When majority rules, minority loses: bias amplification of gradient descent

François Bachoc — University of Lille
Jérôme Bolte — Toulouse School of Economics
Ryan Boustany — Toulouse School of Economics
Jean-Michel Loubes — Regalia INRIA

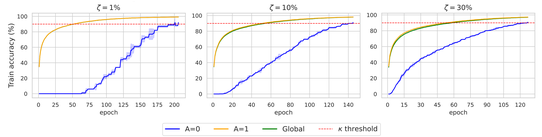
The setting: majority-minority model and generic losses

- n observations of a variable $Z := (X, Y) \in \mathbb{R}^d \times \mathbb{R}$ $(d > 0)$.
- Divided into two groups following the values of a binary variable $A \in \{0, 1\}$.
  - Majority: $A = 1$.
  - Minority: $A = 0$.
- Predictor $f_\theta(X)$ (e.g. deep neural network).
- Loss function $\ell(f_\theta(X), Y)$ (e.g. square error).
- Loss decomposition:

$$L(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i)$$
$$= \underbrace{\frac{1}{n}\sum_{\substack{i=1,\dots,n\\ A_i=1}} \ell(f_\theta(X_i), Y_i)}_{\substack{:=L_1(\theta)\\ \text{Majority loss}}} + \underbrace{\frac{1}{n}\sum_{\substack{i=1,\dots,n\\ A_i=0}} \ell(f_\theta(X_i), Y_i)}_{\substack{:=L_0(\theta)\\ \text{Minority loss}}}.$$

Numerical experiments with deep learning

Imbalance level ζ.



Catch-up overcost (in %): extra time needed for minority loss to reach threshold κ, after majority loss does.

Model	Size	$\kappa = 90\%$				$\kappa = 99\%$			
		$\zeta = 1\%$	10%	30%	80%	1%	10%	30%	80%
MobileNetV2	543K	450	275	166	0	62	52	42	0
SqueezeNet	727K	270	203	150	0	55	53	32	0
VGG11	9M	291	171	114	0	53	44	25	0
ResNet18	11M	292	164	113	0	61	49	31	0
VGG19	20M	280	152	112	0	70	65	50	0
ResNet50	25M	157	86	68	0	50	37	37	0
ResNet101	42M	145	90	64	0	34	36	26	0

Related reference

- Song Mei, Yu Bai and Andrea Montanari, The landscape of empirical risk for nonconvex losses, *The Annals of Statistics 46(6A) (2018) 171–203.*

Perturbation results for critical points

Setting:

- Functions L_1 and L_0 from $\mathbb{R}^d$ to $\mathbb{R}$ and a subset $K \subset \mathbb{R}^d$.
- Constants δ, c, M, τ.

Assumption: Strong Morse property. For all $\theta \in K$,

$$\|\nabla L_1(\theta)\| \le c \implies \rho_{\min}(\nabla^2 L_1(\theta)) \ge \delta.$$

Assumption: Strong Lipschitz regularity. For all $\theta_1, \theta_2 \in K$

$$\rho_{\max}\left(\nabla^2 L_1(\theta_1) - \nabla^2 L_1(\theta_2)\right) \le M\|\theta_1 - \theta_2\|,$$
$$\rho_{\max}\left(\nabla^2 L_0(\theta_1) - \nabla^2 L_0(\theta_2)\right) \le M\|\theta_1 - \theta_2\|.$$

Assumption: Bounds on the minority loss.

$$\sup_{\theta \in K} \|\nabla L_0(\theta)\| \le \tau,$$
$$\sup_{\theta \in K} \rho_{\max}\left(\nabla^2 L_0(\theta)\right) \le \tau.$$

Assumption: minority small enough.

$$\tau < \min\left\{\frac{c}{2}, \frac{\delta}{8}, \frac{\delta^2}{32M}\right\}.$$

Theorem: critical points are close. For each $\widehat{\theta}_1 \in \operatorname{crit} L_1 \cap K$ (resp. $\widehat{\theta} \in \operatorname{crit} L \cap K$) there exists a unique corresponding $\widehat{\theta} \in \operatorname{crit} L \cap K$ (resp. $\widehat{\theta}_1 \in \operatorname{crit} L_1 \cap K$) such that

$$\|\widehat{\theta}_1 - \widehat{\theta}\| \le \frac{4\tau}{\delta}.$$

Learning unbalanced data with the gradient method

Continuous approximation of gradient descent for loss $L(\theta)$:

$$\frac{d}{dt}\theta(t) = -\nabla L(\theta(t)) \text{ with } \theta(0) = \theta_{\text{init}} \in \mathbb{R}^d.$$

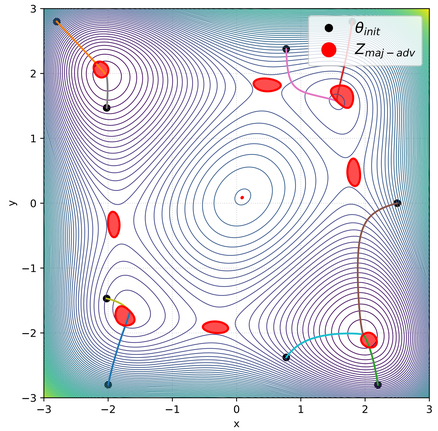
Majority-training zone. For losses $L = L_1 + L_0$:

$$Z_{\text{maj}} = \{\theta \in \mathbb{R}^d : \langle \nabla L(\theta), \nabla L_1(\theta)\rangle > 0\}.$$

- Gradient descent on L also decreases L_1.

Theorem: Majority-training zone is large. Under the previous setting,

$$Z_{\text{maj-adv}} := K \backslash Z_{\text{maj}} \subset \bigcup_{\widehat{\theta}_1 \in \operatorname{crit} L_1 \cap K} B\left(\widehat{\theta}_1, \frac{2\tau}{\delta}\right).$$



Paper published in Proceedings of NeurIPS 2025 and available at https://arxiv.org/abs/2505.13122