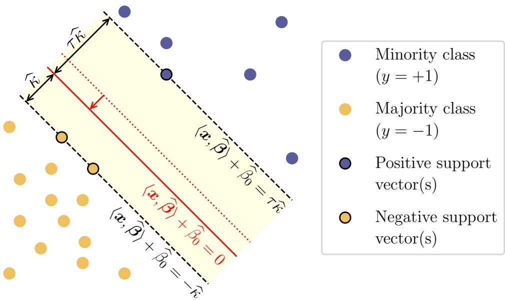
Contents

Rebalancing margin

Rebalancing margin is crucial in separable regime.

Consider **margin-rebalanced SVM**:

$$
\begin{aligned}
\underset{\boldsymbol{\beta}\in\mathbb{R}^d,\beta_0\in\mathbb{R},\kappa\in\mathbb{R}}{\text{maximize}} \quad & \kappa, \\
\text{subject to} \quad & y_i(\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle + \beta_0) \ge {\color{red}\tau}\kappa, \quad \forall\, i : y_i = +1 \\
& y_i(\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle + \beta_0) \ge \ \ \kappa, \quad \forall\, i : y_i = -1 \\
& \|\boldsymbol{\beta}\|_2 \le 1.
\end{aligned}
$$



Margin ratio: $\tau > 0$.

- **Note:** $\widehat{\beta}$ does not depend on τ.
- Question: what is the optimal τ?

Classification errors

		Exact testing error		Asymptotic testing error
	(minority error)	$\mathrm{Err}_+ = \mathbb{P}(\widehat{y}(\boldsymbol{x}) \neq y \mid y = +1)$	$\to$	$\mathrm{Err}_+^* = \Phi(-\rho^* \|\boldsymbol{\mu}\| - \beta_0^*)$
	(majority error)	$\mathrm{Err}_- = \mathbb{P}(\widehat{y}(\boldsymbol{x}) \neq y \mid y = -1)$	$\to$	$\mathrm{Err}_-^* = \Phi(-\rho^* \|\boldsymbol{\mu}\| + \beta_0^*)$
✗	(total error)	$\mathrm{Err} = \mathbb{P}(\widehat{y}(\boldsymbol{x}) \neq y)$	$\to$	$\mathrm{Err}^* = \pi \mathrm{Err}_+^* + (1-\pi)\mathrm{Err}_-^*$
✓	(balanced error)	$\mathrm{Err}_\mathrm{b} = \frac{1}{2}\mathrm{Err}_+ + \frac{1}{2}\mathrm{Err}_-$	$\to$	$\mathrm{Err}_\mathrm{b}^* = \frac{1}{2}\mathrm{Err}_+^* + \frac{1}{2}\mathrm{Err}_-^*$

Setting 1: proportional regime

Simulations

Setup: sample size $n = 100$, dimension $d = 200$. Run SVM, report errors over 100 runs.

Figure: **Effects of margin rebalancing on test errors**.

Setting 1: proportional regime
Simulations

Setup: sample size $n = 100$, dimension $d = 200$. Run SVM, report errors over 100 runs.

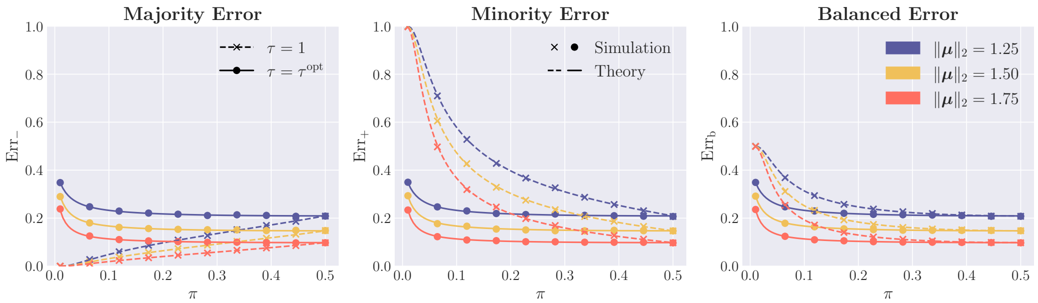


Figure: **Impact of imbalance on test errors**.

Proposition (Proportional regime)

Define τ^{opt} as the optimal margin ratio which minimizes the asymptotic balanced error

$$\tau^{\mathrm{opt}} := \operatorname*{arg\,min}_{\tau \geq 1} \mathrm{Err}_{\mathrm{b}}^{*} = \operatorname*{arg\,min}_{\tau \geq 1} \left\{ \Phi(-\left\|\boldsymbol{\mu}\right\|_2 \rho^{*} - \beta_0^{*}) + \Phi(-\left\|\boldsymbol{\mu}\right\|_2 \rho^{*} + \beta_0^{*}) \right\}.$$

(a) *When $\tau = \tau^{\mathrm{opt}}$, we have $\beta_0^{*} = 0$ and $\mathrm{Err}_{+}^{*} = \mathrm{Err}_{-}^{*} = \mathrm{Err}_{\mathrm{b}}^{*}$. In particular,*

$$\tau^{\mathrm{opt}} = \frac{g_1^{-1}\left(\dfrac{\rho^{*}}{2\pi \left\|\boldsymbol{\mu}\right\|_2 \delta}\right) + \rho^{*} \left\|\boldsymbol{\mu}\right\|_2}{g_1^{-1}\left(\dfrac{\rho^{*}}{2(1-\pi) \left\|\boldsymbol{\mu}\right\|_2 \delta}\right) + \rho^{*} \left\|\boldsymbol{\mu}\right\|_2}, \qquad \text{where} \qquad \begin{gathered} g_1(t) = \mathbb{E}[(G+t)_+] \\ G \sim \mathcal{N}(0,1),\ (t)_+ = 0 \vee t \end{gathered}$$

(b) *When $\tau = \tau^{\mathrm{opt}}$, the testing error $\mathrm{Err}_{\mathrm{b}}^{*}$ is a* **decreasing** *function of $\left\|\boldsymbol{\mu}\right\|_2$ (signal strength), δ (aspect ratio) and $\pi \in (0, 1/2)$ (imbalance ratio).*

- When π is small, roughly speaking $\tau^{\mathrm{opt}} \asymp 1/\sqrt{\pi}$.

Setting 2: high imbalance

$\pi \to 0$, $\|\boldsymbol{\mu}\| \to \infty$, $\delta = n/d \to \infty$

- Motivation: in overparametrized model, the imbalance ratio (π) is **vanishingly small** relative to dimension (d) and sample size (n).

Under Gaussian mixture model, consider $(a, b, c > 0)$

$$\boxed{\pi \asymp d^{-a}, \qquad \|\boldsymbol{\mu}\|^2 \asymp d^b, \qquad n \asymp d^{c+1}.}$$

- Such high imbalance dataset is **always separable** (with high probability).
- Feature distribution can be generalized to **sub-Gaussian**.

Setting 2: high imbalance: phase transition

Theorem (High imbalance regime, sub-Gaussian mixture model)

Suppose that $a - c < 1$ *(i.e.* $n\pi \to \infty$*).*

(a) **High signal** *(no need for margin rebalancing):* $\boxed{a - c < b}$. *If* $1 \leq \tau_d \ll d^{b/2}$, *then*

$$\mathrm{Err}_+^* = o(1), \qquad \mathrm{Err}_-^* = o(1).$$

(b) **Moderate signal** *(margin rebalancing is crucial):* $\boxed{b < a - c < 2b}$. *If we choose* $d^{a-b-c} \ll \tau_d \ll d^{(a-c)/2}$, *then*

$$\mathrm{Err}_+^* = o(1), \qquad \mathrm{Err}_-^* = o(1).$$

However, if we naively choose $\tau_d \asymp 1$, *then*

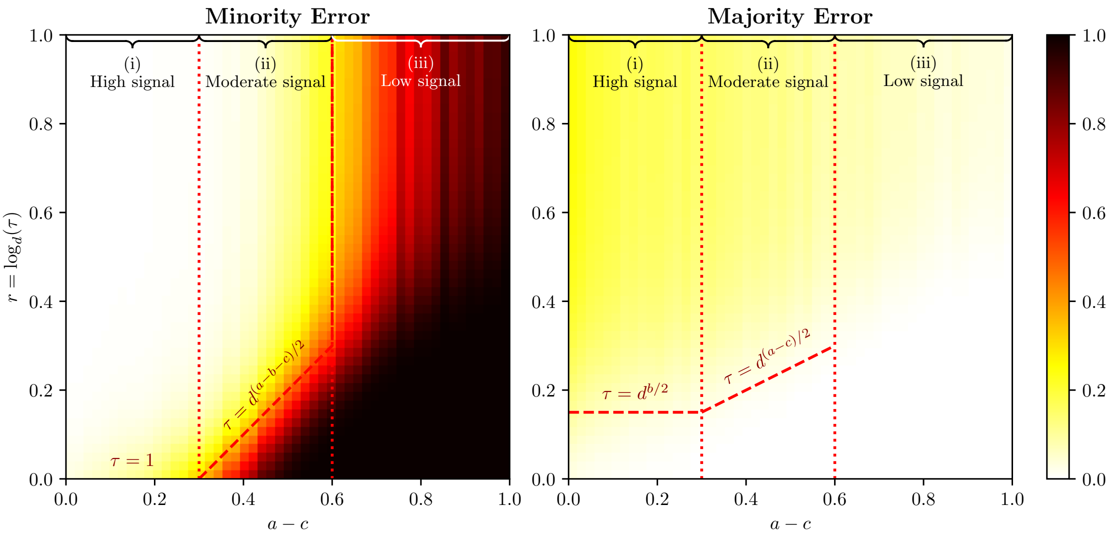
$$\mathrm{Err}_+^* = 1 - o(1), \qquad \mathrm{Err}_-^* = o(1).$$

(c) **Low signal** *(no better than random guess):* $\boxed{a - c > 2b}$. *For any* τ_d, *we have*

$$\mathrm{Err}_{\mathrm{b}}^* \geq \frac{1}{2} - o(1).$$

Simulation: $\tau = d^r$

$\pi \asymp d^{-a}$, $\|\boldsymbol{\mu}\| \asymp d^{b/2}$, $n \asymp d^{c+1}$ (fix $b = 0.3$, $c = 0.1$, $d = 2000$)



Contents

Confidence and Calibration

Confidence (predicted probability)

- Multiclass classification: `softmax`
- Binary classification: sigmoid transformation $p(\boldsymbol{x}) = 1/\big[1 + \exp\big(-f(\boldsymbol{x})\big)\big]$.

Ideally, we expect $p(\boldsymbol{x}) \approx \mathbb{P}(y = 1 \mid \boldsymbol{x})$. But the RHS is often intractable in high dim.

Definition (calibration)

A function $p : \mathcal{X} \to [0, 1]$ is *(perfectly) calibrated* if

$$p(\boldsymbol{x}) = \mathbb{P}(y = 1 \mid p(\boldsymbol{x})) \qquad \text{a.s.}$$

Intuition: Given 1,000 predictions, each with confidence of 0.2, we expect that about 200 should be classified as positive.

- Most informative example: $p(\boldsymbol{x}) = \mathbb{P}(y = 1 \mid \boldsymbol{x})$.
- Least informative example: $p(\boldsymbol{x}) \equiv \mathbb{P}(y = 1) = \pi$.

Calibration and other uncertainty measurements

Calibration error (CE).

$$\mathrm{CE}(p) = \mathbb{E}\left[\Big(\mathbb{P}\big(y = 1 \mid p(\boldsymbol{x})\big) - p(\boldsymbol{x})\Big)^2\right]$$

- Calibration itself does not guarantee a useful predictor, e.g., $p(\boldsymbol{x}) = \pi$.
- The variance in y explained by prediction $p(\boldsymbol{x})$ shouldn't be too small (**Sharpness**).

Mean squared error (MSE).

$$\mathrm{MSE}(p) = \mathbb{E}\left[\big(\mathbb{1}\{y = 1\} - p(\boldsymbol{x})\big)^2\right]$$

Confidence estimation error (ConfErr).

$$\mathrm{ConfErr}(p) := \mathbb{E}\left[\big(\mathbb{P}(y = 1 \mid \boldsymbol{x}) - p(\boldsymbol{x})\big)^2\right] .$$

Calibration: simulation

Setup: 2-GMM, $n = 1,000$, $d = 500$, $\pi = 0.05$, $\|\boldsymbol{\mu}\| = 1$, train SVM with $\tau = \tau^{\mathrm{opt}}$.

Reliability diagrams: For each p (x-axis), calculate $\mathbb{P}(y = 1 \mid \widehat{p}(\boldsymbol{x}) = p)$ (y-axis) on test set.

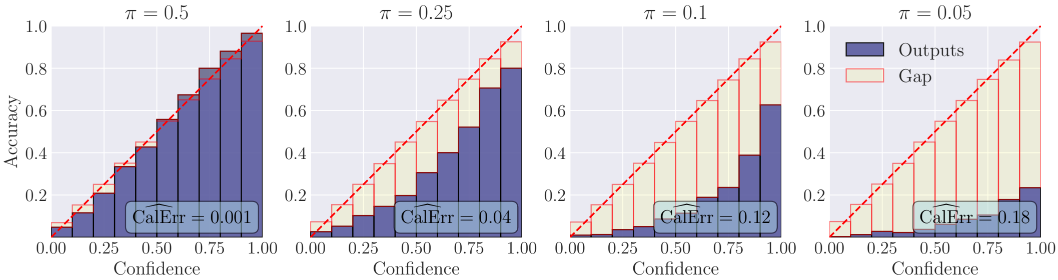


Figure: **Imbalance worsens calibration**.

Confidence and calibration: Theoretical foundations

Under proportional regime $n, d \to \infty$, $n/d \to \delta$, we show:

	$\mathrm{Err}_+^*, \mathrm{Err}_-^*, \mathrm{Err}_\mathrm{b}^*$	CE^*	MSE^*	$\mathrm{ConfErr}^*$
imbalance ratio $\pi \uparrow$	$\downarrow$		$\downarrow$	$\downarrow$
signal strength $\|\boldsymbol{\mu}\|_2 \uparrow$	$\downarrow$	$\downarrow$	$\downarrow$	
aspect ratio $n/d \to \delta \uparrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$

Table: Monotonicity of test errors and confidence/calibration metrics

Qualitatively, the effects of imbalance is similar to signal strength and effective sample size.

Takeaway message

$\boxed{\text{increasing dimension}}$ $(\delta = \frac{n}{d} \downarrow)$

$\boxed{\text{more severe imbalance}}$ $(\pi \downarrow)$

$\boxed{\text{low signal strength}}$ $(\|\boldsymbol{\mu}\| \downarrow)$

$\Longrightarrow$ $\boxed{\textbf{More severe overfitting}}$ $\Longrightarrow$

$\boxed{\text{test accuracy} \downarrow}$

$\boxed{\text{poor calibrated}}$

Contents

Generalization

- **Non-isotropic covariance.**
  - We obtained a variational form based on formal calculation.
  - Dependence on the covariance spike and direction is complicated.
- **Multiclass classification.**
  - Truncation for 2-dim Gaussian can be observed for empirical logit distribution.

Multiclass classification: CIFAR-10

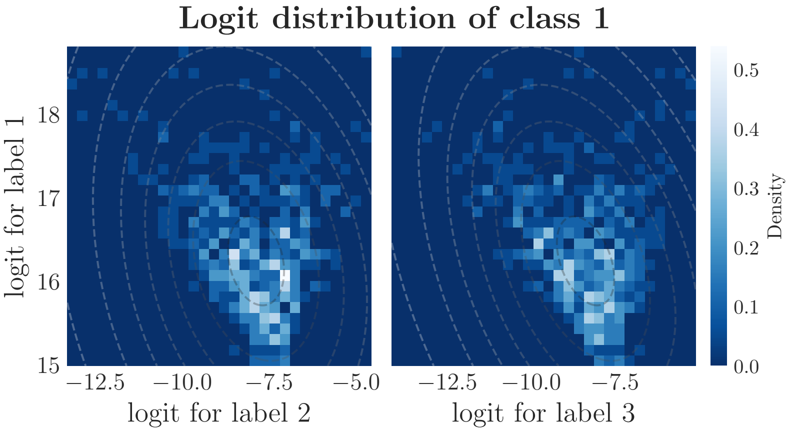


Figure: Joint logit distribution

Multiclass classification: GMM

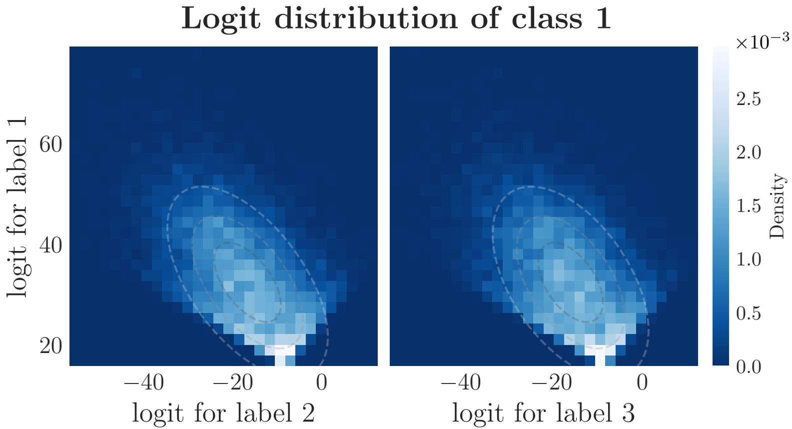


Figure: Joint logit distribution

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Thank you for listening.

ArXiv paper

GitHub page

References

- El Karoui, Noureddine, et al. "On robust regression with high-dimensional predictors." Proceedings of the National Academy of Sciences 110.36 (2013): 14557-14562.
- Donoho, David, and Andrea Montanari. "High dimensional robust m-estimation: Asymptotic variance via approximate message passing." Probability Theory and Related Fields 166 (2016): 935-969.
- Sur, Pragya, and Emmanuel J. Candès. "A modern maximum-likelihood theory for high-dimensional logistic regression." Proceedings of the National Academy of Sciences 116.29 (2019): 14516-14525.
- Belkin, Mikhail, et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." Proceedings of the National Academy of Sciences 116.32 (2019): 15849-15854.
- Bartlett, Peter L., et al. "Benign overfitting in linear regression." Proceedings of the National Academy of Sciences 117.48 (2020): 30063-30070.

References

- Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.
- Zhang, Hongyi, et al. "mixup: Beyond Empirical Risk Minimization." International Conference on Learning Representations. 2018.
- Cao, Kaidi, et al. "Learning imbalanced datasets with label-distribution-aware margin loss." Advances in neural information processing systems 32 (2019).
- Montanari, Andrea, and Kangjie Zhou. "Overparametrized linear dimensionality reductions: From projection pursuit to two-layer neural networks." arXiv preprint arXiv:2206.06526 (2022).