Impact of Class Imbalance on Gradient Descent Dynamics and Generalization

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 neural network learning dynamics and generalization. Consider a simple example where each datapoint $x = (x_1, x_2) \in \mathbb{R}^2$ has a label y that depends only on x_2, i.e., $y = g(x_2)$. Here, x_2 is the relevant variable while x_1 is noise that the model should learn to ignore.

Dataset-Related Assumptions

For simplicity, we assume $x_i \in \{-1, +1\}$ for $i = 1, 2$, and $y = x_2$.

We construct two groups:

- **Majority group** G (bias-aligned): Contains $x = (1, 1)$ and $x = (-1, -1)$, where both variables align. Each vector is duplicated n times, so $|G| = 2n$.
- **Minority group** G' (bias-conflicting): Contains $x = (-1, 1)$ and $x = (1, -1)$, where the variables oppose each other. Each vector is duplicated n' times, so $|G'| = 2n'$.

The complete training dataset is $D = G \cup G'$ with $|D| = 2(n + n')$. We denote $\varepsilon = \frac{n'}{n+n'}$ as the proportion of the minority group.

Model and Training Setup

We consider a one-layer neural network:

$$f(x, \theta) = \begin{bmatrix} \theta_1 x_1 + \theta_2 x_2 \\ \theta_3 x_1 + \theta_4 x_2 \end{bmatrix}$$

We use the binary cross-entropy loss $l[f(x, \theta), y] = -\log p_y(x, \theta)$ where

$$\begin{bmatrix} p_{-1}(x, \theta) \\ p_1(x, \theta) \end{bmatrix} = \text{Softmax}[f(x, \theta)]$$

The empirical loss is:

$$L(D,\theta) = \frac{1}{|D|} \sum_{(x,y)\in D} l[f(x,\theta), y]$$

We train using gradient descent with learning rate h:

$$\theta^{t+1} = \theta^t - h\nabla_\theta L \tag{1}$$

2. Symmetry Properties

Lemma 1. (i) *If* $\theta_3^0 - \theta_4^0 = -(\theta_1^0 - \theta_2^0)$, *then*

$$\theta_3^t - \theta_4^t = -(\theta_1^t - \theta_2^t) \quad \textit{for all } t \geq 0$$

(ii) *If* $\theta_3^0 + \theta_4^0 = -(\theta_1^0 + \theta_2^0)$, *then*

$$\theta_3^t + \theta_4^t = -(\theta_1^t + \theta_2^t) \quad \textit{for all } t \geq 0$$

Proof. These follow from the gradient formula:

$$\nabla_\theta l[f(x,\theta), y] = y(1 - p_y(x,\theta)) \cdot [x_1, x_2, -x_1, -x_2]^T \tag{2}$$

Summing over D yields:

$$\nabla_{\theta_1} L = -\nabla_{\theta_3} L \quad \text{and} \quad \nabla_{\theta_2} L = -\nabla_{\theta_4} L \tag{3}$$

The result follows by induction using Eq (1). □

Throughout this work, we assume θ^0 satisfies the conditions in Lemma 1, which holds trivially when $\theta^0 = 0$.

3. Learning Dynamics

Proposition 1. *The parameter difference evolves according to:*

$$\boxed{\theta_3^{t+1} - \theta_4^{t+1} = \theta_3^t - \theta_4^t - 2\varepsilon h \phi(\theta_3^t - \theta_4^t)} \tag{4}$$

where $\phi(x) = \frac{e^x}{e^x + e^{-x}} = \frac{1}{1+e^{-2x}}$ *with* $\phi'(x) = \frac{1}{2\cosh^2(x)} > 0$.

Proof. We first compute $\nabla_{\theta_1} L^t - \nabla_{\theta_2} L^t$:

$$\begin{aligned}
\nabla_{\theta_1} L^t - \nabla_{\theta_2} L^t &= \frac{1}{|D|} \sum_{(x,y)\in D} \left[\nabla_{\theta_1} l[f(x,\theta^t), y] - \nabla_{\theta_2} l[f(x,\theta^t), y]\right] \\
&= \frac{1}{|D|} \sum_{(x,y)\in D} y[1 - p_y(x,\theta^t)](x_1 - x_2) \quad \text{(by Eq (2))} \\
&= \frac{1}{|D|} \sum_{(x,y)\in G'} y[1 - p_y(x,\theta^t)](-2y) \quad \text{(since } x_1 = x_2 \text{ for } x \in G\text{)}
\end{aligned}$$

Using $y^2 = 1$, we obtain:

$$\nabla_{\theta_1} L^t - \nabla_{\theta_2} L^t = \frac{-2}{|D|} \sum_{(x,y)\in G'} [1 - p_y(x, \theta^t)] \tag{5}$$

For $(x, y) \in G'$, we analyze both cases:
Case 1: If $y = 1$, then $x = (-1, 1)$ (since $x_2 = y$ and $x_1 x_2 < 0$):

$$\begin{aligned} 1 - p_y(x, \theta^t) &= \frac{e^{\theta_1 x_1 + \theta_2 x_2}}{e^{\theta_1 x_1 + \theta_2 x_2} + e^{\theta_3 x_1 + \theta_4 x_2}} \\ &= \frac{e^{-(\theta_1 - \theta_2)}}{e^{-(\theta_1 - \theta_2)} + e^{-(\theta_3 - \theta_4)}} \\ &= \frac{e^{(\theta_3 - \theta_4)}}{e^{(\theta_3 - \theta_4)} + e^{-(\theta_3 - \theta_4)}} \quad \text{(by Lemma 1)} \end{aligned}$$

Case 2: If $y = -1$, then $x = (1, -1)$, and similarly:

$$1 - p_y(x, \theta^t) = \frac{e^{(\theta_3 - \theta_4)}}{e^{(\theta_3 - \theta_4)} + e^{-(\theta_3 - \theta_4)}}$$

In both cases:

$$1 - p_y(x, \theta^t) = \phi(\theta_3^t - \theta_4^t) \tag{6}$$

Summing over G' (which contains $2n'$ points), Eq (5) gives:

$$\nabla_{\theta_1} L^t - \nabla_{\theta_2} L^t = \frac{-2 \cdot 2n'}{|D|} \phi(\theta_3^t - \theta_4^t) = -2\varepsilon\phi(\theta_3^t - \theta_4^t) \tag{7}$$

Finally:

$$\begin{aligned} \theta_3^{t+1} - \theta_4^{t+1} &= \theta_3^t - \theta_4^t - h(\nabla_{\theta_3} L^t - \nabla_{\theta_4} L^t) \quad \text{(by Eq (1))} \\ &= \theta_3^t - \theta_4^t + h(\nabla_{\theta_1} L^t - \nabla_{\theta_2} L^t) \quad \text{(by Eq (3))} \\ &= \theta_3^t - \theta_4^t - 2\varepsilon h\phi(\theta_3^t - \theta_4^t) \quad \text{(by Eq (7))} \end{aligned}$$

□

Remark 1. (i) *Equation* (4) *shows that* $(\theta_3^t - \theta_4^t)_{t\geq 0}$ *is a decreasing sequence. If* ℓ *denotes its limit in* $\mathbb{R} \cup \{\pm\infty\}$*, then* $\ell = \ell - 2\varepsilon h\phi(\ell)$*, which implies* $\phi(\ell) = 0$ *and thus* $\ell = -\infty$*.*

(ii) *Equation* (4) *is the explicit Euler method with step size* h *for the differential equation:*

$$u'(t) = -2\varepsilon\phi(u(t)), \quad u(0) = \theta_3^0 - \theta_4^0 \tag{8}$$

Assumption 1. *From Remark (i), we assume for simplicity that* $\theta_3^0 - \theta_4^0 \leq -1$*. Then the solution of Eq* (4) *can be approximated by the solution* u *of Eq* (8)*:*

$$\boxed{\forall t \geq 0, \quad \theta_3^t - \theta_4^t = u(th) + \delta_t}, \quad \text{with } |\delta_t| \leq 5 \cdot 10^{-5}$$

This assumption is justified by *Sauer, Numerical Analysis (2nd ed.), Corollary 6.5*. Using standard parameters:

$$h = 10^{-4}, \quad \varepsilon = 0.1, \quad M = \max_{s\leq -1} |2\varepsilon\phi' \cdot 2\varepsilon\phi| \leq 7.4 \times 10^{-3}, \quad L = \max_{x\leq -1} |2\varepsilon\phi'| = 4.2 \times 10^{-2}$$

Corollary 6.5 ensures:

$$|\theta_3^t - \theta_4^t - u(th)| \leq \frac{Mh}{2L}(e^{Lth} - 1) \leq 3.7 \times 10^{-5} \quad \forall t \leq 10^6$$

4. Convergence Analysis

Proposition 2. *Under Assumption 1:*

$$\forall t \geq 0, \quad \forall (x,y) \in G', \quad p_y(x, \theta^t) = 1 - \frac{1}{4\varepsilon t h} + \Delta_t + \mathcal{O}\Big(\frac{\log t}{t^2}\Big). \tag{9}$$

with $|\Delta_t| \leq 5 \cdot 10^{-5}$.

Proof. We solve the Cauchy problem in Eq (8). From $\frac{du}{dt} = -2\varepsilon\phi(u)$, we obtain:

$$\frac{du}{\phi(u)} = -2\varepsilon dt$$

Integrating gives:

$$u - \frac{e^{-2u}}{2} = -2\varepsilon t + K$$

where $K = \theta_3^0 - \theta_4^0 - \frac{e^{-2(\theta_3^0 - \theta_4^0)}}{2}$. Thus:

$$u(th) - \frac{e^{-2u(th)}}{2} = -2\varepsilon t h + K \tag{10}$$

For large t, $u(th) \sim -\frac{1}{2}\log(t)$ and:

$$\begin{aligned} e^{-2u(th)} &= 4\varepsilon t h + 2u(th) - 2K \\ \phi(u(th)) &= \frac{1}{e^{-2u(th)} + 1} = \frac{1}{4\varepsilon t h + 2u(th) - 2K + 1} \\ &= \frac{1}{4\varepsilon t h - \frac{1}{2}\log(t) + o(\log(t))} \\ &= \frac{1}{4\varepsilon t h} + \mathcal{O}\Big(\frac{\log t}{t^2}\Big). \end{aligned}$$

Finally:

$$\begin{aligned} 1 - p_y(x, \theta^t) &= \phi(\theta_3^t - \theta_4^t) \quad \text{(by Eq (6))} \\ &= \phi(u(th) + \delta_t) \quad \text{(by Assumption 1)} \\ &= \phi(u(th)) + \Delta_t \quad (|\Delta_t| < 5 \cdot 10^{-5} \text{ since } \phi \text{ is a contraction for } x \leq -1) \\ &= \frac{1}{4\varepsilon t h - \frac{1}{2}\log(t) + o(\log(t))} + \Delta_t \end{aligned}$$

□

Corollary 1. *For the majority group:*

$$\forall (x,y) \in G, \quad p_y(x, \theta^t) = 1 - \frac{1}{4(1-\varepsilon) t h} + \Delta'_t + \mathcal{O}\Big(\frac{\log t}{t^2}\Big). \tag{11}$$

with $|\Delta'_t| \leq 5 \cdot 10^{-5}$.

Proof. The proof follows the same steps as Proposition 2, replacing ε with $1 - \varepsilon$. □

5. Generalization Gap

Equations (9) and (11) reveal that model performance improves faster on G than on G', since $\varepsilon \ll 1 - \varepsilon$.

Let θ^* denote the final parameters after $T \simeq 10^6$ training steps (to comply with Assumption 1). 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:

$$
\begin{aligned}
L(\tilde{D}, \theta^*) &= \frac{1}{|\tilde{D}|} \sum_{(x,y)\in\tilde{G}} l[f(x,\theta^*), y] + \frac{1}{|\tilde{D}|} \sum_{(x,y)\in\tilde{G}'} l[f(x,\theta^*), y] \\
&= \frac{1}{|\tilde{D}|} \sum_{(x,y)\in\tilde{G}} -\log p_y(x,\theta^*) + \frac{1}{|\tilde{D}|} \sum_{(x,y)\in\tilde{G}'} -\log p_y(x,\theta^*) \\
&\approx -\frac{1}{2}\log\left(1 - \frac{1}{4(1-\varepsilon)Th}\right) - \frac{1}{2}\log\left(1 - \frac{1}{4\varepsilon Th}\right) \quad \text{(by Eqs (9), (11))} \\
&\approx \frac{1}{2}\cdot\frac{1}{4(1-\varepsilon)Th} + \frac{1}{2}\cdot\frac{1}{4\varepsilon Th} \quad \text{(Taylor expansion)} \\
&= \frac{1}{8\varepsilon(1-\varepsilon)Th}
\end{aligned}
$$

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.

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