
Impact of Spurious Correlation on Gradient Descent Dynamics and Generalization

Boammani Aser Lompo

École de Technologie Supérieure
Montreal, Canada

boammani.lompo.1@ens.etsmtl.ca

Patrik Kenfack

École de Technologie Supérieure, MILA

Abstract

We provide a mathematical analysis of how class imbalance affects the generalization of a softmax classifier trained with gradient descent. By explicitly characterizing the evolution of the logits during training, we show that the classification error for each class converges at a rate proportional to $\frac{1}{\varepsilon h T}$, where ε is the class proportion, h the learning rate, and T the number of training steps. These theoretical predictions are confirmed by controlled experiments in realistic settings. The article formalizes, in closed form, the empirical observation that imbalance implicitly biases the optimization dynamics toward the majority class.

1 Problem Statement

We study how imbalanced training data affects the learning dynamics and generalization of neural networks. Consider a training dataset $\mathcal{D}_{\text{train}}$ composed of pairs (\mathbf{x}, y) , where $\mathbf{x} = (\mathbf{r}, \mathbf{s}) \in \mathbb{R}^d$ denotes the input features, decomposed into two parts: $\mathbf{r} \in \mathbb{R}^{d_1}$ represents the *relevant features* that truly determine the label, and $\mathbf{s} \in \mathbb{R}^{d_2}$ represents a *spurious signal*, a nuisance factor that should ideally be ignored. The binary label $y \in \{-1, 1\}$ depends only on \mathbf{r} , i.e., $y = g(\mathbf{r})$. Each feature vector \mathbf{x} is drawn from a random variable with joint distribution $p_{\mathbf{r}, \mathbf{s}}(\cdot)$. The training set can be partitioned into two disjoint subgroups: $\mathcal{D}_{\text{train}} = \mathcal{G}_{\text{maj}} \cup \mathcal{G}_{\text{min}}$.

The *majority group* \mathcal{G}_{maj} corresponds to *bias-aligned* samples, defined by

$$p(b(\mathbf{r}, \mathbf{s}) > \alpha \mid \mathcal{G}_{\text{maj}}) = 1,$$

where $b(\cdot, \cdot)$ is a continuous function that measures the alignment between relevant and spurious features (e.g., an inner product). The *minority group* \mathcal{G}_{min} consists of *bias-conflicting* samples, defined analogously by

$$p(b(\mathbf{r}, \mathbf{s}) < -\alpha \mid \mathcal{G}_{\text{min}}) = 1,$$

where $\alpha > 0$ is a fixed margin ensuring a clear separation between the two subgroups in the feature space. The minority fraction is denoted by $\varepsilon = |\mathcal{G}_{\text{min}}| / |\mathcal{D}_{\text{train}}|$. We train a neural network $f(\mathbf{x}, \theta)$ using the binary cross-entropy loss

$$\ell(f(\mathbf{x}, \theta), y) = -\log p_y(\mathbf{x}, \theta),$$

where $p_y(\mathbf{x}, \theta)$ denotes the model's predicted probability (logit) for the ground-truth class y . The empirical loss over the training dataset is

$$\mathcal{L}(\mathcal{D}_{\text{train}}, \theta) = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(\mathbf{x}, y) \in \mathcal{D}_{\text{train}}} \ell(f(\mathbf{x}, \theta), y),$$

and parameters are updated via gradient descent with step size h :

$$\theta^{t+1} = \theta^t - h \nabla_{\theta} \mathcal{L}(\mathcal{D}_{\text{train}}, \theta^t). \quad (1)$$

Our main theoretical result characterizes how the imbalance ratio ε determines the asymptotic learning pace of each subgroup:

Theorem 1.1. *Assume (\mathbf{r}, \mathbf{s}) has a finite second moment and a symmetric joint distribution $p_{\mathbf{r}, \mathbf{s}}(\cdot)$. Then, for sufficiently large t ,*

$$\mathbb{E}_{\mathcal{G}_{\text{maj}}} [1 - p_y(\mathbf{x}, \theta^t)] \propto \frac{1}{(1 - \varepsilon) t}, \quad (2)$$

$$\mathbb{E}_{\mathcal{G}_{\text{min}}} [1 - p_y(\mathbf{x}, \theta^t)] \propto \frac{1}{\varepsilon t}. \quad (3)$$

This result formalizes how the learning speed of each subgroup depends on its prevalence in the training set. Since $\varepsilon < 1 - \varepsilon$, the majority (bias-aligned) group converges faster than the minority (bias-conflicting) group, resulting in a systematic imbalance in training dynamics. In the remainder of this paper, we present a detailed proof for a simplified setting of Theorem 1.1 and provide empirical results in realistic neural network setups that confirm these theoretical predictions.

Simplified Statement

We now study a simplified setting that captures the essential dynamics of class imbalance. Each input is represented by a two-dimensional feature vector $\mathbf{x} = (\mathbf{r}, \mathbf{s}) \in \mathbb{R}^2$ and the label y is defined as $y = \text{sign}(\mathbf{r})$. The two subgroups are defined as follows:

$$p(\mathbf{r} = \mathbf{s} \wedge \mathbf{r}\mathbf{s} > \alpha^2 \mid \mathcal{G}_{\text{maj}}) = 1, \quad (4)$$

$$p(\mathbf{r} = -\mathbf{s} \wedge \mathbf{r}\mathbf{s} < -\alpha^2 \mid \mathcal{G}_{\text{min}}) = 1, \quad (5)$$

where $\alpha > 0$ denotes a fixed margin separating bias-aligned and bias-conflicting samples in feature space. In words, the *majority group* \mathcal{G}_{maj} consists of samples whose relevant and spurious features are positively correlated, while the *minority group* \mathcal{G}_{min} contains samples where the two features are perfectly anti-correlated. We define the four subgroup partitions:

$$\mathcal{G}_{\text{maj}}^+ = \mathcal{G}_{\text{maj}} \cap \{y = 1\}, \quad \mathcal{G}_{\text{maj}}^- = \mathcal{G}_{\text{maj}} \cap \{y = -1\}$$

$$\mathcal{G}_{\text{min}}^+ = \mathcal{G}_{\text{min}} \cap \{y = 1\}, \quad \mathcal{G}_{\text{min}}^- = \mathcal{G}_{\text{min}} \cap \{y = -1\}.$$

Equations (4) and (5) guarantee that

$$p(\mathbf{r} > \alpha \mid \mathcal{G}) = 1 \quad \text{for each subgroup } \mathcal{G} \in \{\mathcal{G}_{\text{maj}}^+, \mathcal{G}_{\text{min}}^+\}.$$

$$p(\mathbf{r} < -\alpha \mid \mathcal{G}) = 1 \quad \text{for each subgroup } \mathcal{G} \in \{\mathcal{G}_{\text{maj}}^-, \mathcal{G}_{\text{min}}^-\}.$$

We consider a one-layer neural network defined by

$$f(\mathbf{x}, \theta) = \begin{bmatrix} \theta_1 \mathbf{r} + \theta_2 \mathbf{s} \\ \theta_3 \mathbf{r} + \theta_4 \mathbf{s} \end{bmatrix},$$

whose outputs serve as logits. The corresponding class probabilities are obtained via the softmax function:

$$\begin{bmatrix} p_{-1}(\mathbf{x}, \theta) \\ p_1(\mathbf{x}, \theta) \end{bmatrix} = \text{Softmax}(f(\mathbf{x}, \theta)). \quad (6)$$

This simplified setting enables us to state an analytically tractable version of Theorem 1.1.

Assumption 1.2. We make the following mild assumptions:

- (1) The feature variable \mathbf{r} has the following conditional distributions:

$$p_{\mathbf{r}}(z \mid \mathcal{G}) = \frac{1}{\ln(\beta/\alpha)} \cdot \frac{1}{z} \mathbf{1}_{[\alpha, \beta]}(z), \quad \text{for each subgroup } \mathcal{G} \in \{\mathcal{G}_{\text{maj}}^+, \mathcal{G}_{\text{min}}^+\}, \quad (7)$$

$$p_{\mathcal{G}}(z | \mathcal{G}) = \frac{1}{\ln(\beta/\alpha)} \cdot \frac{1}{z} \mathbf{1}_{[-\beta, -\alpha]}(z), \quad \text{for each subgroup } \mathcal{G} \in \{\mathcal{G}_{\text{maj}}^-, \mathcal{G}_{\text{min}}^-\}, \quad (8)$$

with $\beta > \alpha > 1$. Here, $\mathbf{1}_A(z)$ denotes the indicator function that equals 1 if $z \in A$ and 0 otherwise.

(2) The model initialization satisfies

$$\theta_1^0 - \theta_2^0 = -(\theta_3^0 - \theta_4^0) \leq -1.$$

Assumption (1) imposes symmetric feature distributions within each subgroup and ensures that high-magnitude features occur with probability inversely proportional to their magnitude, therefore penalizing too large features. Assumption (2) is very reasonable as discussed in Remark 2.3 and guarantees numerical stability as required in Lemma C.1.

Theorem 1.3. *Under Assumptions 1.2, for all $t \geq 0$ such that*

$$h \leq \eta \frac{\ln t \ln(\beta/\alpha) e^{2\alpha}}{2\epsilon(2\alpha+1)t} \quad \text{with } 0 < \eta < 1, \quad (9)$$

there exists two positive constants κ_{maj} and κ_{min} such that :

$$\mathbb{E}_{\mathcal{G}_{\text{maj}}} [1 - p_y(\mathbf{x}, \theta^t)] = \frac{\kappa_{\text{maj}}}{(1-\varepsilon)t} + o\left(\frac{1}{t}\right), \quad (10)$$

$$\mathbb{E}_{\mathcal{G}_{\text{min}}} [1 - p_y(\mathbf{x}, \theta^t)] = \frac{\kappa_{\text{min}}}{\varepsilon t} + o\left(\frac{1}{t}\right), \quad (11)$$

Theorem 1.3 precisely quantifies the difference in convergence speeds between the two subgroups in this minimal model. Specifically, the expected prediction error decays at rates inversely proportional to each subgroup's relative size, reproducing the asymptotic behavior described in Theorem 1.1.

2 Proof of Statement

2.1 Symmetry Properties

Lemma 2.1. *We have the following symmetry properties, for all $t \geq 0$:*

$$(i) \theta_3^t - \theta_4^t = -(\theta_1^t - \theta_2^t), \quad (ii) \theta_3^t + \theta_4^t = -(\theta_1^t + \theta_2^t)$$

Proof. These follow from the gradient formula:

$$\nabla_{\theta} \ell(f(\mathbf{x}, \theta), y) = y(1 - p_y(\mathbf{x}, \theta)) \cdot [\mathbf{r}, \mathbf{s}, -\mathbf{r}, -\mathbf{s}]^T \quad (12)$$

$$\text{Summing over } \mathcal{D}_{\text{train}} \text{ yields:} \quad \nabla_{\theta_1} \mathcal{L} = -\nabla_{\theta_3} \mathcal{L} \quad \text{and} \quad \nabla_{\theta_2} \mathcal{L} = -\nabla_{\theta_4} \mathcal{L} \quad (13)$$

The result follows by induction using Eq (1). \square

2.2 Learning Dynamics

Proposition 2.2. *The parameter gap evolves according to*

$$\theta_1^{t+1} - \theta_2^{t+1} = \theta_1^t - \theta_2^t + h \Psi(\theta_1^t - \theta_2^t) + h \xi(\theta_1^t - \theta_2^t), \quad (14)$$

where

$$\Psi(x) = \frac{\varepsilon}{\ln(\beta/\alpha) \cdot x} \ln(1 + e^{2\alpha x}) \quad (15)$$

$$\xi(x) = -\frac{\varepsilon}{\ln(\beta/\alpha) \cdot x} \ln(1 + e^{2\beta x}). \quad (16)$$

Proof. We proceed in three steps.

Step 1: Computing the gradient difference. We begin by expanding the gradient difference:

$$\begin{aligned}
\nabla_{\theta_1} \mathcal{L}^t - \nabla_{\theta_2} \mathcal{L}^t &= \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(\mathbf{x}, y) \in \mathcal{D}_{\text{train}}} \left(\nabla_{\theta_1} \ell(f(\mathbf{x}, \theta^t), y) - \nabla_{\theta_2} \ell(f(\mathbf{x}, \theta^t), y) \right) \\
&= \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(\mathbf{x}, y) \in \mathcal{D}_{\text{train}}} y(1 - p_y(\mathbf{x}, \theta^t))(\mathbf{r} - \mathbf{s}) \quad (\text{by Eq. (12)}) \\
&= \mathbb{E}[y(1 - p_y)(\mathbf{r} - \mathbf{s})] \\
&= \mathbb{E}_{\mathcal{G}_{\text{maj}}} [y(1 - p_y)(\mathbf{r} - \mathbf{s})] p(\mathcal{G}_{\text{maj}}) + \mathbb{E}_{\mathcal{G}_{\text{min}}} [y(1 - p_y)(\mathbf{r} - \mathbf{s})] p(\mathcal{G}_{\text{min}}) \\
&= 2\varepsilon \mathbb{E}_{\mathcal{G}_{\text{min}}} [y(1 - p_y)\mathbf{r}] \quad (\text{using Eqs. (4)–(5)}) \\
&= 2\varepsilon \left(\mathbb{E}_{\mathcal{G}_{\text{min}}^+} [(1 - p_1)\mathbf{r}] p(y=1 | \mathcal{G}_{\text{min}}) - \mathbb{E}_{\mathcal{G}_{\text{min}}^-} [(1 - p_{-1})\mathbf{r}] p(y=-1 | \mathcal{G}_{\text{min}}) \right)
\end{aligned}$$

Since $y = \text{sign}(\mathbf{r})$ for all $((\mathbf{r}, \mathbf{s}), y) \in \mathcal{D}_{\text{train}}$, the above quantity is positive, implying that $(\theta_1^t - \theta_2^t)$ decreases over training by Eq. (1). Given the initialization $\theta_1^0 - \theta_2^0 \leq -1$ (Assumption 1.2), we have:

$$\forall t \geq 0, \quad \theta_1^t - \theta_2^t \leq -1. \quad (17)$$

Step 2: Computing the expectations. We write $\theta = \theta^t$ (the superscript t is omitted for clarity), and introduce the mappings:

$$\phi(x) := \frac{1}{1 + e^{-2x}}, \quad \phi'(x) = \frac{1}{2 \cosh^2(x)} > 0, \quad (18)$$

$$\begin{aligned}
g_+(\mathbf{r}) &:= (1 - p_1((\mathbf{r}, -\mathbf{r}), \theta^t)) \mathbf{r}, & \text{for } \mathbf{r} > 0, \\
g_-(\mathbf{r}) &:= (1 - p_{-1}((\mathbf{r}, -\mathbf{r}), \theta^t)) \mathbf{r}, & \text{for } \mathbf{r} < 0.
\end{aligned}$$

Using Eq. (6) and Lemma 2.1(i), we can rewrite

$$g_+(\mathbf{r}) = \mathbf{r} \left(1 - \frac{e^{(\theta_3 - \theta_4)\mathbf{r}}}{e^{(\theta_1 - \theta_2)\mathbf{r}} + e^{(\theta_3 - \theta_4)\mathbf{r}}} \right) = \mathbf{r} \phi((\theta_1 - \theta_2)\mathbf{r}),$$

This allows us to compute the expectation

$$\mathbb{E}_{\mathcal{G}_{\text{min}}^+} [(1 - p_1)\mathbf{r}] = \frac{1}{2 \ln(\beta/\alpha) \cdot (\theta_1 - \theta_2)} \ln \left(\frac{1 + e^{2\beta(\theta_1 - \theta_2)}}{1 + e^{2\alpha(\theta_1 - \theta_2)}} \right) \quad (\text{using Eq (7)})$$

For g_- , we similarly have

$$g_-(\mathbf{r}) = \mathbf{r} \left(1 - \frac{e^{(\theta_1 - \theta_2)\mathbf{r}}}{e^{(\theta_1 - \theta_2)\mathbf{r}} + e^{(\theta_3 - \theta_4)\mathbf{r}}} \right) = \mathbf{r} \phi(-(\theta_1 - \theta_2)\mathbf{r}) = -g_+(-\mathbf{r}),$$

which directly gives $\mathbb{E}_{\mathcal{G}_{\text{min}}^+} [(1 - p_1)\mathbf{r}] = -\mathbb{E}_{\mathcal{G}_{\text{min}}^-} [(1 - p_{-1})\mathbf{r}]$.

Step 3: Final update. Substituting these two results into our earlier expression yields:

$$\nabla_{\theta_1} \mathcal{L}^t - \nabla_{\theta_2} \mathcal{L}^t = \frac{\varepsilon}{\ln(\beta/\alpha) \cdot (\theta_1 - \theta_2)} \ln \left(\frac{1 + e^{2\beta(\theta_1 - \theta_2)}}{1 + e^{2\alpha(\theta_1 - \theta_2)}} \right)$$

Using the gradient descent rule Eq. (1), we finally obtain:

$$\begin{aligned}
\theta_1^{t+1} - \theta_2^{t+1} &= \theta_1^t - \theta_2^t - h(\nabla_{\theta_1} \mathcal{L}^t - \nabla_{\theta_2} \mathcal{L}^t) \\
&= \theta_1^t - \theta_2^t - \frac{\varepsilon h}{\ln(\beta/\alpha) \cdot (\theta_1^t - \theta_2^t)} \ln \left(\frac{1 + e^{2\beta(\theta_1 - \theta_2)}}{1 + e^{2\alpha(\theta_1 - \theta_2)}} \right) \\
&= \theta_1^t - \theta_2^t + h \Psi(\theta_1^t - \theta_2^t) + h \xi(\theta_1^t - \theta_2^t)
\end{aligned}$$

□

Remark 2.3. We highlight two useful observations:

- (i) The sequence $(\theta_1^t - \theta_2^t)_{t \geq 0}$ is decreasing and remains upper bounded by -1 . Moreover, the convergence condition $\nabla_{\theta_1} \mathcal{L}^t - \nabla_{\theta_2} \mathcal{L}^t = 0$ implies that this sequence diverges toward $-\infty$.
- (ii) Equation (14) corresponds to the explicit Euler scheme with step size h for the differential equation

$$v'(z) = \Psi(v(z)), \quad v(0) = \theta_1^0 - \theta_2^0 \quad (19)$$

where we added perturbations ξ to the *drift-function*:

$$w'(z) = (\Psi + \xi)(w(z)), \quad w(0) = \theta_1^0 - \theta_2^0 \quad (20)$$

Hence, the sequence $(\theta_1^t - \theta_2^t)_{t \geq 0}$ can be viewed as a discrete-time approximation of the continuous solution $v(t)$. We will provide a more accurate formulation of this result below.

2.3 Error Bound Between the Discrete and Continuous Dynamics

Proposition 2.4. Let $(\theta_1^t - \theta_2^t)$ evolve according to Eq. (14), and let v denote the solution of the continuous ODE (19). Then, for all $t \geq 0$ satisfying condition (9),

$$\theta_1^t - \theta_2^t = v(t h) + \mathcal{O}(1).$$

Proof. The proof proceeds in three steps.

Step 1: Bounding the perturbation error. Recall that Ψ and ξ are defined in Proposition 2.2. We aim to bound the deviation $|v(t) - w(t)|$, where w solves the perturbed ODE (20). Since $\Psi < \Psi + \xi < 0$ on \mathbb{R}^- , Lemma A.1 implies that $v(z) < w(z) < 0$ for all $z > 0$. Hence,

$$\begin{aligned} (w - v)' &= (\Psi + \xi)(w) - \Psi(v) = ((\Psi + \xi)(w) - (\Psi + \xi)(v)) + \xi(v) \\ &\leq \xi(v), \quad \text{since } (\Psi + \xi) \text{ is decreasing on } \mathbb{R}^- \text{ and } v < w. \end{aligned}$$

Applying Grönwall's inequality (Emmrich, 1999, *Proposition 2.2*) gives

$$0 < (w - v)(z) \leq \int_0^z \xi(v(x)) dx \leq \int_0^{+\infty} \xi(v(x)) dx,$$

where the last inequality follows from Lemma C.1.

Step 2: Discretization error of the Euler method. We now bound the error $|w(t h) - (\theta_1^t - \theta_2^t)|$ due to the Euler discretization. By Sauer (2018, *Corollary 6.5*), the global error depends on the Lipschitz constants of $(\Psi + \xi)$ and its derivative. Since both $|\Psi + \xi|$ and $|(\Psi + \xi)'|$ are increasing on \mathbb{R}^- ,

$$\sup_{x \leq -1} |\Psi + \xi| = |(\Psi + \xi)(-1)|, \quad \sup_{x \leq -1} |(\Psi + \xi)'| = |(\Psi + \xi)'(-1)|.$$

Following the notation of Sauer (2018), set

$$L := \sup_{x \leq -1} |(\Psi + \xi)'|, \quad M := \sup_{x \leq -1} |(\Psi + \xi)'(\Psi + \xi)|.$$

Then,

$$L \leq \frac{2\varepsilon}{\ln(\beta/\alpha)} (2\alpha + 1) e^{-2\alpha}, \quad (21)$$

$$\frac{M}{L^2} = \frac{|(\Psi + \xi)(-1)|}{|(\Psi + \xi)'(-1)|} < 1, \quad (22)$$

since $|(\Psi + \xi)'(-1)| = |(\Psi + \xi)(-1)| + \frac{\varepsilon}{\ln(\beta/\alpha)} \left(\frac{2\alpha}{e^{2\alpha}+1} - \frac{2\beta}{e^{2\beta}+1} \right)$.

Under condition (9), Eq (21) gives $h \leq \frac{\eta \ln t}{L t}$, and thus by Sauer (2018, *Corollary 6.5*), for all $z \leq t$:

$$|\theta_1^z - \theta_2^z - w(z h)| \leq \frac{M h}{2L} (e^{L z h} - 1) \leq \frac{M t h^2}{2} e^{L t h} \leq \frac{M (\eta \ln t)^2}{2L^2 t^{1-\eta}} \leq \frac{(\ln t)^2}{2t^{1-\eta}},$$

where the last inequality follows from (22).

Step 3: Combining both bounds. We finally combine the two sources of error:

$$\begin{aligned} |\theta_1^t - \theta_2^t - v(t h)| &\leq |\theta_1^t - \theta_2^t - w(t h)| + |w(t h) - v(t h)| \\ &\leq \frac{(\ln t)^2}{2t^{1-\eta}} + \int_0^{+\infty} \xi(v(x)) dx \\ &= \mathcal{O}(1), \end{aligned}$$

which concludes the proof. \square

Remark 2.5 (On the Practical Validity of Condition (9)). The constraint imposed by condition (9) is mild in practice. Using standard parameter values:

$$\frac{\beta}{\alpha} = e, \quad \alpha = 1, \quad \varepsilon = 0.1, \quad \eta = 0.5, \quad h = 10^{-4},$$

we find that the condition remains valid up to $t \approx 8.4 \times 10^5$, which corresponds to a large enough training horizon in these type of settings.

Moreover, it is standard in optimization theory to require the product hT to exceed a minimal threshold to ensure convergence and near-optimality, as discussed in Zadeh et al. (2020, *Theorems 8.3–8.7*). This requirement is fully compatible with our condition (9).

Indeed, letting T denote the total training horizon and setting

$$h_T = \eta \frac{\ln T \ln(\beta/\alpha) e^{2\alpha}}{2\varepsilon(2\alpha+1)T},$$

we obtain: (i) condition (9) holds for all $2 \leq t \leq T$, and (ii) $h_T T = \eta \frac{\ln T \ln(\beta/\alpha) e^{2\alpha}}{2\varepsilon(2\alpha+1)} \rightarrow \infty$ as $T \rightarrow +\infty$. Hence, both conditions can be satisfied simultaneously under realistic training regimes.

2.4 Conclusion of the Proof

Proof. We now complete the proof of Theorem 1.3.

From Eq. (18), we have

$$\mathbb{E}_{\mathcal{G}_{\min}^+}[(1-p_1)(\mathbf{x}, \theta^t)] = \mathbb{E}_{\mathcal{G}_{\min}^+}[\phi((\theta_1^t - \theta_2^t)\mathbf{r})] = \frac{1}{\ln(\beta/\alpha)} I(\theta_1^t - \theta_2^t),$$

where

$$I(x) = \int_{\alpha}^{\beta} \frac{1}{z(1 + e^{-2xz})} dz. \tag{23}$$

By Lemma D.1,

$$\mathbb{E}_{\mathcal{G}_{\min}^+}[(1-p_1)(\mathbf{x}, \theta^t)] \sim -\frac{1}{\ln(\beta/\alpha)} \cdot \frac{e^{2\alpha(\theta_1^t - \theta_2^t)}}{2\alpha(\theta_1^t - \theta_2^t)}.$$

But,

$$\begin{aligned} \frac{e^{2\alpha(\theta_1^t - \theta_2^t)}}{2\alpha(\theta_1^t - \theta_2^t)} &= \frac{e^{2\alpha v(th) + \mathcal{O}(1)}}{2\alpha v(th) + \mathcal{O}(1)} && \text{(using Proposition 2.4)} \\ &= \frac{e^{-\ln t + \ln \ln t + \mathcal{O}(1)}}{-\ln t + o(\ln t)} && \text{(using Proposition B.1)} \\ &= e^{\mathcal{O}(1)} \cdot \frac{1}{-t} (1 + o(1)) \end{aligned}$$

From the proofs of Propositions 2.4 and B.1, the term $\mathcal{O}(1)$ corresponds to a convergent quantity whose limit can be expressed as

$$-\ln \varepsilon + o(-\ln \varepsilon) + \gamma(\alpha, \beta, \varepsilon),$$

where $\gamma(\alpha, \beta, \varepsilon) = o(-\ln \varepsilon)$. Specifically, Proposition 2.4 introduces a linear dependence on ε through the integral term $\int \xi(v)$, while Proposition B.1 contributes a constant term $-\ln \varepsilon$ via the constant C in its proof. Therefore, there exists $\kappa > 0$ such that

$$\mathbb{E}_{\mathcal{G}_{\min}^+}[(1-p_1)(\mathbf{x}, \theta^t)] \sim -\frac{1}{\ln(\beta/\alpha)} \cdot \frac{e^{2\alpha(\theta_1^t - \theta_2^t)}}{2\alpha(\theta_1^t - \theta_2^t)} \sim \frac{\kappa}{\varepsilon t}.$$

By model's weights symmetry (Proposition 2.2) and feature symmetry (Assumption 1.2),

$$\mathbb{E}_{\mathcal{G}_{\min}^-}[(1-p_{-1})(\mathbf{x}, \theta^t)] = \mathbb{E}_{\mathcal{G}_{\min}^+}[(1-p_1)(\mathbf{x}, \theta^t)] \sim \frac{\kappa}{\varepsilon t}.$$

Combining the two symmetric subgroups, we obtain:

$$\begin{aligned} \mathbb{E}_{\mathcal{G}_{\min}}[(1-p_y)(\mathbf{x}, \theta^t)] &= \mathbb{E}_{\mathcal{G}_{\min}^+}[(1-p_1)(\mathbf{x}, \theta^t)] p(y=1 \mid \mathcal{G}_{\min}) + \mathbb{E}_{\mathcal{G}_{\min}^-}[(1-p_{-1})(\mathbf{x}, \theta^t)] p(y=-1 \mid \mathcal{G}_{\min}) \\ &\sim \frac{\kappa}{\varepsilon t} \quad \text{as } t \text{ grows large enough.} \end{aligned}$$

By symmetry, replacing ε with $(1 - \varepsilon)$ and $\theta_1 - \theta_2$ with $\theta_1 + \theta_2$ yields the analogous result for the majority group \mathcal{G}_{maj} . \square

3 Generalization Gap

Equations (11) and (10) reveal that model performance improves faster on G than on G' , since $\varepsilon < 1 - \varepsilon$. Let θ^* denote the final parameters after $T \simeq 10^6$ training steps (to comply with Assumption ??). To assess generalization, we evaluate the loss on a *balanced* test dataset \tilde{D} where both groups have equal size:

- \tilde{G} : bias-aligned points $\{(1, 1), (-1, -1)\}$ duplicated n times, $|\tilde{G}| = 2n$
- \tilde{G}' : bias-conflicting points $\{(-1, 1), (1, -1)\}$ duplicated n times, $|\tilde{G}'| = 2n$
- $\tilde{D} = \tilde{G} \cup \tilde{G}'$ with $|\tilde{D}| = 4n$

The test loss is:

$$\begin{aligned} L(\tilde{D}, \theta^*) &= \frac{1}{|\tilde{D}|} \sum_{(x,y) \in \tilde{G}} l[f(x, \theta^*), y] + \frac{1}{|\tilde{D}|} \sum_{(x,y) \in \tilde{G}'} l[f(x, \theta^*), y] \\ &= \frac{1}{|\tilde{D}|} \sum_{(x,y) \in \tilde{G}} -\log p_y(x, \theta^*) + \frac{1}{|\tilde{D}|} \sum_{(x,y) \in \tilde{G}'} -\log p_y(x, \theta^*) \\ &\approx -\frac{1}{2} \log \left(1 - \frac{1}{4(1-\varepsilon)Th} \right) - \frac{1}{2} \log \left(1 - \frac{1}{4\varepsilon Th} \right) \quad (\text{by Eqs (11), (10)}) \\ &\approx \frac{1}{2} \cdot \frac{1}{4(1-\varepsilon)Th} + \frac{1}{2} \cdot \frac{1}{4\varepsilon Th} \quad (\text{Taylor expansion}) \\ &= \frac{1}{8\varepsilon(1-\varepsilon)Th} \end{aligned}$$

Conclusion: The smaller ε (i.e., the more imbalanced the training set), the larger the test loss, demonstrating poor generalization. The model overfits to the spurious correlation in the majority group, failing to learn that only x_2 is relevant for prediction.

4 Experiments

After rescaling by $4\varepsilon ht$, all curves converge to a constant plateau, confirming the predicted $\frac{1}{t}$ decay of $1 - p_y$. The plateau height varies with ε , reflecting higher-dimensional feature interactions absent from the idealized analytic model

References

- Etienne Emmrich. *Discrete versions of Gronwall's lemma and their application to the numerical analysis of parabolic problems*. Techn. Univ., 1999.
- Eric Jourdain. Intégrales généralisées : Théorème d'intégration des équivalents. <https://perso.univ-rennes1.fr/eric.jourdain/AP3/Chapitre4.pdf>, 2018. Théorème 4.11 — *Intégration des équivalents*.
- Timothy Sauer. *Numerical analysis*. Pearson, 2018.
- Matt Visser. Primes and the lambert w function. *Mathematics*, 6(4):56, 2018.
- Roderick Wong. *Asymptotic approximations of integrals*. SIAM, 2001.
- Reza Zadeh, Matroid, and Stanford University. Cme 323: Distributed algorithms and optimization, lecture 8 notes. https://stanford.edu/~rezab/dao/notes/L08/cme323_lec8.pdf, 2020. Course lecture notes, Spring 2020.

Appendix

A Solution Comparison with Respect to Drift Functions

Lemma A.1. Let $f, g : \mathbb{R}^- \rightarrow \mathbb{R}^-$ be continuous and negative functions such that

$$f(x) < g(x), \quad \forall x \in \mathbb{R}^-.$$

Consider the autonomous ODEs

$$\begin{aligned} u'(z) &= f(u(z)), & u(0) &= a, \\ v'(z) &= g(v(z)), & v(0) &= a, \end{aligned}$$

with the same negative initial condition $a < 0$. Then, for all $z \geq 0$,

$$u(z) < v(z).$$

Proof. Let F and G be primitives (antiderivatives) of $\frac{1}{f}$ and $\frac{1}{g}$, respectively. Integrating both ODEs yields

$$\begin{aligned} F(u(z)) - F(a) &= z, \\ G(v(z)) - G(a) &= z. \end{aligned}$$

We choose the integration constants such that $F(a) = G(a)$, hence

$$F(u(z)) = G(v(z)).$$

Since $f < g < 0$, it follows that

$$\begin{aligned} \frac{1}{g} < \frac{1}{f} < 0 &\Rightarrow G'(x) < F'(x) < 0. \\ &\Rightarrow G(a) - G(x) < F(a) - F(x) \quad \text{for all } x \leq a \end{aligned}$$

Thus, both F and G are strictly decreasing, with $G(x) > F(x)$ for all $x \leq a$.

Evaluating at $v(z)$ gives

$$F(u(z)) = G(v(z)) > F(v(z)),$$

and since F is strictly decreasing, we deduce $u(z) < v(z)$ for all $z > 0$. \square

B Asymptotic Behavior of the Solution to the ODE (19)

Proposition B.1. *The solution v of the differential equation (19) satisfies*

$$v(z) \sim -\frac{\ln z}{2\alpha} \quad \text{as } z \rightarrow +\infty. \quad (24)$$

More precisely

$$v(z) = -\frac{\ln z}{2\alpha} + \frac{\ln \ln z}{2\alpha} + \mathcal{O}(1) \quad \text{as } z \rightarrow +\infty. \quad (25)$$

Proof. We solve the Cauchy problem associated with Eq. (19). From $\frac{dv}{dz} = \Psi(v)$, we can write

$$\frac{dv}{\Psi(v)} = dz.$$

Integrating both sides yields

$$F(v(z)) = z + K,$$

where F is any primitive of $\frac{1}{\Psi}$, and $K = F(\theta_1^0 - \theta_2^0)$.

Asymptotic behavior of $1/\Psi$. As $x \rightarrow -\infty$, we have

$$\frac{1}{\Psi(x)} \sim C x e^{-2\alpha x}, \quad \text{with } C = \frac{\ln(\beta/\alpha)}{\varepsilon}.$$

Since $x e^{-2\alpha x}$ is not integrable over $(-\infty, 0]$, we can apply the *theorem of integration of asymptotic equivalents* Jourdain (2018, Theorem 4.11), which gives

$$F(x) = \int^x \frac{1}{\Psi(z)} dz \sim C \int^x z e^{-2\alpha z} dz = -\frac{C e^{-2\alpha x}}{2\alpha} \left(x + \frac{1}{2\alpha} \right) \quad \text{as } x \rightarrow -\infty.$$

Asymptotic inversion. Since $v(z) \rightarrow -\infty$ (see Remark 2.3(i)), we have

$$F(v(z)) - K = -\frac{C v e^{-2\alpha v}}{2\alpha} + o(v e^{-2\alpha v}) = z.$$

Hence,

$$-2\alpha v e^{-\alpha v} = \frac{4\alpha^2 z}{C(1+o(1))} = \frac{4\alpha^2 z}{C} (1+o(1)).$$

Using the principal real branch of Lambert W function we get

$$\begin{aligned} -2\alpha v &= W\left(\frac{4\alpha^2 z}{C}(1+o(1))\right) \\ &= \ln\left(\frac{4\alpha^2 z}{C}(1+o(1))\right) - \ln \ln\left(\frac{4\alpha^2 z}{C}(1+o(1))\right) + o(1), \quad (\text{using Visser (2018, Eq A2)}) \\ &= \ln z - \ln \ln z + \mathcal{O}(1). \end{aligned}$$

Therefore,

$$v(z) = -\frac{\ln z}{2\alpha} + \frac{\ln \ln z}{2\alpha} + \mathcal{O}(1) \quad \text{as } z \rightarrow +\infty.$$

□

C Bounding the Solution Error Induced by Perturbations

Lemma C.1. *Let v be the solution of the ODE (19). Then*

$$\int_0^{+\infty} \xi(v(x)) dx < +\infty$$

Proof. From Eq. (24), we have

$$\begin{aligned} e^{2\beta v(x)} &= e^{\frac{-\beta \ln x}{\alpha} + o(\ln x)} = x^{-\beta/\alpha} e^{o(\ln x)} = x^{-\beta/\alpha} o(x^\eta) \quad \text{since } o(\ln x) - \eta \ln x \rightarrow -\infty \text{ for all } \eta > 0, \\ &= o\left(x^{-\frac{\beta}{\alpha} + \eta}\right). \end{aligned}$$

Hence,

$$\begin{aligned} \xi(v(x)) &= -\frac{\varepsilon}{\ln(\beta/\alpha) v(x)} \left(e^{2\beta v(x)} + o(e^{2\beta v(x)}) \right) \quad (\text{by the Taylor expansion of } \ln(1+x) \text{ around 0}) \\ &= \frac{2\varepsilon\alpha}{\ln(\beta/\alpha) \ln x} \cdot o\left(x^{-\frac{\beta}{\alpha} + \eta}\right) \quad (\text{using Eq. (24)}) \\ &= o\left(\frac{x^{-\frac{\beta}{\alpha} + \eta}}{\ln x}\right) \end{aligned}$$

Therefore, $\xi(v(x))$ is integrable near $+\infty$. On another hand, $\xi(v(x)) \underset{x \rightarrow 0}{\sim} \xi(\theta_1^0 - \theta_2^0) < \xi(-1)$ using Assumption 1.2-(2). This guarantees integrability near 0.

Finally,

$$\int_0^{+\infty} \xi(v(x)) dx < +\infty$$

which establishes the claim. \square

D Asymptotic Behavior of the Misclassification Probability

Lemma D.1. *Let I be defined as in Eq. (23). Then,*

$$I(x) \sim -\frac{e^{2\alpha x}}{2\alpha x} \quad \text{as } x \rightarrow -\infty.$$

Proof. We proceed in two steps.

Step1: Approximation of $I(x)$. For $x < 0$, we estimate the deviation between $I(x)$ and the simplified integral:

$$\begin{aligned} \left| I(x) - \int_\alpha^\beta \frac{e^{2xz}}{z} dz \right| &= \left| \int_\alpha^\beta \frac{e^{2xz}}{z} \left(\frac{-e^{2xz}}{1+e^{2xz}} \right) dz \right| \\ &\leq e^{2x\alpha} \int_\alpha^\beta \frac{e^{2xz}}{z} dz = o\left(\int_\alpha^\beta \frac{e^{2xz}}{z} dz\right) \quad \text{as } x \rightarrow -\infty. \end{aligned}$$

Hence, $I(x) \sim \int_\alpha^\beta \frac{e^{2xz}}{z} dz$ as $x \rightarrow -\infty$.

Step 2: Asymptotic evaluation of the integral. Let $x = -u$ with $u \rightarrow +\infty$. Then

$$\int_\alpha^\beta \frac{e^{2xz}}{z} dz = \int_\alpha^\beta \frac{e^{-2uz}}{z} dz,$$

which is a standard Laplace-type integral. By Wong (2001, Chapter 3, Theorem 1),

$$\int_\alpha^\beta \frac{e^{-2uz}}{z} dz \sim \frac{e^{-2\alpha u}}{2\alpha u} = -\frac{e^{2\alpha x}}{2\alpha x} \quad \text{as } x \rightarrow -\infty.$$

Combining Steps 1 and 2 yields

$$I(x) \sim -\frac{e^{2\alpha x}}{2\alpha x}.$$

\square