WORKING PAPERS

N° 1641

May 2025

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

François Bachoc, Jérôme Bolte, Ryan Boustany and Jean-Michel Loubes

Toulouse School of Economics

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

François Bachoc
Institut de Mathématiques de Toulouse
Institut universitaire de France (IUF)
francois.bachoc@math.univ-toulouse.fr

Jérôme Bolte
Toulouse School of Economics
France
jerome.bolte@tse-fr.eu

Ryan Boustany
Toulouse School of Economics
France
ryan.boustany@tse-fr.eu

Jean-Michel Loubes
Université de Toulouse
ANITI, Regalia INRIA
France
jean-michel.a.loubes@inria.fr

Abstract

Despite growing empirical evidence of bias amplification in machine learning, its theoretical foundations remain poorly understood. We develop a formal framework for majority-minority learning tasks, showing how standard training can favor majority groups and produce stereotypical predictors that neglect minority-specific features. Assuming population and variance imbalance, our analysis reveals three key findings: (i) the close proximity between "full-data" and stereotypical predictors, (ii) the dominance of a region where training the entire model tends to merely learn the majority traits, and (iii) a lower bound on the additional training required. Our results are illustrated through experiments in deep learning for tabular and image classification tasks.

1 Introduction

Machine learning systems deployed in high-stakes domains — such as credit scoring, healthcare, hiring, and predictive policing — are expected to deliver consistent performance across the populations they serve, without favoring one group of individuals over another. Yet, in many situations, there are observable tendencies toward bias, which may distort predictions and sometimes compromise fundamental rights. Addressing this concern is crucial for the future of AI in our societies, particularly under regulatory frameworks such as the European Union's AI Act, which places strong emphasis on ensuring fairness and non-discrimination. We refer to [12, 19, 27] for seminal papers on fairness in machine learning, and to [2, 13, 17, 40] for recent reviews on the topic. Many approaches have been proposed for bias detection [16, 21] and mitigation [11, 14, 22, 24, 25, 34, 50].

When discussing algorithmic bias in machine learning algorithms, we are often conflating three distinct notions: the bias inherent in the training data, the bias present in the data used for evaluation, and the bias introduced by the training process itself. This last source of bias is particularly significant to understand how so-called AI algorithms are so sensitive to bias. As a matter of fact, in many studies, an algorithm built using machine learning is said not only to learn and reproduce the bias present in the data, but also to amplify it. This idea has been developed in most of the research papers with many empirical proofs for the so-called *bias amplification* phenomenon using simulated or even real-life data, see e.g., [6, 26, 49, 51, 52]. While this amplification is well-documented experimentally,

Preprint.

there are few theoretical results attempting to elucidate its causes [44] and, to our knowledge, the question is still largely open.

In practice, this situation generally occurs when the training data are not balanced in particular when there is *little or no prior knowledge* of this imbalance. Mathematically, this makes the empirical risk dominated by majority features, so that a careless training process may distort the subtleties of minority phenomena or simply overlook them. The problem of class imbalance arises in many real-world classification tasks where the minority class, often representing the target of interest (e.g., fraudulent transactions or rare diseases), contains significantly fewer samples than the majority class [9, 29]. This skewed distribution can lead to suboptimal performance in conventional machine learning algorithms, which tend to focus on the majority class. As a result, extensive research has explored data-level techniques such as random oversampling, under-sampling, and synthetic data generation (e.g., SMOTE [10]) to mitigate the disproportionate impact of the majority class [9, 48]. Beyond these methods, algorithm-level strategies like cost-sensitive learning and threshold-moving have also shown promises, since they effectively penalize misclassifications of the rare class [20, 29]. With the advent of deep learning, the imbalance challenge becomes even more pronounced due to the large capacity and data requirements of deep neural networks [36]. We also refer to [38, 42] where some specific strategies using data augmentation schemes, combined with specialized loss functions can partially overcome gradient dominance by the majority class. These diverse approaches highlight a growing consensus that effectively handling imbalanced data in deep learning requires a tailored combination of sampling, cost adjustment, and architectural innovations to balance minority-class performance with overall accuracy, in particular when subgroups are not easily detectable.

Our objective focuses on the learning process: why would it lead to decisions that favor the majority group at the expense of the minority, as highlighted in [5]? We consider scenarios with significant imbalance in the training data, where the majority group is overrepresented in sample size and variability. In that case, under a typical training budget (e.g., 200–300 epochs in deep learning), models often amplify existing biases, producing stereotypical predictions disregarding minority populations. In contrast, prolonged and careful training on well-dimensioned architectures can detect minority-specific features more faithfully. We develop a mathematical framework to capture these phenomena and identify the core mechanisms behind bias amplification—specifically, population and variability imbalance, along with their geometric and dynamical implications.

Contributions. Our contributions are as follows.
— We first formalize the problem as a generic majority-minority learning task $\min L := L_1 + L_0$, with $L_0 \ll L_1$ using second-order differentiability domination. We prove that each critical point of L, which corresponds to a predictor, can be paired with a critical point of L_1, termed *stereotypical predictor*. We bound their distance: it is what we call the *stereotype gap*. It depends on a ratio measuring population and variance imbalance.
— The proximity of L and L_1 implies that the region where minimizing L is equivalent to minimizing L_1 (and vice versa) occupies nearly the entire parameter space. In linear regression, this results in a close overlap between the stereotypical gradient path and the actual training path, illustrating how standard training neglects minority-specific characteristics.
— We prove that gradient descent may need a fairly long training time to merely identify stereotypical predictors, ignoring minority-specific aspects. Debiasing the model requires additional training; we derive a lower bound on this extra training duration. The corresponding ratio is called the *fairness overcost ratio*; it quantifies the additional training time required to achieve unbiased predictions.
— We illustrate our theoretical findings through numerical experiments on several tabular and image classification tasks with deep neural networks.

Notations can be found in Appendix A.1.

2 Predictions for majority-minority problems in machine learning

We first present a majority-minority scenario in Section 2.1 as a minimization problem: $\min L := L_1 + L_0$. Given this setting, we aim to estimate the distance between a predictor obtained by minimizing the total loss L and a neighboring majority-based predictor obtained by minimizing L_1. In practice, the latter may represent a biased or stereotyped view that a user holds about the underlying problem. We show that a small population and low variance for the minority group lead to proximity between the predictor and the majority-based predictor, making them difficult to distinguish. Our

results are first presented for abstract equations (Proposition 4) and general variational problems (Theorem 1); discussions on fairness appear in Sections 2.3 and 2.4.

2.1 The setting: majority-minority model and generic losses

A majority-minority model. We consider n observations of a variable $Z := (X, Y) \in \mathbb{R}^d \times \mathbb{R}$ ($d > 0$) that can be divided into two groups following the values of a binary variable $A \in \{0, 1\}$. In our scenario, the data are unbalanced: there is a majority group $A = 1$ and a minority group $A = 0$, typically with $n_0 \ll n_1$. This heterogeneity, i.e., the variable A, may be unknown to the user. In the fairness literature, this variable is referred to as the *sensitive attribute*, such as gender or ethnic origin, as it may introduce or reflect discriminatory biases within the population.

Consider a collection of models or predictors $f_\theta : \mathbb{R}^d \mapsto \mathbb{R}$ indexed by parameters or weights $\theta \in \mathbb{R}^d$ that are learned by minimizing some empirical loss function over the learning set.

Given a discrepancy measure $\ell : \mathbb{R}^2 \to \mathbb{R}_+$ we may define the total, majority and minority losses as: for $\theta \in \mathbb{R}^d$,

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

In the training phase of a learning process, the parameters are often computed through first order methods and thus eventually through vanishing gradients. Assuming both ℓ and f_θ are differentiable, we are thus led to consider equations of the form: $\nabla L(\theta) = 0, \quad \nabla L_j(\theta) = 0$, for $j \in \{0, 1\}$. In a strongly imbalanced scenario, L_0 may become negligible with respect to L_1, so that the equations $\nabla L_1 = 0$ and $\nabla L = 0$ have very close solutions. On the other hand, this proximity does not prevent solutions to the equation $\nabla L_1(\theta) = 0$ from producing biased or stereotyped predictors as it ignores, by definition, the influence of data underlying L_0.

The aim of the following sections is to study this phenomenon and provide a set of assumptions for estimating the distance between full-data and stereotypical predictors.

Generic losses. For the rest of the article, we adopt a genericity perspective on loss functions by assuming that their critical points are non-degenerated. For $G : \mathbb{R}^d \to \mathbb{R}$ twice differentiable this means that

$$\nabla G(\theta) = 0 \Rightarrow \nabla^2 G(\theta) \text{ is invertible.}$$

In other words, G is a *Morse function*. These functions are generic in the sense that they form an open dense subset in $C^k(\mathbb{R}^d, \mathbb{R})$ for the C^2 topology whenever $k \geq 2$, see e.g., [23].

In the machine learning perspective, this is not extremely demanding as, for a fixed C^2 function G, perturbations of the form $\mathbb{R}^n \ni x \mapsto G_{\gamma,\epsilon}(x) = G(x) + \gamma\|x - \epsilon\|^2$ with $\gamma > 0$ are Morse for almost all couple $(\gamma, \epsilon) \in \mathbb{R}_+ \times \mathbb{R}^n$ – actually it holds true with linear perturbations, see e.g., [46]. This approach aligns with statistical and learning practices, both through ridge regularization (pioneered in [32] whose use in data science is developed for instance in [28], and references therein) and the weight decay approach in deep learning [8].

2.2 Perturbation results for critical points of generic losses

Assume $L = L_1 + L_0$ is a general cost. The spirit of the following results is that L_1 corresponds to a majority behavior while L_0 is attached to minority features, for instance as in the scenario of Section 2.1. In an analytical setting, it translates into a property of the type: L_0 is negligible w.r.t L_1 (see the assumptions below). We then aim at comparing $\operatorname{crit} L$ and $\operatorname{crit} L_1$; $\operatorname{argmin-loc} L$ and $\operatorname{argmin-loc} L_1$. Note that the theorem below is a general-purpose perturbation result, it will be applied in the fairness setting in the remaining sections (see Appendix A.1 for notations).

Theorem 1 (Distances between critical points)**.** *Consider two functions L_1 and L_0 from $\mathbb{R}^d$ to $\mathbb{R}$ that are two times continuously differentiable.*

Assume that there are strictly positive numbers δ, c, M, τ such that

- *Strong Morse property: For all $\theta \in \mathbb{R}^d$,*

$$\|\nabla L_1(\theta)\| \leq c \implies \rho_{\min}(\nabla^2 L_1(\theta)) \geq \delta, \tag{1}$$

- *Lipschitz regularity: for all $\theta_1, \theta_2 \in \mathbb{R}^d$*

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

$$\rho_{\max}\left(\nabla^2 L_0(\theta_1) - \nabla^2 L_0(\theta_2)\right) \leq M\|\theta_1 - \theta_2\|, \tag{3}$$

- *Bounds on the "minority loss":*

$$\sup_{\theta \in \mathbb{R}^d} \|\nabla L_0(\theta)\| \leq \tau, \tag{4}$$

$$\sup_{\theta \in \mathbb{R}^d} \rho_{\max}\left(\nabla^2 L_0(\theta)\right) \leq \tau. \tag{5}$$

Assume further that

$$\tau < \min\left\{\frac{c}{2}, \frac{\delta}{8}, \frac{\delta^2}{32M}\right\}. \tag{6}$$

Then we have the following conclusions.

(i) For each $\widehat{\theta}_1 \in \operatorname{crit} L_1$ (resp. $\widehat{\theta} \in \operatorname{crit} L$) there exists a unique corresponding $\widehat{\theta} \in \operatorname{crit} L$ (resp. $\widehat{\theta}_1 \in \operatorname{crit} L_1$) such that

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

and $\widehat{\theta}, \widehat{\theta}_1$ have the same indexes, that is the same number of strictly negative eigenvalues of the Hessian matrices $\nabla^2 L_1(\widehat{\theta}_1)$ and $\nabla^2 L(\widehat{\theta})$.

(ii) For each pair of distinct elements $\theta, \theta' \in \operatorname{crit} L_1$ (resp. $\operatorname{crit} L$), we have

$$\|\theta - \theta'\| \geq \frac{\delta}{32M}.$$

(iii) For each bounded set K, $\operatorname{crit} L_1 \cap K$ and $\operatorname{crit} L \cap K$ are finite sets.

Corollary 1 (Distances between critical and local minimizer sets)**.** *In the context of Theorem 1, if $\operatorname{crit} L_1$ is non-empty, then $\operatorname{crit} L$ is non-empty and we have*

$$\operatorname{dist}_{\mathrm{H}}(\operatorname{crit} L_1, \operatorname{crit} L) \leq \frac{4\tau}{\delta}. \tag{7}$$

Also, if $\operatorname{argmin-loc} L_1$ is non-empty then $\operatorname{argmin-loc} L$ is non-empty and we have

$$\operatorname{dist}_{\mathrm{H}}(\operatorname{argmin-loc} L_1, \operatorname{argmin-loc} L) \leq \frac{4\tau}{\delta}. \tag{8}$$

Finally, for each $\theta \in \operatorname{argmin-loc} L_1$, there is $\theta' \in \operatorname{argmin-loc} L$ such that the ball $B(\theta', \frac{6\tau}{\delta})$ contains θ, and L is $\delta/8$ strongly convex on this ball.

We note that [39] provides results similar to Theorem 1 and Corollary 1, for the different problem of comparing theoretical and empirical risks with random independent and identically distributed data.

2.3 A machine learning view: the representative and stereotypical predictions

Let us interpret the above within a fairness perspective. Under the premises of Theorem 1, we consider a machine learning model with loss $L : \theta \mapsto L(\theta)$ decomposed into a sum $L = L_1 + L_0$ where L_1 and L_0 respectively correspond to some majority and minority phenomena.

A critical point of L is called a *representative prediction*, as it takes into account all available data encoded within L, i.e. both those in L_1 and L_0[1]. In the majority-minority model, the critical points

[1]It would be more natural to reserve that name for local minimizers, as those are generally obtained after training, but we do so for simplicity.

of L_1 ignore data corresponding to the case when $A = 0$, we thus call them *stereotypical predictions*. The quantity $\operatorname{dist}_{\mathrm{H}}(\operatorname{crit} L, \operatorname{crit} L_1)$ is called the *stereotype gap*.

Roughly speaking Theorem 1 tells us, in particular, that each representative prediction corresponds to one and only one stereotypical prediction and that these predictions are close whenever the ratio

$$\Delta = \rho_{\max}\left(\nabla^2 L_0(\theta)\right) / \rho_{\min}(\nabla^2 L_1(\theta))$$

is uniformly small. This ratio is the key quantity that governs the stereotype gap.

The result is even more accurate, as Corollary 1 shows that the minimizers of L and L_1 actually come by pairs as well, so that the stereotypical and representative predictors obtained in practice are "dangerously" close in a majority-minority scenario. As we will see through theoretical and numerical experiments, this renders the training phase delicate and potentially biased. Using the well-known fact that gradient descent converges to critical points in the Morse case (see next section and Appendix A.4), we may empirically estimate the stereotypical gaps and the associated "debiasing training time" in our imbalanced setting (see also the following sections).

Protocol (Table 1 opposite): find a stereotypical predictor $\widehat{\theta}_1$ via the gradient flow $-\nabla L_1$ with Kaiming random initialization. Initialize from this predictor $\widehat{\theta}_1$ and follow the flow of $-\nabla L$, with the guarantee (see Corollary 1) of reaching the corresponding representative predictor $\widehat{\theta}$. Use these values to estimate the gap $\operatorname{dist}_{\mathrm{H}}(\operatorname{crit} L, \operatorname{crit} L_1)$ via proxies like $\|\hat{\theta} - \hat{\theta}_1\|$, and to define a debiasing time from $\widehat{\theta}_1$ to its representative $\widehat{\theta}$ using gradient descent on L with stopping criterion $\|\theta_{k+1} - \widehat{\theta}_1\| \geq 0.99\|\theta_k - \widehat{\theta}_1\|$.

Table 1: Stereotypical and representative predictions for imbalanced CIFAR-2 ($n_0/n \approx 3\%$, see Appendix D.1) with ResNet 18. We report the average and standard deviation over 30 runs.

Metric	**Mean**	**± Std**
Debiasing time	469 epochs	± 9.4
$\|\hat{\theta} - \hat{\theta}_1\|$	0.6723	± 0.0083
$\|\hat{\theta} - \hat{\theta}_1\|_\infty$	0.0353	± 0.0047
$\frac{\|\hat{\theta} - \hat{\theta}_1\|}{\|\hat{\theta}\|}$	0.00602	± 0.00007

2.4 A case study: linear regression

To illustrate further our result, consider a multidimensional regression model with loss

$$L(\theta) = \frac{1}{2n}\|X\theta - Y\|^2 = \underbrace{\frac{1}{2n}\|X^1\theta - Y^1\|^2}_{L_1(\theta)} + \underbrace{\frac{1}{2n}\|X^0\theta - Y^0\|^2}_{L_0(\theta)}, \tag{9}$$

where X is $n \times d$ with rows $X_1^\top, \ldots, X_n^\top$, and where X^1 (respectively X^0) contains the rows of X from the majority (respectively minority) class. Similarly, Y is n-dimensional with components $Y_1, \ldots, Y_n$ and Y^1 (respectively Y^0) contains the components of Y from the majority (respectively minority) class. Letting $X^{j\top} = (X^j)^\top$ for $j = 0, 1$, we define the corresponding empirical covariance matrices $S = X^\top X/n$, $S_0 = X^{0\top}X^0/n_0$, $S_1 = X^{1\top}X^1/n_1$. Assume that the covariance matrices are invertible, which may be granted through a ridge regression model in the generic/regularized spirit presented in Section 2.1. In this setting, our results become:

Theorem 2 (Representative-stereotypical gap: linear regression case)**.** *Assume that S_0 and S_1 are invertible, and denote by $\widehat{\theta}$, $\widehat{\theta}_1$, and $\widehat{\theta}_0$, respectively, the unique global minimizers of L, L_1, and L_0, respectively. Then*

$$\|\widehat{\theta} - \widehat{\theta}_1\| \leq \frac{2\rho_{\max}(n_0 S_0)}{\rho_{\min}(n_1 S_1)}\left(1 + \|\widehat{\theta}_1 - \widehat{\theta}_0\|\right).$$

The key quantity behind the stereotypical gap is the ratio

$$\frac{\rho_{\max}(n_0 S_0)}{\rho_{\min}(n_1 S_1)} = \frac{\rho_{\max}(\nabla^2 L_0)}{\rho_{\min}(\nabla^2 L_1)},$$

where we have omitted the dependence on θ in the Hessians, which are constant. Two statistical effects drive this ratio:

— Population size ratio: when the majority is much larger than the minority, then n_0/n_1 is small, this tends to increase the risk of stereotypical predictions.

— Min-max variability ratio: if the smallest variability of the majority is much bigger than the largest variability of the minority group, then stereotypical predictions are more likely.

3 Learning unbalanced data with the gradient method

In this section, we study how gradient descent procedures may bias predictions in the sense that a "careless training" may yield a stereotypical predictor rather than a representative one. Gradient descent training on a C^2 loss L is modeled through the ODE (see Appendix A.4 for the representation of ODE curves):

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

3.1 The majority-training and the majority-adverse zones

For C^2 smooth losses $L = L_1 + L_0$, the *majority-training zone* is defined by

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

In this region, descending along the gradient of L also decreases L_1, and vice versa. In other words, Z_{maj} is a zone where training L with gradient descent implies training the majority L_1.The *majority-adverse* zone is defined as

$$Z_{\text{maj-adv}} = \{\theta \in \mathbb{R}^d : \langle \nabla L(\theta), \nabla L_1(\theta)\rangle \leq 0\} \text{ so that } \mathbb{R}^d \setminus Z_{\text{maj}} = Z_{\text{maj-adv}}\,. \tag{10}$$

We can similarly consider the minority-training and the minority-adverse zones. One easily sees that, under the Morse assumption, critical points of L or L_1 lie in between Z_{maj} and $Z_{\text{maj-adv}}$, (see Proposition 5 in Appendix B.3 for details). In other words, the stereotypical and representative predictors lie on the boundary of Z_{maj}.

We now establish two major facts: first, the majority zone is typically large, meaning that training the entire model often results in learning only the majority traits (see also the illustration of Figure 1); second, the majority-adverse zone promotes the training of the minority loss.

Theorem 3 (Majority adverse zone)**.** *Under Theorem 1 assumptions:*

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

Remark 1 (Minority-training zone)**.** Little can be said about the size of the minority-training zone; it can be huge or empty (e.g., when $L_0 = aL_1$ with $|a|$ small, see also Proposition 6 in Appendix B.3).

Lemma 1 (The majority adverse zone favors minority)**.** *For $\theta \in Z_{\text{maj-adv}}$, we have*

$$\langle \nabla L(\theta), \nabla L_0(\theta)\rangle \geq 0.$$

Thus a training trajectory $\theta : I \to \mathbb{R}^d$ evolving within $Z_{\text{maj-adv}}$ is such that $L_0(\theta(t))$ is non-increasing over the interval I.

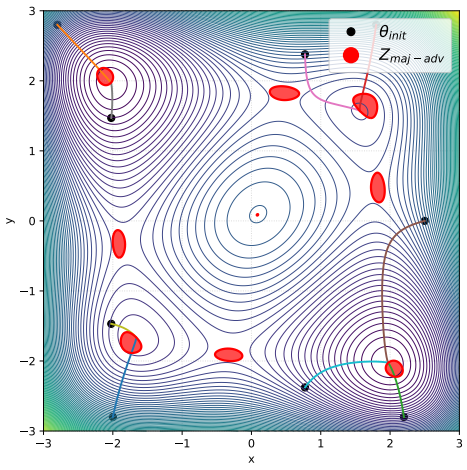

Figure 1: Majority training (white), majority adverse (red) zones, and a few gradient curves.

3.2 A case study: the training phase for linear regression

Consider the quadratic setting of Section 2.4, where L, L_1, L_0 from (9), have unique minimizers $\widehat{\theta}, \widehat{\theta}_1, \widehat{\theta}_0$.

Stereotypical and representative training curves. In this section, we compare training L, which provides a *representative training curve*, with training on the majority group, which provides a *stereotypical training curve* as if the minority did not exist. The stereotypical training curve is:

$$\frac{d}{dt}\theta_1(t) = -\nabla L_1(\theta_1(t)).$$

Our estimate depends once more on the ratio $\Delta = \rho_{\max}(n_0 S_0)/\rho_{\min}(n_1 S_1)$.

Proposition 1 (Distance between stereotypical and representative training curves)**.** *We have, for any* $t > 0$,

$$\|\theta(t) - \theta_1(t)\| \leq \|\widehat{\theta} - \widehat{\theta}_1\| + t\rho_{\max}\left(\frac{n_0}{n}S_0\right) e^{-t\rho_{\min}\left(\frac{n_1}{n}S_1\right)} \left(\|\widehat{\theta}_1\| + \|\theta_{\text{init}}\|\right),$$

$$\|\theta - \theta_1\|_\infty := \sup_{t>0} \|\theta(t) - \theta_1(t)\| \leq \|\widehat{\theta} - \widehat{\theta}_1\| + \frac{\|\widehat{\theta}_1\| + \|\theta_{\text{init}}\|}{e} \frac{\rho_{\max}(n_0 S_0)}{\rho_{\min}(n_1 S_1)}.$$

On fair training duration. In order to estimate debiasing duration, we consider in this paragraph an "unlucky gradient curve" $t \to \theta(t)$ which somehow ignores minority until it detects the majority predictor: from 0 to t_1 the curve brings its initial condition θ_{init} to the stereotype $\theta(t_1) = \widehat{\theta}_1$, as if only L_1 was trained; then $\theta(t)$ travels towards $\widehat{\theta}$, the representative predictor. By the Cauchy-Lipschitz existence theorem, this trajectory exists. This curve and neighboring ones may be quite detrimental to fair predictions as shown in Figure 2:

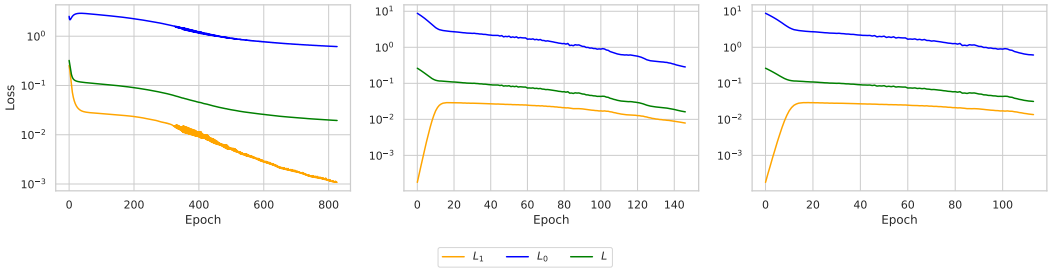

Figure 2: Left to right: an unlucky trajectory with stopping criterion (s.c.) $\text{Acc}_0 > 99\%$ (see Appendix A.2), a random trajectory with s.c. $\text{Acc}_0 > 99\%$, a minority agnostic trajectory with s.c. $\text{Acc} > 99\%$. One observes important disparities in training time: unlucky initialization has 800 epochs while "careless training" yields a 100 epochs training and a much higher final L_0 value. Middle picture: the "fair" and luckier training has 140 epochs. Due to randomness and numerical precision, these results are, of course, very sensitive to initialization and discretization.

Consider an "unlucky gradient curve", i.e. a L gradient curve θ with $\theta(t_1) = \widehat{\theta}_1$. The next proposition shows that t_1 is typically large as $\|\widehat{\theta}_1 - \widehat{\theta}\|$ is typically very small (see Theorem 2 and Table 2).

Proposition 2 (Training duration)**.** *Assume* $\widehat{\theta} \neq \widehat{\theta}_1$*, then* $t_1 \geq \dfrac{1}{\rho_{\max}(S)} \log\left(\dfrac{\|\theta_{\text{init}} - \widehat{\theta}\|}{\|\widehat{\theta}_1 - \widehat{\theta}\|}\right)$.

In view of understanding "fair training" better, let us provide a lower bound on the extra-time $t_\epsilon - t_1$ needed to achieve the relative ϵ precision:

$$\frac{\|\theta(t_\epsilon) - \widehat{\theta}\|}{\|\widehat{\theta}_1 - \widehat{\theta}\|} \leq \epsilon \quad \text{where } \epsilon \in (0, 1).$$

Proposition 3 (Debiasing duration[2])**.** *Assume* $\widehat{\theta} \neq \widehat{\theta}_1$*, then* $t_\epsilon - t_1 \geq \dfrac{1}{\rho_{\max}(S)} \log\left(\dfrac{1}{\epsilon}\right)$.

Remark 2 (Extensions to the nonlinear case)**.** Propositions 2 and 3 are based on Lemma 4 in Appendix B.4, which holds for general loss functions, beyond linear regression.

4 Numerical experiments

We study the effect of subgroup imbalance in supervised deep learning using image (CIFAR-10 [37], EuroSAT [31]) and tabular (Adult [4]) datasets. Each dataset is denoted by $\mathcal{D} = \{(X_i, Y_i, A_i)\}_{i=1}^n$,

[2]See also Proposition 7 in Appendix B.4 for complementary results on relative values of L_0.

where (X_i, Y_i) is an input-label pair and $A_i \in \{0, 1\}$ is a binary attribute. While A is not used during training, it enables evaluation of model performance across imbalanced subgroups. In each experiment, we report the global loss $L = L_0 + L_1$, and average loss per sample in each group, i.e., $(nL_0)/n_0$ and $(nL_1)/n_1$. For details on the implementation setup, see Appendix D.

Fairness metrics. We evaluate fairness using training accuracy, denoted by $\mathrm{Acc}, \mathrm{Acc}_0$ (see Appendix A.2), as our focus is on optimization under imbalance. Assessing fairness on the test set would require extensive tuning and is left for future work. To measure the time cost of fairness, we track the number of epochs t needed to reach a threshold accuracy level $\kappa \in [0, 1]$:

$$T_{\text{early}} := \min_{t \in \mathbb{N}} \{\mathrm{Acc}(\theta_t) \geq \kappa\}, \quad T_{\text{final}} := \min_{t \in \mathbb{N}} \{\mathrm{Acc}_0(\theta_t) \geq \kappa\}, \quad T_{\text{debias}} := T_{\text{final}} - T_{\text{early}}.$$

We define the *fairness overcost* as the relative delay to reach a satisfying minority accuracy:

$$\text{Fairness Overcost} := \frac{T_{\text{debias}}}{T_{\text{early}}}.$$

Imbalanced CIFAR-10. We investigate the effect of class imbalance on CIFAR-10 using models from 100K to 25M parameters (see Table 2). The original dataset has 10 classes with 5000 samples each. To create imbalance, we subsample one class (denoted $A = 0$) to retain n_0 samples, and keep the others ($A = 1$) unchanged with $n_1 = 9 \times 5000$. As in [36], we define the imbalance ratio as $\zeta = n_0/5000$, which gives a group proportion $n_0/(n_0 + n_1) = \zeta/(\zeta + 9)$, and consider four imbalance levels: $\zeta \in \{1\%, 10\%, 30\%, 80\%\}$. In Figure 3, we show the results for ResNet-18 (see also Appendix C for more). For $\zeta = 1\%$, Acc_0 remains close to zero for about 60 epochs, following a stereotypical training curve (see Section 3.2).

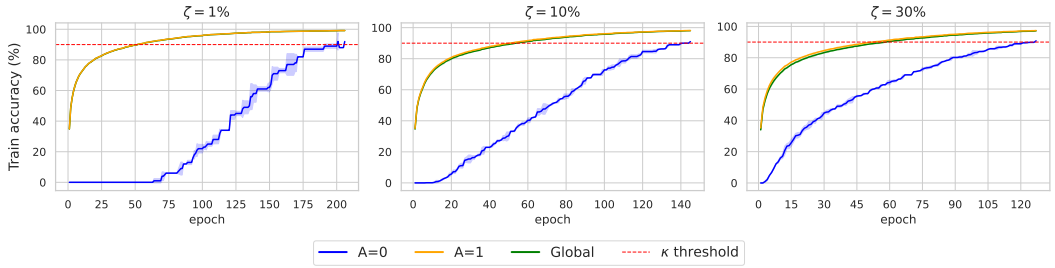

Figure 3: Training accuracy for different subgroup imbalance scenarios (1%, 10%, and 30%) using ResNet18 on CIFAR-10 and threshold $\kappa = 90\%$.

Table 2: Fairness overcost (in %) for each model across imbalance levels $\zeta \in \{1\%, 10\%, 30\%, 80\%\}$. We report means over 3 runs with thresholds $\kappa \in \{90\%, 99\%\}$, and model parameter counts.

Models	**Number of parameters**	$\kappa = 90\%$				$\kappa = 99\%$			
		1%	10%	30%	80%	1%	10%	30%	80%
MobileNetV2 [45]	543K	450	275	166	0	62	52	42	0
SqueezeNet [35]	727K	270	203	150	0	55	53	32	0
VGG11 [47]	9M	291	171	114	0	53	44	25	0
ResNet18 [30]	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

EuroSAT. We use a ResNet18 model and evaluate its behavior on a binary classification task derived from the EuroSAT dataset. Images are labeled according to a binary attribute A, where $A = 0$ corresponds to bluish images ($n_0/(n_0 + n_1) \approx 0.03$) and $A = 1$ to all others. We do not modify the class proportions and use the imbalance present in the original dataset. Figure 4 displays losses and accuracies for both subgroups evidencing a fairness overcost of 45% for a threshold $\kappa = 90\%$.

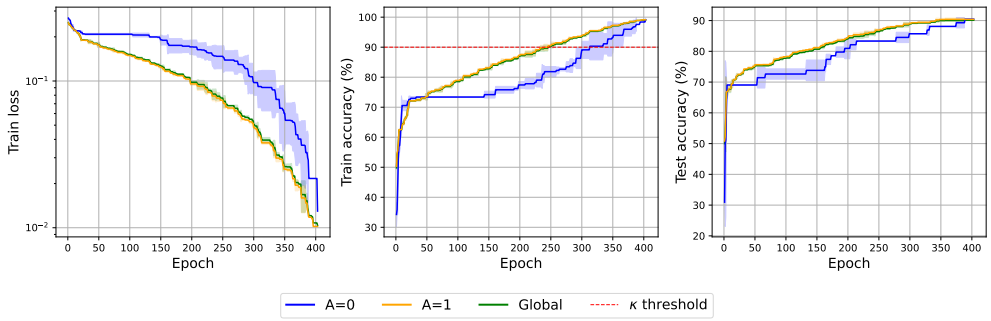

Figure 4: Training and test loss and accuracy using ResNet18 on EuroSAT (mean over 3 runs).

Adult income census. We train a TabNet classifier [1] on a binary task from the Adult dataset [4]. The minority group ($A = 0$) includes high-income women, representing only 3% of the training set. The majority group ($A = 1$) includes all others. We preserve the original class distribution and train with cross-entropy loss, tracking subgroup metrics. Results are shown in Figure 5 and we have a fairness overcost of 416% for a threshold $\kappa = 90\%$.

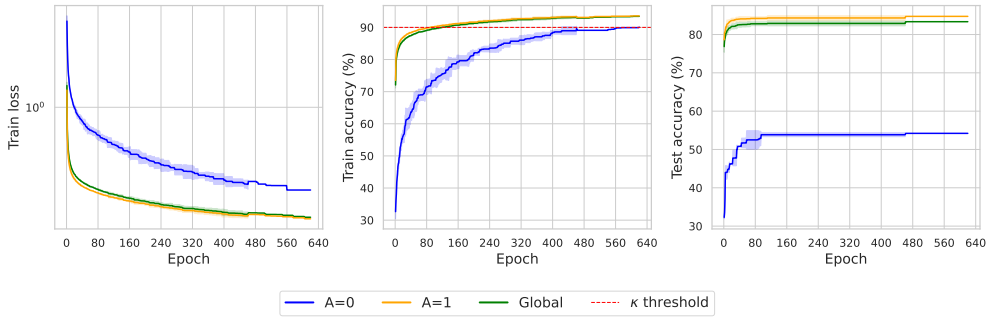

Figure 5: Loss and accuracy using TabNet on Adult (mean over 3 runs).

Results and discussion. Fairness under imbalance requires much longer training: the minority group ($A = 0$) consistently reaches κ much later than the global accuracy. The fairness overcost is particularly high under strong imbalance, exceeding 400% on Adult and CIFAR-10. Empirically, larger models reduce this overcost but do not eliminate the necessity of longer well-tailored training. These results support our theoretical findings on debiasing duration in imbalanced settings (see Section 3.2), the overwhelming dominance of the majority-training zone, and the difficulty of distinguishing a representative predictor from a stereotypical one.

5 Conclusion

Although our goals are primarily theoretical and future research should explore more refined training protocols, we can draw several conclusions supported by both theory and numerics. These conclusions may also serve as recommendations for practitioners. Two key quantities emerge as critical in our study: the stereotype gap and the training duration. Additionally, we have empirical evidence that the model size may be a determining factor in achieving budget frugality.

— In a majority-minority scenario, population and variability imbalances are determining factors influencing the stereotype gap (Theorem 1 and the subsequent subsections). This gap, between stereotypes and representative predictors, can be very small in severely imbalanced cases.

— For convex or deep learning problems, gradient training generally leads to a "satisfying predictor" in the sense of a low-value loss L, see e.g., [3] or [18]. However, in our majority-minority scenario, the action of L_0 is generally almost indetectable, as shown in Figure 1 and Section 3.2, thus early stopping and under-dimensioned models are prone to produce stereotypes.

— To obtain a representative predictor, it is advisable to use larger networks and extend the training

duration, as supported by Propositions 2 and 3, and the numerical section. The corresponding fairness overcost ratio can take considerable values, e.g., from 25% to 450% for the imbalanced CIFAR-10.

Acknowledgments

The authors acknowledge the support from the AI Interdisciplinary Institute ANITI, TRIAL and UQPhysAI chairs. J.B. and J.-M.L. are supported by ANR REGULIA. JB is also supported by the Air Force Office of Scientific Research FA8655-22-1-7012, ANR Chess (ANR-17-EURE-0010), ANR ESRE (ANR-21-ESRE-0051). Access to MesoNET resources in Toulouse was granted under allocation m23038.

References

[1] Sercan Ö Arik and Tomas Pfister. TabNet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687, 2021.

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[3] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[4] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[5] Samuel James Bell and Levent Sagun. Simplicity bias leads to amplified performance disparities. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 355–369, 2023.

[6] Philippe Besse, Eustasio del Barrio, Paula Gordaliza, Jean-Michel Loubes, and Laurent Risser. A survey of bias in machine learning through the prism of statistical parity. *The American Statistician*, 76(2):188–198, 2022.

[7] FF Bonsall, LV Kantorovich, and GP Akilov. Functional analysis in normed spaces; translated from the russian by DE Brown; Robertson, AP, Ed, 1964.

[8] Nathaniel Bottman, Y. Cooper, and Antonio Lerario. How regularization affects the geometry of loss functions. *arXiv preprint arXiv:2307.15744*, 2023.

[9] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. In *Journal of Artificial Intelligence Research*, volume 16, pages 321–357, 2002.

[10] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W.P̃hilip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[11] Silvia Chiappa, Ray Jiang, Tom Stepleton, Aldo Pacchiano, Heinrich Jiang, and John Aslanides. A general approach to fairness with optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3633–3640, 2020.

[12] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[13] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

[14] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with Wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33:7321–7331, 2020.

[15] Philippe G Ciarlet and Cristinel Mardare. On the Newton-Kantorovich theorem. *Analysis and Applications*, 10(03):249–269, 2012.

[16] Eustasio del Barrio, Paula Gordaliza, and Jean-Michel Loubes. A central limit theorem for lp transportation cost on the real line with application to fairness assessment in machine learning. *Information and Inference: A Journal of the IMA*, 8(4):817–849, 2019.

[17] Eustasio Del Barrio, Paula Gordaliza, and Jean-Michel Loubes. Review of mathematical frameworks for fairness in machine learning. *arXiv preprint arXiv:2005.13755*, 2020.

[18] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018.

[19] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[20] Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, volume 17, pages 973–978, 2001.

[21] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[22] Solenne Gaucher, Nicolas Schreuder, and Evgenii Chzhen. Fair learning with Wasserstein barycenters for non-decomposable performance measures. In *International Conference on Artificial Intelligence and Statistics*, pages 2436–2459. PMLR, 2023.

[23] Martin Golubitsky and Victor Guillemin. *Stable mappings and their singularities*, volume 14. Springer Science & Business Media, 2012.

[24] Paula Gordaliza, Eustasio del Barrio, Fabrice Gamboa, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2357–2365. PMLR, 2019.

[25] Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

[26] Melissa Hall, Laurens van der Maaten, Laura Gustafson, Maxwell Jones, and Aaron Adcock. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022.

[27] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[28] Trevor Hastie. Ridge regularization: An essential concept in data science. *Technometrics*, 62(4):426–433, 2020.

[29] Haibo He and Edward A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[31] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[32] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[33] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[34] Max Hort, Zhenpeng Chen, Jie M Zhang, Mark Harman, and Federica Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 1(2):1–52, 2024.

[35] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *arXiv preprint arXiv:1602.07360*, 2016.

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

[37] Alex Krizhevsky and Geoffrey Hinton. The CIFAR-10 dataset, 2010.

[38] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[39] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[40] Luca Oneto and Silvia Chiappa. Fairness in machine learning. In *Recent trends in learning from data: Tutorials from the INNS big data and deep learning conference (INNSBDDL2019)*, pages 155–196. Springer, 2020.

[41] Shalin Parekh. The KPZ limit of ASEP with boundary. *Communications in Mathematical Physics*, 365:569–649, 2019.

[42] Samira Pouyanfar, Yao Tao, Hao Tian, Jing Shang, Shu-Ching Chen, S. Sitharama Iyengar, Ahmed S. Kaseb, and Mei-Ling Shyu. Dynamic sampling in convolutional neural networks for imbalanced data classification. *arXiv preprint arXiv:1810.00889*, 2018.

[43] Laurent Risser, Agustin Martin Picard, Lucas Hervier, and Jean-Michel Loubes. Detecting and processing unsuspected sensitive variables for robust machine learning. *Algorithms*, 16(11):510, 2023.

[44] 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.

[45] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018.

[46] Anant R Shastri. *Elements of differential topology*. CRC Press, 2011.

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[48] Qiang Sun, Hongwei Xu, and Yufeng Yang. Cost-sensitive boosting for classification of imbalanced data. In *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 2168–2173, 2007.

[49] Angelina Wang and Olga Russakovsky. Directional bias amplification. In *International Conference on Machine Learning*, pages 10882–10893. PMLR, 2021.

[50] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR, 2017.

[51] Dora Zhao, Jerone Andrews, and Alice Xiang. Men also do laundry: Multi-attribute bias amplification. In *International Conference on Machine Learning*, pages 42000–42017. PMLR, 2023.

[52] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

This is the appendix for "When majority rules, minority loses: bias amplification of gradient descent".

Contents

A Notations and auxiliary results

A.1 Notations.

For a matrix A, we write $\rho_{\min}(A)$ and $\rho_{\max}(A)$ for its smallest and largest singular value. If the matrix A is square symmetric, we write $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ for its smallest and largest eigenvalue. For a vector $x \in \mathbb{R}^d$ we write $\|x\|$ for its Euclidean norm and for $\epsilon > 0$, we write $B(x, \epsilon) = \{y \in \mathbb{R}^d; \|x - y\| \leq \epsilon\}$.

For a function $f : \mathbb{R}^d \to \mathbb{R}^d$ and for $x \in \mathbb{R}^d$, we write $\operatorname{Jac} f(x)$ for the Jacobian matrix of f at x. For a function $f : \mathbb{R}^d \to \mathbb{R}$ and for $x \in \mathbb{R}^d$, we write $\nabla f(x)$ for the gradient vector of f at x and $\nabla^2 f(x)$ for the Hessian matrix of f at x. For a function $G : \mathbb{R}^d \to \mathbb{R}$, we write

$$\operatorname{crit} G = \{\theta \in \mathbb{R}^d : \nabla G(\theta) = 0\}$$
$$\operatorname{argmin-loc} G = \{\theta \in \mathbb{R}^d : \theta \text{ is a local minimizer of } G \text{ over } \mathbb{R}^d\}.$$

For two non-empty subsets A and B of $\mathbb{R}^d$, the Hausdorff distance between A and B is denoted by

$$\operatorname{dist}_{\mathrm{H}}(A, B) = \max\left(\sup_{x \in A} \inf_{y \in B} \|x - y\|, \sup_{y \in B} \inf_{x \in A} \|x - y\|\right),$$

when A and B are nonempty and bounded this quantity is finite.

The topological boundary of A is written $\operatorname{bdry} A$.

A.2 Training metrics

Let $f_\theta : \mathbb{R}^d \to \mathbb{R}^C$ be a neural network, parameterized by θ, that maps an input $x \in \mathbb{R}^d$ to a vector of C class scores. We denote $[C] = \{1, \ldots, C\}$ the set of class indices, and define the predicted label as $\hat{y}(x) = \arg\max_{c \in [C]} f_\theta(x)_c$. We compute accuracy separately for each group $j \in \{0, 1\}$ as the proportion of correct predictions in $\mathcal{D}_{A=j}$, and define the *global accuracy* as the weighted average across groups. For each $j \in \{0, 1\}$:

$$\mathrm{Acc}_j(\theta) = \frac{1}{n_j} \sum_{(x_i, y_i) \in \mathcal{D}_{A=j}} \mathbb{1}[\hat{y}(x_i) = y_i], \text{ and } \mathrm{Acc}(\theta) = \frac{n_0}{n}\mathrm{Acc}_0(\theta) + \frac{n_1}{n}\mathrm{Acc}_1(\theta),$$

where $\mathbb{1}[\hat{y}(x_i) = y_i]$ denotes the indicator function, equal to 1 if the predicted label matches the true label and 0 otherwise.

A.3 Lemma

The next lemma is well-known but stated here for convenience.

Lemma 2. *Let E be an open set of $\mathbb{R}^k$ for some $k \in \mathbb{N}$. Let $f : E \to \mathbb{R}^k$ have Jacobian $\operatorname{Jac} f$. Let $x, y \in E$ so that the segment between x and y is in E. Then*

$$\|f(y) - f(x)\| \leq \left(\sup_{u \in E} \rho_{\max}(\operatorname{Jac} f(u))\right) \|y - x\|.$$

A.4 Discretization of ODE curves

In various parts of this paper, we refer to or represent ODE curves in our experiments. Unless otherwise specified, this refers to a discretization of the ODE using small step sizes. For instance, given the dynamics

$$\dot{\theta}(t) = F(\theta(t)), \quad \theta(0) = \theta_{\text{init}},$$

with $F : \mathbb{R}^p \to \mathbb{R}^p$ a locally Lipschitz field, the discretization we use is of the form

$$\theta_{k+1} = \theta_k - s_k F(\theta_k),$$

where the step size $s_k \ll 1$; in practice, we typically use $s_k = O(10^{-3})$.

Note however that for the numerical section, we proceed differently as our objective is rather training through the gradient method. We thus use larger steps and mini-batches.

B Proofs and extra results

B.1 Proofs and extra results of Section 2.2

Kantorovich theorem. A great part of Section 2.2 relies on a theorem of Kantorovich type for Newton's method [15, Theorem 5] whose proof is based on [7]. This result is recalled below:

Theorem 4 (Newton–Kantorovich Theorem "with only one constant" (existence))**.** *Let $\theta^\star \in \mathbb{R}^d$ and $\widetilde{R} > 0$. Let Ω be an open set containing the closed ball $B(\theta^\star, \widetilde{R})$. Let $G : \Omega \to \mathbb{R}^d$ be a continuously differentiable mapping. Suppose that the following conditions are satisfied:*

(K1) $\operatorname{Jac} G(\theta^\star)$ *is invertible with* $\|\operatorname{Jac} G(\theta^\star)^{-1} G(\theta^\star)\| \leq \frac{\widetilde{R}}{2}$.

(K2) For all $\theta, \theta' \in B(\theta^\star, \widetilde{R})$, $\rho_{\max}\left(\operatorname{Jac} G(\theta^\star)^{-1} \left(\operatorname{Jac} G(\theta) - \operatorname{Jac} G(\theta')\right)\right) \leq \frac{\|\theta - \theta'\|}{\widetilde{R}}$.

Then there exists a unique $\widetilde{\theta} \in B(\theta^\star, \widetilde{R})$ such that $G(\widetilde{\theta}) = 0$.

We need beforehand abstract results on equation perturbations. Let $\theta^\star \in \mathbb{R}^d$. We consider a function $F : \mathbb{R}^d \to \mathbb{R}^d$ such that $F(\theta^\star) = 0$. Let $p : \mathbb{R}^d \to \mathbb{R}^d$ and consider the equation defined for $\theta \in \mathbb{R}^d$,

$$F(\theta) = p(\theta). \tag{11}$$

If the function p is negligible, in a certain sense, with respect to the dominant term $F(\theta)$, (11) becomes a perturbed version of equation $F(\theta) = 0$. Its solution will be close to the solution of the non perturbed equation, $\theta^\star$. Proposition 4 quantifies partly this phenomenon.

Proposition 4 (Distance to a perturbed solution)**.** *Assume F and p are continuously differentiable and that there are strictly positive numbers δ, M, τ such that:*

- *Conditioning of F and $F - p$*

$$\rho_{\min}(\operatorname{Jac} F(\theta^\star)) \geq \delta \quad \textit{and} \quad \rho_{\min}(\operatorname{Jac}(F - p)(\theta^\star)) \geq \delta, \tag{12}$$

- *Differential regularity of the nonlinear equation*

$$\rho_{\max}\left(\operatorname{Jac} F(\theta) - \operatorname{Jac} F(\theta')\right) \leq M\|\theta - \theta'\|, \qquad \theta, \theta' \in \mathbb{R}^d, \tag{13}$$

$$\rho_{\max}\left(\operatorname{Jac} p(\theta) - \operatorname{Jac} p(\theta')\right) \leq M\|\theta - \theta'\|, \qquad \theta, \theta' \in \mathbb{R}^d, \tag{14}$$

- *Perturbation bounds*

$$\|p(\theta^\star)\| \leq \tau, \tag{15}$$

$$\rho_{\max}(\operatorname{Jac} p(\theta^\star)) \leq \tau. \tag{16}$$

If the perturbation ratio τ/δ satisfies

$$\tau/\delta < \frac{\delta}{4M}, \tag{17}$$

then, there is a unique θ_p solution to $F(\theta_p) = p(\theta_p)$, which is close to the solution of $F(\theta^*) = 0$, in the sense that

$$\theta_p \in B\left(\theta^\star, \frac{2\tau}{\delta}\right).$$

Proof of Proposition 4. For $\theta \in \mathbb{R}^d$, let $G(\theta) = F(\theta) - p(\theta)$. We apply Kantorovich's Theorem above (Theorem 4) to the function G. The quantity $\widetilde{R}$ is taken as

$$\widetilde{R} = \frac{2\tau}{\delta}.$$

Let us check Assumption (K1). We have, using (12),

$$\|\mathrm{Jac}\, G(\theta^\star)^{-1} G(\theta^\star)\| \leq \frac{\|G(\theta^\star)\|}{\rho_{\min}(\mathrm{Jac}\, G(\theta^\star))} \leq \frac{\|p(\theta^\star)\|}{\delta} \leq \frac{\tau}{\delta}.$$

Hence (K1) holds since $\frac{\widetilde{R}}{2} = \frac{\tau}{\delta}$.

Let us check Assumption(K2). For all $\theta, \theta' \in B(\theta^\star, \widetilde{R})$, we have

$$\begin{aligned}\rho_{\max}\left(\mathrm{Jac}\, G(\theta^\star)^{-1}\left(\mathrm{Jac}\, G(\theta) - \mathrm{Jac}\, G(\theta')\right)\right) \leq & \frac{1}{\rho_{\min}(\mathrm{Jac}\, G(\theta^\star))}\rho_{\max}\left(\mathrm{Jac}\, G(\theta) - \mathrm{Jac}\, G(\theta')\right) \\ \text{from (12), (13) and (14)} \quad \leq & \frac{2M}{\delta}\|\theta - \theta'\|.\end{aligned}$$

From (17), we have $\frac{\tau}{\delta} < \frac{\delta}{4M}$ and thus $\frac{2M}{\delta} \leq \frac{\delta}{2\tau} = \frac{1}{\widetilde{R}}$. Hence (K2) holds.

Hence we can indeed apply Theorem 4 and we obtain that there is a unique $\theta_p \in B(\theta^\star, \frac{2\tau}{\delta})$ such that $G(\theta_p) = 0$, that is $F(\theta_p) = p(\theta_p)$ □

Proof of Theorem 1.

Proof of the first conclusion. Set $\theta^\star \in \mathrm{crit}\, L_1$. We apply Proposition 4 with $F = \nabla L_1$, $p = -\nabla L_0$ and with $\theta^\star, M, \tau$ there given by the same notation here. We take δ there as $\delta/2$ here. Then indeed $F(\theta^\star) = 0$.

We have $\rho_{\min}(\nabla^2 L_1(\theta^\star)) \geq \delta$ from (1). Hence, using (5) and (6),

$$\rho_{\min}(\nabla^2(L_1 + L_0)(\theta^\star)) \geq \delta - \tau \geq \delta - \frac{\delta}{8} \geq \frac{\delta}{2}.$$

Hence (12) holds.

The conditions (13) to (16) hold by the assumptions (2) to (5). The condition (17) holds from (6).

Hence all the assumptions of Proposition 4 are verified. We conclude that there is a unique $\theta_p \in B(\theta^\star, \frac{4\tau}{\delta})$ such that $F(\theta_p) = p(\theta_p)$, that is $\nabla(L_1 + L_0)(\theta_p) = 0$.

Assume now that there is a different number of strictly negative eigenvalues between $\nabla^2 L_1(\theta^\star)$ and $\nabla^2 L(\theta_p)$. Write $\lambda_1(Q) \leq \cdots \leq \lambda_m(Q)$ for the m ordered eigenvalues of a symmetric $m \times m$ matrix Q. The eigenvalues of $\nabla^2 L_1(\theta^\star)$ are in $\mathbb{R}\backslash[-\delta, \delta]$ since we have observed that $\rho_{\min}(\nabla^2 L_1(\theta^\star)) \geq \delta$. Hence, if $\nabla^2 L_1(\theta^\star)$ and $\nabla^2(L_1 + L_0)(\theta_p)$ do not have the same number of strictly negative eigenvalues, there would exist $i \in \{1, \ldots, d\}$ such that $\left|\lambda_i(\nabla^2 L_1(\theta^\star)) - \lambda_i(\nabla^2(L_1 + L_0)(\theta_p))\right| \geq \delta$. However from Problem 4.3.P1 in [33], we have

$$\begin{aligned}\left|\lambda_i(\nabla^2 L_1(\theta^\star)) - \lambda_i(\nabla^2(L_1 + L_0)(\theta_p))\right| \leq & \rho_{\max}\left(\nabla^2 L_1(\theta^\star) - \nabla^2(L_1 + L_0)(\theta_p)\right) \\ \leq & \rho_{\max}\left(\nabla^2 L_1(\theta^\star) - \nabla^2 L_1(\theta_p)\right) + \rho_{\max}\left(\nabla^2 L_0(\theta_p)\right) \\ \text{from (2) and (5)} \quad \leq & \frac{4\tau M}{\delta} + \tau \\ \text{from (6)} \quad \leq & \frac{\delta}{8} + \frac{\delta}{8} \\ < & \delta.\end{aligned}$$

This is a contradiction and thus $\nabla^2 L_1(\theta^\star)$ and $\nabla^2(L_1 + L_0)(\theta_p)$ have the same number of strictly negative eigenvalues.

Consider now $\theta^\star \in \mathrm{crit}\,(L_1 + L_0)$. We will apply Proposition 4 with $F = \nabla L_1 + \nabla L_0$ and $p = \nabla L_0$. In Proposition 4 we will take for τ the same value as here. The quantity M in Proposition 4 will

be taken as $2M$ here. The quantity δ in Proposition 4 will be taken as $\delta/2$ here. Let us check the conditions of Proposition 4.

We have, from (4), and since $\nabla(L_1 + L_0)(\theta^\star) = 0$,

$$\|\nabla L_1(\theta^\star)\| \leq \tau \leq \frac{c}{2}.$$

Hence from (1), $\rho_{\min}(\nabla^2 L_1(\theta^\star)) \geq \delta$. Hence, from (5),

$$\rho_{\min}(\nabla^2(L_1 + L_0)(\theta^\star)) \geq \delta - \tau \geq \frac{\delta}{2} \tag{18}$$

because by assumption $\tau \leq \delta/8$. Hence (12) holds (with δ in (12) taken as $\delta/2$ here). Next, for all $\theta_1, \theta_2 \in \mathbb{R}^d$, from (2) and (3),

$$\rho_{\max}\left(\nabla^2(L_1 + L_0)(\theta_1) - \nabla^2(L_1 + L_0)(\theta_2)\right) \leq M\|\theta_1 - \theta_2\| + M\|\theta_1 - \theta_2\| = 2M\|\theta_1 - \theta_2\|. \tag{19}$$

Hence (13) holds (with M in (13) taken as $2M$ here).

The conditions (14), (15) and (16) hold by assumption from (3), (4) and (5). Equation (17) in Proposition 4 holds from (6) in Theorem 1.

Hence the conclusion of Proposition 4 holds and there is $\theta_p \in B(\theta^\star, \frac{4\tau}{\delta})$ such that $\nabla(L_1 + L_0)(\theta_p) = \nabla L_0(\theta_p)$, that is $\nabla L_1(\theta_p) = 0$.

Similarly as above, assume now that there is a different number of strictly negative eigenvalues between $\nabla^2(L_1 + L_0)(\theta^\star)$ and $\nabla^2 L_1(\theta_p)$. Since we have established $\rho_{\min}(\nabla^2(L_1 + L_0)(\theta^\star)) \geq \delta/2$ from (18), there would exist $i \in \{1, \ldots, d\}$ such that $\left|\lambda_i(\nabla^2(L_1 + L_0)(\theta^\star)) - \lambda_i(\nabla^2 L_1(\theta_p))\right| \geq \delta/2$. However from Problem 4.3.P1 in [33], we have

$$\begin{aligned}
\left|\lambda_i(\nabla^2(L_1 + L_0)(\theta^\star)) - \lambda_i(\nabla^2 L_1(\theta_p))\right| \leq & \rho_{\max}\left(\nabla^2(L_1 + L_0)(\theta^\star) - \nabla^2 L_1(\theta_p)\right) \\
\leq & \rho_{\max}(\nabla^2 L_0(\theta^\star)) + \rho_{\max}\left(\nabla^2 L_1(\theta^\star) - \nabla^2 L_1(\theta_p)\right) \\
\text{from (2) and (4)} \quad \leq & \tau + \frac{4\tau M}{\delta} \\
\text{from (6)} \quad \leq & \frac{\delta}{8} + \frac{\delta}{8} \\
< & \frac{\delta}{2}.
\end{aligned}$$

This is a contradiction and thus $\nabla^2 L_1(\theta^\star)$ and $\nabla^2(L_1 + L_0)(\theta_p)$ have the same number of strictly negative eigenvalues.

Hence we have established the first conclusion of the theorem.

Proof of the second conclusion. To establish the second conclusion, consider $\theta^\star \in \operatorname{crit} L_1$. Let us apply Proposition 4 with $F = \nabla L_1$, p taken as the zero function, δ there equal to δ here, M there taken as M here and τ there taken as a quantity that we write τ' and that is arbitrarily close to but strictly smaller than $\frac{\delta^2}{4M}$. With similar arguments as above, we can check that (12) holds, and that (13) holds. Trivially, (14), (15) and (16) hold. Finally (17) holds because

$$\frac{\tau'}{\delta} < \frac{\delta^2}{4M\delta} = \frac{\delta}{4M}.$$

Hence the conclusion of Proposition 4 is that for $\theta' \in \operatorname{crit} L_1$, $\theta' \neq \theta^\star$, we have

$$\|\theta^\star - \theta'\| \geq \frac{2\tau'}{\delta}.$$

Thus, letting τ' arbitrarily close to $\frac{\delta^2}{4M}$, we get

$$\|\theta^\star - \theta'\| \geq \frac{2}{\delta}\frac{\delta^2}{4M} = \frac{\delta}{2M}.$$

Conversely, consider $\theta^\star \in \operatorname{crit}(L_1 + L_0)$. Let us apply Proposition 4 with $F = \nabla(L_1 + L_0)$, p taken as the zero function, δ there equal to $\delta/2$ here, M there taken as $2M$ here and τ there taken

as a quantity that we write τ' and that is arbitrarily close but strictly smaller to $\frac{\delta^2}{64M}$. With similar arguments as above, we can check that (12) holds, and that (13) holds. Trivially, (14), (15) and (16) hold. Finally (17) holds because

$$\frac{\tau'}{\frac{\delta}{2}} < \frac{2}{\delta}\frac{\delta^2}{64M} = \frac{\delta}{32M} \leq \frac{\frac{\delta}{2}}{4(2M)}.$$

Hence the conclusion of Proposition 4 is that for $\theta' \in \operatorname{crit}(L_1 + L_0)$ with $\theta' \neq \theta^\star$, we have

$$\|\theta^\star - \theta'\| \geq 2\frac{\tau'}{\delta}.$$

Hence, letting τ' arbitrarily close to $\frac{\delta^2}{64M}$, we have

$$\|\theta - \theta'\| \geq \frac{2}{\delta}\frac{\delta^2}{64M} = \frac{\delta}{32M}.$$

Proof of the third conclusion. Since $\operatorname{crit} L_1 \cap K$ is bounded, and since to each $\theta \in \operatorname{crit} L_1 \cap K$ we can associate a ball of fixed radius containing no other points of $\operatorname{crit} L_1 \cap K$ (second conclusion), we deduce that $\operatorname{crit} L_1 \cap K$ is a finite set. Similarly, $\operatorname{crit}(L_1 + L_0) \cap K$ is a finite set. □

Proof of Corollary 1. From Theorem 1, for $\theta \in \operatorname{crit} L_1$, there is $\theta' \in \operatorname{crit} L$ such that $\|\theta - \theta'\| \leq \frac{4\tau}{\delta}$. Conversely for $\theta \in \operatorname{crit} L$, there is $\theta' \in \operatorname{crit} L_1$ such that $\|\theta - \theta'\| \leq \frac{4\tau}{\delta}$. Hence $\operatorname{dist}_{\mathrm{H}}(\operatorname{crit} L_1, \operatorname{crit} L) \leq \frac{4\tau}{\delta}$.

Also from Theorem 1, for $\theta \in \operatorname{argmin-loc} L_1$, since $\theta \in \operatorname{crit} L_1$, there is $\theta' \in \operatorname{crit} L$ such that $\|\theta - \theta'\| \leq \frac{4\tau}{\delta}$. Also $\nabla^2 L_1(\theta)$ has no strictly negative eigenvalues since $\theta \in \operatorname{argmin-loc} L_1$. Hence from Theorem 1, $\nabla^2 L(\theta')$ has no strictly negative eigenvalues. As observed in (18) in the proof of Theorem 1, $\nabla^2 L(\theta')$ has no zero eigenvalues. Hence $\theta' \in \operatorname{argmin-loc} L$. Similarly, for $\theta \in \operatorname{argmin-loc} L$, we can show that there is $\theta' \in \operatorname{argmin-loc} L_1$ with $\|\theta - \theta'\| \leq \frac{4\tau}{\delta}$. Hence

$$\operatorname{dist}_{\mathrm{H}}(\operatorname{argmin-loc} L_1, \operatorname{argmin-loc} L) \leq \frac{4\tau}{\delta}.$$

Finally, from (8), for each $\theta \in \operatorname{argmin-loc} L_1$, there is indeed $\theta' \in \operatorname{argmin-loc} L$ with $\|\theta - \theta'\| \leq \frac{4\tau}{\delta}$. For each $\widetilde{\theta} \in B(\theta', \frac{6\tau}{\delta})$, we have, using (18) and (19) from the proof of Theorem 1, and then (6),

$$\lambda_{\min}\left(\nabla^2 L(\widetilde{\theta})\right) \geq \lambda_{\min}\left(\nabla^2 L(\theta')\right) - \rho_{\max}\left(\nabla^2 L(\widetilde{\theta}) - \nabla^2 L(\theta')\right) \geq \frac{\delta}{2} - \frac{12\tau M}{\delta} \geq \frac{\delta}{8}.$$

This concludes the proof.

□

B.2 Proof of Section 2.4

Proof of Theorem 2. Let us apply Proposition 4 to the linear model. We let, for $i = 0, 1$, $f_i(\theta) = \|Y^i - X^i\theta\|^2$. We can apply Proposition 4 to

$$F = \nabla f_1, \;\; p = -\nabla f_0, \;\; \theta^\star = \widehat{\theta}_1$$

and with constants δ, M, τ to be specified later.

We have

$$\nabla f_i(\theta) = -2X^{i\top}Y^i + 2X^{i\top}X^i\theta$$

and

$$\nabla^2 f_i(\theta) = 2X^{i\top}X^i.$$

Hence taking

$$\delta = 2\rho_{\min}(X^{1\top}X^1)$$

we obtain that (12) holds in Proposition 4. Furthermore, the Hessian matrices of f_0 and f_1 are constant and thus we can take $M = 0$ in Proposition 4 while still having that (13) and (14) hold.

Next,

$$\begin{aligned}\nabla f_0(\widehat{\theta}_1) =& -2X^{0\top}Y^0 + 2X^{0\top}X^0\widehat{\theta}_1 \\ =& -2X^{0\top}Y^0 + 2X^{0\top}X^0\widehat{\theta}_0 + 2X^{0\top}X^0(\widehat{\theta}_1 - \widehat{\theta}_0) \\ =& 2X^{0\top}X^0(\widehat{\theta}_1 - \widehat{\theta}_0).\end{aligned}$$

Hence we take

$$\tau = 2\rho_{\max}(X^{0\top}X^0)\left(1 + \|\widehat{\theta}_1 - \widehat{\theta}_0\|\right)$$

to ensure that $\rho_{\max}(\nabla^2 f_0(\widehat{\theta}_1)) \leq \tau$ and $\|\nabla f_0(\widehat{\theta}_1)\| \leq \tau$. Thus, (15) and (16) hold in Proposition 4. Hence, we can apply Proposition 4 that yields

$$\|\widehat{\theta} - \theta_1\| \leq \frac{2\tau}{\delta} = \frac{2\rho_{\max}(n_0 S_0)}{\rho_{\min}(n_1 S_1)}\left(1 + \|\widehat{\theta}_1 - \widehat{\theta}_0\|\right).$$

Note that the constraint (17) becomes vacuous since $M = 0$. □

B.3 Proofs and extra results of Section 3.1

Proof of Theorem 3. Consider $\theta \in Z_{\text{maj-adv}}$. We have

$$\begin{aligned}\langle \nabla L(\theta), \nabla L_1(\theta)\rangle =& \|\nabla L_1(\theta)\|^2 + \langle \nabla L_1(\theta), \nabla L_0(\theta)\rangle \\ \geq& \|\nabla L_1(\theta)\|^2 - \|\nabla L_1(\theta)\| \cdot \|\nabla L_0(\theta)\| \\ =& \|\nabla L_1(\theta)\|\left(\|\nabla L_1(\theta)\| - \|\nabla L_0(\theta)\|\right).\end{aligned}$$

Note that, by (4), $\|\nabla L_0(\theta)\| \leq \tau$. Hence, since $\theta \in Z_{\text{maj-adv}}$, we have $\|\nabla L_1(\theta)\| \leq \tau$.

We then apply Theorem 4 with $\theta^\star$ there equal to θ here, with G equal to ∇L_1 and with $\widetilde{R}$ equal to $\frac{2\tau}{\delta}$. Since $\tau \leq c$ by (6), then from (1), we have $\rho_{\min}(\nabla^2 L_1(\theta)) \geq \delta$. Hence

$$\|(\nabla^2 L_1(\theta))^{-1}\nabla L_1(\theta)\| \leq \frac{\tau}{\delta}$$

and thus (K1) holds in Theorem 4. Also, for all $\theta', \theta'' \in B(\theta, \widetilde{R})$, we have from (2),

$$\rho_{\max}\left((\nabla^2 L_1(\theta))^{-1}\left(\nabla^2 L_1(\theta') - \nabla^2 L_1(\theta'')\right)\right) \leq \frac{M\|\theta' - \theta''\|}{\delta} \leq \frac{\|\theta' - \theta''\|}{\widetilde{R}},$$

because $\frac{M}{\delta} \leq \frac{1}{\widetilde{R}}$ since $\frac{2\tau}{\delta} \leq \frac{\delta}{M}$ since $\tau \leq \frac{\delta^2}{2M}$ from (6). Hence (K2) holds in Theorem 4.

Hence Theorem 4 implies that there exists $\widetilde{\theta}$ such that $\nabla L_1(\widetilde{\theta}) = 0$ and $\|\widetilde{\theta} - \theta\| \leq \widetilde{R} = \frac{2\tau}{\delta}$. Hence we have

$$\theta \in \bigcup_{\widehat{\theta}_1 \in \text{crit } L_1} B\left(\widehat{\theta}_1, \frac{2\tau}{\delta}\right)$$

which concludes the proof. □

Proof of Lemma 1. We have

$$\begin{aligned}\langle \nabla L(\theta), \nabla L_0(\theta)\rangle =& \langle \nabla L(\theta), \nabla L(\theta)\rangle - \langle \nabla L(\theta), \nabla L_1(\theta)\rangle \\ \geq& 0,\end{aligned}$$

because $\theta \in Z_{\text{maj-adv}}$ means by definition that $\langle \nabla L(\theta), \nabla L_1(\theta)\rangle \leq 0$. The rest is the classical Lyapunov computation. □

Proposition 5. *Consider that $L_0, L_1 : \mathbb{R}^d \to \mathbb{R}$ are twice continuously differentiable with the properties that*

$$\nabla(L_1(\theta)) = 0 \implies \rho_{\min}(\nabla^2 L_1(\theta)) > 0, \qquad \theta \in \mathbb{R}^d \tag{20}$$

and, using $L = L_1 + L_0$,

$$\nabla(L(\theta)) = 0 \implies \rho_{\min}(\nabla^2 L(\theta)) > 0, \qquad \theta \in \mathbb{R}^d. \tag{21}$$

Then, recalling the symmetric difference notation

$$\operatorname{crit} L_1 \,\Delta\, \operatorname{crit} L := (\operatorname{crit} L_1 \cup \operatorname{crit} L) \backslash (\operatorname{crit} L_1 \cap \operatorname{crit} L),$$

we have

$$\operatorname{crit} L_1 \,\Delta\, \operatorname{crit} L \subset \operatorname{bdry} Z_{\text{maj-adv}}\,.$$

Proof of Proposition 5. Consider $\widehat{\theta}_1 \in \operatorname{crit} L_1 \backslash \operatorname{crit} L$. Then $\nabla L(\widehat{\theta}_1) \neq 0$. By continuity, there exists $\epsilon_0 > 0$ such that for $\|\theta - \widehat{\theta}_1\| \leq \epsilon_0$,

$$\|\nabla L(\theta) - \nabla L(\widehat{\theta}_1)\| \leq \frac{1}{2}\|\nabla L(\widehat{\theta}_1)\|.$$

Consider $0 < \epsilon < \epsilon_0$. From (20) and from the local inversion theorem, there are neighborhoods U of $\widehat{\theta}_1$ and V of $0 \in \mathbb{R}^d$ such that $U \subset B(\widehat{\theta}_1, \epsilon)$ and such that ∇L_1 is bijective from U to V.

Hence, there is $t_\epsilon > 0$ (small enough) and there is $\widetilde{\theta} \in U$ such that

$$\nabla L_1(\widetilde{\theta}) = t_\epsilon \nabla L(\widehat{\theta}_1) \in V. \tag{22}$$

Hence

$$\begin{aligned}
\left\langle \nabla L_1(\widetilde{\theta}), \nabla L(\widetilde{\theta}) \right\rangle =& t_\epsilon \left\langle \nabla L(\widehat{\theta}_1), \nabla L(\widetilde{\theta}) \right\rangle \\
=& t_\epsilon \left\langle \nabla L(\widehat{\theta}_1), \nabla L(\widehat{\theta}_1)) \right\rangle + t_\epsilon \left\langle \nabla L(\widehat{\theta}_1), \nabla L(\widetilde{\theta}) - \nabla L(\widehat{\theta}_1) \right\rangle \\
\geq& t_\epsilon \left\| \nabla L(\widehat{\theta}_1) \right\|^2 - t_\epsilon \left\| \nabla L(\widehat{\theta}_1) \right\| \cdot \left\| \nabla L(\widetilde{\theta}) - \nabla L(\widehat{\theta}_1) \right\| \\
\geq& t_\epsilon \left\| \nabla L(\widehat{\theta}_1) \right\|^2 - t_\epsilon \left\| \nabla L(\widehat{\theta}_1) \right\| \cdot \frac{\left\| \nabla L(\widehat{\theta}_1) \right\|}{2} \\
>& 0.
\end{aligned}$$

We can proceed similarly as from (22) but this time with $\widetilde{\theta}' \in U$, $t'_\epsilon < 0$ and

$$\nabla L_1(\widetilde{\theta}') = t'_\epsilon \nabla L(\widehat{\theta}_1) \in V.$$

This yields

$$\left\langle \nabla L_1(\widetilde{\theta}), \nabla L(\widetilde{\theta}) \right\rangle < 0.$$

Since this holds for any $\epsilon > 0$ there are two sequences $(\widetilde{\theta}_k)_k$ and $(\widetilde{\theta}'_k)_k$ that converge to $\widehat{\theta}_1$ with $\widetilde{\theta}_k \in Z^c_{\text{maj-adv}}$ and $\widetilde{\theta}'_k \in Z_{\text{maj-adv}}$. Hence $\widehat{\theta}_1 \in \operatorname{bdry} Z_{\text{maj-adv}}$. Hence

$$\operatorname{crit} L_1 \backslash \operatorname{crit} L \subset \operatorname{bdry} Z_{\text{maj-adv}}\,.$$

We can show symmetrically

$$\operatorname{crit} L \backslash \operatorname{crit} L_1 \subset \operatorname{bdry} Z_{\text{maj-adv}}\,.$$

□

The minority-adverse zone can be large. The minority-adverse zone is defined as

$$Z_{\text{min-adv}} = \{\theta \in \mathbb{R}^d : \langle \nabla L(\theta), \nabla L_0(\theta) \rangle \leq 0\}$$

and is the counterpart to the majority-adverse zone in (10). In the linear regression setting, the next proposition exhibits a ball of radius R that is contained in the minority-adverse zone. This radius R is large whenever $\|\widehat{\theta} - \widehat{\theta}_0\|$ is large and S and S_0 are well-conditioned. Hence, roughly speaking, while Theorem 3 states that the majority-adverse zone is always small, the next proposition states that the minority-adverse zone can be large. Hence, gradient descents on L may not decrease L_0 over long training times, which is a conclusion of our numerical experiments in Section 4.

Proposition 6. *Assume $\widehat{\theta} \neq \widehat{\theta}_0$. Let*

$$R = \frac{\rho_{\min}(S_0)\rho_{\min}(S)}{33\rho_{\max}(S_0)\rho_{\max}(S)}\|\widehat{\theta} - \widehat{\theta}_0\|. \tag{23}$$

Then there exists $\overline{\theta} \in \mathbb{R}^d$ such that $B(\overline{\theta}, R) \subset Z_{\text{min-adv}}$.

Proof of Proposition 6. Let

$$\theta(t) = \widehat{\theta} + e^{-tS}\left(\widehat{\theta}_0 - \widehat{\theta}\right).$$

Then as in Lemma 3, we have

$$\begin{aligned}
\frac{d}{dt}L_0(\theta(t)) =& \Big\langle \left(\nabla L_0\right)(\theta(t)), \frac{d}{dt}\theta(t)\Big\rangle \\
=& \Big\langle \left(\nabla L_0\right)(\theta(t)), -(\nabla L)(\theta(t))\Big\rangle \\
=& -\mathcal{S}(\theta(t)),
\end{aligned} \tag{24}$$

defining

$$\mathcal{S}(\theta) = \Big\langle \nabla L_0(\theta), \nabla L(\theta)\Big\rangle.$$

Let

$$T = \frac{1}{\rho_{\min}(S)}.$$

Then

$$\|\theta(T) - \widehat{\theta}\| \leq e^{-T\rho_{\min}(S)}\|\widehat{\theta}_0 - \widehat{\theta}\| \leq \frac{\|\widehat{\theta}_0 - \widehat{\theta}\|}{2}.$$

Then, using Lemma 2, for any $\widetilde{\theta}$ in the segment between $\theta(T)$ and $\widehat{\theta}$,

$$\begin{aligned}
\|\nabla L_0(\widetilde{\theta})\| =& \|\frac{n_0}{n}S_0(\widetilde{\theta} - \widehat{\theta}_0)\| \\
\text{(convexity of Euclidean norm:)} \quad \leq& \frac{n_0}{n}\rho_{\max}(S_0)\left(\|\theta(T) - \widehat{\theta}_0\| + \|\widehat{\theta} - \widehat{\theta}_0\|\right) \\
\leq& \frac{n_0}{n}\rho_{\max}(S_0)\left(\|\theta(T) - \widehat{\theta}\| + 2\|\widehat{\theta} - \widehat{\theta}_0\|\right) \\
\leq& 3\frac{n_0}{n}\rho_{\max}(S_0)\|\widehat{\theta} - \widehat{\theta}_0\|.
\end{aligned}$$

Then, by convexity,

$$\begin{aligned}
L_0(\theta(T)) - L_0(\widehat{\theta}_0) \geq& \frac{n_0}{2n}\rho_{\min}(S_0)\|\theta(T) - \widehat{\theta}_0\|^2. \\
\geq& \frac{n_0}{8n}\rho_{\min}(S_0)\|\widehat{\theta} - \widehat{\theta}_0\|^2,
\end{aligned}$$

since $\|\theta(T) - \widehat{\theta}\| \leq \|\widehat{\theta} - \widehat{\theta}_0\|/2$. Also, using (24),

$$L_0(\theta(T)) - L_0(\widehat{\theta}_0) = \int_0^T \frac{dL_0(\theta(t))}{dt}dt = \int_0^T -S(\theta(t))dt \leq -T\min_{t\in[0,T]} S(\theta(t)).$$

Combining the two last displays,

$$\begin{aligned}
\min_{t\in[0,T]} \mathcal{S}(\theta(t)) \leq& \frac{L_0(\widehat{\theta}_0) - L_0(\theta(T))}{T} \\
\leq& -\frac{\frac{n_0}{n}\rho_{\min}(S_0)\|\widehat{\theta} - \widehat{\theta}_0\|^2}{8T} \\
=& -\frac{n_0}{8n}\rho_{\min}(S_0)\rho_{\min}(S)\|\widehat{\theta} - \widehat{\theta}_0\|^2.
\end{aligned}$$

Let $\overline{\theta} = \theta(\overline{t})$ with $\overline{t} \in \operatorname*{argmin}_{t\in[0,T]} S(\theta(t))$. For $\theta \in B(\overline{\theta}, R)$, we have

$$\begin{aligned}
\left|\mathcal{S}(\theta) - \mathcal{S}(\overline{\theta})\right| &= \left|\left\langle \nabla L(\theta), \nabla L_0(\theta)\right\rangle - \left\langle \nabla L(\overline{\theta}), \nabla L_0(\overline{\theta})\right\rangle\right| \\
&= \left|\left\langle \nabla L(\theta), \nabla L_0(\theta) - \nabla L_0(\overline{\theta})\right\rangle + \left\langle \nabla L(\theta) - \nabla L(\overline{\theta}), \nabla L_0(\overline{\theta})\right\rangle\right| \\
\text{(Lemma 2:)} \quad &\leq \|\nabla L(\theta)\| \frac{n_0}{n}\rho_{\max}(S_0)\|\theta - \overline{\theta}\| + \rho_{\max}(S)\|\theta - \overline{\theta}\|\|\nabla L_0(\overline{\theta})\| \\
&\leq R\frac{n_0}{n}\rho_{\max}(S_0)\|\nabla L(\theta)\| + R\rho_{\max}(S)\|\nabla L_0(\overline{\theta})\| \\
\text{(Lemma 2:)} \quad &\leq R\frac{n_0}{n}\rho_{\max}(S_0)\rho_{\max}(S)\|\theta - \widehat{\theta}\| + R\rho_{\max}(S)\frac{n_0}{n}\rho_{\max}(S_0)\|\overline{\theta} - \widehat{\theta}_0\| \\
&\leq R\frac{n_0}{n}\rho_{\max}(S_0)\rho_{\max}(S)\left(R + \|\overline{\theta} - \widehat{\theta}\| + \|\overline{\theta} - \widehat{\theta}_0\|\right).
\end{aligned}$$

We recall

$$\theta(t) = \widehat{\theta} + e^{-tS}\left(\widehat{\theta}_0 - \widehat{\theta}\right).$$

Hence, $\|\theta(t) - \widehat{\theta}\| \leq \|\widehat{\theta}_0 - \widehat{\theta}\|$ and $\|\theta(t) - \widehat{\theta}_0\| \leq 2\|\widehat{\theta}_0 - \widehat{\theta}\|$. Thus we have

$$\left|\mathcal{S}(\theta) - \mathcal{S}(\overline{\theta})\right| \leq R\frac{n_0}{n}\rho_{\max}(S_0)\rho_{\max}(S)\left(R + 3\|\widehat{\theta}_0 - \widehat{\theta}\|\right).$$

Hence, let us take R as in (23), with in particular $R \leq \|\widehat{\theta}_0 - \widehat{\theta}\|$. Then to satisfy $\langle \nabla L(\theta), \nabla L_0(\theta)\rangle \leq 0$ for all $\theta \in B(\overline{\theta}, R)$, it is sufficient that

$$R\frac{n_0}{n}\rho_{\max}(S_0)\rho_{\max}(S)\left(R + 3\|\widehat{\theta}_0 - \widehat{\theta}\|\right) < \frac{1}{8}\frac{n_0}{n}\rho_{\min}(S_0)\rho_{\min}(S)\|\widehat{\theta} - \widehat{\theta}_0\|^2.$$

For this it is sufficient that

$$32R\frac{n_0}{n}\rho_{\max}(S_0)\rho_{\max}(S)\|\widehat{\theta}_0 - \widehat{\theta}\| < \frac{n_0}{n}\rho_{\min}(S_0)\rho_{\min}(S)\|\widehat{\theta} - \widehat{\theta}_0\|^2$$

which is implied by

$$R < \frac{\rho_{\min}(S_0)\rho_{\min}(S)}{32\rho_{\max}(S_0)\rho_{\max}(S)}\|\widehat{\theta} - \widehat{\theta}_0\|.$$

This concludes the proof. □

B.4 Proofs and extra results of Section 3.2

The following lemma provides the expression of the (well-known) solutions of

$$\frac{d}{dt}\theta(t) = -\nabla L(\theta(t)), \quad \frac{d}{dt}\theta_i(t) = -\nabla L_i(\theta_i(t)),\ i = 1, 2, \quad \theta(0) = \theta_0(0) = \theta_1(0) = \theta_{\text{init}}.$$

Note that, with the uniqueness assumption, we have $\widehat{\theta} = (nS)^{-1}X^\top Y$ and $\widehat{\theta}_i = (n_i S_i)^{-1}X^{i\top}Y^i$.

Lemma 3. *We have, for $t \geq 0$,*

$$\theta(t) = \widehat{\theta} + e^{-tS}\left(\theta_{\text{init}} - \widehat{\theta}\right)$$

and for $i = 0, 1$ and $t \geq 0$,

$$\theta_i(t) = \widehat{\theta}_i + e^{-t(n_i/n)S_i}\left(\theta_{\text{init}} - \widehat{\theta}_i\right).$$

Proof of Lemma 3. We have

$$\begin{aligned}
\frac{d}{dt}\theta(t) &= -(\nabla L)(\theta(t)) \\
&= -\frac{1}{n}X^\top X\theta(t) + \frac{1}{n}X^\top Y \\
&= -\frac{1}{n}X^\top X\theta(t) + \frac{1}{n}X^\top X(X^\top X)^{-1}X^\top Y \\
&= -S\theta(t) + S\widehat{\theta}.
\end{aligned}$$

Hence,

$$\frac{d}{dt}\left(\theta(t)-\widehat{\theta}\right)=-S\left(\theta(t)-\widehat{\theta}\right)$$

and $\theta(0)-\widehat{\theta}=\theta_{\text{init}}-\widehat{\theta}$. Hence

$$\theta(t)=\widehat{\theta}+e^{-tS}\left(\theta_{\text{init}}-\widehat{\theta}\right).$$

We then provide a similar proof for $\theta_i(t)$. We have

$$\begin{aligned}\frac{d}{dt}\theta_i(t)=&-(\nabla L_i)(\theta(t))\\ =&-\frac{1}{n}X^{i\top}X^i\theta(t)+\frac{1}{n}X^{i\top}Y^i\\ =&-\frac{1}{n}X^{i\top}X^i\theta(t)+\frac{1}{n}X^{i\top}X^i(X^{i\top}X^i)^{-1}X^{i\top}Y^i\\ =&-\frac{n_i}{n}S_i\theta(t)+\frac{n_i}{n}S_i\widehat{\theta}_i.\end{aligned}$$

Hence,

$$\frac{d}{dt}\left(\theta_i(t)-\widehat{\theta}_i\right)=-\frac{n_i}{n}S_i\left(\theta(t)-\widehat{\theta}_i\right)$$

and $\theta_i(0)-\widehat{\theta}_i=\theta_{\text{init}}-\widehat{\theta}_i$. Hence

$$\theta_i(t)=\widehat{\theta}_i+e^{-t\frac{n_i}{n}S_i}\left(\theta_{\text{init}}-\widehat{\theta}_i\right).$$

This concludes the proof. □

Proof of Proposition 1. We have, using Lemma 3,

$$\begin{aligned}\|\theta(t)-\theta_1(t)\|\leq&\left\|\left(I_d-e^{-tS}\right)\widehat{\theta}-\left(I_d-e^{-t\frac{n_1}{n}S_1}\right)\widehat{\theta}_1\right\|+\left\|\left(e^{-tS}-e^{-t\frac{n_1}{n}S_1}\right)\theta_{\text{init}}\right\|\\ =&\left\|\left(I_d-e^{-tS}\right)\left(\widehat{\theta}-\widehat{\theta}_1\right)+\left(e^{-t\frac{n_1}{n}S_1}-e^{-tS}\right)\widehat{\theta}_1\right\|+\left\|\left(e^{-tS}-e^{-t\frac{n_1}{n}S_1}\right)\theta_{\text{init}}\right\|\\ \leq&\rho_{\max}\left(I_d-e^{-tS}\right)\|\widehat{\theta}-\widehat{\theta}_1\|+\rho_{\max}\left(e^{-tS}-e^{-t\frac{n_1}{n}S_1}\right)\left(\|\widehat{\theta}_1\|+\|\theta_{\text{init}}\|\right).\end{aligned}$$

Since I_d-e^{-tS} has eigenvalues between 0 and 1 and from [41, Lemma 3.24], we obtain

$$\begin{aligned}\|\theta(t)-\theta_1(t)\|\leq&\|\widehat{\theta}-\widehat{\theta}_1\|+t\rho_{\max}\left(S-\frac{n_1}{n}S_1\right)e^{-t\rho_{\min}\left(\frac{n_1}{n}S_1\right)}\left(\|\widehat{\theta}_1\|+\|\theta_{\text{init}}\|\right)\\ =&\|\widehat{\theta}-\widehat{\theta}_1\|+t\rho_{\max}\left(\frac{n_0}{n}S_0\right)e^{-t\rho_{\min}\left(\frac{n_1}{n}S_1\right)}\left(\|\widehat{\theta}_1\|+\|\theta_{\text{init}}\|\right).\end{aligned}$$

The maximizer (over t) of $te^{-t\rho_{\min}\left(\frac{n_1}{n}S_1\right)}$ is $t_{\max}=1/\rho_{\min}\left(\frac{n_1}{n}S_1\right)$ which yields

$$\sup_{t>0}\|\theta(t)-\theta_1(t)\|\leq\|\widehat{\theta}-\widehat{\theta}_1\|+\frac{\rho_{\max}(n_0S_0)(\|\widehat{\theta}_1\|+\|\theta_{\text{init}}\|)}{e\cdot\rho_{\min}(n_1S_1)}.$$

This concludes the proof. □

Lemma 4 (Duration for proximity to a local minimizer)**.** *Consider a function $L:\mathbb{R}^d\to\mathbb{R}$ that is twice continuously differentiable and satisfies, for some $M<\infty$ and for all $\theta\in\mathbb{R}^d$,*

$$\rho_{\max}\left(\nabla^2L(\theta)\right)\leq M. \tag{25}$$

Assume that there exists a trajectory $[0,\infty)\ni t\mapsto\theta(t)$ satisfying

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

Consider a critical point $\widehat{\theta}$ of L such that $\widehat{\theta}\neq\theta_{\text{init}}$. Consider $\epsilon\in(0,1)$ and $t_\epsilon\in(0,\infty)$ satisfying

$$\|\theta(t_\epsilon)-\widehat{\theta}\|\leq\epsilon\|\theta_{\text{init}}-\widehat{\theta}\|.$$

Then we have

$$t_\epsilon\geq\frac{1}{M}\log\left(\frac{1}{\epsilon}\right).$$

Proof of Lemma 4. Without loss of generality, we can consider that

$$t_\epsilon = \inf\left\{t \geq 0; \|\theta(t) - \widehat{\theta}\| \leq \epsilon \|\theta_{\text{init}} - \widehat{\theta}\|\right\} < \infty.$$

Consider the function $[0, t_\epsilon) \ni u \mapsto g(u) = \|\theta(t_\epsilon - u) - \widehat{\theta}\|$. Note that this function is strictly positive and differentiable on $[0, t_\epsilon]$ (since $\|\theta(t_\epsilon - u) - \widehat{\theta}\| \geq \epsilon \|\theta_{\text{init}} - \widehat{\theta}\| > 0$ for $u \in [0, t_\epsilon]$). The derivative at $u \in [0, t_\epsilon]$ satisfies

$$\begin{aligned}
g'(u) =& \left\langle \frac{d}{du}\theta(t_\epsilon - u), \frac{\theta(t_\epsilon - u) - \widehat{\theta}}{\|\theta(t_\epsilon - u) - \widehat{\theta}\|} \right\rangle \\
=& \left\langle (\nabla L)(\theta(t_\epsilon - u)), \frac{\theta(t_\epsilon - u) - \widehat{\theta}}{\|\theta(t_\epsilon - u) - \widehat{\theta}\|} \right\rangle \\
\leq& \|(\nabla L)(\theta(t_\epsilon - u))\| \\
=& \|(\nabla L)(\theta(t_\epsilon - u)) - (\nabla L)(\widehat{\theta})\| \\
\text{Lemma 2:} \quad \leq& M\|\theta(t_\epsilon - u) - \widehat{\theta}\| \\
=& M g(u).
\end{aligned}$$

Hence we can apply Grönwall's inequality, yielding

$$g(t_\epsilon) \leq g(0) e^{M t_\epsilon} = \epsilon \|\theta_{\text{init}} - \widehat{\theta}\| e^{M t_\epsilon}.$$

On the other hand $g(t_\epsilon) = \|\theta_{\text{init}} - \widehat{\theta}\|$ and thus

$$\epsilon \|\theta_{\text{init}} - \widehat{\theta}\| e^{M t_\epsilon} \geq \|\theta_{\text{init}} - \widehat{\theta}\|.$$

This yields

$$t_\epsilon \geq \frac{1}{M} \log\left(\frac{1}{\epsilon}\right).$$

This concludes the proof. □

Proof of Proposition 3. We apply Lemma 4 with L in the lemma equal to L here, with M in the lemma equal to $\rho_{\max}(S)$ here and with ϵ in the lemma equal to ϵ here. The lemma yields

$$t_\epsilon - t_1 \geq \frac{1}{M} \log\left(\frac{1}{\epsilon}\right)$$

which directly concludes the proof. □

Proof of Proposition 2. We apply Lemma 4 with L in the lemma equal to L here, with M in the lemma equal to $\rho_{\max}(S)$ here and with ϵ in the lemma equal to

$$\frac{\|\widehat{\theta}_1 - \widehat{\theta}\|}{\|\theta_{\text{init}} - \widehat{\theta}\|}$$

here. Then we have

$$\|\widehat{\theta}_1 - \widehat{\theta}\| = \epsilon \|\theta_{\text{init}} - \widehat{\theta}\|$$

and so the lemma yields

$$t_1 \geq \frac{1}{M} \log\left(\frac{1}{\epsilon}\right)$$

which concludes the proof. □

Proposition 7 (Debiasing duration measured with the loss L_0)**.** *In the context of Proposition 3, assume that $\widehat{\theta}$, $\widehat{\theta}_0$ and $\widehat{\theta}_1$ are two-by-two distinct. Consider a representative training curve $t \mapsto \theta(t)$, such that for some t_1, $\theta(t) = \widehat{\theta}_1$. Assume that for $t \geq t_1$, $\theta(t) \in Z_{\text{maj-adv}}$.*

For $0 < \epsilon < 1$, consider t_ϵ such that

$$\frac{L_0(\theta(t_\epsilon)) - L_0(\widehat{\theta})}{L_0(\widehat{\theta}_1) - L_0(\widehat{\theta})} \leq \epsilon. \tag{26}$$

Then we have

$$t_\epsilon - t_1 \underset{\epsilon \to 0}{\longrightarrow} \infty.$$

Proof of Proposition 7. Without loss of generality, we can consider that $t_1 = 0$ and $\theta_{\text{init}} = \widehat{\theta}_1$. Because for $t \geq 0$, $\theta(t) \in Z_{\text{maj-adv}}$, the function $t \mapsto L_1(\theta(t))$ is non-decreasing. Assume that there exists $t < \infty$ such that $L_0(\theta(t)) = L_0(\widehat{\theta})$. Then $L(\theta(t)) \leq L(\widehat{\theta})$ and thus $\theta(t) = \widehat{\theta}$. From Lemma 3, this is a contradiction because $\theta_{\text{init}} \neq \widehat{\theta}$. Hence, because $L_0(\theta(0)) > L_0(\widehat{\theta})$, the function $t \mapsto L_0(\theta(t)) - L_0(\widehat{\theta})$ is strictly positive on $[0, \infty)$ by continuity.

Finally, if t_ϵ does not go to infinity, then there is a subsequence $(\epsilon_\ell)_{\ell \in \mathbb{N}}$ and a constant $T < \infty$ such that $t_{\epsilon_\ell} \leq T$. By compacity, we can extract a further convergent subsequence $(t_{\epsilon_{\ell_k}})_{k \in \mathbb{N}}$ with $t_{\epsilon_{\ell_k}} \to t^\star \in [0, T]$. We have

$$\frac{L_0(\theta(t_{\epsilon_{\ell_k}})) - L_0(\widehat{\theta})}{L_0(\widehat{\theta}_1) - L_0(\widehat{\theta})} \leq \epsilon_{\ell_k} \underset{k \to \infty}{\longrightarrow} 0$$

and thus by continuity $L_0(\theta(t^\star)) = L_0(\widehat{\theta})$. This is a contradiction, which concludes the proof. □

C Additional experiments on CIFAR-10

We provide detailed figures for additional architectures used in our CIFAR-10 experiments in Section 4.

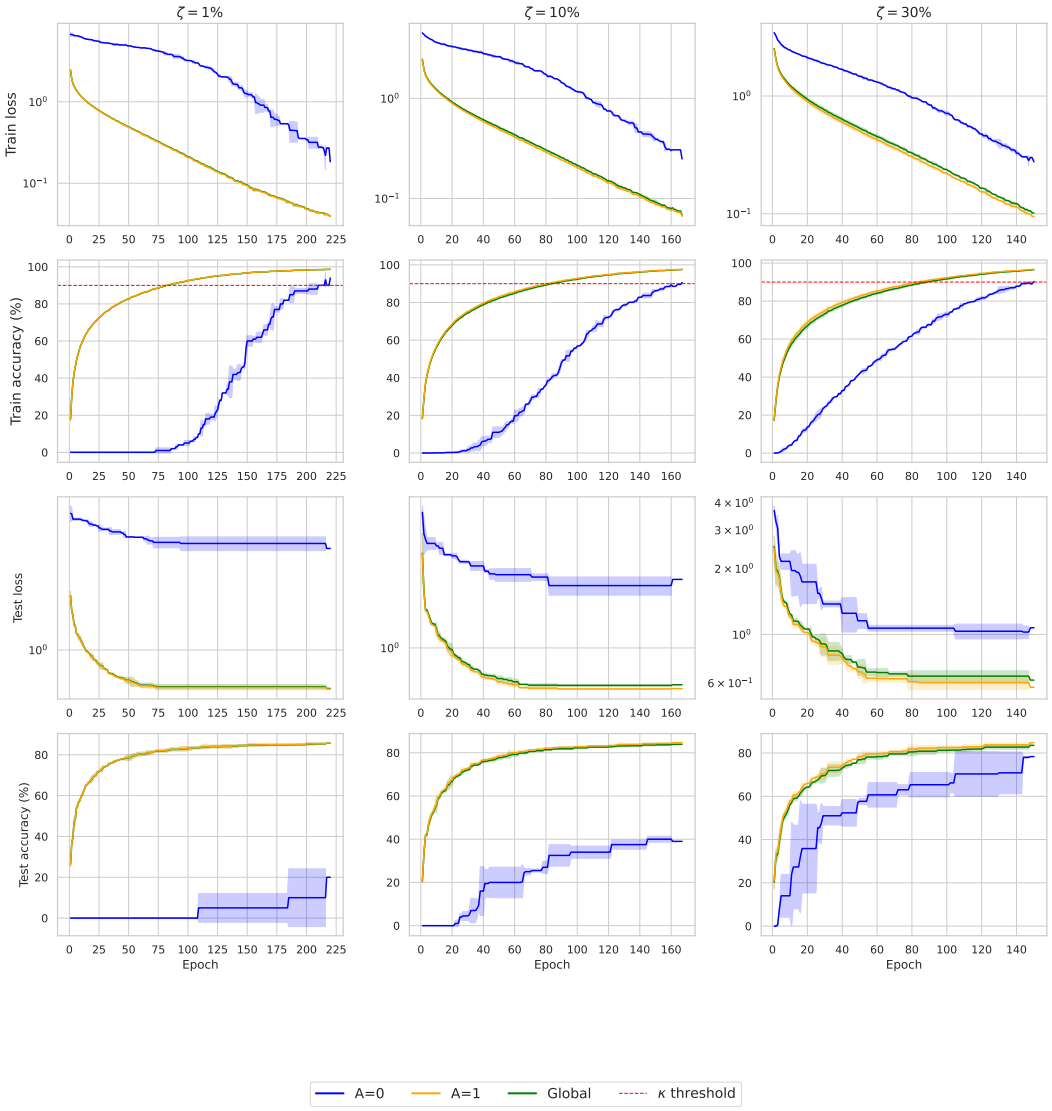

Figure 6: Training and test loss and accuracy for different subgroup imbalance scenarios (1%, 10%, and 30%) using ResNet50 on CIFAR-10 and threshold $\kappa = 90\%$.

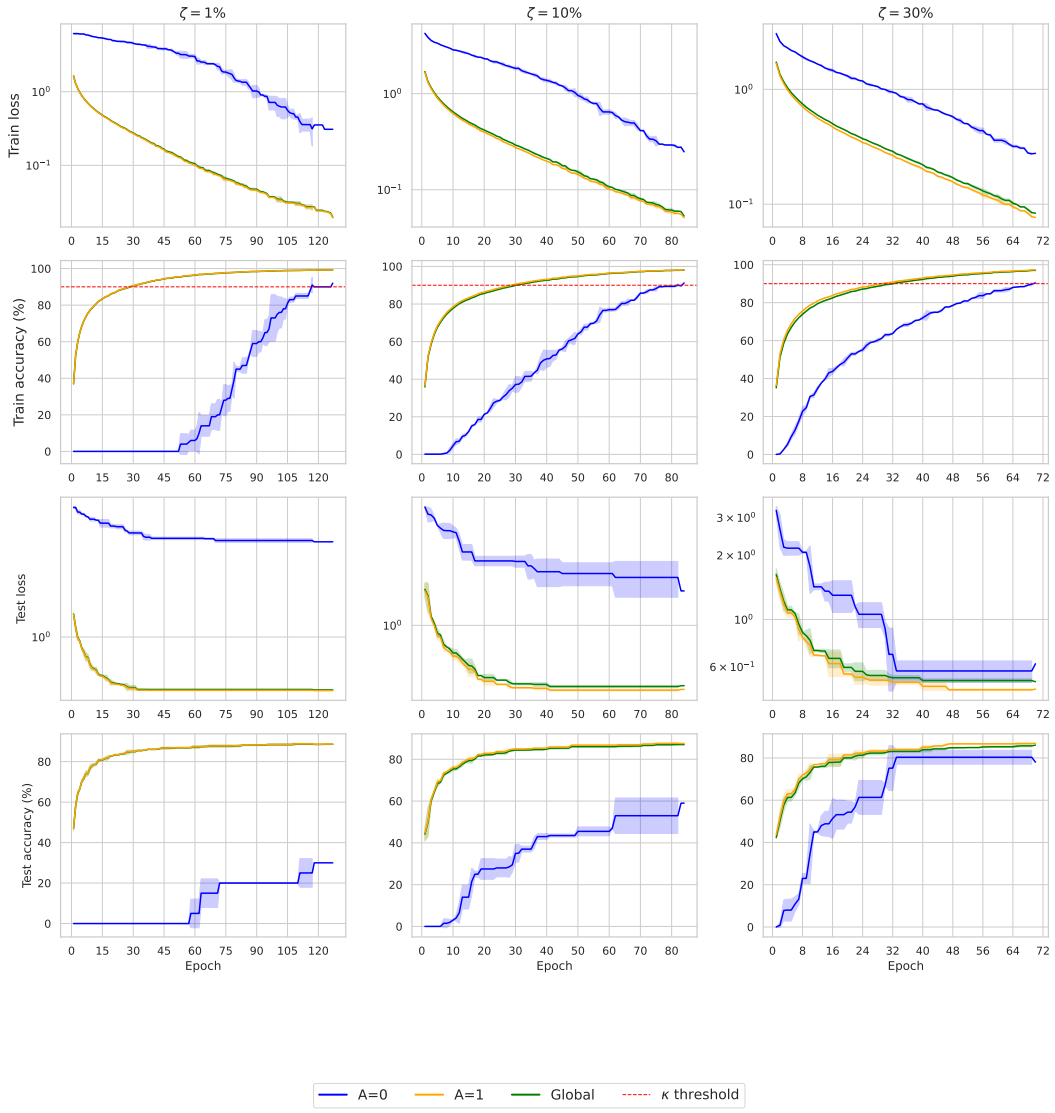

Figure 7: Training and test loss and accuracy for different subgroup imbalance scenarios (1%, 10%, and 30%) using VGG19 on CIFAR-10 and threshold $\kappa = 90\%$.

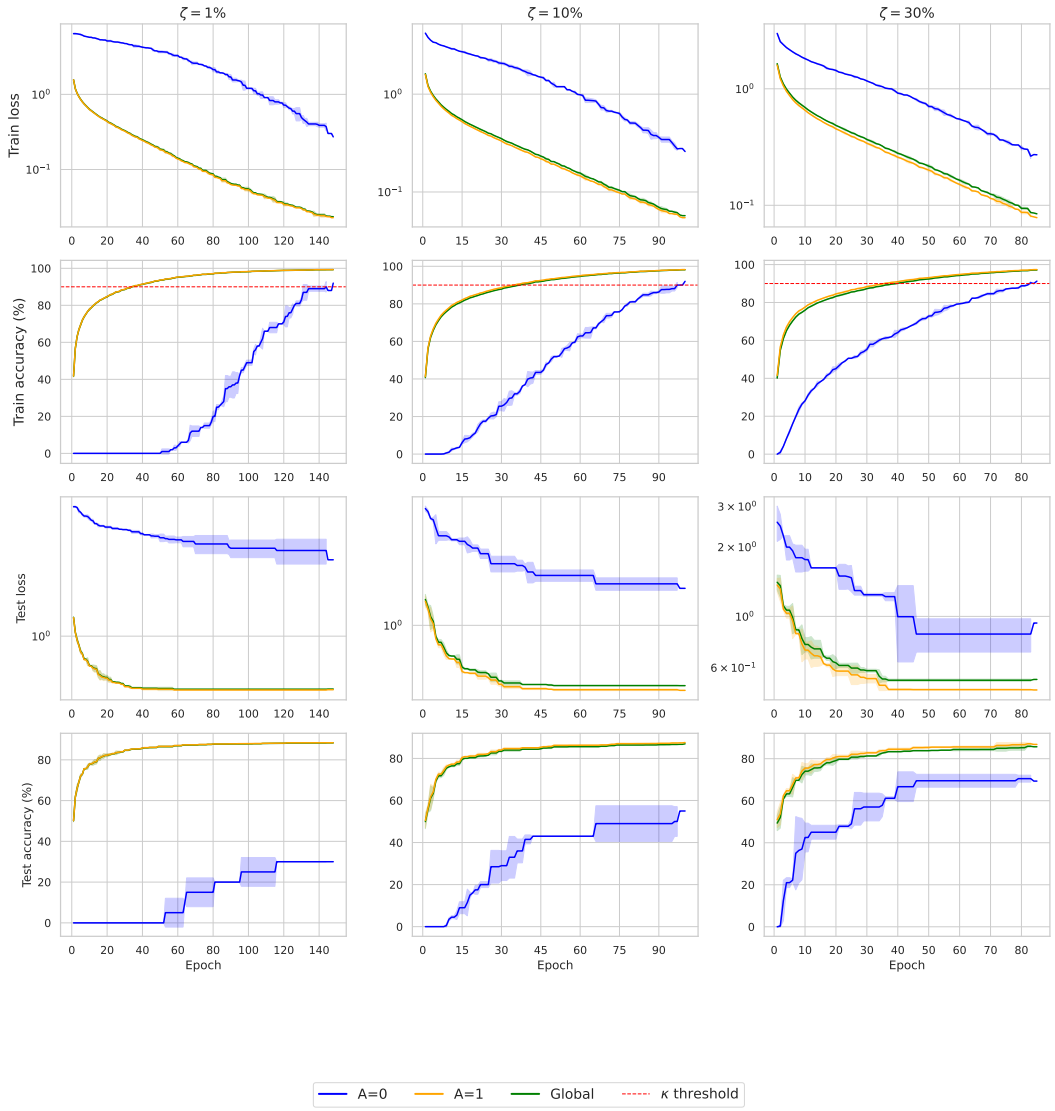

Figure 8: Training and test loss and accuracy for different subgroup imbalance scenarios (1%, 10%, and 30%) using VGG11 on CIFAR-10 and threshold $\kappa = 90\%$.

D Experimental details

This section provides full details to ensure reproducibility of our experiments. We describe hardware specifications, training hyperparameters, implementation details, and evaluation protocol for each dataset and model used.

D.1 Datasets

We conduct experiments on a mix of image and tabular datasets with varying levels of class imbalance. Below, we describe the construction and preprocessing steps for each dataset used in our study.

CIFAR-10. We use the standard CIFAR-10 dataset, consisting of 60,000 color images (32×32 pixels) in 10 classes, with 50,000 training and 10,000 test samples. To induce group imbalance, we define a binary sensitive attribute $A \in 0, 1$, following the approach detailed in Section 4.

CIFAR-2. We consider a binary classification task derived from CIFAR-10 by selecting the two vehicle-related classes "automobile" and "truck". We refer to this subset as CIFAR-2. To simulate a highly imbalanced scenario, we drastically reduce the number of "automobile" (car) samples to a small fraction of their original count (e.g., retaining only 3%), while keeping all "truck" examples. This creates a pronounced majority-minority setting, suitable for studying bias amplification under imbalance.

EuroSAT. EuroSAT is a land use and land cover classification dataset based on Sentinel-2 satellite images. We use the RGB version comprising 27,000 labeled images across 10 classes. For our binary classification task, we select two visually distinct classes: *Highway* and *River*. The input images are resized to 64×64 pixels and normalized. We define a binary sensitive attribute A by thresholding the average blue-channel intensity to distinguish between "bluish" and "non-bluish" images, following the approach of [43].

Adult. The Adult dataset is a standard benchmark for fairness and tabular learning. It contains approximately 48,000 examples with demographic and income information. We treat the binary income variable as the label and use "gender" (male vs. female) as the sensitive attribute A.

D.2 Hardware and runtime

Experiments were conducted on a computing cluster equipped with NVIDIA A100 40GB GPUs. Each experiment ran on a single GPU unless otherwise specified. Average runtime per training run is reported in Table 3.

Table 3: Average training time per run across datasets and models.

Dataset	Model	Runtime (h)	GPU
CIFAR-10	ResNet-18	0.26	A100
EuroSAT	ResNet-18	0.1	A100
Adult	TabNet	0.16	A100

D.3 Optimization and training

We use SGD with a constant learning rate for image models and tabular data. In order to match our theoretical setting, no weight decay or learning rate decay schedule was applied. Models were trained from scratch without pertaining. Refer to Table 4 for more details.

Table 4: Optimization hyperparameters for each task.

Dataset	Model	Optimizer	Learning rate
CIFAR-10	ResNet-18	SGD	1×10^{-2}
CIFAR-10	VGG19	SGD	1×10^{-2}
EuroSAT	ResNet-18	SGD	1×10^{-4}
Adult	TabNet	SGD	2×10^{-2}

D.4 Evaluation and reporting

All results are averaged over 3 random seeds. We report mean and standard deviation of accuracy and loss metrics across groups. Class imbalance ratios ζ are detailed in the main text (Section 4).

D.5 Reproducibility

All code and configuration files (including seed control, training logs, and plotting scripts) will be made publicly available upon publication. We follow best practices for reproducible research and ensure all experimental figures can be regenerated with a single command.